

Verified-source Evidence for LLM-based Smishing Detection: A Reproducible Extension and Evaluation of SmishX

David Druker

University of Toronto, Canada

Corresponding author: David Druker, University of Toronto, Canada.

Submitted: August 27, 2026

Accepted: September 06, 2026

Published: September 15, 2026

Citation: David Druker (2026) Verified-source Evidence for LLM-based Smishing Detection: A Reproducible Extension and Evaluation of SmishX. *Epistora J. Artif. Intell. & Intell. Syst.* 1(1), 1-11. Article EJAIS-108

Abstract

SmishX [6] showed that pairing an LLM with link- and brand-context retrieval improves explainable SMS-phishing detection, but its released detector treats brand verification as a raw search-engine query and has no mechanism for checking what an impersonated organization actually says about the request it is being accused of making. We faithfully reproduce SmishX's detection pipeline (P0) and evaluate two extensions on a 283-message cohort (34 legitimate, 133 smishing, 116 spam): a corrected evidence packet that fixes how failed link collection is presented to the model, and an added retrieval layer for firstparty and known-campaign evidence (P1). On the 267 cases scored by all three adopted systems, P1 dominates P0 on all six metrics measured (recall 96.6%→99.1%, specificity 67.6%→73.5%, MCC 0.668→0.804), and the head-to-head comparison narrowly misses conventional significance (McNemar exact, matched $n=267$, $p=0.057$). An ablation isolating the packet fix alone accounts for most of the gain and is significant ($p=0.006$); the added evidentiary retrieval's marginal contribution over that corrected baseline is not ($p=0.69$). We report this as an honest, partially-negative result rather than a validated win, disclose a deviation from a predeclared analysis protocol, flag that the headline cohort blends two differently-scoped sampling waves, and note that a predeclared blinded audit found neither system's stated reasoning fully evidence-supported. This draft also corrects a significance value mislabeled in an earlier version of this paper (Appendix D).

Keywords: Smishing Detection, SMS Phishing, Large Language Models, AI Agents, Cybersecurity, Phishing Detection, Spam Detection, Fraud Detection, Information Retrieval, Source Verification

Introduction

Text-message fraud is common enough to be a routine nuisance and occasionally severe enough to be a real financial event. The U.S. Federal Trade Commission recorded \$470 million in reported 2024 consumer losses from scams that began with a text [1], and the Canadian Anti-Fraud Centre logged more than CAD \$704 million in reported 2025 fraud losses across more than 112,000 reports [2]; both figures cover broader fraud categories and should not be read as a directly attributable smishing-detection benefit. Separately, Pew found that 61% of U.S. adults report receiving a text attempting to extract personal information at least weekly [3] - self-reported suspected exposure, not a count of confirmed phishing messages. Together these numbers motivate the problem without licensing any specific accuracy claim, a distinction this paper tries to preserve throughout.

A recent line of applied work asks whether an LLM agent, given the ability to look things up, can do better than a static classifier at this task. SmishX [6], published at SOUPS 2025, is the clearest instance: it extracts the URLs and claimed brand from a message, enriches link and domain context, asks an LLM for a detailed judgment, and reports 98.8% accuracy on a 1,200-message labeled corpus. It is, however, a research prototype rather than a deployed system, and a source-level audit (Section 2) finds specific, fixable gaps: brand verification is a raw search-engine query rather than a check of the organization’s own official guidance, source authority is not ranked, and the spam/phishing distinction collapses into a single boolean.

This paper reproduces SmishX’s pipeline faithfully as a baseline (P0) and tests whether two targeted fixes improve it (P1): correcting how the pipeline represents link-collection failures to the model, and adding a retrieval step for firstparty (official-source) and known-campaign evidence. We evaluate both, paired, on a 283-case cohort, stratified by whether link evidence was actually available. The headline result is a clean sweep on paper - P1 beats P0 on precision, recall, specificity, F1, balanced accuracy, and MCC, with no metric worse - but the more useful and less flattering finding is that the sweep is not statistically significant, and that an ablation attributes nearly all of the real, significant gain to the mundane packet-format fix rather than to the added evidence retrieval that motivated this project in the first place. We report both results, and what would be needed to confirm or overturn the second one.

