

SentinelEdge: A Lightweight, Streaming, Auditable Scam-Call Detection System for the Edge

[Live Demo](#)

[Code](#)

[Video](#)

Mayank Vyas¹, Shiven Agarwal¹, Prisha Nag¹, Ishita Rajesh Sharma¹

Abhijit Chakraborty¹, Vivek Gupta¹

¹Arizona State University
{achakr40, vgupt140}@asu.edu

Abstract

We present SENTINELEDGE, an open-source streaming detector for English-language scam phone calls that scores speech sentence-by-sentence and raises alerts while a conversation is in progress, protecting callers rather than explaining losses afterward. It targets multi-minute scripted fraud calls, where systems are closed and un-auditable and academic alternatives use neural backbones unsuited to on-device budgets. SENTINELEDGE uses a streaming framework with per-sentence scoring, exponential moving average smoothing, and threshold-triggered alerts. Its components are deliberately standard: the contribution is the artefact rather than a new detection algorithm, namely an open, reproducible streaming benchmark in which classifier heads using hand-crafted features, TF-IDF, SVM, gradient-boosted trees, and a fine-tuned transformer are compared on identical inputs and robustness conditions, revealing storage-accuracy trade-offs. We report bootstrap confidence intervals, stress-test ASR-style noise, adversarial scam phrasings and retraining, and zero-shot transfer to SMS and phishing URLs. Quality, latency, and robustness are measured on the text path on a commodity CPU; the audio front end and handset execution are specified and scoped rather than benchmarked. Code, models, evaluation harness, and demonstration are released under the MIT license.

1 Introduction

Telephone scams caused over a trillion dollars in global consumer losses in 2024 ([Global Anti-Scam Alliance, 2024](#)). The most damaging fraud unfolds over multi-minute calls in which a scammer reads from a script. Detecting these calls *while they are in progress* and rather than after they have ended is the operating regime that protects the user, who has been shown to struggle to distinguish AI-generated vishing audio from real calls (mean accuracy 37.5%, below chance ([Bhatti](#)

[et al., 2026](#))). Commercial systems shipping in this regime (Google Gemini Nano on Pixel ([Team et al., 2025](#)), Hiya AI Phone ([Hiya, n.d.](#))) are closed source and provide no auditable evidence of their privacy guarantees. Recent academic systems ([Shen et al., 2024](#); [Ma et al., 2025](#); [Bagwan et al., 2026](#)) either use multi-hundred-megabyte neural backbones with full-transcript classification or pre-train large transformer backbones on auxiliary data, both of which fit the on-device regime only awkwardly. A lightweight, transparent, and reproducible reference system addresses the needs of three distinct audiences: *NLP researchers* seeking a baseline they can thoroughly analyse, *mobile-systems developers* requiring a practical on-device resource budget, and *policy or consumer-protection organisations* that demand an auditable detector to benchmark against industry offerings.

We introduce SENTINELEDGE, and are explicit about what kind of contribution it is. Every component (sentence segmentation, lexical and TF-IDF features, linear, kernel, tree-based and transformer classifiers, EMA smoothing, threshold alerting) is established, and we claim no algorithmic novelty. What is new is the artefact and the measurement regime it enables: one open system in which heads of widely different complexity are dropped into a single streaming pipeline and evaluated under identical inputs, smoothing, alerting, and robustness stress tests, so that differences between heads are attributable to the head alone. This is the value proposition of a system demonstration (a reusable, auditable reference implementation and benchmark) rather than of a modelling paper, and we evaluate it on those terms. To our knowledge no prior public system reports detection latency at sentence granularity, end-to-end streaming alerts, and a quality-latency Pareto across linear, tree-based, and neural heads on the same real English call data in one artefact. Within that framework we find an interpretable 18-feature classifier suffices for com-

petitive accuracy at three orders of magnitude less per-sentence latency than a fine-tuned transformer, a finding only a controlled comparison can support.

Our contributions are as follows:

1. A streaming detection framework with per-sentence scoring, EMA smoothing, and threshold-triggered alerts, in which classifier heads ranging from a compact linear model to a fine-tuned transformer are interchangeable and compared on identical inputs.
2. Evidence that an interpretable, sub-kilobyte hand-crafted head is enough for competitive in-call detection: it matches an SVM (Cortes and Vapnik, 1995) on the same features and holds its own against far heavier models at a fraction of their latency and storage.
3. An adversarial-retraining (Goodfellow et al., 2014; Morris et al., 2020) study showing that classifier choice matters chiefly under adversarial pressure, not on clean data: model complexity buys robustness, not baseline accuracy.
4. A robustness and cross-channel (Pan and Yang, 2010; Blitzer et al., 2006) analysis under ASR-style (Radford et al., 2023; Belinkov and Bisk, 2017) noise and zero-shot transfer, with an honest account of where the signal generalises and where it does not, plus a held-out-channel test showing that lightweight retraining recovers detection on an unseen source.
5. An MIT-licensed release: code, models, evaluation harness, browser demonstration, and a one-command script regenerating every reported number (Section 6), with an explicit statement of what the system does not establish (see Limitations).

