



ARTICLE

Auditable and Robust LLM-Based Phishing Detection via Provenance-Guided Evidence Contracts

Yue Xu¹ and Xiong Luo^{2,*}

¹Computer Science, Northeastern University, Vancouver, BC, Canada

²Department of Information Technology, Uppsala University, Uppsala, Sweden

*Corresponding Author: Xiong Luo. Email: xiong.luo.se@gmail.com

Received: 08 May 2026; Accepted: 28 July 2026; Published: 22 October 2026

ABSTRACT: Large language model (LLM)-based phishing detectors can produce decisions and explanations, but adversarial webpages violate the assumption that all input text is trustworthy evidence. Hidden comments, metadata, scripts, Cascading Style Sheets (CSS), and retrieved snippets can act as instruction-injection channels, while free-form rationales are hard to audit mechanically. We argue that robust LLM phishing detection should separate interface-attributed invariance from learned adversarial robustness. We introduce an evidence contract that binds model conditioning, output provenance, and deterministic validation to the same admissible evidence scope. A deterministic extractor exposes rendered-visible text, interaction-relevant artifacts, and Uniform Resource Locator (URL) tokens when available, while excluding non-admissible channels. The model emits atomic findings: fixed-schema intent-artifact records with mandatory provenance pointers. Validators check schema, provenance admissibility, value-span matching, and phishing-specific intent-artifact consistency. The formal claim is narrow: a predictor that conditions only on serialized admissible evidence is invariant to transformations that preserve that evidence. Learning is used for the non-guaranteed regime. We propose Intent-Artifact Constrained Regularized Learning (IACRL), which penalizes validator violations and paired-output instability under admissible perturbations. A four-corpus protocol separates interface-attributed invariance from learned empirical stability. Across four corpora, IACRL changes clean area under the receiver operating characteristic curve (AUROC) by -0.004 to $+0.006$ while improving validator pass rates by 0.096 – 0.181 , intent-artifact consistency by 0.073 – 0.151 , and empirical finding stability.

KEYWORDS: Phishing detection; large language models; evidence contract; atomic findings; provenance; deterministic auditing; instruction injection; adversarial robustness; intent-artifact consistency

1 Introduction

Phishing and scam webpages are adversarial evidence environments. Large language models (LLMs) can turn heterogeneous webpage artifacts into decisions and rationales, but raw pages also contain hidden comments, metadata, scripts, Cascading Style Sheets (CSS)-hidden text, and copied snippets that may be irrelevant to user-facing evidence yet visible to the model.

The usual framing of LLM robustness is too coarse for this setting. A phishing detector can fail because it learned a brittle decision boundary, because it followed non-evidence instructions embedded in the page, or because its rationale cites content that cannot be inspected or validated. Treating all three failures as “model robustness” obscures a stronger design principle: invariance to non-admissible content should

be attributed to the evidence interface, while robustness under admissible changes must be learned and measured empirically.

This paper therefore treats LLM phishing detection as an adversarial evidence-use problem. The central object is an *evidence contract*: a deterministic interface that specifies which evidence units the model may condition on, which evidence units outputs may cite, and which claims deterministic validators may check. The contract intentionally trades semantic breadth for deterministic auditability. It does not attempt to verify every possible analyst explanation; it verifies a focused family of phishing indicators whose support can be checked from admissible webpage evidence.

The framework has three components. First, a deterministic admissible-evidence function $E(x)$ extracts rendered-visible text, interaction artifacts such as forms and links, and Uniform Resource Locator (URL) tokens when present, while excluding comments, metadata, scripts/styles, hidden text, and untrusted snippets. Second, the LLM emits *atomic findings*: fixed-schema records with intent, artifact, optional target, and mandatory provenance pointers. Third, validators check schema form, admissible citation, value–span matching, and intent–artifact compatibility.

The key technical distinction is between interface-attributed invariance and learned robustness. If a transformation modifies only the non-admissible complement of $E(x)$, then an implementation that truly conditions only on the serialized admissible view $S(E(x))$ is invariant to that transformation by construction. This is a scoping property of the deployed interface, not evidence that phishing pages themselves are semantically robust. If a transformation modifies admissible evidence itself, no such guarantee applies; in that regime, the proposed constrained learning objective improves empirical supportability and output stability.

The evaluation uses four public webpage corpora: LNU-Phish [1,2], the Zenodo phishing website corpus associated with Van Dooremaal et al. [3], a second artifact-rich Zenodo phishing website dataset [4], and the Mendeley phishing websites dataset [5]. These corpora stress different parts of the contract: URL+HyperText Markup Language (HTML) archives, raw webpage bundles, screenshot/text/CSS/HTML-structure artifacts, and large-scale URL+HTML variability.

The contributions are:

- **Evidence contracts for adversarial webpages.** We define a contract that binds model conditioning, output provenance, and deterministic validation to the same admissible evidence scope.
- **Auditable atomic findings.** We replace free-form rationales with provenance-linked intent–artifact records and validators for schema, provenance, value–span, and intent–artifact support.
- **Robustness attribution.** We separate interface-guaranteed invariance under non-admissible perturbations from learned empirical stability under admissible perturbations.
- **Constrained learning and four-corpus evaluation.** We introduce Intent–Artifact Constrained Regularized Learning (IACRL) and a heterogeneous public-corpus protocol that measures utility, auditability, robustness, transfer, and diagnostic failure modes.

Fig. 1 summarizes the evidence boundary and robustness attribution used throughout the paper.

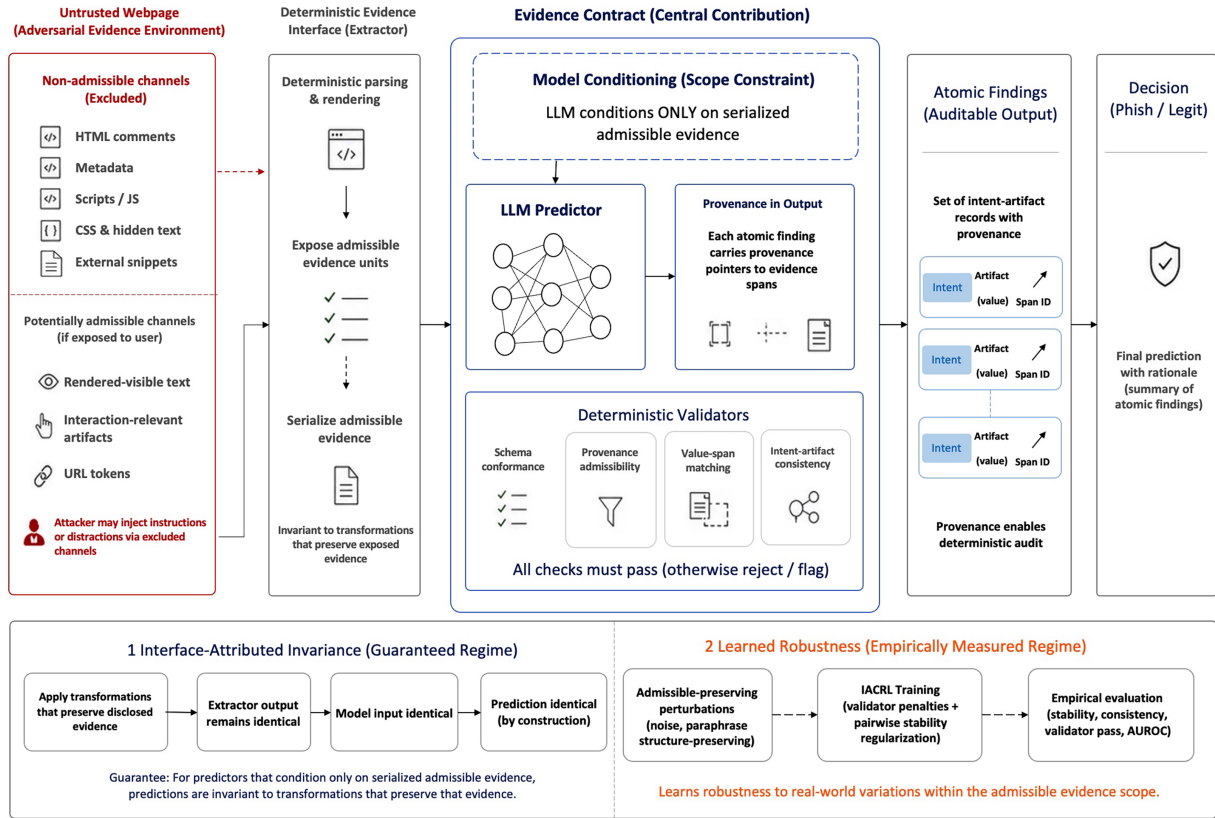


Figure 1: Evidence contract and robustness attribution. Invariance applies only when scoped serialization exposes admissible evidence; admissible-evidence changes remain empirical.

2 Related Work

2.1 Phishing Detection from URLs, HTML, and Page Artifacts

Phishing detection spans URL-only models, handcrafted webpage features, text encoders, markup-aware transformers, and multimodal Document Object Model (DOM)/URL methods. URLNet [6] and URLTran [7] are representative URL baselines; RoBERTa [8], Longformer [9], BigBird [10], and DeBERTa-v3 [11] provide text-encoding references; and MarkupLM [12], PhishSim [13], DOM-graph integration [14], WebPhish [15], dual-layer URL/HTML modeling [16], and hybrid feature-selection/LightGBM detection [17] cover structural or feature-optimized detectors. These systems optimize detection utility but do not specify which webpage evidence an LLM may condition on, cite, or validate deterministically.

2.2 LLM Security and Web-Mediated Prompt Injection

Recent LLM-centered phishing and abuse studies motivate a stronger evidence boundary. KnowPhish combines LLM reasoning and multimodal brand knowledge for reference-based phishing detection [18]. Li et al. analyze an adversarial asymmetry in which semantically plausible LLM-generated phishing pages can become harder to detect [19]. Measurements of web-enabled LLM threats show that tool-using or browser-connected systems can be steered by content encountered on the web [20]. These studies demonstrate the risk of exposing instruction-following models to adversarial web content, but they do not provide a deterministic contract that separates trusted evidence from non-evidence channels inside a webpage.