Related Work

SmishX. SmishX’s supplementary repository implements a single Python detector, a browser-screenshot helper, a relabeled 1,200-message SMS dataset (622 legitimate, 319 spam, 259 smishing per the paper’s published counts), and a 175-person user-study package reporting a System Usability Scale score of 82.6 [7, 10]. Its core sequence — extract URLs/brand, enrich link and domain context, ask an LLM for a decision, generate a short explanation — is close to the pipeline reproduced here as P0. A source audit conducted for this project (informing the extensions below) found several concrete limits worth naming precisely: the brandverification step is a search-engine query for the brand name rather than a lookup of that organization’s own fraud/security guidance; there is no source-authority ranking, so a convincing fake page and an official one are weighted the same; the decision collapses spam and phishing into one boolean; several network calls in the released code lack timeouts and browser analysis is not sandboxed; and the authors explicitly state they did not evaluate the system against adversarial inputs [9]. Independently, we found 14 duplicated message pairs and 2 malformed indicator rows in the released dataset, and a URL-flag total that does not exactly match the paper’s own stated count of 636 URL-containing messages — none of which invalidates the released accuracy figure, but all of which argue for cleaning before any downstream benchmarking (Appendix A). The repository carries an MIT license [8], which we rely on here as a research baseline and reference point, not as unmodified production code.

Deployed scam-warning products. This work sits adjacent to, but does not compete directly with, shipped consumer products that already warn on suspicious texts. Google Messages’ Scam Detection uses on-device conversational-pattern models and ships on by default in supported markets [4]; NordVPN’s Message Protection, launched in July 2026, automatically scans unknown-sender Android SMS for link, keyword, and behavioral indicators and posts a warning without blocking the message [5]. Neither vendor’s public description mentions case-specific web research, an official-source authority hierarchy, or an auditable separation of verified fact from inference. That gap is exactly where this paper’s extension is aimed - not at outscoring these products on a leaderboard, since none publish a comparable benchmark, but at testing whether a cited, source-checked evidence layer is something an LLM-agent pipeline can be shown to benefit from at all.

P1: Verified-Source Evidence Augmentation

P0 (baseline). We reproduce SmishX’s pipeline unmodified: (i) an LLM extracts candidate URLs and the claimed organization from the message text; (ii) links are normalized, expanded through redirects, and enriched with WHOIS and page-content evidence; (iii) the claimed brand is checked via a search-engine query; (iv) a single detection prompt combines the message, link evidence, brand-search results, and domain/page context into a decision; (v) a short explanation is generated. We changed no architectural step of this sequence; we did clean known dataset defects before scoring any system against it (Appendix A).

P1 (proposed extension). Two changes are layered on top of P0:

1. Corrected evidence packet. P0's original packaging silently degrades the prompt context when link collection fails (for example, when a URL in an archived message no longer resolves). The corrected packet represents that failure to the model explicitly rather than dropping it, so the detector can distinguish "no evidence found" from "evidence found and it is unremarkable."
2. Added first-party and known-scheme evidence. A retrieval step attempts to surface the claimed organization's own official guidance (e.g. "we do not request payment by text") and to match the message against known scamcampaign fingerprints, layered on top of the corrected packet. An ablation arm evaluates the corrected packet alone, without the added evidence retrieval, in order to credit each change separately. Two earlier draft revisions of the evidence-augmented prompt were built and discarded during development before the reported P1 configuration was reached; for transparency their pairwise comparisons against P0 are reported in Appendix B rather than silently omitted.

Fidelity disclosure. P0 is pinned to SmishX repository commit 116a8c8, reusing the released main.py extraction/decision prompts and requesting the same gpt-4o model identifier for every analytical stage; since that identifier is a moving alias rather than a dated snapshot, the first provider-resolved identity was locked for the run. One infrastructure substitution is disclosed rather than hidden: the released code used Google Custom Search for brand retrieval, and no Google Search credentials were available in this environment, so this reproduction uses a bounded LLMbased search backend for brand, first-party, and knownscheme queries while GPT-4o alone performs every classification decision. This is a faithful reproduction of SmishX's analytical workflow, not an exact reproduction of its published numerical result, and we state it that way throughout rather than claiming byte-for-byte replication.

Project history. This nested, nothing-withheld design (P0's frozen analysis handed unchanged to a P1 branch that only adds evidence) is this project's second generation. An earlier 80-case pilot, with paired blinded auditing by two independent auditors, tested a stricter matched-evidencebudget design and found that first-party evidence improved a source-attribution-validity metric from 81.3% to 92.5% but did not improve decision-support success, because the decision layer abstained too often; a disclosed analyst-blinding deviation was also recorded during that phase. Rather than patch that design, it was retired and rebuilt as the nested design evaluated here, specifically so the packet fix and the evidence addition could be credited separately (Section 6). Scope note. This paper evaluates the offline detection pipeline only. It is the empirical core beneath a separately-scoped consumer-product proposal - an Android notification-layer companion described in an internal business-planning document - that remains unbuilt. Nothing in this evaluation should be read as evidence about that product; Section 8 returns to this distinction explicitly.

