ScamWeb: A Multimodal Benchmark for Grounded Understanding of Cyber-enabled Fraud Webpages

Anonymous submission

Abstract

Rampant cyber-enabled fraud has escalated into a severe global public-safety crisis, causing massive economic losses and grave threats to social stability. Existing public datasets primarily focus on narrow forms of online fraud, such as phishing or brand impersonation, and typically provide coarse labels, which hinders comprehensive understanding of cybercrime in the real world. We present ScamWeb, a high-quality multimodal and multilingual benchmark for cyber-enabled fraud webpages. It contains over 10,000 curated webpages covering 25 cyber-enabled fraud subtypes, each including full-page screenshots and audited MHTML evidence. In particular, 3,416 fraudulent webpages are annotated with bounding boxes with human-written rationales for explainable analysis. We further propose ScamDet, a unified framework designed for joint tasks of fraud detection, multiclass classification, evidence grounding, and rationale generation. It is a training-free LLM–VLM baseline that integrates URL, textual, visual, and structural webpage information through staged multimodal reasoning. Experimental results show that, although binary fraud detection has become increasingly achievable, fine-grained subtype recognition and grounded evidence localization remain challenging. These results demonstrate the notable gap between current capabilities and the requirements of practical cybercrime understanding. By providing fine-grained annotations, unified evaluation tasks, and baselines, ScamWeb lays a reproducible foundation for developing transparent anti-scam tools for online platforms and public-safety practitioners. The dataset and code will be made publicly available upon acceptance.

Introduction

Cyber-enabled fraud has become one of the fastest-growing forms of cybercrime, causing unprecedented economic losses and posing a growing threat to public safety. According to the Internet Crime Complaint Center (IC3), cybercrime losses in the United States reached $20.9 billion in 2025, with cyber-enabled fraud accounting for more than 85% of the total (Federal Bureau of Investigation 2025). Globally, scams caused an estimated $442 billion in losses in the same year (Global Anti-Scam Alliance and Feedzai 2025). Beyond direct financial damage, cyber-enabled fraud has become increasingly intertwined with underground finance, illegal gambling, human trafficking, and other forms of organized cybercrime (Button and Cross 2017; Chen et al. 2025). Combating cyber-enabled fraud therefore requires not only detecting fraudulent content but also understanding the diverse deceptive strategies through which online scams are carried out.

Fraudulent webpages have become one of the primary delivery mechanisms for cyber-enabled fraud. Advances in generative AI and commodity webpage-generation toolkits allow attackers to produce realistic, multilingual scam webpages at scale with minimal cost. Once an effective template is developed, numerous visually similar variants can be generated through minor modifications, enabling scam campaigns to evolve rapidly and evade existing detection systems (Begou et al. 2023). Moreover, many fraudulent webpages remain online for only a short time before disappearing, leaving a narrow window for forensic collection and making large-scale, representative benchmark construction particularly challenging (Kotzias et al. 2025; Oest et al. 2020).

Despite the growing importance of cyber-enabled fraud, existing public benchmarks remain limited. Most datasets focus on phishing detection, malicious webpage identification, or illicit content classification, each capturing only a narrow aspect of online abuse. In contrast, cyber-enabled fraud encompasses a broad spectrum of deceptive schemes, including investment scams, fake marketplaces, romance scams, impersonation, and extortion. Fraudulent webpages often combine social engineering, visual persuasion, and multimodal evidence, requiring models to understand not only that a webpage is fraudulent but also how it implements a scam. Existing datasets typically provide only URLs or coarse page-level labels, preventing the evaluation of fine-grained fraud taxonomy recognition, evidence localization, and grounded explanations.