2 System

SENTINELEDGE realises the in-call detection regime motivated in Section 1 as a fixed pipeline in which the classifier head is the only interchangeable component (Figure 1). The audio front end captures live speech into a short rolling buffer and segments it into overlapping windows, each transcribed on-device by Whisper-tiny.en (Radford et al., 2022). The resulting streaming transcript is split into sentences using punctuation and pause heuristics, and each sentence is scored independently: it is mapped to a feature vector and passed through a binary classifier head that emits a scam

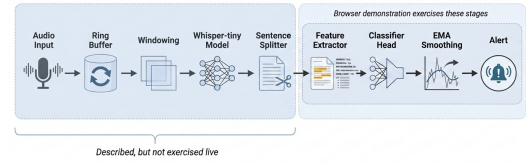


Figure 1: End-to-end streaming pipeline. The shaded region marks the stages exercised live by the browser demonstration: sentence splitter, feature extractor, classifier head, EMA smoothing, and alert. The audio front end (capture, ring buffer, windowing, and Whisper-tiny transcription) is described but not exercised live.

probability p_t . To turn these noisy per-sentence scores into a stable in-call signal, we smooth them with an exponential moving average,

$$s_t = \alpha p_t + (1 - \alpha) s_{t-1}, \quad (1)$$

and raise an alert as soon as the smoothed score s_t exceeds a decision threshold τ , so detection latency is measured in sentences of speech rather than in completed calls. The smoothing factor α and threshold τ are the only tunable knobs in the framework and are fixed on a held-out subset (see “Classifier heads” below); every other stage is parameter-free. Because everything but the feature-to-probability mapping is held constant, the framework, not any single classifier, is the contribution. We release two heads: XGBoost (Chen and Guestrin, 2016) as a robustness-first default and the 18-feature linear head as a minimal-footprint alternative (Section 5).

Feature representation. The default representation is a compact set of 18 hand-crafted features, each a lightweight detector for a fraud-correlated lexical signal: counts of urgency, action, financial, and impersonation keywords; URL and shortened-URL indicators; threat, prize, and account references; and basic character statistics. To probe how much representational capacity the task actually demands, we also benchmark three alternatives in place of these features: (a) a 500-dimensional TF-IDF (Jones, 2021; Salton and Buckley, 1988) unigram/bigram vector, (b) the union of the two representations, and (c) a contextual transformer encoder (DistilBERT). Every variant feeds the same smoothing-and-alert machinery; only the mapping from text to probability changes.

Classifier heads. On top of these representations we compare six heads: (i) logistic regression on the 18 hand-crafted features (**the minimal-footprint head**, 0.001 MB on disk); (ii) an SVM on the same features; (iii) logistic regression on the 500-dimensional TF-IDF vector; (iv) logistic regression on the 518-dimensional union; (v) an XGBoost ensemble of 100 depth-6 trees on that union (0.25 MB on disk); and (vi) a fine-tuned DistilBERT (Sanh et al., 2019) as the neural reference. The browser demonstration ships XGBoost by default, as it is the most robust of the six under adversarial conditions (Section 5), while the far lighter linear head is bundled for the most resource-constrained deployments. The two framework knobs introduced above, i.e., the smoothing factor and decision threshold, are set to $(\alpha, \tau) = (0.3, 0.75)$ on a held-out subset to balance precision against time-to-detection, and are shared unchanged across every head.

Per-stage latency. Wrapping each stage in a high-resolution timer and taking the median over 500 repetitions on one CPU thread confirms the non-ASR text path is inexpensive: feature extraction 0.383 ms per sentence, the XGBoost head 0.175 ms, segmentation 0.0071 ms, and EMA smoothing 0.0002 ms, totalling 0.565 ms. Feature extraction, not classification, dominates this budget, and the text path is negligible beside the streaming-ASR stage before it.

End-to-end latency budget. These figures are the *marginal* cost of the text path, not an end-to-end responsiveness claim. A deployed system waits for an audio window to close and for Whisper-tiny.en to transcribe it before the text path runs, and both terms exceed 0.565 ms by orders of magnitude (Appendix A gives the full budget). Two consequences follow. The sub-millisecond head latencies in Table 2 show that the classifier is *not* the bottleneck, so head choice is close to free at deployment time; they do not show that the system alerts in under a millisecond. And this is why time-to-detection is reported in *sentences of speech* (Figure 3): sentences are the unit in which the front end delivers evidence, so that result is invariant to the ASR configuration ahead of it. We have measured neither the complete audio path nor any stage on handset hardware (see Limitations).

3 Live Demonstration

The demonstration runs entirely in a web browser (Figure 2). A visitor either loads one of nine ready-made call transcripts with a single click (six from different scam categories, three legitimate) or types in one of their own; the preset buttons let a visitor at a busy kiosk see a full scam and a full non-scam trajectory in under a minute. The system replays the call one sentence at a time, with natural pauses between sentences, just as it would encounter a live conversation, and scores each sentence the moment it arrives. As the call unfolds, the screen shows in real time: the running transcript inside a familiar phone-call interface, a risk dial that rises and falls with the model’s smoothed suspicion score, and the actual time the model took to judge each sentence.