2.3 Grounded Generation and Provenance

Grounded-generation and provenance-aware systems typically ask whether generated text is supported by retrieved documents. Phishing webpages invert the trust assumption: the source is attacker controlled, and hidden text, comments, scripts, and metadata should not support user-facing phishing indicators. Our contract therefore first defines an admissible evidence boundary, then constrains model input and output citations to that boundary.

2.4 Auditable Explanations for Security Decisions

Security explanations must support downstream action, not merely persuade a reader. A detector with high area under the receiver operating characteristic curve (AUROC) may still be operationally weak if it fabricates an identifier, cites hidden prompt text, or asserts a credential request without a matching form/input artifact. Conversely, a deterministic extractor can be auditable but brittle under page variation. The present work sits between these traditions: it uses learning for detection and empirical stability, but uses deterministic evidence contracts and validators for accountability. The resulting claims are narrower than open-ended semantic explanation, but stronger with respect to mechanical audibility.

3 Evidence Contracts for Adversarial Web Inputs

3.1 Inputs and Evidence Views

Each sample is a webpage instance with a dataset label and public artifacts such as URL, raw HTML, rendered or extracted text, screenshot, CSS, or HTML-structure metadata. Because corpora expose different artifacts, the protocol converts records into a common evidence-unit representation.

For a raw sample x , the deterministic admissible-evidence function $E(x)$ returns an ordered collection of admissible evidence units. Depending on artifact availability, admissible units include rendered-visible text spans with deterministic IDs, interaction-relevant elements such as `form`, `input`, `button`, `a`, download links, contact links, and payment/contact identifiers, URL token spans when URLs are present, and screenshot/optical character recognition (OCR) tokens or released text artifacts when screenshots or text views are available.

Formally, let $\mathcal{U}(x)$ be all evidence units extractable from the released artifact bundle and let $A(x) \subseteq \mathcal{U}(x)$ be the admissible subset. We write

$$E(x) = \{e_i = (\text{id}_i, \text{type}_i, \text{value}_i, \text{span}_i)\}_{i=1}^m, \quad E(x) \equiv A(x). \quad (1)$$

The non-admissible complement is $N(x) = \mathcal{U}(x) \setminus A(x)$. An evidence contract is the tuple

$$\mathcal{C} = (E, S, V), \quad (2)$$

where E extracts admissible evidence, $S(E(x))$ serializes only admissible units for the model, and V is a deterministic validator family for decoded outputs.

The non-admissible complement includes scripts, styles, metadata, comments, hidden text, off-screen text, fully CSS-clipped regions, and attached untrusted snippets. These excluded regions may be useful for other forensic tasks, but they are not valid support for the claims evaluated in this paper.

Evidence-unit IDs are assigned after admissibility filtering. Otherwise, hidden-node insertion could shift later IDs even when admissible content is unchanged. We derive IDs from an admissible-only traversal and stable local signatures; URL tokens and released text/OCR tokens use separate namespaces so validators can distinguish URL, visible text, element, and screenshot-derived citations.

3.2 Deterministic Rendering and Visibility Rules

When raw HTML is available, rendered-visible spans are extracted with a fixed browser build, viewport, font configuration, JavaScript time budget, and network policy. The reported runs used Playwright v1.49 with its bundled Chromium 131.0.6778.33, a 1365×768 viewport, device-scale factor 1.0, default sans/serif/monospace fonts, a 5 s JavaScript budget, and an archive-local network policy that blocks external retrieval after released artifacts are loaded. The DOM is snapshotted after network-idle or timeout. A text node is admissible only when visible by computed style and when its bounding box has nonzero viewport intersection. Nodes under `script`, `style`, `noscript`, `meta`, `comments`, `input type = hidden`, fully clipped regions, and far off-screen elements are excluded.

For datasets that release screenshots or derived text artifacts rather than complete browser-replayable HTML, the contract is applied to the released artifact view. Screenshot/OCR tokens are not treated as interchangeable with DOM spans; they form a separate admissible view with separate IDs. This allows the four-corpus protocol to be honest about heterogeneity: a dataset can support the contract without pretending that every corpus exposes the same raw evidence.

3.3 Threat Families and Scope Boundaries

The threat model separates attacks by whether they preserve $E(x)$. Guaranteed-regime attacks modify hidden comments, metadata, scripts, styles, attached snippets, or fully clipped CSS regions while leaving admissible units unchanged. Empirical-regime attacks modify admissible evidence directly, including visible prompt-injection text, keyword splitting, homoglyph substitutions, DOM fragmentation, form-attribute obfuscation, and CSS camouflage when the transformation changes rendered visibility, OCR tokens, element geometry, or interaction-relevant artifacts.

This split prevents an overclaim. The contract can guarantee invariance only when admissible evidence is preserved. If an attacker changes visible wording or interaction artifacts, the system may still be robust in practice, but that behavior must be measured empirically rather than proven from the interface alone.

Fig. 1 summarizes this boundary: guaranteed-regime perturbations are outside $E(x)$, while empirical-regime perturbations cross the admissible evidence boundary.

3.4 The Contract

The evidence contract binds three rules. The *conditioning rule* states that scoped model input is built only from admissible units $E(x)$ and their stable IDs. The *citation rule* states that every atomic finding must cite admissible evidence IDs. The *validation rule* states that deterministic validators evaluate output supportability only with respect to admissible evidence. Hidden prompt text cannot affect a scoped model unless it changes admissible evidence, and a finding cannot be treated as supported unless it cites admissible evidence that satisfies the relevant validator.

The scoped predictor is therefore written as

$$f_{\theta}^C(x) = h_{\theta}(S(E(x))), \quad (3)$$

not as a function of the full artifact bundle. This notation makes the attribution boundary explicit: changes to $N(x)$ are outside the scoped input, while changes to $A(x)$ may affect both the decision and the findings.

The implementation must preserve this boundary end to end. In the reported guaranteed-regime path, tokenization occurs only after $S(E(x))$ is constructed; raw HTML, hidden nodes, CSS, and detached snippets never enter the model context. An interface-leakage diagnostic replaced non-admissible channels with high-entropy canaries and adversarial instructions before serialization. Exact-match and hash checks found zero

canaries in serialized prompts, and scoped deterministic decoding was unchanged on all guaranteed-regime pairs. This documents the evaluated implementation, not every future integration.

3.5 Interface-Level Invariance

Let $E(x)$ be deterministic and let a scoped predictor satisfy $f(x) = g(E(x))$. Let $\mathcal{T}_G$ denote transformations that modify only non-admissible content, so that $E(T(x)) = E(x)$.

Proposition 1 (Scoped-interface identity): *For any $T \in \mathcal{T}_G$ and any input x , if $f(x) = g(E(x))$, then $f(T(x)) = f(x)$.*

Proof: Since $T \in \mathcal{T}_G$, $E(T(x)) = E(x)$. Therefore $f(T(x)) = g(E(T(x))) = g(E(x)) = f(x)$. $\square$

This identity is intentionally simple: it is a contract-compliance check rather than a discovery about all real phishing pages. Its role is to state what must be true of the implemented scoped interface before any robustness attribution is meaningful. It holds only when the implemented pipeline enforces scoped serialization before tokenization and prevents raw-page side channels from reaching the model. It does not cover attacks that modify visible or otherwise admissible evidence. Those attacks are evaluated empirically. The paper therefore reports two regimes: *guaranteed* perturbations that preserve $E(x)$ and *empirical* perturbations that modify admissible evidence under semantic-preservation guards.

4 Atomic Findings and Deterministic Validation

4.1 Schema

An atomic finding is a fixed-slot record:

$$r = (\text{intent}, \text{artifact}, \text{target?}, \text{provenance}). \quad (4)$$

The intent slot is drawn from a finite set such as `credential_request`, `payment_request`, `contact_request`, `install_request`, and `personal_info_request`. The artifact slot identifies the concrete page mechanism, such as `form_password`, `form_payment`, `mailto`, `tel`, `external_link`, or `download_link`. The optional `target?` slot records an impersonation target or identifier only when it is derivable from admissible evidence. The `provenance` slot is mandatory.

The schema is intentionally smaller than analyst-written phishing explanations. It captures indicators that can be grounded in local webpage evidence and checked with deterministic rules, avoiding a conflation of persuasive prose with verifiable evidence.

We selected the ontology from phishing evidence patterns that can be locally supported by released webpage artifacts: credential entry, payment request, contact solicitation, software installation/download, and personal-information collection. Table 1 reports the primary weak-target intent distribution after canonicalization. The “outside ontology” row counts phish-labeled records for which the extractor found local suspicious cues but none mapped to the five evaluated intents; these cases are excluded from validator-positive target generation and analyzed as missed-cue risk rather than forced into the schema.

To understand the residual outside-ontology bucket, we manually grouped the excluded phish-labeled records that had suspicious local cues but no supported mapping to the five intent classes. The largest groups were brand-only impersonation without a local requested action (31.4%), redirect or interstitial-only pages (24.7%), script- or login-gated content outside the fixed rendering budget (18.9%), social-engineering-only text without a structured requested action (15.2%), and insufficient or ambiguous local support (9.8%). Thus, most outside-ontology cases are pages where the suspicious behavior is contextual, deferred, or not locally expressible as an auditable request.

Table 1: Intent ontology coverage in weak target generation, computed over phishing-labeled records with at least one local extractor cue.