To address these limitations, we introduce **ScamWeb**, the first large-scale multimodal benchmark for grounded understanding of cyber-enabled fraud webpages. ScamWeb comprises more than 10,000 curated webpages collected from both live crawling and historical web archives, spanning 25 fraud subtypes across multiple languages. Each webpage is paired with a full-page screenshot and an audited MHTML file, while 3,416 fraudulent webpages are further annotated with bounding boxes and human-written rationales, enabling grounded evaluation beyond fraud classification. We further formulate GroundScam, a unified benchmark task that jointly evaluates fraud detection, fraud subtype classification,

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

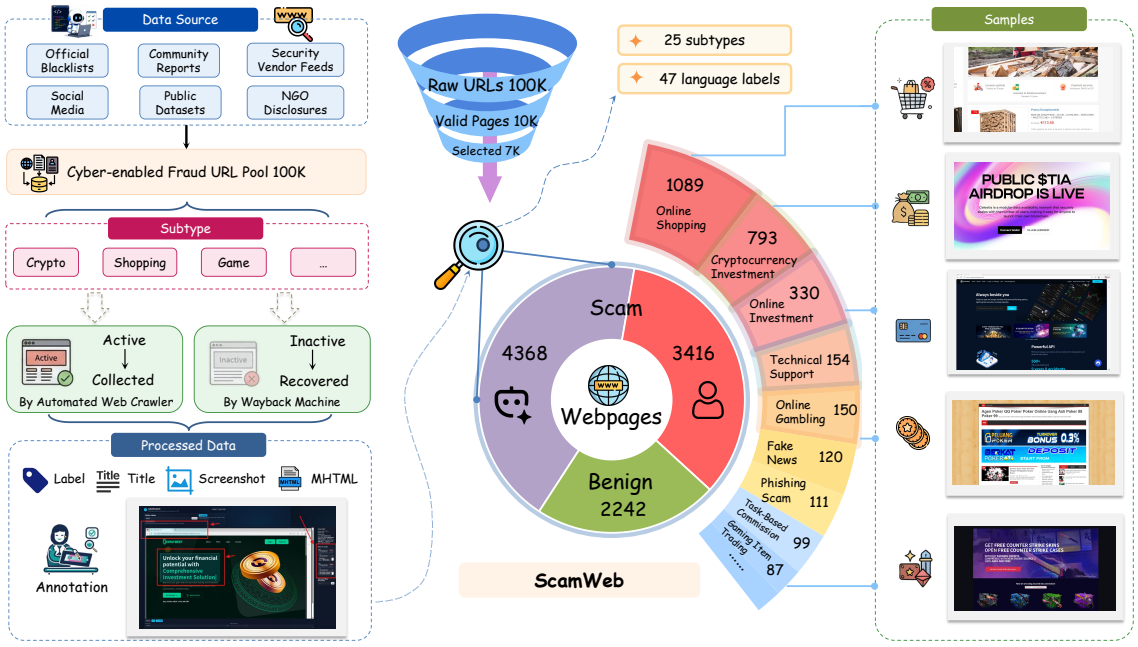


Figure 1: Overview of ScamWeb. **Left:** Data collection, webpage retrieval, and annotation pipeline. **Middle:** Dataset statistics, including the data filtering pipeline and fraud subtype distribution. **Right:** Examples of fraudulent webpages.

evidence localization, and rationale generation. Finally, we introduce **ScamDet**, a training-free LLM–VLM framework that integrates URL, textual, visual, and structural webpage information through staged multimodal reasoning to establish reference performance on GroundScam.

Our contributions are summarized as follows:

- We introduce **ScamWeb**, the first large-scale multimodal benchmark for grounded understanding of cyber-enabled fraud webpages. It comprises more than 10,000 multilingual webpages spanning 25 fraud subtypes, each paired with full-page screenshots and audited MHTML files. A subset of 3,416 fraudulent webpages is further annotated with bounding boxes and human-written rationales to support grounded evaluation.
- We formulate GroundScam, a unified benchmark task that jointly evaluates fraud detection, fraud subtype classification, evidence localization, and rationale generation, enabling comprehensive assessment of grounded multimodal understanding for cyber-enabled fraud. We further introduce ScamDet, a training-free LLM–VLM framework that establishes strong reference performance by integrating URL, textual, visual, and structural webpage information through staged multimodal reasoning.
- Experiments with frontier MLLMs reveal a clear capability gap: although binary fraud detection has become increasingly reliable, performance degrades substantially on fine-grained fraud subtype classification and grounded evidence localization, highlighting the challenges of comprehensive cyber-enabled fraud understanding.

Related Work

Cyber-enabled fraud has evolved from traditional phishing into a diverse ecosystem of online scams. Correspondingly, detection methods have progressed from handcrafted URL features to multimodal reasoning with MLLMs.