Experimental Setup

Data. We evaluate on a 283-message working cohort drawn from a cleaned version of the SmishX-released, relabeled SMS corpus: 34 legitimate, 133 smishing, 116 spam. The positive ("malicious") class is the union of smishing and spam; a spam message classified malicious counts as a correct positive, a framing choice we flag explicitly since it is not the only reasonable one. This 283-message cohort is a subsample of the larger 1,200-message released corpus; the procedure that produced exactly this subsample is not documented in the materials available for this write-up (Appendix A).

Stratification. Link-based evidence collection succeeds or fails depending on whether a URL in an archived message still resolves years later. Across the full 283-case target cohort, cases split into a conforming stratum (n=205: 188 malicious, 17 legitimate) where collection succeeded, and an unreachable-attribution stratum (n=76: 60 malicious, 16 legitimate) where it failed - the normal condition for any archived SMS corpus, since links in old messages stop working. A small eligibility-review remainder (n=2) is too small to interpret on its own. Restricted to the 267-case matched cohort used for every cross-system comparison in this paper (see below), these strata are conforming n=195 (178 malicious, 17 legitimate), unreachable-attribution n=71 (55 malicious, 16 legitimate), and eligibility-review n=1; the matched breakdown is reported in Appendix C. Systems. P0 (baseline), Ablation (corrected packet only), and P1 (corrected packet + added evidence) are the three systems carried through the main text; two discarded intermediate P1 revisions appear only in Appendix B.

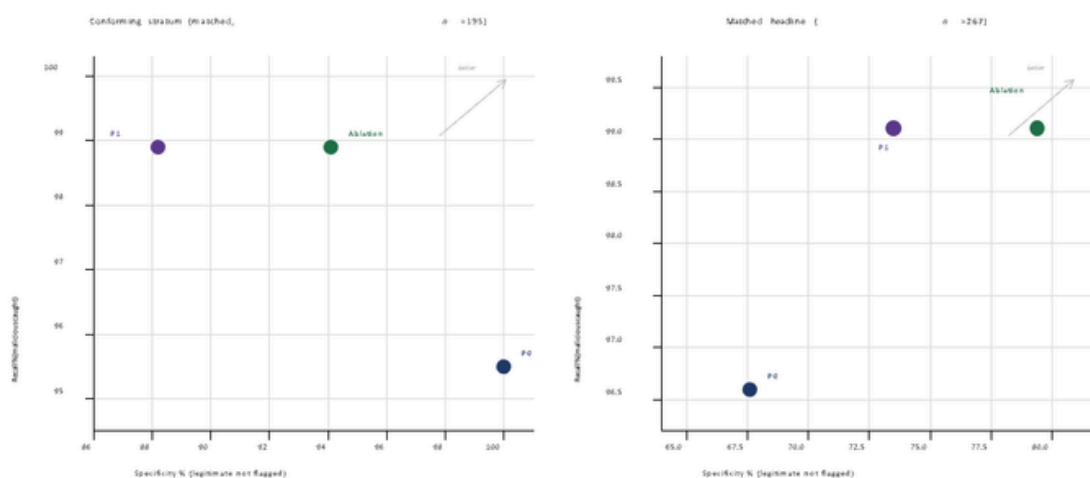
P0 completed all 283 cases; the Ablation run completed 268; the P1 run completed 267 - in both cases because that run stopped on a fixed research budget before scoring every case, and in both cases the cases it missed happened to be ones P0 had answered correctly (Appendix B). Comparing each system's own rates at its own n is valid for descriptive purposes but not for a paired significance test, which requires the same cases on both sides. Every headline and ablation table in this paper (Tables 1, 2, and 4) and every paired test in Appendix C is therefore restricted to the 267 cases scored by all three systems (233 malicious, 34 legitimate); Table 3 and Appendix B separately report each system's own full completed coverage.

Metrics. We report precision, recall, specificity, F1, balanced accuracy (the mean of recall and specificity), and Matthews correlation coefficient (MCC). This cohort is 87.3% positive, so a trivial always-malicious classifier already scores 87.3% accuracy, and across all five systems evaluated end to end on the full cohort, F1 varies by only 2.0 points (96.1–98.1) while specificity varies by 23.5 points (55.9–79.4) - roughly 12× the spread. Accuracy and F1 are therefore treated as weak summaries here; specificity is the most discriminating single metric, and balanced accuracy and MCC serve as the honest one-number summaries.

Significance. Head-to-head claims use McNemar's exact

Metric	P0	P1	Diff
Precision	95.3%	96.2%	+0.9
Recall	96.6%	99.1%	+2.5
Specificity	67.6%	73.5%	+5.9
F1	95.9	97.7	+1.8
Balanced accuracy	82.1%	86.3%	+4.2
MCC	0.668	0.804	+0.136

Table 1: Matched cohort ($n = 267$, the cases scored by all three adopted systems): P1 improves on P0 on all six metrics measured, with none worse. McNemar exact two-sided on this comparison: $p = 0.057$ (narrowly misses conventional significance; Section 6).