Intent or Bucket	LNU	Z492	Z804	M80K	Aggregate
credential_request	54.8%	49.6%	51.2%	57.9%	54.4%
payment_request	7.1%	5.8%	8.4%	6.5%	6.9%
contact_request	13.4%	18.7%	14.1%	11.8%	13.9%
install_request	5.6%	4.9%	7.8%	5.2%	5.8%
personal_info_request	13.9%	14.8%	12.1%	12.7%	13.2%
Outside ontology/insufficient local support	5.2%	6.2%	6.4%	5.9%	5.8%

Two formatting conventions reduce spurious variance. First, findings are canonicalized by intent, artifact, normalized target, and provenance IDs before set-level comparison. Second, extractable values such as email addresses, phone numbers, and payment identifiers are normalized before value-span matching. Canonicalization is not used to make decoding easier; it is used only for training targets, set-distance metrics, and deterministic evaluation.

4.2 Validators

Validators map decoded outputs and admissible evidence to pass/fail flags for schema conformance, provenance admissibility, intent-artifact consistency, and value-span matching. To prevent empty outputs from passing vacuously, validator pass (ValPass) is conditioned on coverage; coverage (Cov) and auditable yield (AY) retain empty outputs as failures.

Schema conformance checks that every output parses into the required fields and uses only ontology members. Provenance admissibility checks that each citation resolves to an evidence unit in $E(x)$. Intent-artifact consistency encodes local phishing support rules; for example, `credential_request` should cite a credential-related form/input artifact, while `contact_request` should cite a mailto, telephone, contact form, or equivalent interaction artifact. Value-span matching applies when the model asserts an identifier: the normalized value must match the cited admissible text or artifact attribute.

These validators deliberately avoid a learned judge. A learned judge may be useful for broader semantic review, but it would blur the paper's central claim because robustness and auditability would then depend on another model's behavior. The deterministic validators define exactly what is being measured.

Let $R(x)$ be the set of decoded findings. For each finding $r \in R(x)$, the validator family is

$$V(r, E(x)) = (V_{\text{schema}}, V_{\text{prov}}, V_{\text{IA}}, V_{\text{span}}). \quad (5)$$

For the N test records used in auditability reporting, define

$$C_i = \mathbb{1}[\hat{d}(x_i) = \text{phish} \wedge |R(x_i)| \geq 1], \quad (6)$$

$$P_i = \mathbb{1}\left[C_i = 1 \wedge \forall r \in R(x_i), \bigwedge_{v \in V} v(r, E(x_i)) = 1\right]. \quad (7)$$

The aggregate metrics are

$$\text{Cov} = \frac{1}{N} \sum_i C_i, \quad \text{ValPass} = \frac{\sum_i P_i}{\sum_i C_i}, \quad \text{AY} = \frac{1}{N} \sum_i P_i = \text{Cov} \times \text{ValPass}. \quad (8)$$

Thus AY is the joint covered-and-valid proportion, not a product-of-marginals approximation. Empty outputs contribute zero to Cov and AY and do not enter the conditional ValPass denominator. Intent-artifact consistency is averaged over emitted findings as $IA(x) = |R(x)|^{-1} \sum_{r \in R(x)} V_{IA}(r, E(x))$. For paired robustness, finding-set drift is computed after canonicalization:

$$d_R(R, R') = 1 - \frac{|\text{canon}(R) \cap \text{canon}(R')|}{|\text{canon}(R) \cup \text{canon}(R')|}. \quad (9)$$

Fig. 2 illustrates the scoped inference and validation path from raw webpage artifacts to validator-checkable atomic findings and a final decision.

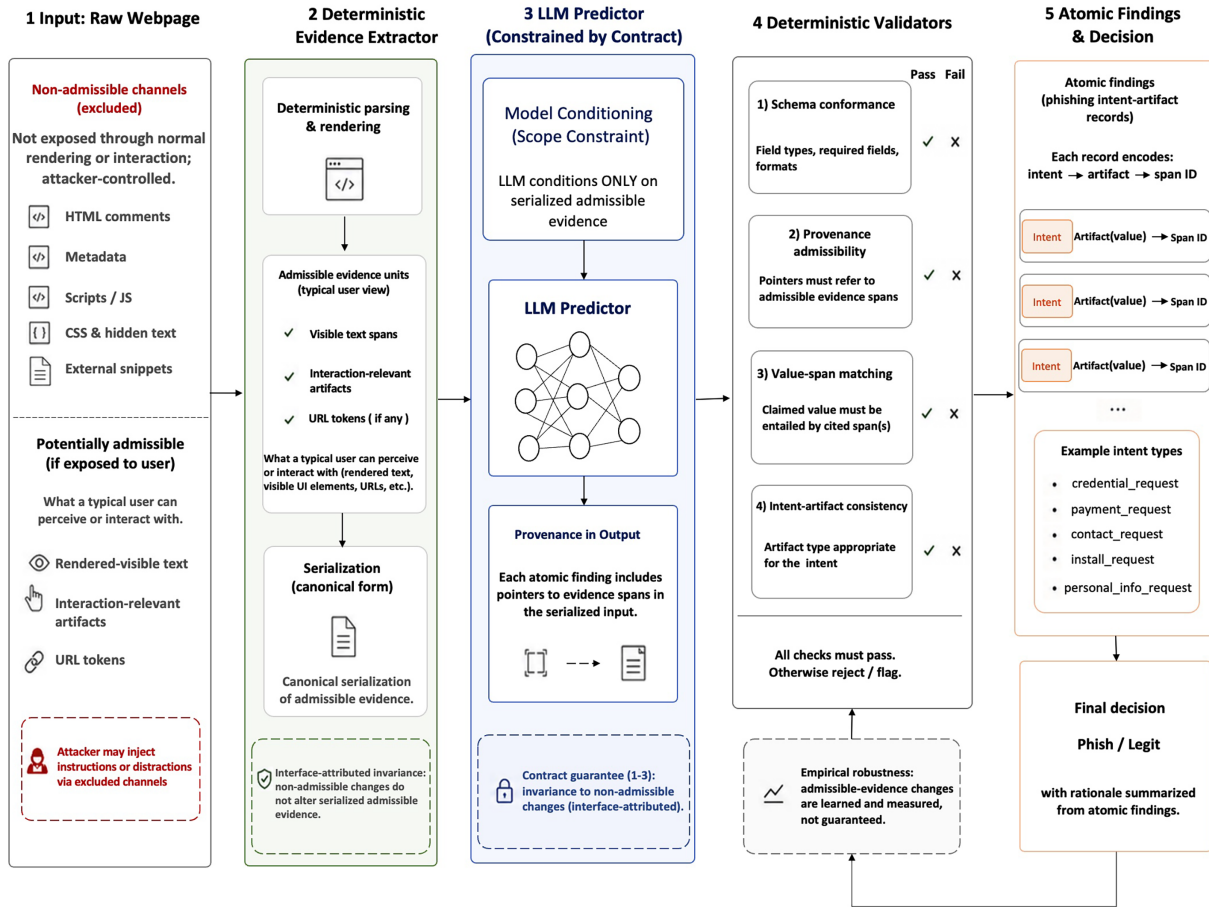


Figure 2: Scoped evidence-contract inference and validation path from raw artifacts to validator-checkable findings and decision.

5 Intent-Artifact Constrained Regularized Learning

5.1 Output and Detection Score

The model emits a structured sequence with a decision field and a findings array. The decision score is the logit difference between two verbalizer tokens, e.g., phish and benign, at the decision position. This keeps detection and explanation in a single text-native interface without adding a separate classification head.

The finding loss is token-level cross-entropy on canonicalized target findings. Canonicalization sorts findings by intent, artifact, and provenance IDs, and normalizes extractable values. Decoding remains unconstrained beyond the fixed schema.

The input prompt contains only admissible evidence units, their short type labels, and stable IDs. It does not contain hidden comments, scripts, metadata, or untrusted attached snippets in the scoped condition. For unscoped diagnostic variants, the same output schema is used while the input includes full HTML-like context. This matched-schema contrast isolates the effect of evidence scoping without confusing it with the difficulty of parsing free-form rationales.

Joint output is important for deployment. A standalone classifier can decide that a page is suspicious but leave no checkable path for review. A post-hoc explanation module can produce plausible indicators that do not correspond to the classifier's decision. The sequence format forces the model to emit a decision and a set of candidate indicators under one contract, while validators determine whether those indicators are supportable.

5.2 Objective

IACRL fine-tunes an instruction-tuned LLM to jointly output a decision and atomic findings:

$$\min_{\theta} \mathcal{L}_{\text{det}}(\theta) + \mathcal{L}_{\text{find}}(\theta) + \lambda_{\text{IA}} \mathbb{E}[C_{\text{IA}}(\hat{R}, x)] + \lambda_{\text{INV}} \mathbb{E}[C_{\text{INV}}(\theta; x)]. \quad (10)$$

$\mathcal{L}_{\text{det}}$ is binary cross-entropy on dataset labels. $\mathcal{L}_{\text{find}}$ is teacher-forced sequence loss on weakly generated targets. C_{IA} penalizes schema, provenance, value-span, and intent-artifact validator failures. C_{INV} penalizes output discrepancy on paired inputs:

$$C_{\text{INV}} = \mathbb{I}[\hat{d}(x) \neq \hat{d}(T(x))] + (1 - \text{Jaccard}(\hat{R}(x), \hat{R}(T(x)))) . \quad (11)$$

In fully scoped deterministic guaranteed-regime pairs, this term is zero by interface design; in practice, it regularizes mixed-interface settings and stochastic output drift.

Weak finding targets are generated by deterministic extractors that identify admissible forms, inputs, contact/payment identifiers, download links, and URL/text spans, then map them to the intent ontology in Table 1. These targets are used only as scalable supervision for supportable local indicators; they are not treated as complete analyst labels. The loss weights were selected on validation splits by a small grid that prioritized validator pass (ValPass) and empirical finding stability subject to a maximum clean AUROC drop of 0.005 relative to the schema-only scoped model. The selected setting was $\lambda_{\text{IA}} = 0.15$ and $\lambda_{\text{INV}} = 0.10$ for all corpora. Larger λ_{IA} values increased ValPass but reduced coverage, and larger λ_{INV} values over-smoothed finding sets under admissible perturbations. Because this selection criterion includes AUROC and coverage constraints, the validator penalty is not optimized in isolation.