Cyber-Enabled Fraud Detection. Early approaches relied on lexical and host-based URL features with classical machine learning (Sahingoz et al. 2019) or deep neural networks (Le et al. 2018; Bahnsen et al. 2017). However, these methods are vulnerable to domain rotation, homoglyph substitution, and multi-stage redirection. Visual and content-based approaches (Abdelnabi, Krombholz, and Fritz 2020; Lin et al. 2021; Liu et al. 2022) leverage webpage layout and structural cues, but remain largely tailored to brand impersonation and phishing. More recently, LLM-based systems have introduced multimodal reasoning for cyber-enabled fraud detection (Liu et al. 2024; Li et al. 2024; Jiang 2024). For example, ScamFerret employs an autonomous LLM agent for evidence collection and webpage classification (Nakano, Koide, and Chiba 2025), while ScamNet adopts a two-stage LoRA framework for explainable fraud detection (Bitaab et al. 2025). Despite these advances, existing methods primarily target specific fraud categories and rarely support fine-grained evidence grounding, largely due to the lack of comprehensive benchmark datasets.

Cyber-Enabled Fraud Datasets. Existing public datasets remain limited in scenario coverage and annotation granularity (Table 1). Early resources focus on phishing URLs or brand impersonation (Lin et al. 2021), while subsequent datasets expand to individual domains such as e-commerce fraud (Bitaab et al. 2023; Cho et al. 2025), cryptocurrency

Attribute	Phish-pedia	Beyond-Phish	Scam-Magnifier	Scam-Ferret	LOKI	Double&Nothing	ScamNet	ScamWeb (Ours)
Scam Size	∼29.5K	∼6.1K	∼47K	∼2.4K	∼1.7K	∼10.1K	∼5.3K	∼10K
Year Range	2020	2019–21	2023–24	2024–25	2024	2022	2024	**2025–26**
URL	✓	✓	✓	✓	✓	✓	✓	✓
Screenshot	✓	–	✓	✓	–	✓	–	✓
HTML / Source	✓	✓	✓	✓	–	✓	✓	✓
Multilingual	–	–	–	✓(3)	–	–	–	**✓(47)**
Fine-grained Types	–	✓(8)	○	✓(4)	✓(10)	–	✓	**✓(25)**
Region Annotations	–	–	–	–	–	–	–	✓
Expert Rationales	○	–	–	–	–	–	○	○
Scam Focus	○	✓	✓	○	✓	✓	✓	✓

Table 1: Comparison of representative datasets for cyber-enabled fraud. ✓ = full support; ○ = partial support; – = not available.

scams (Li, Yepuri, and Nikiforakis 2023), toxic query discovery (Paudel and Stringhini 2026), and LLM-assisted fraud explanation (Nakano, Koide, and Chiba 2025; Bitaab et al. 2025; Yang et al. 2025). In contrast, ScamWeb is the first large-scale benchmark for grounded understanding of cyber-enabled fraud webpages, combining persistent multimodal artifacts, 47-language coverage, 25 fraud subtypes, region-level evidence grounding, and human-written rationales.

Dataset Construction

Collection and Recovery

Figure 1 illustrates the construction pipeline of ScamWeb. We collected approximately 100,000 candidate URLs between 2025-11-30 and 2026-02-26 from public anti-scam platforms, scam directories, public reports, and existing cyber-enabled fraud datasets (Janavicute and Liutkevicius 2023). For each URL, we archived a full-page screenshot and a single-file MHTML archive (Palme, Hopmann, and Shelness 1999) to preserve both visual appearance and DOM structure independently of live web availability.

Because cyber-enabled fraud webpages are often short-lived, only 10,041 webpages were accessible or recoverable, including 4,693 retrieved from the Internet Archive Wayback Machine[1]. After manual verification and domain-level deduplication, the final dataset contains 7,784 unique fraudulent webpages spanning 25 fraud subtypes and 47 languages. We further incorporated 2,242 benign webpages from LegitPhish (Potpelwar, Kulkarni, and Waghmare 2025). Supplementary Material B provides additional details on the collection sources and archive recovery protocol.