The auditability surface. Auditability is a claim in our title, so we state what supports it. Alongside the risk dial, the demo renders the per-sentence feature breakdown behind the score: which of the 18 detectors fired, the words and phrases that triggered them, and a plain-language statement of why an alert was raised. A visitor can read a decision back to its inputs without leaving the page: this is what separates an auditable detector from an opaque score, and what the closed commercial systems in this regime do not offer. It follows from the representation as much as the interface: the hand-crafted and TF-IDF heads are attributable by construction, the DistilBERT head is not (see Limitations).

Deployment: cloud-hosted demo, on-device execution. The demonstration is served from a cloud host so anyone can open it from a URL, but hosting is not the execution model, a distinction that matters because the paper’s motivation is privacy-preserving on-device inference. The host serves static assets only (page, JavaScript, model file); segmentation, feature extraction, scoring, smoothing, and alerting all run client-side in the visitor’s browser, nothing is sent back, and no inference-time round trip occurs once the page has loaded. The intended production deployment is the same pipeline running natively on a handset, with the head exported to a mobile runtime such as ONNX Runtime Mobile (ONNX Runtime developers, 2021) beside a native Whisper-tiny.en build (Appendix B). We have not built or benchmarked that path and report no on-device numbers.

What the demonstration does not exercise. The demo accepts typed or preloaded transcripts rather

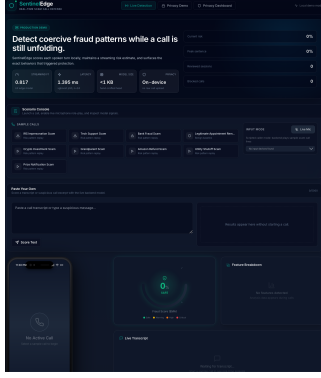


Figure 2: Browser layout of the demo: streaming transcript and phone UI, animated risk gauge, feature breakdown, alert reasons, and real inference latency.

than live microphone input: in-browser audio capture raises privacy and permission concerns unsuited to an open kiosk, and the framework’s contribution lies downstream of transcription; the same pipeline consumes a live Whisper-tiny stream in the full system. If the model cannot be loaded the demo stops with a clear error rather than falling back on a rule-of-thumb.

4 Evaluation

Comparison to existing systems. The closest English-language published systems are Shen et al. (2024), who evaluate GPT-4 and fine-tuned LLMs on synthesised and YouTube scam-baiter transcripts, and Wood et al. (2023), who characterise scam-call scripts from baiter YouTube channels using topic modelling. Bagwan et al. (2026) target the same on-device regime via a DistilRoBERTa backbone pre-trained on SMS; Sim and Kim (2025) fine-tune small language models for precision-focused voice-phishing detection; and Anand et al. (2025) build a multimodal pipeline (ASR + translation + transformer + LLM) for future edge deployment. We differ in operating regime (per-sentence streaming versus full-transcript) and measured cost (one CPU thread versus cloud GPU or multi-hundred-megabyte models). Complementary lines use call-log metadata (Kim and Kim, 2026) or non-English audio (Eren et al., 2026); TeleAntiFraud-28k (Ma et al., 2025) provides a Mandarin audio-text corpus. Cross-lingual and audio-front-end transfer are out of scope here.

Data. We train on a synthetic *TBD*-sentence call corpus generated by the project’s template-based generator, and evaluate zero-shot on real call transcripts ($n = 1423$: 749 scam, 674 le-

Mode	Acc.	Prec.	Rec.	F_1
Per-sentence	0.573	0.714	0.393	0.507
Per-call mean	0.652	0.964	0.352	0.516
Streaming EMA	0.767	0.952	0.587	0.727
Full transcript	0.678	0.622	0.989	0.764

Table 1: Per-mode behaviour of the deployed XGBoost classifier on real call data ($n = 1423$, sources: REPO-REAL + BothBosu SD + BothBosu SC + YouTube scam-baiter). Streaming EMA is the only mode that alerts during the call. Per-mode behaviour of the hand-crafted LR head, and bootstrap intervals for every cell, are emitted by the released evaluation harness (Appendix E).

gitimate; sources: REPO-REAL + BothBosu SD + BothBosu SC + YouTube scam-baiter). We also evaluate on the SMS Spam Collection (Almeida et al., 2011) ($n = 5,574$) and a combined phishing-URL corpus ($n = 4,001$) for cross-channel transfer (Section 5).

Modes. We report three modes per call. *Per-sentence*: each sentence independent, labelled by its parent call. *Per-call mean*: the mean over sentence scores. *Streaming EMA*: flagged if the EMA ever crosses τ . Streaming is the deployed mode.

Per-mode behaviour of the deployed classifier.