5.3 Optimization

Because validator costs are defined on discrete decoded outputs, penalty expectations are optimized with a self-critical sequence-level estimator. All runs use $K = 4$ samples, an exponential moving-average baseline with decay 0.9, gradient clipping 1.0, Quantized Low-Rank Adaptation (QLoRA) rank $r = 32$, $\alpha = 64$, dropout 0.05, AdamW learning rate 2×10^{-4} , cosine scheduling, warmup 3%, and seeds $\{13, 17, 23, 42, 57\}$.

For each input, sampled outputs are compared with a greedy baseline and the resulting advantages are normalized and clipped within the mini-batch. Parse failures receive maximal validator cost. For invariance training, a paired transformation $T(x)$ is sampled, both members of the pair are decoded with

matched sampling settings, and the loss penalizes decision disagreement plus canonicalized finding-set distance. This makes the regularizer sensitive to semantically meaningful output drift rather than superficial record ordering.

In the fully scoped deterministic guaranteed regime, C_{INV} should be zero because paired inputs serialize to the same admissible evidence. Its practical value is therefore not to prove the guarantee; the contract already does that. Its value is to reduce drift in stochastic decoding, mixed-view settings, and empirical perturbations where admissible evidence is changed.

Algorithm 1 gives the inference-time contract path used by all scoped evaluations.

Algorithm 1: Evidence-contract inference and validation

Require: Webpage artifact bundle x ; deterministic extractor E ; model p_θ ; validator set $\mathcal{V}$

- 1: Extract admissible evidence units $E(x)$ and stable evidence IDs
 - 2: Serialize only admissible units into the scoped model input
 - 3: Decode structured output $\hat{o} = (\hat{d}, \hat{R})$ with decision $\hat{d}$ and findings $\hat{R}$
 - 4: **for** each finding $\hat{r} \in \hat{R}$ **do**
 - 5: Check schema conformance, provenance admissibility, value–span match, and intent–artifact consistency
 - 6: **end for**
 - 7: Return decision, canonicalized findings, validator report, and supportability metrics
-

Table 2 separates the training-only signals from the scoped inference path, complementing the conceptual contract in Fig. 1.

Table 2: Training and inference roles of intent–artifact constrained regularized learning (IACRL) components.

Component	Training Role	Inference Role	Reviewer-Facing Clarification
Evidence extractor $E(x)$	Builds admissible evidence and weak finding targets from local forms, links, identifiers, and URL/text spans	Builds admissible evidence for the test page	Deterministic; hidden comments, metadata, scripts, and non-admissible snippets are removed before tokenization
Scoped serialization $S(E(x))$	Constructs the scoped prompt with evidence IDs	Constructs the scoped prompt with evidence IDs	The model sees only admissible evidence in both training and inference
LLM decoder	Produces a decision and atomic findings under teacher forcing or sampling	Produces a decision and atomic findings	The output schema is the same in training and inference
$\mathcal{L}_{\text{det}}$	Supervised by dataset labels	Not used	Preserves phishing-detection utility
$\mathcal{L}_{\text{find}}$	Supervised by canonicalized weak finding targets	Not used	Weak targets are scalable extractor-derived supervision, not complete analyst labels
$\lambda_{\text{IA}} C_{\text{IA}}$	Penalizes schema, provenance, value–span, and intent–artifact failures	Not used as a loss; validator still reports failures	Validator pass is not optimized alone; selection is constrained by AUROC and coverage

(Continued)

Table 2 (continued)

Component	Training Role	Inference Role	Reviewer-Facing Clarification
$\lambda_{INV}C_{INV}$	Compares outputs on x and $T(x)$ for decision and finding-set drift	Not used	Stabilizes outputs under paired perturbations without defining the interface-level guarantee
Deterministic validator V	Supplies validator penalties and evaluation metrics	Returns the validator report	Auditability is deterministic local support checking, not a learned judge

6 Four-Corpus Evaluation Protocol

6.1 Corpus-to-Contract Design

The evaluation is organized around the evidence contract rather than a single raw webpage format. Table 3a,b records artifacts, admissible views, excluded channels, split accounting, and supported evidence-dependent evaluations. For Z804, the official release has 10,395 websites; the retained unit count is larger because multiple artifact views are expanded after site/template-level grouping.

Table 3: Four-corpus evidence-contract matrix. (a): released artifacts and admissible views. (b): Excluded channels, usable splits, supported evaluations, and contract stress.

(a)				
Corpus	Source	Released Artifacts	Admissible View	
LNU	LNU-Phish [1,2]	URL, HTML; screenshots optional	URL, visible text, forms/links	
Z492	Zenodo 4922598 [3]	HTML bundles, screenshots; no URLs	visible text, forms/links, optional OCR	
Z804	Zenodo 8041387 [4]	URL, HTML snapshots, screenshots, text, CSS, structure	text, structure, screenshot units	
M80K	Mendeley n96ncsr5g4 [5]	URLs, relevant HTML pages	URL, visible text, forms/links	
(b)				
Corpus	Excluded Channels	Usable/Split	Supported Eval.	Contract Stress
LNU	scripts, metadata, comments, hidden text	23,364; 16,284/2364/4716	clean, G, E, transfer	URL+HTML scoping
Z492	scripts, metadata, comments, hidden text	11,204; 7790/1133/2281	clean, G, E, transfer	no-URL robustness
Z804	non-visible/unsupported artifacts	10,395 sites/18,452 views; 12,916/1845/3691 views	clean, G, E, transfer	artifact-rich release
M80K	scripts, metadata, comments, hidden text	78,102; 54,671/7810/15,621	clean, G, E, transfer	large-scale URL+HTML

All preprocessing is deterministic and scripted. Exact duplicates use URL normalization and raw-artifact hashes where available; near duplicates use rendered-text fingerprints and DOM/artifact signatures; and train/validation/test assignment is performed at template-family level with a 70/10/20 target split. For Z804, site/domain/template grouping precedes artifact-view expansion, so views of the same website cannot cross splits.

6.2 Duplicate, Template, and Family Leakage Controls

Phishing corpora often contain cloned kits, mirrored benign templates, and repeated landing pages. To reduce leakage before splitting, exact duplicates are removed using normalized URLs, raw-artifact hashes, and released file identifiers where available; near duplicates are grouped with rendered-text fingerprints and DOM/form/link signatures; and split assignment is performed at the family level.

Empirical attack candidates are also filtered by label-preservation guards. For a candidate pair $(x, T(x))$, the guard requires unchanged dataset label metadata, preserved presence/absence of core credential/payment/contact/download primitives, and no insertion or removal of core interaction elements outside the configured perturbation scope. This guard does not prove semantic equivalence; it defines a deterministic minimum standard for scoring admissible-evidence perturbations.

6.3 Attack Regimes and Measurement Grid

[Table 4](#) defines the attack grid and the accounting reported for each perturbation family. This table is designed to prevent a common ambiguity: a robustness number is only meaningful if the paper states whether the perturbation preserves admissible evidence, which records were eligible, and which metric attributes the observed behavior.

Table 4: Perturbation and measurement grid. Eligibility is the fraction of test records with at least one guard-passing transformed candidate.

Family	Regime	Changes $E(x)$?	Severity Schedule	Primary Metrics	LNU elig.	Z492 elig.	Z804 elig.	M80K elig.
hidden-text-instr	G	no	single	Chg, FChg, R-Val	100.0%	100.0%	100.0%	100.0%
meta-instr	G	no	single	Chg, FChg, R-Val	100.0%	100.0%	100.0%	100.0%
untrusted-snippet	G	no	single	Chg, FChg, R-Val	100.0%	100.0%	100.0%	100.0%
css-clipped-hidden	G	no	single	Chg, FChg, R-Val	100.0%	100.0%	100.0%	100.0%
visible-text-instr	E	yes	8/16/24 tokens	R-AUROC, R-Val, Chg	89.1%	82.4%	91.3%	93.5%
form-attr-obfus	E	yes	15/30/45% attrs	R-AUROC, R-Val, FChg	84.7%	68.8%	87.5%	90.2%
keyword-split	E	yes	10/20/30% tokens	R-AUROC, R-Val, FChg	92.3%	78.9%	93.0%	94.6%
homoglyph	E	yes	10/20/30% tokens	R-AUROC, R-Val, FChg	90.4%	81.5%	91.8%	93.1%
dom-fragment	E	yes	10/20/30% nodes	R-AUROC, R-Val, FChg	86.0%	72.1%	88.4%	91.0%
css-camouflage-visible	E	yes	10/20/30% elems	R-AUROC, R-Val, FChg	83.7%	65.9%	85.6%	89.4%

6.4 Baselines, Metrics, and Implementation

We compare IACRL with schema-only scoped and schema-compatible unscoped LLMs, deterministic extractor-only outputs, and detection-only baselines grouped by evidence family. DeBERTa-v3 is included as a recent encoder baseline, but the encoders are not parameter- or pretraining-matched to Qwen2.5-7B-Instruct and serve as detection context rather than model-family superiority claims. Auditability metrics are reported only for methods that emit atomic findings.

Detection utility is reported with area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (PR-AUC), false positive rate at true positive rate 95% ($\text{FPR@TPR} = 95\%$), true positive rate at false positive rate 1% ($\text{TPR@FPR} = 1\%$), expected calibration error (ECE), and Brier score. Auditability is reported with joint validator pass (ValPass), intent–artifact pass (IA), coverage (Cov), categorical validator failures, and finding-set stability. Robustness is reported with decision change, finding-set change, robust AUROC, robust PR-AUC, and robust ValPass under guaranteed and empirical regimes.