Annotation and Taxonomy

To ensure annotation quality while minimizing selection bias, webpages were annotated under two protocols. Specifically, 3,416 webpages were independently annotated by two domain experts from a public-security research institute, with disagreements resolved by a third expert. Supplementary Material A provides the annotation interface, guidelines, and adjudication procedure. Each annotation includes the scam subtype, language, summary, and evidence regions, where each region is associated with a bounding box ($xyxy$) and a human explanation. The remaining 4,368 fraudulent webpages were annotated using our semi-automatic pipeline following the same expert-defined schema. Although the dataset follows a long-tailed distribution, with Online Shopping $(1,089)$ and Cryptocurrency (793) Scams as the largest categories, we retain all fraud subtypes, including rare, high-impact cases such as Romance Scams and Website Hijacking. Supplementary Material A provides subtype descriptions and class distributions.

[1]https://web.archive.org

Scam Evidence Modality	Representative Scam Types (25 Types)	Grounding Modality and Content
Text (URL)	Phishing; Website Hijacking Scam; Business Impersonation Scam	**Vision:** deceptive screenshot regions (bounding boxes) **Text:** expert rationale for each region
Text (Title)	Fake News Website Scam; Technical Support Scam; Lottery/Prize Scam	
Vision	Online Shopping Scam; Cryptocurrency Investment Scam; Online Investment Scam; Online Gambling Scam	
Text–Vision	Task-Based Commission Scam; Romance Scam; Sextortion; Fake Streaming Service Scam	
Code (MHTML)	Spoofed Login Page; Bank Impersonation Scam; Online Loan Scam	

Table 2: Evidence modalities, representative scam types, and grounding targets in ScamWeb.