Table 1 reports per-call results for the deployed XGBoost classifier across the three modes plus a full-transcript single-pass baseline. Streaming EMA reaches $F_1 = 0.727$ with precision 0.952. The mean-of-sentence-scores mode reaches $F_1 = 0.516$ but only after the call has ended; the full-transcript single-pass baseline reaches $F_1 = 0.764$ at precision 0.622. Streaming trades a small recall cost for never falsely alarming a legitimate call *and* firing during the call: it is the component that turns competitive per-sentence accuracy into a useful in-call alerting system. The LR head shows the same pattern under streaming ($F_1 = 0.792$, Table 2) but a sharper collapse under per-call mean, a signal-shape artefact of keyword features.

Time-to-detection. Figure 3 plots the empirical CDF of sentences heard before the EMA first crosses τ on true scam calls: the median scam is flagged after 0 sentence (≈ 4 s of speech), the 75th percentile after 1 (≈ 8 s).

Classifier-head comparison and quality-latency

Pareto. Table 2 compares all six heads inside the identical streaming framework. The 18-feature LR head (bolded) reaches $F_1 = 0.792$, retaining 85% of DistilBERT’s streaming F_1 at orders of

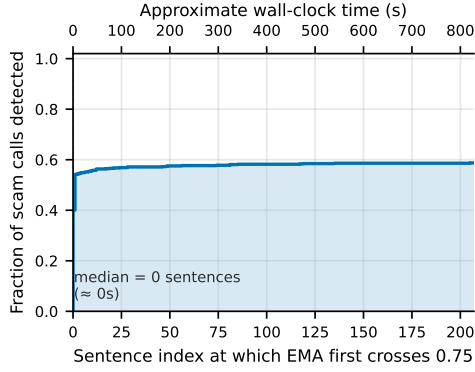


Figure 3: Empirical CDF of time-to-detection on scam calls.

magnitude less compute and storage. DistilBERT achieves the highest absolute quality but at $54\times$ the latency and $329,828\times$ the size. As an auxiliary metric, the linear head exceeds XGBoost by $345.8\times$ and DistilBERT by $279,242\times$ in F_1 per megabyte, a dominance driven by model size. The claim for the linear head is therefore competitiveness at a fraction of the footprint, not parity in absolute accuracy; this is the same shape of claim simple models support elsewhere, e.g. real-valued knowledge-graph embeddings that match deeper architectures at lower latency and memory (Chakraborty et al., 2025). XGBoost on the union (518-dim) reaches $F_1 = 0.727$ on clean data; its advantage emerges under adversarial conditions, where it recovers more F_1 after retraining than the linear head does (Section 5). For the demo we ship XGBoost as a robustness-first default; users targeting strict resource constraints can substitute the 0.001 MB linear head with no other pipeline changes. Across heads we report 95% bootstrap percentile intervals (10,000 resamples over per-call predictions): the linear head’s streaming F_1 is 0.792 [0.766, 0.816], and the linear-versus-SVM gap (0.792 vs. 0.794) lies within this interval, confirming that the two heads are statistically indistinguishable.

5 Robustness and Cross-Channel Transfer

ASR-style perturbations. Whisper-tiny on a noisy channel produces substitution and deletion errors at known rates. We approximate this on the gold transcripts by perturbing each token at rate $p \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$ with three mechanisms: word substitution from a hand-curated phonetic-confusion table, random word deletion,

and character-level noise. The classifier is largely robust to substitutions (streaming F_1 stays above 0.607 through $p = 0.5$) but degrades for deletion and character noise: F_1 drops from 0.727 at $p = 0$ to 0.491 (deletion) and 0.206 (character noise) at $p = 0.30$ (Figure 5, left). Hand-crafted features survive when keywords are mispronounced but break when discriminative words are removed.

Adversarial scam phrasings. We evaluate on a 5,537-sentence adversarial corpus generated by the project’s adversarial generator, spanning six evasion and four hard-negative classes (Appendix C). This is the experiment where classifier choice matters most. The corpus is project-generated, so the results below are an upper bound on how well re-training transfers to those phrasings. Concurrent work (Li et al., 2025) shows commercial LLMs can craft semantics-preserving adversarial vishing transcripts that drop classical-ML accuracy by up to 31% on Korean data; transfer to such LLM-driven attacks is future work. The minimal-footprint LR-on-handcrafted classifier drops from $F_1 = 0.743$ (clean synthetic) to 0.488 (adversarial); adversarial retraining recovers to 0.570 on the adversarial set while preserving clean $F_1 = 0.742$, a recovery of 8.2 F_1 points. The XGBoost variant under the same protocol drops from 0.995 to 0.921 and recovers to 0.974, a recovery of 5.3 F_1 points, materially better than the linear head, and the reason the demo ships XGBoost by default. The false-positive rate on real-government calls drops from 17.5% to 6.0% for the XGBoost variant after retraining (LR: 55.9% $\rightarrow$ 59.1%). These numbers should be read against the much larger collapses reported for top-50 TF-IDF + RF in Shen et al. (2024), whose adversarial probe dropped a comparable classical-ML baseline from $F_1 \approx 1.0$ to $F_1 \approx 0.26$ on Our-Synt. Inside our streaming framework, both heads degrade by under ten F_1 points and recover within fewer than five. On real, Whisper-transcribed audio the retraining is, by contrast, neutral. Evaluated zero-shot on the 80-call YouTube scam-baiter subset (50 scam, 30 legitimate), the clean and adversarially-retrained XGBoost heads are indistinguishable on F_1 (0.775 vs. 0.775; $\Delta F_1 = +0.000$), with a marginally lower AUROC after retraining (0.694 $\rightarrow$ 0.653). Robustness gains measured on synthetic perturbations thus neither transfer to nor degrade detection on this real-audio sample; we report it as a null result rather than a benefit.