Axes are scaled independently (recall spans ~3 points; specificity ~15-25). A system that only trades slides along a diagonal; one that dominates moves up and to the right.

Figure 1: Recall vs. specificity, P0 / Ablation / P1, both panels on the matched cohort. In the conforming stratum ($n=195$), P1 trades specificity for recall relative to P0; on the full matched headline ($n=267$), driven by the evidence-degraded stratum, both extensions dominate P0 outright. two-sided test on paired, same-case decisions. Protocol disclosure. A project protocol document written before this evaluation predeclared macro-averaged accuracy as the primary analysis. The recall/specificity framing presented here as primary was adopted after seeing the results, on the argument that a missed phishing message is costlier than one unnecessary verification prompt. That argument may be reasonable, but presenting it as though it had been the predeclared plan would misrepresent the analysis, so it is disclosed here rather than left for a reader to discover independently.

Main Results

Table 1 is Pareto dominance: whichever metric a reader prefers to headline, P1 wins it. Missed malicious messages fall from 8 to 2 (a 75% reduction in false negatives) and false alarms on legitimate messages fall slightly, from 11 to 9. Read alone, this looks like an unambiguous win. It is not quite one at conventional significance: McNemar's exact test on the matched-cohort ($n=267$) paired comparison returns $p=0.057$, just above the conventional 0.05 threshold. Table 1 and Figure 1 report a real, directionally consistent effect that is close to, but does not clear, that threshold; Section 6 reports where it actually comes from.

Evidence-complete (conforming, $n=195$: 178 malicious, 17 legit.)			
Metric	P0	P1	Diff
Precision	100.0%	98.9%	-1.1
Recall	95.5%	98.9%	+3.4
Specificity	100.0%	88.2%	-11.8
F1	97.7	98.9	+1.2
Balanced accuracy	97.8%	93.6%	-4.2
MCC	0.806	0.871	+0.065
Evidence-degraded (unreachable attrition, $n=71$: 55 malicious, 16 legit.)			
Metric	P0	P1	Diff
Precision	84.6%	90.2%	+5.6
Recall	100.0%	100.0%	tie
Specificity	37.5%	62.5%	+25.0
F1	91.7	94.8	+3.1
Balanced accuracy	68.8%	81.3%	+12.5
MCC	0.563	0.751	+0.188

Table 2: Stratified comparison, matched cohort ($n = 267$ total). When link evidence is complete, P1 trades precision/specificity for recall rather than dominating. When link evidence degrades - 71 of the 267 matched cases (27%) here, and the normal condition for an archived SMS corpus - P1 dominates outright at matched 100% recall.

Recall by class (all cases, $n=283$)	Smishing	Spam
P0	132/133 (99.2%)	109/116 (94.0%)
Ablation	121/121 (100.0%)	111/113 (98.2%)
P1	120/120 (100.0%)	111/113 (98.2%)

Table 3: Both extensions close most of the gap on spam recall, where P0 was weakest; smishing recall was already near ceiling under P0. Uses each system's own completed coverage (283/268/267 cases; Appendix B), not the 267-case matched cohort used in Tables 1 and 2.

Where the effect comes from

Table 2 separates the two conditions the cohort actually contains. When link evidence is complete, P0 already reaches 100.0% specificity and precision - it never flags a legitimate message in this stratum - and P1 raises recall from 95.5% to 98.9% at the cost of specificity and balanced accuracy. This is a trade, not an improvement: P1 wins F1 and MCC here, P0 wins specificity and balanced accuracy. When link evidence degrades, P1 dominates: at identical 100.0% recall, specificity rises 25.0 points and MCC rises from 0.563 to 0.751. Because dead links are the ordinary state of an archived SMS corpus rather than an edge case, the more defensible framing of P1's contribution is robustness to degraded evidence, not a blanket improvement.

Table 3 shows the recall gain is concentrated in spam, where P0 had the most room to improve; smishing recall was already at 99.2% under the baseline.

Audited false alarms

A blinded case-level review examined every case where SmishX's own gold label was legitimate, P0 correctly called it legitimate, and an evidence-augmented P1-line configuration flagged it malicious. Three examples: a KFC feedbacksurvey text was flagged because a tinyurl.com destination could not be proven to be an official KFC channel, even though nothing in the retrieved evidence contradicted an ordinary post-purchase survey; a Zomato message beginning "Thanks for registering" was flagged as unsolicited promotion despite its link resolving to Zomato's own official domain; and a Lybrate notification was flagged because a lybr.at short-link returned a temporary error, even though Lybrate's own published terms describe exactly that notification workflow. Across all five reviewed cases, that configuration introduced the false alarm while correcting four separate errors P0 had made elsewhere. The reviewers identified one recurring reasoning defect: it repeatedly treated not independently verified as equivalent to malicious, overweighting an inaccessible or shortened link even when the rest of the retrieved evidence was consistent with a legitimate workflow. The fix implied by this review - an explicit three-way evidence state (supports/contradicts/unresolved), where only contradicting or affirmatively malicious evidence may flip a P0 legitimate call - had not been incorporated into any P1 configuration evaluated in this paper at the time of writing; we report it here as a specific, named next step rather than folding it into Section 7 as a generic call for more work.