All runs use Qwen2.5-7B-Instruct with Quantized Low-Rank Adaptation (QLoRA) rank $r = 32$, $\alpha = 64$, dropout 0.05, AdamW with learning rate 2×10^{-4} , weight decay 0.01, cosine scheduling, 3% warmup, gradient clipping 1.0, and seeds {13, 17, 23, 42, 57}. Input/output lengths are 4096/384 for scoped evidence views and 8192/384 for unscoped full-context diagnostic variants. Guaranteed-regime change-rate measurements use deterministic decoding. On a single NVIDIA A100 80 GB graphics processing unit (GPU), scoped deterministic inference averaged 1.38 s per page at batch size 8, with 95th-percentile latency of 2.61 s; unscoped 8192-token diagnostics averaged 2.44 s per page. The deterministic extractor averaged 38 ms per HTML page and 91 ms when screenshot/OCR tokens were available. These measurements are intended as deployment-cost context, not as optimized serving results.

6.5 Statistical Testing and Reporting

Method comparisons are paired by split, seed, and test instance whenever possible. We report seed-level mean $\pm$ standard deviation and paired bootstrap confidence intervals for effects such as ΔAUROC , $\Delta\text{ValPass}$, and ΔFChg . Thresholds are selected only on validation splits. Cross-corpus experiments tune on the source validation split or held-out validation mixture, then evaluate once on the target test corpus.

6.6 Human Audit and Error Analysis Protocol

To reduce rule-closure risk, we add a manual audit not used for training, validation, or threshold selection. From each corpus test split, we sample 120 records stratified by label, validator outcome, and empirical-attack decision change. Reviewers see the released artifact view, model decision, and cited finding, but not the model variant, validator flag, or aggregate metrics. Two reviewers inspect each packet, and disagreements are adjudicated by a third reviewer. [Table 5](#) defines the audit fields.

The audit reports human-supported finding precision, validator–human agreement, major missed-cue rate, qualitative error categories, and inter-reviewer agreement before adjudication. It does not convert the full benchmark into analyst-labeled ground truth. Its purpose is narrower: to test whether validator gains correspond to human-inspectable support and to surface failure cases that deterministic metrics compress away.

Table 5: Manual audit rubric and reported uses.

Audit Field	Reviewer Judgment/Operational Definition	Reported Use
Finding support	Supported/partial/unsupported/insufficient evidence; cited artifact locally supports the intent–artifact claim	Human-supported precision
Citation admissibility	Admissible/non-admissible/unresolved; evidence ID points to a visible or interaction-relevant released artifact	Validator–human agreement
Missed local cue	None/minor/major; salient credential, payment, contact, download, or brand cue absent from findings	Major missed-cue rate
Adjudication	Agree/adjudicated; third reviewer resolves support and missed-cue disagreements	Inter-reviewer agreement

7 Results

7.1 Finding 1: Supportability Improves without Utility Collapse

Across the four corpora, IACRL changes AUROC by -0.004 to $+0.006$ relative to schema-only scoped LLM, while improving ValPass by $+0.096$ to $+0.181$ and IA by $+0.073$ to $+0.151$. Table 6 shows that supportability improves substantially without systematic clean-utility collapse.

Table 7 summarizes a corpus-level diagnostic for the utility/auditability trade-off. With only four corpora, this should be read as descriptive analysis rather than a reliable regression. The two AUROC regressions occur on Z492 and M80K, where either URLs are absent (Z492) or many records have sparse short cues (M80K). Z804, the richest artifact release, shows the largest empirical decision-change reduction. This pattern suggests that IACRL helps most when the admissible evidence view contains enough artifact diversity for the validator penalty to improve support without suppressing useful cues.

Fig. 3 visualizes detection utility against AY, the joint proportion of records that are covered and pass every deterministic support check. Because ValPass is conditional on coverage, $AY = Cov \times ValPass$ exactly. IACRL increases AY by 0.065 – 0.145 across the four corpora while changing AUROC by only -0.004 to $+0.006$.

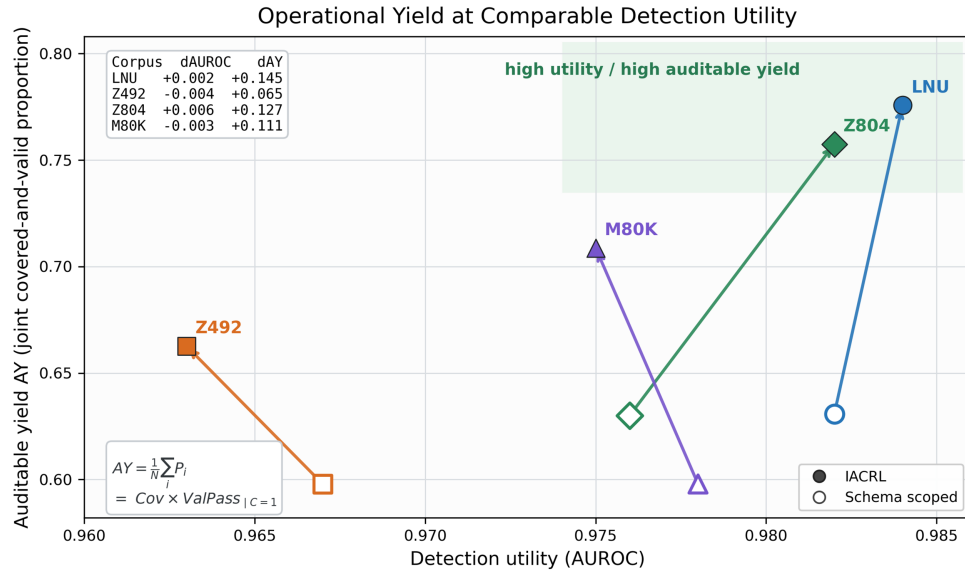
The delta rows make the intended claim explicit: IACRL is not primarily an AUROC optimizer; it is a mechanism for producing more supportable phishing indicators at broadly comparable utility. Coverage is mostly stable, but the M80K drop shows why ValPass should be interpreted jointly with coverage rather than as a standalone success metric. A focused decomposition of the M80K $\Delta Cov = -0.027$ attributes 0.019 to suppressed unsupported findings, 0.006 to no supportable finding after canonicalization, and 0.002 to parse/schema failures, so the drop mainly reflects conservative suppression rather than pipeline breakage. Likewise, the lower Z492 empirical attack coverage is driven by artifact eligibility: text/homoglyph guards retain about 82% of candidates, form/DOM guards about 70%, visible CSS guards 65.9%, and URL-oriented perturbations are unavailable because Z492 has no URLs.

Table 6: Main within-corpus results. ValPass is conditional on Cov = 1; IA and Cov apply only to atomic-finding methods. Detection-only baselines are reported separately below.

Corpus	Method	AUROC	PR-AUC	FPR@95 (%) ↓	TPR@1 (%) ↑	ECE (%) ↓	Brier ↓	ValPass	IA	Cov
LNU	IACRL scoped	0.984 ± 0.005	0.980 ± 0.006	5.8 ± 0.9	78.6 ± 2.4	1.50 ± 0.24	0.051 ± 0.004	0.901 ± 0.014	0.935	0.861
LNU	Schema scoped	0.982 ± 0.006	0.978 ± 0.007	6.9 ± 1.0	73.4 ± 3.1	1.54 ± 0.21	0.053 ± 0.004	0.720 ± 0.026	0.784	0.876
LNU	IACRL-Schema Δ	+0.002	+0.002	-1.1	+5.2	-0.04	-0.002	+0.181	+0.151	-0.015
Z492	IACRL scoped	0.963 ± 0.008	0.968 ± 0.009	8.9 ± 1.4	67.2 ± 3.5	2.05 ± 0.30	0.070 ± 0.006	0.811 ± 0.024	0.874	0.817
Z492	Schema scoped	0.967 ± 0.006	0.970 ± 0.008	8.2 ± 1.2	70.4 ± 3.0	1.74 ± 0.26	0.064 ± 0.006	0.715 ± 0.030	0.801	0.836
Z492	IACRL-Schema Δ	-0.004	-0.002	+0.7	-3.2	+0.31	+0.006	+0.096	+0.073	-0.019
Z804	IACRL scoped	0.982 ± 0.004	0.981 ± 0.005	4.7 ± 0.8	82.1 ± 2.3	1.43 ± 0.22	0.050 ± 0.004	0.887 ± 0.015	0.925	0.854
Z804	Schema scoped	0.976 ± 0.006	0.976 ± 0.007	7.1 ± 1.1	72.5 ± 2.8	1.50 ± 0.24	0.054 ± 0.004	0.743 ± 0.022	0.807	0.848
Z804	IACRL-Schema Δ	+0.006	+0.005	-2.4	+9.6	-0.07	-0.004	+0.144	+0.118	+0.006
M80K	IACRL scoped	0.975 ± 0.006	0.974 ± 0.006	6.8 ± 1.1	73.6 ± 3.0	1.72 ± 0.25	0.056 ± 0.005	0.858 ± 0.018	0.904	0.826
M80K	Schema scoped	0.978 ± 0.005	0.973 ± 0.005	7.8 ± 1.0	69.5 ± 2.7	1.54 ± 0.23	0.057 ± 0.005	0.701 ± 0.028	0.776	0.853
M80K	IACRL-Schema Δ	-0.003	+0.001	-1.0	+4.1	+0.18	-0.001	+0.157	+0.128	-0.027

Table 7: Corpus-level effect diagnostic for IACRL. Deltas are relative to the schema-only scoped model.

Corpus	URL Available	Main Artifact Stress	Δ AUROC	Empirical Chg Reduction/ Δ ValPass
LNU	yes	URL+HTML forms/links	+0.002	0.062/+0.181
Z492	no	no-URL webpage bundles	−0.004	0.051/+0.096
Z804	yes	text, screenshot, CSS, structure	+0.006	0.092/+0.144
M80K	yes	large-scale sparse URL+HTML	−0.003	0.063/+0.157

**Figure 3:** Detection utility vs. auditable yield (AY), the joint covered-and-valid proportion. Values use the conditional ValPass and Cov in Table 6.