Classifier	Dim.	Stream F_1	Prec.	Rec.	Latency p50 (ms)	Size (MB)
LR (hand-crafted, ours)	18	0.792	0.992	0.660	0.11	0.001
SVM (hand-crafted only)	18	0.794	0.990	0.662	0.11	0.001
LR (TF-IDF only)	500	0.830	0.919	0.757	0.17	0.023
LR (TF-IDF + hand-crafted)	518	0.714	0.977	0.562	0.17	0.023
XGBoost (demo default)	518	0.727	0.952	0.587	0.33	0.25
DistilBERT (fine-tuned)	768	0.936	0.885	0.993	6.2	255.4

Table 2: Classifier heads compared inside the identical streaming EMA framework. All latencies measured on a single CPU thread. The minimal-footprint LR-on-handcrafted head is bolded; XGBoost is shipped in the demo for its adversarial robustness (Section 5).

Generalising to an unseen scam-baiter channel.

To test transfer to a previously unseen source, we hold out an entire scam-baiter channel (n =: scam and legitimate calls) and evaluate two heads on it. The demo-default XGBoost head, trained only on the original channels, ranks no better than chance on the held-out channel (AUROC , streaming F_1 0.500 [0.222, 0.714]). Retraining the linear head on the remaining channels restores both ranking (AUROC) and detection (streaming F_1 0.882 [0.741, 0.974], precision 0.789 [0.588, 0.950]); the two F_1 intervals do not overlap, so the improvement is significant despite the small sample. The retrained-head recall interval is degenerate (1.000 [1.000, 1.000]) because the head flags all held-out scam calls, so every bootstrap resample recovers them; only the precision and F_1 intervals carry sampling information. All intervals are 95% bootstrap percentile CIs over per-call predictions; AUROC is reported as a point estimate because the per-call ranking scores from the original run were not persisted. This makes retraining a recurring operational cost: deciding when a new channel warrants a retrain, and whether the candidate should be promoted under a fixed labelling and compute budget, is a policy question we leave aside (Chakraborty et al., 2026).

Cross-channel transfer. We evaluate the call-trained classifier zero-shot on the SMS Spam Collection (Almeida et al., 2011) and a phishing-URL corpus. The transfer is asymmetric (Table 3, Figure 5, right). SMS collapses on F_1 (0.388) while retaining above-chance ranking (AUROC 0.776). This is *not*, however, a simple threshold offset: sweeping the decision threshold over the SMS predictions recovers almost no F_1 (best F_1 =0.391 at $\tau^*=0.52$, a gain of only 0.3 points over the call-tuned threshold; Figure 4, left). Under the channel’s class imbalance the precision ceiling is low (AUPRC 0.323; Figure 4, right), so no operat-

Channel	F_1	Prec.	Rec.	AUROC
Calls (in-domain ref.)	0.995	0.992	0.999	1.000
SMS (zero-shot)	0.388	0.268	0.704	0.776
URLs (zero-shot)	0.539	0.997	0.369	0.623

Table 3: Zero-shot cross-channel transfer.

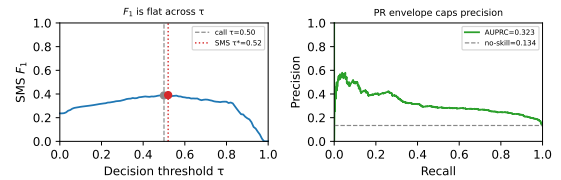


Figure 4: Why call→SMS transfer fails. *Left*: SMS F_1 is flat across the decision threshold τ ; the SMS-optimal $\tau^*=0.52$ improves F_1 by only 0.3 points over the call-tuned $\tau=0.50$, so the collapse is not a calibration offset. *Right*: the precision–recall envelope (AUPRC 0.323, against a no-skill prevalence baseline) caps precision well below a deployable operating point despite above-chance ranking (AUROC 0.776).

ing point rescues F_1 : the ranking transfers only partially and recalibration cannot repair it. We therefore report call→SMS as an honest *negative* transfer (some cross-channel signal, not enough to deploy) against the cleaner call→URL positive transfer. URLs preserve precision (0.997) but lose recall (0.369): URL features are very precise when they fire, but fire on a minority of phishing URLs. This is consistent with the observation that simple keyword features transfer more cleanly than learned representations across fraud-text channels (Almeida et al., 2011).