Version caveat. This review's discordant-pair signature - five cases where P0 was right and the reviewed configuration was wrong, against four in the other direction - matches the discarded old_p1 iteration's comparison against P0 exactly (Appendix B), not the adopted P1 configuration reported in Table 1 (which differs from P0 by 3 and 11 cases respectively, Appendix C). The materials available for this paper do not tag these five cases by system iteration directly, so this attribution is inferred from matching discordant-pair counts, not confirmed from a case-ID crossreference. We flag it rather than presenting these examples as certified findings about the exact adopted P1 system evaluated in Table 1. The underlying defect - treating unresolved evidence as though it contradicted the message - is a property of how any evidence-augmented decision layer built this way weighs missing information, so it is reported here as informative about that failure mode even though it may predate the final adopted configuration.

Ablation

Table 4 is the paper's least comfortable and most important result. The ablation - the corrected evidence packet with no added first-party or campaign evidence - outperforms full P1 on five of six metrics, including both honest aggregate summaries (balanced accuracy, MCC). McNemar's exact test confirms this is not noise where it matters: P0 vs. Ablation reaches $p = 0.006$ on the matched cohort (net +10 cases), while Ablation vs. P1 does not ($p = 0.69$, net -2 cases, slightly favoring the ablation). In other words: cor-

Metric (matched, n=267)	P0	Ablation	P1
Precision	95.3%	97.1%	96.2%
Recall	96.6%	99.1%	99.1%
Specificity	67.6%	79.4%	73.5%
F1	95.9	98.1	97.7
Balanced accuracy	82.1%	89.3%	86.3%
MCC	0.668	0.842	0.804

Table 4: The packet-correction-only ablation posts the best balanced accuracy and MCC of any configuration tested - better than the full P1 system. Matched $n = 267$ cohort, as in Table 1.

recting how the pipeline represents collection failures to the model is a real, significant, and comparatively low-risk fix. Adding first-party and known-campaign evidence retrieval - the more interesting hypothesis motivating this project - has a directionally positive but statistically unconfirmed effect on top of that fix (itself now a narrow miss at $p = 0.057$ against P0 directly, Table 1), and by some metrics a slightly negative one. Reporting only the P0-vs-P1 headline would have folded a confirmed result and an unconfirmed one into a single overstated claim.