7.2 Finding 2: The Contract Operates in a High-Utility Detection Regime

Table 8 gives detection-only context for methods that do not necessarily emit atomic findings. The table is grouped by evidence family so that URL-only, markup-aware, URL+HTML, and scoped-evidence methods are not misread as consuming the same inputs. The table shows that the audibility gains in Finding 1 occur in a high-utility region relative to strong text and markup baselines.

The baseline table supports a narrow conclusion: the audibility gains occur in a competitive detection regime, without treating all modalities as directly comparable or assigning audibility metrics to methods that cannot emit validator-checkable findings.

7.3 Finding 3: Guaranteed-Regime Robustness Is Interface-Attributed

Table 9 separates interface-attributed invariance from learned empirical stability. In the guaranteed regime, scoped methods show zero decision and finding-set change because admissible evidence is unchanged; the schema-compatible unscoped variant remains sensitive to non-admissible content. In the empirical regime, admissible evidence changes, and IACRL reduces decision/finding volatility while retaining higher robust ValPass.

Table 8: Detection context by evidence family. “Atomic” indicates validator-checkable findings; dashes mark unsupported evidence views.

Family	Method	Evidence Input	Atomic	LNU AUROC	Z492 AUROC	Z804 AUROC	M80K AUROC
Scoped LLM	IACRL (Qwen)	Admissible vis-text	yes	0.984 ± 0.005	0.963 ± 0.008	0.982 ± 0.004	0.975 ± 0.006
Scoped LLM	Schema-only LLM	Admissible vis-text	yes	0.982 ± 0.006	0.967 ± 0.006	0.976 ± 0.006	0.978 ± 0.005
HTML/text	RoBERTa [8]	Rendered text	no	0.972 ± 0.003	0.959 ± 0.004	0.965 ± 0.004	0.961 ± 0.005
HTML/text	Longformer [9]	Rendered text	no	0.973 ± 0.004	0.958 ± 0.005	0.967 ± 0.003	0.963 ± 0.004
HTML/text	BigBird [10]	Rendered text	no	0.969 ± 0.005	0.954 ± 0.006	0.962 ± 0.005	0.957 ± 0.006
HTML/text	DeBERTa-v3 [11]	Rendered text	no	0.976 ± 0.004	0.961 ± 0.005	0.970 ± 0.004	0.966 ± 0.005
Markup	MarkupLM [12]	Markup	no	0.970 ± 0.004	0.956 ± 0.006	0.964 ± 0.005	0.958 ± 0.007
URL+HTML	DOM-graph [14]	URL+HTML	no	0.964 ± 0.006	–	–	0.966 ± 0.005
URL+HTML	PhishSim [13]	URL+HTML	no	0.958 ± 0.006	–	–	0.960 ± 0.006
URL+HTML	WebPhish [15]	URL+HTML	no	0.949 ± 0.007	–	–	0.952 ± 0.007
URL-only	URLTran [7]	URL	no	0.933 ± 0.009	–	–	0.938 ± 0.008
URL-only	URLNet [6]	URL	no	0.916 ± 0.011	–	–	0.921 ± 0.010
URL-only	TF-IDF+linear	URL	no	0.904 ± 0.012	–	–	0.908 ± 0.011
Rule-only	Extractor-only	Deterministic rules	yes	0.891 ± 0.015	0.876 ± 0.016	0.883 ± 0.014	0.887 ± 0.015

Table 9: Robustness under guaranteed and empirical regimes. Scoped invariance is reported separately from learned empirical stability.

Corpus	Regime	Method	AtkCov	Chg	FCkg	R-AUROC	R-PR-AUC	R-Val	Cov
LNU	Guaranteed	IACRL scoped	100.0%	0.000	0.000	0.984	0.980	0.901	0.861
LNU	Guaranteed	Schema scoped	100.0%	0.000	0.000	0.982	0.978	0.720	0.876
LNU	Guaranteed	Schema unscoped	100.0%	0.081	0.304	0.960	0.956	0.560	0.884
LNU	Empirical	IACRL scoped	88.4%	0.084	0.221	0.944	0.938	0.787	0.842
LNU	Empirical	Schema scoped	88.4%	0.146	0.296	0.943	0.936	0.579	0.855
LNU	Empirical	Schema unscoped	88.4%	0.236	0.445	0.931	0.924	0.421	0.884
Z492	Guaranteed	IACRL scoped	100.0%	0.000	0.000	0.963	0.968	0.811	0.817
Z492	Guaranteed	Schema scoped	100.0%	0.000	0.000	0.967	0.970	0.715	0.836
Z492	Guaranteed	Schema unscoped	100.0%	0.092	0.318	0.942	0.939	0.531	0.874
Z492	Empirical	IACRL scoped	76.5%	0.132	0.284	0.902	0.899	0.688	0.792
Z492	Empirical	Schema scoped	76.5%	0.183	0.346	0.908	0.904	0.543	0.811
Z492	Empirical	Schema unscoped	76.5%	0.251	0.468	0.894	0.887	0.382	0.867
Z804	Guaranteed	IACRL scoped	100.0%	0.000	0.000	0.982	0.981	0.887	0.854
Z804	Guaranteed	Schema scoped	100.0%	0.000	0.000	0.976	0.976	0.743	0.848
Z804	Guaranteed	Schema unscoped	100.0%	0.064	0.258	0.958	0.953	0.552	0.872
Z804	Empirical	IACRL scoped	89.6%	0.059	0.169	0.949	0.945	0.811	0.849
Z804	Empirical	Schema scoped	89.6%	0.151	0.302	0.936	0.932	0.591	0.843
Z804	Empirical	Schema unscoped	89.6%	0.205	0.419	0.927	0.921	0.442	0.874
M80K	Guaranteed	IACRL scoped	100.0%	0.000	0.000	0.975	0.974	0.858	0.826
M80K	Guaranteed	Schema scoped	100.0%	0.000	0.000	0.978	0.973	0.701	0.853
M80K	Guaranteed	Schema unscoped	100.0%	0.077	0.291	0.956	0.950	0.542	0.876
M80K	Empirical	IACRL scoped	92.7%	0.076	0.236	0.936	0.931	0.742	0.807
M80K	Empirical	Schema scoped	92.7%	0.139	0.287	0.944	0.937	0.554	0.839
M80K	Empirical	Schema unscoped	92.7%	0.223	0.451	0.928	0.921	0.407	0.870

Fig. 4 visualizes the guaranteed and empirical regimes as separate attribution regions. The right-hand empirical region reports learned changes under admissible-evidence perturbations, while the guaranteed region attributes zero-change behavior to scoped serialization rather than to IACRL.

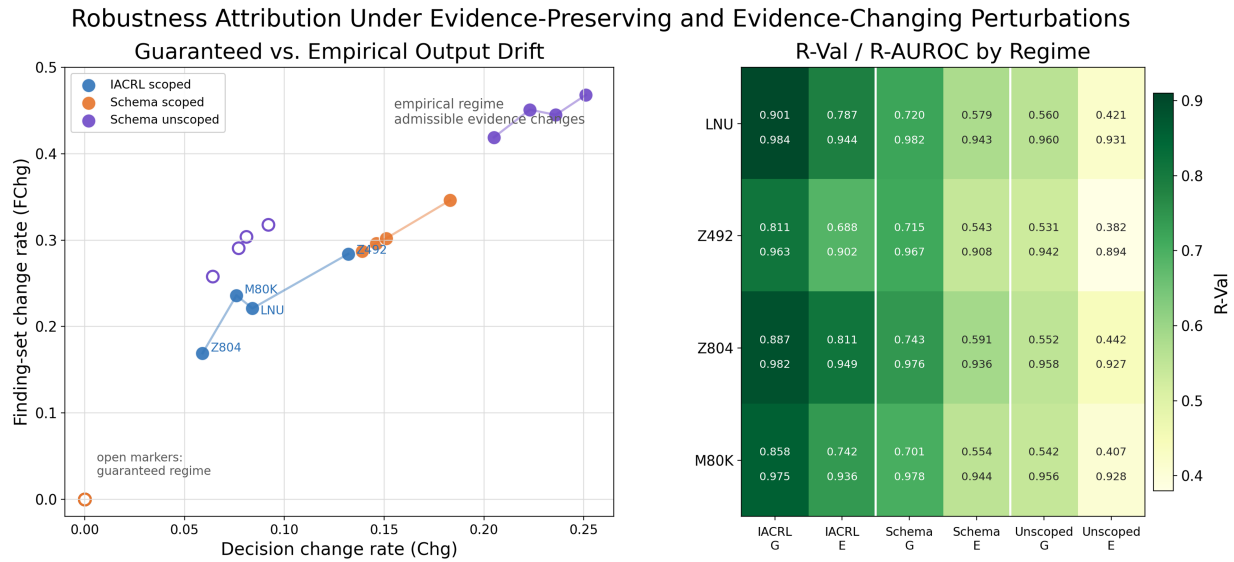


Figure 4: Guaranteed vs. learned robustness. Scoped serialization accounts for guaranteed-regime invariance; IACRL reduces empirical decision drift.

The rows make the attribution visible: the guarantee is not produced by the learning objective, but by the scoped interface. The learning objective matters when the guarantee no longer applies, namely under admissible-evidence perturbations.

7.4 Finding 4: Cross-Corpus Transfer Tests Contract Reusability

Table 10 reports source-to-target transfer as rows rather than a compact matrix, because the four-corpus version includes both utility and auditability fields. Single-source transfer incurs target-dependent ValPass gaps, while the three-corpus mixture tests corpus exclusion rather than equal-data transfer and narrows, but does not eliminate, the remaining target gaps.