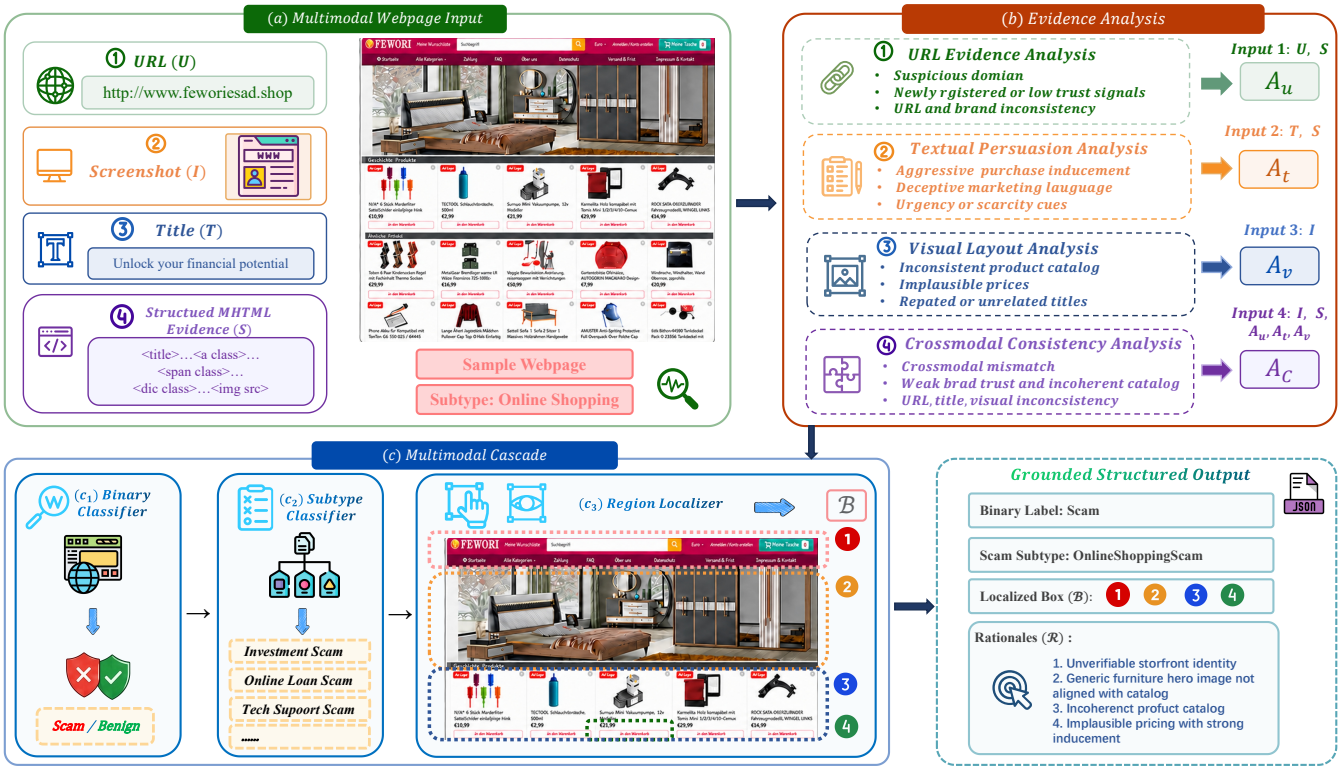


Figure 2: Overview of ScamDet. (a) Multimodal webpage inputs, including the screenshot, URL, page title, and archived MHTML content. (b) Parallel analysis of URL, textual, visual, and cross-modal evidence. (c) Staged reasoning pipeline for fraud detection, fraud subtype classification, evidence localization, and rationale generation, producing a structured prediction.

Quality Control and Input Modalities

To ensure benchmark reliability and fair evaluation, we perform two quality-control procedures before inference. First, all benchmark metadata, ground-truth labels, bounding boxes, and expert rationales are excluded from model inputs to prevent information leakage. Second, all annotation boxes are automatically validated for geometric consistency, while an independent MHTML audit verifies the correspondence between archived content and captured webpages. Only verified archives, or those with explainable redirects, are provided as textual and structural inputs during evaluation.

Table 2 summarizes the input modalities, representative fraud subtypes, and their corresponding grounding targets in ScamWeb. Additional implementation details and audit statistics are provided in Supplementary Material B.

Detection Method

We introduce ScamDet, a training-free multimodal framework that establishes reference performance on ScamWeb. As illustrated in Figure 2, ScamDet progressively integrates URL, textual, visual, and structural webpage information through staged reasoning to detect fraudulent webpages, classify fraud subtypes, localize deceptive evidence, and generate grounded rationales.

Problem Formulation

As illustrated in Figure 2(a), each webpage is represented as $x = (I, U, T, S)$, where I is the full-page screenshot, U is the URL, T is the page title, and S is the trusted structured content extracted from the archived MHTML file. The corresponding ground truth is defined as (y, c, G), where $y \in \{0, 1\}$ denotes the binary fraud label, $c \in \{1, \ldots, 25\} \cup \{\varnothing\}$ is the fraud subtype, and G is the set of grounded evidence annotations. Formally,

$$(y, c, G) = \begin{cases} \left(1, c, \{(b_j, r_j)\}_{j=1}^{N}\right), & \text{if } x \text{ is fraudulent,} \\ (0, \varnothing, \varnothing), & \text{otherwise,} \end{cases}$$

where N is the number of annotated evidence regions, $b_j \in \mathbb{R}^4$ denotes a bounding box in the screenshot coordinate system, and r_j is the corresponding expert rationale.

GroundScam requires a model to predict $(\hat{y}, \hat{c}, \hat{G})$ through four coupled tasks: (i) binary scam detection; (ii) 25-class single-label subtype classification for scam pages; (iii) spatial grounding of one or more deceptive regions supporting the subtype decision; and (iv) generation of a concise rationale for each grounded region.

Multimodal Analysis Modules

As illustrated in Figure 2(b), ScamDet comprises four analysis modules, $\mathcal{A} = \{\mathcal{A}_u, \mathcal{A}_t, \mathcal{A}_v, \mathcal{A}_c\}$, guided by prompts P_u, P_t, P_v, and P_c.