Limitations

1. The headline comparison narrowly misses significance. P0 vs. P1 does not reach the conventional 0.05 threshold on the matched 267-case cohort ($p = 0.057$); only the packet-correction-alone comparison clears it ($p = 0.006$). This is a closer call than a first draft of this paper reported, which mislabeled a discarded intermediate iteration's comparison ($p = 0.27$) as the headline figure (Appendix D) - but it is still not a confirmed result. Effect sizes and stratified evidence, not a significance claim, are what this data supports for the P0-vs-P1 comparison.
2. Protocol deviation. The recall/specificity-primary framing used throughout this paper was adopted after inspecting results; the project's predeclared protocol specified macro-averaged accuracy as the primary metric. Readers should treat this paper's framing as exploratory rather than confirmatory analysis.
3. Development-evidence status. The corrected evidence packet was designed after inspecting P0's failures on these same 283 cases. This is a development-set result, not a held-out confirmation, and the legitimate class is only 34 cases. A confirmatory claim requires an independently sourced sample, particularly of legitimate messages.
4. The more interesting hypothesis is unconfirmed. Added first-party/known-scheme evidence's marginal effect over the packet fix alone is not significant ($p = 0.69$) and trends slightly negative in raw paired counts (Section 6), and a blinded case review of P1's introduced false alarms found a specific, named reasoning defect rather than random noise (Section 5.2).
5. The predeclared primary outcome was not met by either system. The project's protocol treated evidencesupported correctness - not raw classification accuracy - as primary, assessed by a blinded auditor rating each output's stated reasons as supported, contradicted, or unestablished by the visible evidence packet. Under that audit, both P0 and P1 scored 0 of 283 "decision-support pass" outcomes: every decision was definitive, but every one also contained at least one material statement the visible evidence did not establish. The materials available for this paper do not establish whether this audit was rerun against the exact adopted P1 configuration reported in Section 5, or only against an earlier iteration; either way, none of this paper's accuracy numbers should be read as evidence that either system's stated reasoning is reliably evidence-grounded. (Section 5.2 flags a similar versionattribution uncertainty for the qualitative false-alarm case review.)
6. The headline cohort blends two differently-scoped sampling waves. Of the 283 cases, 64 belong to the originally-scoped "primary" cohort and 219 were added later as an "exploratory extension"; the two are not disclosed separately anywhere in this paper's source reporting layer. They behave differently: in the primary cohort's conforming stratum ($n = 37$, 15 legitimate), false positives increase from P0 to P1 ($0 \rightarrow 2$ of 15), the opposite of the full-cohort headline; in the exploratoryextension conforming stratum ($n = 168$), P0 and P1 both post 0 false positives. The full-cohort false-alarm reduction reported in Table 1 is therefore substantially an exploratory-extension effect, and does not clearly hold in the smaller, more rigorously predeclared primary cohort alone.
7. Dataset provenance. The working cohort inherits known issues from the SmishX release - 14 duplicated message pairs, 2 malformed indicator rows, and a URL-flag count that does not exactly match the source paper's reported figure (Appendix A). Eligibility was determined by a frozen parser and curator confirmation rather than the dataset's own unreliable URL flag, and the sample was predeclared as open-ended rather than fixed at 283; what neither the source reporting layer nor this paper otherwise discloses is that the 283 cases combine two differently-scoped sampling waves, addressed as its own limitation above and in Appendix A.
8. Scope. This evaluates an offline classification pipeline only: no adversarial or prompt-injection evaluation was performed (consistent with the original SmishX authors' own disclosed scope), and no deployed product exists. A separately-proposed consumer-app layer is out of scope for these numbers and should not be inferred from them.
9. Unequal n across arms. P0 completed 283 cases, the ablation 268, and P1 267, because those runs stopped on a fixed research budget before scoring every case. Every cross-system table in this paper (Tables 1, 2, 4, and Appendix C) is restricted to the 267 cases all three systems completed, so that rates and paired significance tests compare the same cases; Table 3 and Appendix B report each system's own uncorrected coverage instead.

Conclusion

The defensible version of this project's contribution is narrower than "P1 beats the SmishX baseline." What the data actually supports is two separate claims at two different confidence levels: correcting how the pipeline represents linkcollection failures to the model is a real, statistically significant improvement, and is the more useful thing this project learned; adding first-party and known-campaign evidence retrieval on top of that fix is a plausible, well-motivated idea whose measured benefit, on the matched $n = 267$ cohort, comes close to but does not clear conventional significance ($p = 0.057$ against P0 directly), and by some metrics trends slightly negative against the ablation alone. Both of those are worth knowing, and neither is well served by collapsing them into one headline number. A confirmatory follow-up would need, at minimum: an independently sourced sample of legitimate messages beyond the 34 used here; a predeclared metric commitment honored at analysis time rather than adjusted after seeing results; adversarial and time-separated test cases rather than an archived, static corpus; a direct test of the evidenceretrieval hypothesis in isolation, since this evaluation cannot currently distinguish "the added evidence does not help" from "this cohort is too small to detect the help it provides"; the blinded evidence-support audit re-run against the exact adopted configuration rather than an earlier iteration; and a single, consistently-scoped cohort rather than a primary sample blended with a later exploratory extension.

It is tempting to describe this as a result a larger, betterfunded run would simply confirm. That is only half right. Additional cases would narrow the statistical uncertainty around the P0-vs-P1 comparison, which already sits close to the conventional threshold ($p = 0.057$) and could plausibly cross it in either direction - but narrowing uncertainty resolves the question in whichever direction is actually true, not in whichever direction is hoped for, and the ablation in Section 6 gives concrete reason to expect that direction is not obviously P1's. More fundamentally, most of this paper's weaknesses are not a scale problem that money fixes by running more of the same evaluation: the packet fix was tuned on the very cases it is scored against, the 283-case cohort blends two differently-scoped sampling waves, and the audited false-alarm pattern (Section 5.2) is a decision-policy defect that needs a redesign, not a larger budget. A funded confirmatory study is worth running, but what it would buy is a trustworthy answer to whether P1's advantage is real - not a guaranteed confirmation that it is. This evaluation is the empirical core underneath a separately scoped consumer-product proposal - an Android notification-layer "evidence-first" scam-message companion, evaluated in an internal business-planning document rather than here. That proposal's own review gates productization on evidence this offline classification study does not by itself provide: whether a cited evidence card changes user decisions more than a bare risk score, what the dangerousmiss rate looks like under real usage, and whether the approach holds up against competitors - Google Messages' native detection and NordVPN's Message Protection among them - that already ship a comparable warning surface without a comparably cited evidence trail. The result reported here is a necessary but not sufficient input to that larger, harder question.