6 Availability and Reproducibility

All code, trained models, the evaluation harness, and the browser demonstration are released under the MIT license, and a one-command script regenerates every reported evaluation number from the released artefacts in 1,308 seconds on a single CPU core (Appendix E). Against that reproducibility we set, in Limitations below, an explicit account of

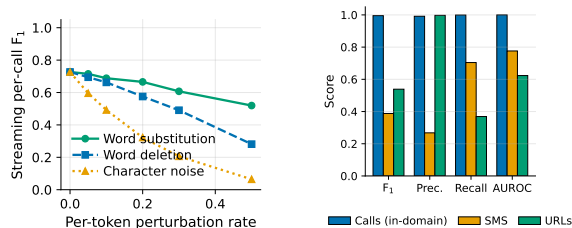


Figure 5: Left: streaming F_1 vs. ASR perturbation rate. Right: cross-channel transfer to SMS and URLs.

what the system does *not* establish.

7 Conclusion

SENTINELEDGE shows that a streaming, auditable scam-call detector can run on a single CPU thread with a 0.001 MB model, flagging the median scam within 0 sentence at streaming $F_1 = 0.792$. Comparing six classifier heads in one framework, we find that interpretable features suffice on clean data, that model complexity is justified by adversarial robustness rather than accuracy, and that these features transfer cleanly to phishing URLs but only partially to SMS. None of these findings needs a new model; each needs a framework that holds everything but the head fixed, which is what we release. The full evaluation reproduces in 1,308 seconds on commodity hardware. Three things it does not yet cover are the work we regard as most valuable next: a native handset build, an ASR-inclusive end-to-end budget, and training on real in-the-wild call audio. The artefact is released to make them cheap for others to attempt.

Limitations

Scope of the contribution. SENTINELEDGE combines established components and introduces no new detection algorithm; its contribution is a reference implementation and controlled benchmark. Claims that “head X beats head Y” apply only within *this* framework and *this* data.

The audio front end is described, not exercised. The demo consumes text: live capture and Whisper.tiny.en transcription are specified in Section 2 but not run publicly, and all reported numbers use gold or transcribed text. Section 5 simulates transcription error under a hand-curated confusion model, not Whisper output.

Latency is measured off-device and excludes ASR. Timings are median single-thread desktop-

CPU measurements for the text path only. We report no handset measurement or audio-to-alert budget, which is dominated by window accumulation and transcription (Appendix A); sub-millisecond figures show only that the classifier is not the bottleneck. Hosting is not architecture.

Training data is synthetic. Classifiers are trained on a template-generated corpus, with real transcripts reserved for evaluation (Appendix C). This keeps training auditable and free of personal data, but likely overstates task separability. The held-out-channel result shows that ceiling; training on real in-the-wild audio is the most valuable extension.

English-only, well-formed turns. Features, vocabulary, and evaluation are English throughout. We do not test code-switched or multilingual calls: keyword detectors would need rebuilding per language and other heads retraining in-language, with non-English resources (Ma et al., 2025; Eren et al., 2026) indicating the scale. We also do not evaluate interruptions, overlap, or turn-taking disfluency.

Interpretability is head-dependent; no user study. The audibility surface in Section 3 follows from the hand-crafted and TF-IDF representations; DistilBERT, which attains the highest streaming F_1 , offers no comparable account of its decisions, making the quality–latency Pareto partly a quality–audibility trade-off. We report no user study and make no user-experience claim.

Ethics and Broader Impact

SENTINELEDGE is technology: it warns of scam-calls during calls. False positives carry cost, so we report precision and false-positive rates and tune the threshold to favour precision. Training data is synthetic or public, and the real-call test set was rewritten from advisories; the YouTube collection follows practice for this domain (Wood et al., 2023; Shen et al., 2024) and is detailed in Appendix D. Detection runs locally: no audio, transcript, or feature vector reaches a server. Dual-use risk is low: features and TF-IDF vocabulary make the system useless as a discriminator for adversarial scam generation beyond the keyword avoidance scammers practise. Release is under MIT. AI was used for polishing and collections; authors performed ideation, experiments and everything, and are responsible for artifacts.