The worst single-source transfer case is Z492→M80K, with ValPass 0.764 and a -0.094 gap. This is consistent with artifact mismatch: Z492 has no URL evidence, while M80K often exposes short URL tokens and sparse HTML cues. The transferred model underuses URL-derived provenance and emits more generic text-supported findings; the three-corpus mixture reduces the held-out M80K gap to -0.012 . Operationally, this suggests a remediation path rather than a change to the main claim: no-URL sources should be paired with URL, redirect-chain, header, or screenshot/OCR evidence when the target deployment expects URL-supported findings.

The transfer stress test is visualized together with the mechanism diagnostics in Fig. 5, so that source–target generalization is interpreted alongside the main shortcut explanations the evaluation probes.

Table 10: Cross-corpus transfer for IACRL. Transfer gap is target ValPass minus target-native ValPass.

Source	Target	AUROC	ValPass	IA	Cov	Transfer Gap vs. within ValPass
LNU	LNU	0.984 ± 0.005	0.901 ± 0.014	0.935	0.861	0.000
LNU	Z492	0.936 ± 0.010	0.742 ± 0.030	0.827	0.778	-0.069
LNU	Z804	0.931 ± 0.011	0.801 ± 0.026	0.858	0.808	-0.086
LNU	M80K	0.957 ± 0.007	0.819 ± 0.022	0.879	0.811	-0.039
Z492	LNU	0.948 ± 0.008	0.834 ± 0.023	0.891	0.826	-0.067
Z492	Z492	0.963 ± 0.008	0.811 ± 0.024	0.874	0.817	0.000
Z492	Z804	0.952 ± 0.009	0.821 ± 0.023	0.872	0.815	-0.066
Z492	M80K	0.929 ± 0.011	0.764 ± 0.031	0.839	0.784	-0.094
3-corpus mixture	held-out LNU	0.983 ± 0.005	0.892 ± 0.016	0.929	0.852	-0.009
3-corpus mixture	held-out Z492	0.954 ± 0.009	0.793 ± 0.027	0.861	0.809	-0.018
3-corpus mixture	held-out Z804	0.985 ± 0.004	0.894 ± 0.014	0.931	0.857	+0.007
3-corpus mixture	held-out M80K	0.973 ± 0.006	0.846 ± 0.020	0.895	0.819	-0.012

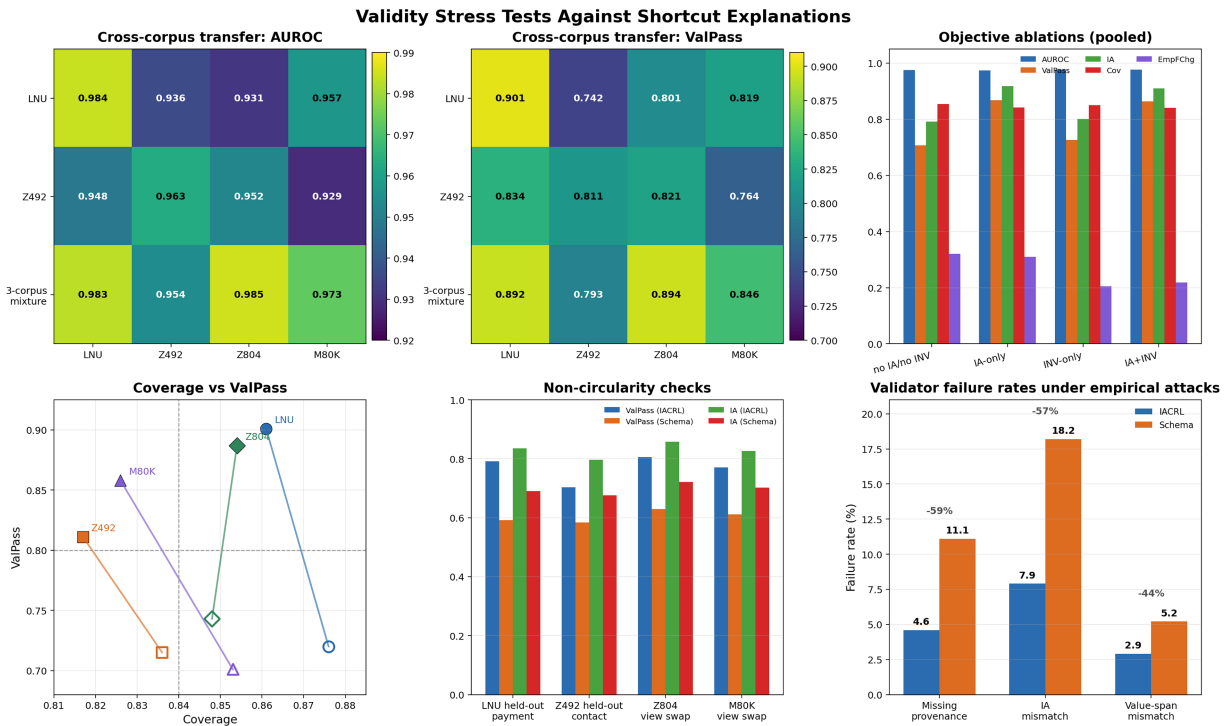
**Figure 5:** Stress tests against shortcut explanations. Top left and center show cross-corpus AUROC and ValPass; top right compares objective ablations. Bottom left plots coverage against ValPass, bottom center reports artifact-family holdouts and view transfer, and bottom right compares validator-failure rates under empirical attacks. Values correspond to Table 11.

Table 11: Mechanism and validity diagnostics. Pooled objective rows average the four diagnostic test sets.

Block	Setting	Variant/Metric	AUROC	ValPass	IA	Cov	EmpFChg	Failure Rate
Objective	Pooled	No IA/no INV	0.975	0.707	0.792	0.854	0.320	–
Objective	Pooled	IA-only	0.974	0.868	0.918	0.842	0.309	–
Objective	Pooled	INV-only	0.976	0.726	0.801	0.850	0.205	–
Objective	Pooled	IA+INV (IACRL)	0.976	0.864	0.910	0.840	0.219	–
Artifact holdout	LNU payment-artifact holdout	IACRL/Schema	–	0.792/0.592	0.836/0.690	0.829/0.842	–	–
Artifact holdout	Z492 contact-artifact holdout	IACRL/Schema	–	0.703/0.584	0.796/0.676	0.792/0.811	–	–
View transfer	Z804 view swap	IACRL/Schema	0.938	0.806/0.629	0.858/0.721	0.834/0.825	–	–
View transfer	M80K view swap	IACRL/Schema	0.943	0.771/0.611	0.826/0.702	0.804/0.832	–	–
Failure mode	Empirical attacks	Missing provenance: IACRL/Schema	–	–	–	–	–	4.6%/11.1%
Failure mode	Empirical attacks	IA mismatch: IACRL/Schema	–	–	–	–	–	7.9%/18.2%
Failure mode	Empirical attacks	Value-span mismatch: IACRL/Schema	–	–	–	–	–	2.9%/5.2%

7.5 Finding 5: Diagnostics Reduce Shortcut Explanations

Table 11 combines objective ablations, artifact holdouts, view swaps, and validator failures. Artifact holdouts remove the designated units from scoped training inputs and weak targets; both variants are retrained, then evaluated with full views on pages bearing the withheld class. LNU withholds `form_payment` and payment identifiers; Z492 withholds `mailto`, `tel`, and contact forms. This tests artifact-level generalization, although rule-closure risk remains.

These diagnostics constrain interpretation. IA-only maximizes clean ValPass, INV-only most directly reduces empirical finding drift, and the combined objective gives the best supportability–stability trade-off. ValPass is not read alone, coverage is not mistaken for supportability, and weak-target training is not presented as analyst-labeled ground truth. The artifact-holdout rows also show that the validator gain survives when a family is withheld from weak target construction, although the remaining dependence on deterministic schema primitives is reported as a limitation.

7.6 Finding 6: Human Audit and Error Analysis

The manual audit in Table 12 provides a small independent check on the rule-grounded metrics. Human-supported finding precision tracks ValPass directionally, but the gap is largest on Z492 and M80K, where released artifacts are sparser and validators miss context that reviewers can see in screenshots or surrounding page structure. Initial reviewer agreement before adjudication is moderate to substantial ($\kappa = 0.74$ for finding support and $\kappa = 0.68$ for major missed-cue labels), which is consistent with the audit’s role as a bounded support check rather than a full semantic phishing adjudication.

Table 12: Manual audit and qualitative error analysis. The audit sample is stratified by corpus, label, validator outcome, and empirical-attack change status.

Corpus	Audited Records	Human-Supported Findings	Validator–Human Agreement	Major Missed-Cue Rate	Most Common Residual Error
LNU	120	0.86	0.82	6.7%	Benign login pages with brand-like language
Z492	120	0.74	0.76	12.5%	no-URL records with ambiguous form intent
Z804	120	0.84	0.81	7.5%	Disagreement between text and structure views
M80K	120	0.79	0.78	10.8%	Sparse pages and short URL-only cues
Aggregate	480	0.81	0.79	9.4%	Artifact availability rather than schema failures

Table 13 separates the primary IACRL audit from a matched schema-only diagnostic audit. The schema-only comparison uses the same 480 record IDs as **Table 12** (120 per corpus), the same stratification variables, and the same two-reviewer-plus-adjudicator rubric; only the model outputs shown in the packets are swapped. Reviewer agreement and human-supported precision are higher for IACRL findings, which argues against the ValPass gain being only an artifact of a simpler or easier-to-pass prediction distribution. The result should still be interpreted cautiously because the audit checks local support and missed cues, not final organizational phishing adjudication.

Table 13: Manual audit by model variant. Both rows use the same 480 records; schema-only packets replace only the model output.

Model	<i>n</i>	Sampling Frame	Human-Supported Findings	Validator–Human Agreement	κ Finding Support	κ Major Missed Cue
Intent–Artifact Constrained Regularized Learning (IACRL) scoped	480	Primary stratified audit	0.81	0.79	0.74	0.68
Schema-only scoped	480	Same record IDs, outputs swapped	0.72	0.73	0.69	0.64

Table 14 summarizes representative de-identified failure patterns from the manual audit.