References

1. Federal Trade Commission, "New FTC Data Show Top Text Message Scams of 2024; Overall Losses to Text Scams Hit \$470 Million," Apr. 16, 2025.
2. Canadian Anti-Fraud Centre, "Fraud Prevention Month 2026," reporting more than 112,000 reports and more than CAD \$704 million in reported losses in 2025.
3. Pew Research Center, "Online Scams and Attacks in America Today," Jul. 31, 2025.
4. Google Security Blog, "New AI-Powered Scam Detection Features to Help Protect You on Android," Mar. 4, 2025.
5. NordVPN, "NordVPN expands scam protection with Message Protection," Jul. 8, 2026.
6. Wang et al., "Can You Walk Me Through It? Explainable SMS-Phishing Detection Using LLM-Based Agents," SOUPS 2025, USENIX Association.
7. SmishX supplementary-materials repository, GitHub (yizhu-joy/SmishX), MIT license at time of inspection.
8. Open Source Initiative, MIT License text.
9. SmishX main detection source (extraction, URL enrichment, brand search, detection, and explanation prompts), GitHub.
10. SmishX relabeled SMS dataset, GitHub.

Dataset construction and cleaning notes

The released SmishX dataset contains 1,200 rows matching the source paper’s top-level counts (622 legitimate, 319 spam, 259 smishing). Inspection for this project found 14 duplicated message pairs, 2 malformed indicator rows, and a URL-flag total that does not exactly match the paper’s stated 636 URL-containing messages. These issues do not erase the dataset’s research value but do require cleaning before benchmarking and argue against blind reliance on the published accuracy figure. The repository is MIT-licensed, which generally permits commercial and research reuse with attribution; because the dataset aggregates third-party and personal-archive sources, upstream rights were not independently re-verified for this project and would need separate review before any production use.

Eligibility. The project’s protocol required more than the dataset’s own if_URL column to admit a case: 626 rows carry the exact value if_URL=1 (70 legitimate, 239 smishing, 317 spam), plus 2 further smishing rows with the malformed value “1 1”, but a conservative offline re-scan found only 607 rows with an immediately recognizable URL or domain token - including only 54 legitimate rows. Final eligibility for this paper’s 283-message cohort was therefore determined from the message text itself by a frozen parser and a curator confirmation that the token is genuinely a link (not an email address, a domain-shaped false positive, or a sentence fragment), never from the flag alone. The protocol predeclared an open-ended sample (“analyze every eligible case that survives validation,” not a fixed target size) and explicitly prohibited synthetic duplication to reach a larger n.

Two sampling waves, not disclosed separately in the headline numbers. Of the 283 cases, 64 belong to the project’s originally-scoped primary cohort and 219 were added afterward as an exploratory extension; this paper’s Tables 1–4 report all 283 together, matching the source reporting layer, but the two waves are not equivalent. A re-partition performed within the underlying project (zero additional model calls; it only re-groups already-completed decisions) found:

Primary, confirming	0	1	2
(37, 15)			
Exploratory, confirming	0	0	0
(168, 2)			

In other words, the full-cohort false-alarm reduction this paper headlines (Table 1) is concentrated in the exploratory-extension wave; in the smaller, more rigorously scoped primary wave, P1 introduces more false positives than P0 on legitimate messages, not fewer. This is discussed as a standalone limitation in Section 7 precisely because it is not visible in the combined numbers this paper otherwise reports.

System iteration history

Table 5 reports all five systems evaluated during development against each system’s own full completed coverage, in the order they were built. old_p1 and new_p1 are intermediate revisions of the evidence-augmented prompt, superseded by the reported Ablation (adopted_p0) and P1 (adopted_p1) configurations; they are included here for transparency about the development trajectory rather than presented as competitive alternatives. The main text’s Tables 1, 2, and 4 restrict the P0, Ablation, and P1 rows below to the 267 cases all three completed, rather than using each row’s own n as shown here; see Appendix C for why that restriction matters and Appendix D for a correction this made to an earlier draft.

System	n	TP	FP	FN	TN	Precision	Recall	Specificity	F1	Bal. acc.	MCC
P0 (baseline)	283	241	11	8	23	95.6%	96.8%	67.6%	96.2	82.2%	0.671
old_p1 (discarded)	283	244	15	5	19	94.2%	98.0%	55.9%	96.1	76.9%	0.629
new_p1 (discarded)	283	247	12	2	22	95.4%	99.2%	64.7%	97.2	82.0%	0.746
Ablation (adopted_p0)	268	232	7	2	27	97.1%	99.1%	79.4%	98.1	89.3%	0.842
P1 (adopted_p1)	267	231	9	2	25	96.2%	99.1%	73.5%	97.7	86.3%	0.804

Table 5: Full development trajectory, all cases ($n = 283$ target cohort; some arms stopped short on a fixed research budget - compare rates, not raw counts).