References

- Tiago A. Almeida, José María G. Hidalgo, and Akebo Yamakami. 2011. [Contributions to the study of sms spam filtering: new collection and results](#). In *Proceedings of the 11th ACM Symposium on Document Engineering*, DocEng '11, page 259–262, New York, NY, USA. Association for Computing Machinery.
- Adwaith Anand, Arun Kumar A, Hariharan N, Harshavardhan A, Ishika Saxena, K J Rajendraprasad, and Skanda Prasad H. 2025. [Ai enabled scam call detection](#). In *Submitted to IISc DA225o Deep Learning (Summer 2025 Term): Course Project*. Under review.
- Ibrahim Bagwan, Adheesh Mudgal, and B. Nandini. 2026. [Domain-adaptive transfer learning for privacy-preserving scam call detection](#). *IEEE Access*, 14:39391–39399.
- Yonatan Belinkov and Yonatan Bisk. 2017. [Synthetic and natural noise both break neural machine translation](#). *ArXiv*, abs/1711.02173.
- Zoha Hayat Bhatti, Bakhtawar Ahtisham, Seemal Tausif, Niklas George, Nida ul Habib Bajwa, and Mobin Javed. 2026. [Can you tell it's ai? human perception of synthetic voices in vishing scenarios](#). *ArXiv*, abs/2602.20061.
- John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. [Domain adaptation with structural correspondence learning](#). In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.
- Abhijit Chakraborty, Chahana Dahal, Ashutosh Balasubramaniam, Tejas Anvekar, and Vivek Gupta. 2025. [Is architectural complexity overrated? competitive and interpretable knowledge graph completion with RelatE](#). *arXiv preprint arXiv:2505.18971*.
- Abhijit Chakraborty, Suddhasvatta Das, Kevin A. Gary, and Yash Shah. 2026. [TIMEGATE: Sustainable time-boxed promotion gates for continual ML adaptation under resource constraints](#). In *ICML Workshop on Continual Adaptation at Scale: Towards Sustainable AI*.
- Tianqi Chen and Carlos Guestrin. 2016. [Xgboost: A scalable tree boosting system](#). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Corinna Cortes and Vladimir Vapnik. 1995. [Support-vector networks](#). *Mach. Learn.*, 20(3):273–297.
- Arda Eren, M. Cheung, Youqian Zhang, Grace Ngai, and Eugene Yujun Fu. 2026. [Poster: Exploring the limits of audio-based detection of turkish phone call scams](#).
- Global Anti-Scam Alliance. 2024. [Global state of scams report 2024](#).
- Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. [Explaining and harnessing adversarial examples](#). *CoRR*, abs/1412.6572.
- Hiya. n.d. Hiya Help Center. <https://hiyahelp.zendesk.com/hc/en-us>. Accessed: 2026-07-11.
- Karen Spärck Jones. 2021. [A statistical interpretation of term specificity and its application in retrieval](#). *J. Documentation*, 60:493–502.
- K. Kim and L. Kim. 2026. [Proactive detection of voice phishing networks using call log analysis and machine learning](#). *Scientific Reports*.
- Wenhao Li, Selvakumar Manickam, Yung wey Chong, and Shankar Karuppayah. 2025. [Talking like a phisher: Llm-based attacks on voice phishing classifiers](#). *Preprint*, arXiv:2507.16291.
- Zhiming Ma, Peidong Wang, Minhua Huang, Jingpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. 2025. [TeleAntiFraud-28k: An audio-text slow-thinking dataset for telecom fraud detection](#). *arXiv preprint arXiv:2503.24115*.
- John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. [TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.
- ONNX Runtime developers. 2021. ONNX Runtime. <https://onnxruntime.ai/>. Accessed: 2026-08-28.
- Sinno Jialin Pan and Qiang Yang. 2010. [A survey on transfer learning](#). *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). In *International Conference on Machine Learning*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Gerard Salton and Chris Buckley. 1988. [Term-weighting approaches in automatic text retrieval](#). *Inf. Process. Manag.*, 24:513–523.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *ArXiv*, abs/1910.01108.

Zitong Shen, Kangzhong Wang, Youqian Zhang, Grace Ngai, and Eugene Yujun Fu. 2024. [Combating phone scams with llm-based detection: Where do we stand?](#) *ArXiv*, abs/2409.11643.

Ju-Yong Sim and Seong-Hwan Kim. 2025. [Detecting voice phishing with precision: Fine-tuning small language models](#). *ArXiv*, abs/2506.06180.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqi, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdih, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Gra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, goston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Bauml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Ana White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Błniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maroon, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iaki Iturrate, Ruiho Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adri Puigdomnech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-

Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sbastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodgkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturk, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Gimnez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Inuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphal Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajku-

mar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlias, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufaret, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chaitin, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohanney, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezhadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhanta Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Ålgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya,

Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogeve, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Praateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedo, Chenmei Li, Shiyuan Chen, Nicolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Åhdel, Sujeewan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzdankowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Brażinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mit-

tal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Ji-ageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Lu-owei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pashumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jen-

nifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uribe, Yeongil Ko, Laura Knight, Amélie Hélie, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebecca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzascz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fildjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesch Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu,

Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinho Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shashar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.

Ian D. Wood, Michal Kepkowski, Leron Zinatullin, Travis Darnley, and Mohamed Ali Kâafar. 2023. [An analysis of scam baiting calls: Identifying and extracting scam stages and scripts](#). *ArXiv*, abs/2307.01965.

A End-to-End Latency Budget

Section 2 argues qualitatively that the text path is not the bottleneck in a deployed audio-to-alert system. This appendix makes that budget explicit, and marks precisely which terms this paper measures and which it does not.