Table 14: Representative de-identified audit failure patterns.

Corpus	Failure Pattern	Model Output Issue	Why Deterministic Metrics Understate It	Revision Implied
LNU	Benign account portal	Credential finding is locally supportable but not phishing-specific	Validator checks local form support, not organizational legitimacy	Separate local support from label correctness
Z492	No-URL form page	Contact or credential intent remains ambiguous	Missing URL and incomplete released context limit both model and reviewer evidence	Report insufficient-evidence bucket
Z804	Text/structure disagreement	Cited text span supports a claim, but structure view points to a different interaction target	View-specific evidence IDs pass independently	Audit cross-view citation consistency

(Continued)

Table 14 (continued)

Corpus	Failure Pattern	Model Output Issue	Why Deterministic Metrics Understate It	Revision Implied
M80K	Sparse artifact bundle	Output cites generic URL or short text cue and misses a stronger local indicator	Coverage and ValPass can both remain acceptable for under-informative findings	Add informativeness and missed-cue review

The error analysis identifies three recurring failures: benign login pages with locally supportable credential evidence, sparse records that yield supportable but under-informative findings, and empirical perturbations that preserve the dataset label while changing which artifact should be cited. These cases mark the boundary between mechanically supportable findings and analyst-level phishing judgment.

8 Discussion

The evidence-contract framing changes the interpretation of LLM phishing detection. The classifier is only one component; the deployment question is whether the system can cite admissible support for each reported indicator and remain insensitive to attacker-controlled content outside the contract.

The four-corpus results show uneven corpus-level behavior: AUROC stays close to schema-only LLM performance, while validator pass and intent–artifact consistency improve. Thus, IACRL is not primarily an AUROC optimizer; it trains supportable findings under a scoped evidence interface.

The attribution split is central. Guaranteed-regime robustness follows from the contract: if admissible evidence is unchanged, a scoped predictor receives the same input. Empirical-regime robustness is a learned property: when visible text, form attributes, or interaction artifacts are modified, IACRL reduces instability and validator failures but does not provide a formal guarantee.

The four-corpus design tests whether these conclusions survive heterogeneous artifacts: URL+HTML evidence in LNU and M80K, no-URL webpage bundles in Z492, and screenshot/text/CSS/HTML-structure views in Z804. The goal is to expose release-format sensitivity rather than treat the corpora as interchangeable.

The paper’s novelty should therefore be positioned as a systems-and-methods contribution rather than a single-model leaderboard entry. The core contribution is the contract that makes evidence use explicit, the validator ontology that makes outputs mechanically checkable, and the constrained objective that trains the model toward that checkable interface. The empirical results indicate that the contract can be implemented while retaining broadly comparable detection utility.

This positioning makes residual gaps interpretable. Single-source ValPass loss can indicate artifact mismatch, weak target extraction, URL limitations, or model failure; utility transfer without ValPass transfer means the decision boundary survived but supportable explanations did not.

9 Limitations and Scope Boundaries

The framework trades semantic breadth for deterministic auditability. It targets static phishing/scam landing pages and rule-grounded indicators, not full semantic understanding of attacker intent, brand

impersonation, campaign attribution, or analyst-level forensics. Validators are operational support checks, not complete truthfulness metrics.

The interface-level guarantee applies only to transformations that preserve admissible evidence. Attacks that modify visible text, forms, links, or interaction artifacts are outside the guarantee and must be evaluated empirically; this prevents learned visible-content robustness from being presented as formal invariance.

Weak supervision remains a limitation. Deterministic extractors provide scalable targets but can introduce rule-closure bias. Cross-view checks, artifact-family holdouts, cross-corpus evaluation, and manual audit reduce this concern but do not replace large-scale independent analyst labels.

Deployment also requires controls outside the theorem: logging, retrieval, debugging, or fallback components must not concatenate the raw page into prompts after scoping. The current 7B structured-output model is best viewed as a server-side batch triage or analyst-assistance component, not as a real-time browser-side blocking engine. A practical deployment would likely use a cascade: a cheap URL/HTML filter screens traffic, cached domain/template decisions remove repeated pages, and the LLM audit path is reserved for suspicious or disputed cases. Inline blocking would require distillation, caching, or a smaller first-stage model.

Finally, public phishing corpora differ in artifact availability and release format. Some records are discarded after deterministic preprocessing, and some attack families are inapplicable without required artifacts; therefore retained counts, guard-retained attack coverage, and view availability are reported in the main evaluation.

10 Conclusion

This paper reframes LLM phishing detection as adversarial evidence use. Non-admissible injection resistance is an interface property only when scoped serialization is enforced, while robustness under admissible perturbations must be learned and measured. Evidence contracts make this separation explicit by binding conditioning, provenance citation, and deterministic validation to the same admissible scope.

Atomic findings make phishing indicators structured and auditable, while IACRL improves supportability and empirical stability through validator-derived penalties across heterogeneous webpage corpora.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Yue Xu: methodology, implementation, experiments, data curation, and writing—original draft. Xiong Luo: conceptualization, supervision, validation, and writing—review and editing. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The public datasets used in this study are available from their official sources: LNU-Phish (no-screenshots format): <https://lnu-phish.github.io/>; Zenodo phishing website corpus (record 4922598): <https://zenodo.org/records/4922598>; Phishing Website Dataset (Zenodo record 8041387): <https://zenodo.org/records/8041387>; Phishing Websites Dataset (Mendeley Data, n96ncsr5g4): <https://data.mendeley.com/datasets/n96ncsr5g4/1>. Data access and reuse follow the licenses and terms specified by the corresponding data hosts. Retained usable counts are reported after deterministic parsing, deduplication, and evidence-artifact validation.

Ethics Approval: Not applicable.

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

References

1. Apruzzese G, Subrahmanian V. Mitigating adversarial gray-box attacks against phishing detectors. *IEEE Trans Dependable Secur Comput.* 2022;20(5):3753–69. doi:10.1109/tdsc.2022.3210029.
2. Apruzzese G, Subrahmanian VS. LNU-Phish dataset. LNU-Phish. 2022 [cited 2026 Jun 21]. Available from: <https://lnu-phish.github.io/>.
3. Van Dooremaal B, Burda P, Allodi L, Zannone N. Combining text and visual features to improve the identification of cloned webpages for early phishing detection. In: *Proceedings of the 16th International Conference on Availability, Reliability and Security*; 2021 Aug 17–20; Vienna, Austria. p. 1–10.
4. Putra R. Phishing website dataset. Zenodo. 2023 [cited 2026 Jun 21]. Available from: <https://zenodo.org/records/8041387>.
5. Mendeley Data. Phishing websites dataset. Mendeley data. 2021 [cited 2026 Jun 21]. Available from: <https://data.mendeley.com/datasets/n96ncsr5g4/1>.
6. Le H, Pham Q, Sahoo D, Hoi SC. URLNet: learning a URL representation with deep learning for malicious URL detection. *arXiv:1802.03162*. 2018.
7. Maneriker P, Stokes JW, Lazo EG, Carutasu D, Tajaddodianfar F, Urltran GA. Improving phishing URL detection using transformers. In: *Proceedings of the Milcom 2021–2021 IEEE Military Communications Conference (Milcom)*; 2021 Nov 29–Dec 2; San Diego, CA, USA. p. 197–204.
8. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized BERT pretraining approach. *arXiv:1907.11692*. 2019.
9. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. *arXiv:2004.05150*. 2020.
10. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, et al. Big bird: transformers for longer sequences. *Adv Neural Inf Process Syst.* 2020;33:17283–97.
11. He P, Gao J, Chen W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv:2111.09543*. 2021.
12. Li J, Xu Y, Cui L, MarkupLM WF. Pre-training of text and markup language for visually rich document understanding. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2022 May 22–27; Dublin, Ireland. p. 6078–87.
13. Purwanto RW, Pal A, Blair A, PhishSim J S. Aiding phishing website detection with a feature-free tool. *IEEE Trans Inf Forensics Secur.* 2022;17:1497–512. doi:10.1109/tifs.2022.3164212.
14. Yoon JH, Buu SJ, Kim HJ. Phishing webpage detection via multi-modal integration of HTML DOM graphs and URL features based on graph convolutional and transformer networks. *Electronics.* 2024;13(16):3344. doi:10.3390/electronics13163344.
15. Opara C, Chen Y, Wei B. Look before you leap: detecting phishing web pages by exploiting raw URL and HTML characteristics. *Expert Syst Appl.* 2024;236(18):121183.
16. Birthriya SK, Ahlawat P, Jain AK. A dual-layer deep learning model for parallel analysis of URL and HTML features in phishing website detection. *Knowl Inf Syst.* 2026;68(1):4. doi:10.1007/s10115-025-02617-w.
17. Birthriya SK, Ahlawat P, Jain AK. Multi-objective feature selection for phishing detection using a hybrid nash equilibrium and LightGBM model. *IETE J Res.* 2025;71(8):2686–99. doi:10.1080/03772063.2025.2498608.
18. Li Y, Huang C, Deng S, Lock ML, Cao T, Oo N, et al. KnowPhish: large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection. In: *Proceedings of the 33rd USENIX Security Symposium (USENIX Security 24)*; 2024 Aug 14–16; Philadelphia, PA, USA. p. 793–810.
19. Li Y, Wang Z, Ren Y, Yang X, Liu Y, Tian Z. When semantic plausibility becomes a liability: LLM-based phishing detection from an adversarial asymmetry perspective. *Cyber Secur Appl.* 2026;4:100126.
20. Kim H, Song M, Na SH, Shin S, Lee K. When LLMs go online: the emerging threat of Web-Enabled LLMs. In: *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*; 2025 Aug 13–15; Seattle, WA, USA. p. 1729–48.