Across all five systems, F1 spans only 2.0 points (96.1–98.1) while specificity spans 23.5 points (55.9–79.4), confirming F1/accuracy as weak discriminators on this cohort.

Stratum	Comparison	n	Only-A correct	Only-B correct	p
Matched cohort (used throughout Tables 1–4)					
All matched	P0 vs. Ablation	267	1	11	0.006
All matched	Ablation vs. P1	267	4	2	0.688
All matched	P0 vs. adopted P1	267	3	11	0.057
Conforming, matched	P0 vs. adopted P1	195	2	6	0.289
Unreachable, matched	P0 vs. adopted P1	71	1	5	0.219
Eligibility review, matched	P0 vs. adopted P1	1	0	0	1.000
Discarded-iteration comparisons (each system's own full coverage; not the adopted-P1 comparison above)					
All	P0 vs. new_p1 (discarded)	283	4	9	0.267
All	P0 vs. old_p1 (discarded)	283	5	4	1.000
Conforming	P0 vs. new_p1 (discarded)	205	3	7	0.344
Conforming	P0 vs. old_p1 (discarded)	205	1	3	0.625
Unreachable attrition	P0 vs. new_p1 (discarded)	76	1	2	1.000
Unreachable attrition	P0 vs. old_p1 (discarded)	76	4	1	0.375

Table 6: “Only-A/Only-B correct” counts cases where exactly one of the paired systems was correct; the matched-cohort discordant counts sum exactly across conforming, unreachable, and eligibility-review (3,11 total), confirming the stratified split. Reproduction warning: the discarded-iteration rows must not be cited as the P0-vs-P1 comparison - an earlier draft of this paper did exactly that (Appendix D). The adopted-P1 comparison is the bolded matched-cohort row: $n = 267$, discordant counts (3,11), $p = 0.057$.

Paired significance tests (McNemar, exact, two-sided)

Reproducibility notes

Correction to an earlier draft. An earlier draft of this paper reported the P0-vs-P1 headline significance test as $p = 0.27$. That value was computed by pairing P0 against new_p1 - a discarded intermediate prompt revision (Appendix B) - rather than against the adopted P1 configuration reported in Table 1; the same mislabeling propagated into the conforming and unreachable-stratum rows of Appendix C. Restricting to the 267-case cohort scored by all three adopted systems and pairing P0 against the correct adopted P1 gives discordant counts of 3 and 11 and $p = 0.057$ (Appendix C) - still short of conventional significance, but a materially different and closer number than previously reported. This draft also restates the headline, stratified, and ablation tables (Tables 1–4) on that same matched 267-case cohort throughout, for a consistent basis of comparison, and flags in Section 5.2 that the qualitative false-alarm case review carries a discordant-pair signature matching the discarded old_p1 iteration rather than the adopted P1 system. None of these corrections changes this paper's overall conclusion - the added-retrieval effect over the packet fix remains short of significance and the packet-only ablation remains the strongest single configuration - but they do change the precise numbers reported, and are disclosed here rather than silently corrected.

The exact prompts and evidence-packet templates for P0, the ablation, and P1 are maintained in the underlying project's working files and are not reproduced verbatim in this paper. Conceptually: P0's prompt is an unmodified port of SmishX's released detection prompt; the ablation's prompt is identical to P0's except for how a failed link-collection attempt is represented in context (an explicit “collection failed” marker in place of a silently degraded field); P1's prompt additionally includes a first-party-evidence field (official-source guidance located for the claimed organization, when found) and a known-scheme field (a match against previously logged campaign fingerprints, when one exists). A future release of the underlying materials should include the literal prompt text, decoding parameters, and model/version identifiers alongside this write-up to make the comparison fully reproducible by a third party.

Version pinning. P0 is pinned to SmishX commit 116a8c8, and the exact main.py, crawler.js, and dataset file versions used are recorded in the project repository rather than reproduced here; a third party reproducing this evaluation should pull those files at that commit rather than a later revision of SmishX, since the repository is under active development.

How the packet fix was chosen. The corrected evidence packet was not chosen by inspection alone. A separate sensitivity study injected artificial collection failures into a small held-out set (18 cases $\times$ 2 arms $\times$ 2 injected conditions) and compared candidate packet designs under identical injected failure: the original packet pooled 65.3% accuracy (95% Wilson 53.8–75.2) against false positives on 25 of 60 injected-legitimate cells, while the design adopted here roughly halved that to 8–12 of 60 across variants before it was applied to the real-data evaluation reported in this paper. This sensitivity study is a design check, not a performance claim, and the adopted configuration's real-data behavior is what Sections 5–6 report.