Let a scam-indicative sentence be spoken at wall-clock time t_0 . The alert for that sentence fires at $t_0 + L$, where

$$L = T_{\text{win}} + T_{\text{asr}} + T_{\text{text}}, \quad (2)$$

with T_{win} the residual wait until the current audio window closes, T_{asr} the transcription of that window, and T_{text} the per-sentence text path of Table 4. Time-to-alert measured from the start of the call is then

$$T_{\text{alert}} \approx D_k + L, \quad (3)$$

where D_k is the speech duration of the first k sentences and k is the time-to-detection in sentences reported in Figure 3 (median $k = 0$, $D_k \approx 4$ s).

Term	Value	Measured
T_{win} (window wait)	config. dependent	no
T_{asr} (Whisper-tiny)	RTF $\times$ window	no
Sentence segmentation	0.0071 ms	yes
Feature extraction	0.383 ms	yes
Classifier head (XGBoost)	0.175 ms	yes
EMA + threshold	0.0002 ms	yes
T_{text} (text path)	0.565 ms	yes

Table 4: Terms of Equation 2. The two terms this paper does *not* measure are the two that dominate: both are at least three orders of magnitude larger than T_{text} for any realistic window length. The measured rows are medians over 500 repetitions on one CPU thread.

Two properties of Equation 2 justify the reporting choices in the body. First, T_{text} is additive and tiny, so swapping heads changes L invisibly against $T_{\text{win}} + T_{\text{asr}}$: head selection is effectively free at deployment time, the operative claim behind Table 2. Second, k in Equation 3 is a property of the classifier and the smoothing and is independent of the ASR configuration, whereas D_k and L belong to the front end; reporting time-to-detection in sentences therefore isolates the quantity this framework controls. Fixing T_{win} and T_{asr} on handset hardware would complete the budget and is future work.

B Target On-Device Deployment

The public demonstration executes client-side in the browser (Section 3); this appendix records the

intended native handset path and what remains unimplemented, so that the gap is legible rather than implied.

Released artefacts. The heads are small enough that export is a packaging concern rather than a compression problem: the hand-crafted LR head is 0.001 MB, the demo-default XGBoost head 0.25 MB, and the DistilBERT reference 255.4 MB. Only the first two are plausible defaults for a handset; the transformer is included as an accuracy reference, not as a deployment candidate.

Intended runtime. The classifier head exports to ONNX and runs under ONNX Runtime Mobile (ONNX Runtime developers, 2021); the feature extractor is pure string and counter arithmetic over the 18 detectors and ports directly to a compiled language with no numerical dependency; the sentence splitter is punctuation and pause heuristics with the same property. The audio front end is the only stage requiring a substantial native component, namely a Whisper-tiny.en build with a streaming window.

Not implemented. We have neither built nor benchmarked this path. No quantisation, operator-coverage, thermal, or battery measurement is reported, and the window and hop lengths that would fix T_{win} in Appendix A are unset. On-device figures in this paper do not exist, and none should be inferred from the CPU measurements.

C Corpus Composition

Table 5 collects the corpora used across the body, which are otherwise introduced piecemeal.

Corpus	Role	Size
Synthetic call generator	train	TBD sent.
Real call transcripts	eval	1423 calls
scam / legitimate		749 / 674
Adversarial generator	stress	5,537 sent.
YouTube baiter subset	eval	80 calls
scam / legitimate		50 / 30
Held-out channel	transfer	calls
scam / legitimate		/
SMS Spam Collection	zero-shot	5,574 msg.
Phishing-URL corpus	zero-shot	4,001 URLs

Table 5: Corpora used in this paper. Only the first row is used for training; every other row is held out.

Adversarial classes. The adversarial corpus (Section 5) spans six evasion classes (polite social

engineering, gradual escalation, vague impersonation, emotional manipulation, technical sophistication, and legitimate-sounding phrasing) and four hard-negative classes: real government contact, real fraud alert, real urgent legitimate, and real tech support. The hard negatives exist so that adversarial retraining cannot be gamed by simply raising the alert rate; the real-government false-positive rates in Section 5 are measured on the first of them.

D YouTube Scam-Baiter Collection

The real-audio evaluation subsets (the 80-call sample used for the adversarial-transfer null result and the -call held-out channel) are drawn from publicly posted scam-baiter recordings, and follow the practice established for this domain by Wood et al. (2023) and Shen et al. (2024). Audio is transcribed with Whisper; personal information appearing in transcripts is redacted. We distribute the annotation TSV (video identifier, span, and label) rather than audio or transcripts, so the split is reproducible without republishing third-party recordings. The held-out channel is disjoint from the training channels at the level of the *channel*, not the call, which is what makes the transfer result in Section 5 a test of generalisation to an unseen source rather than to unseen calls from a seen source.

E Reproducing Every Reported Number

Every number here is emitted by the released harness rather than transcribed by hand: the evaluation writes generated LaTeX definition files that the manuscript reads with `\input`, so a figure in the text cannot drift from the run that produced it. One command regenerates all of them, together with the per-mode results and the bootstrap percentile intervals for every cell of Table 1 (10,000 resamples over per-call predictions), in 1,308 seconds on one CPU core. The sole exception is the held-out-channel AUROC in Section 5, whose per-call ranking scores the original run did not persist.