- **URL Evidence ($\mathcal{A}_u$):** Guided by prompt P_u, $\mathrm{LLM}(P_u; U, S)$ analyzes reputation, TLD anomalies, brand-domain mismatches, and redirect chains.
- **Textual Persuasion ($\mathcal{A}_t$):** $\mathrm{LLM}(P_t; T, S)$ identifies deceptive narratives, unrealistic promises, and urgency cues from the page title and structured webpage content.
- **Visual Layout ($\mathcal{A}_v$):** $\mathrm{VLM}(P_v; I)$ identifies visually deceptive regions, such as fake payment portals, forged credentials, fabricated profit charts, and QR-code lures.
- **Cross-Modal Consistency ($\mathcal{A}_c$):** $\mathrm{VLM}(P_c; I, S, \mathcal{A}_u, \mathcal{A}_t, \mathcal{A}_v)$ integrates evidence across modalities to identify semantic inconsistencies. For example, a webpage may claim to represent a financial institution while using an unrelated domain, or present hyperlinks inconsistent with the advertised service.

Multimodal Cascaded Grounding

As illustrated in Figure 2(c), ScamDet performs grounded fraud understanding through a staged decision cascade that progressively predicts the fraud label, identifies the fraud subtype, localizes deceptive evidence, and generates grounded rationales.

Stage 1 (Fraud Detection). The framework first predicts the binary fraud label:

$$\hat{y} = \mathrm{VLM}(P_y; x, \mathcal{A}) \in \{0, 1\}.$$

If $\hat{y} = 0$, the cascade terminates with $\hat{c} = \varnothing$ and $\hat{G} = \varnothing$.

Stage 2 (Subtype Classification). For webpages predicted as fraudulent, the fraud subtype is inferred as

$$\hat{c} = \mathrm{VLM}(P_{\mathrm{type}}; x, \mathcal{A}, \hat{y}) \in \{1, \ldots, 25\}.$$

Stage 3 (Evidence Grounding). The grounding module predicts view-relative bounding boxes:

$$\widetilde{B} = \mathrm{VLM}(P_g; \mathcal{V}, \mathcal{A}, \hat{y}, \hat{c}),$$

where $\mathcal{V}$ denotes the set of full-page and cropped views. The predicted boxes are mapped back to the original screenshot coordinates to obtain $\hat{B}$.

Stage 4 (Rationale Generation). Conditioned on $\hat{B}$, the rationale generator produces

$$\hat{R} = \mathrm{VLM}(P_r; x, \mathcal{A}, \hat{y}, \hat{c}, \hat{B}),$$

yielding the grounded evidence set $\hat{G} = \{(\hat{b}_j, \hat{r}_j)\}_{j=1}^{M}$.

The final prediction is $\hat{\mathcal{O}} = (\hat{y}, \hat{c}, \hat{G})$, which directly corresponds to the annotation format of ScamWeb.

Experiments

Our experiments are designed to answer the following research questions. **Q1:** How do representative LLM–VLM systems perform on the ScamWeb benchmark? **Q2:** How well can current models recognize fine-grained, long-tailed fraud subtypes? **Q3:** How accurately can current models localize deceptive evidence and generate grounded rationales, and how does ScamDet compare with representative baselines? **Q4:** What insights does ScamWeb reveal about the relative difficulty of different cyber-enabled fraud scenarios?

Experimental Setup

Dataset split. We evaluate ScamDet in a zero-shot setting on the full ScamWeb benchmark. For supervised baselines, the dataset is split into 70% training, 15% validation, and 15% testing.

Implementation details. ScamDet uses GPT-4o-mini (OpenAI 2024a) for URL and textual analysis, and Qwen3-VL-Flash (Alibaba Cloud 2026) for visual analysis and downstream reasoning. The temperature is set to zero, with Qwen reasoning disabled. We compare against representative frontier LLM–VLM systems, including GPT-4o (OpenAI 2024b), GPT-4o-mini, GPT-5 mini (OpenAI 2025), Claude Sonnet 4 (Anthropic 2025), Gemini 2.5 Flash (Comanici et al. 2025), Gemini 3 Flash (Google DeepMind 2025), Qwen2.5-VL (Bai et al. 2025b), and Qwen3-VL-235B-A22B-Thinking (Bai et al. 2025a). Additional implementation details are provided in Supplementary Material D.

Evaluation metrics. We adopt task-specific evaluation metrics. For binary fraud detection, we report accuracy, precision, recall, and F1, treating fraudulent webpages as the positive class. For fraud subtype classification, we report accuracy together with macro-averaged precision, recall, and F1. Evidence grounding is evaluated using page-level mIoU, precision, recall, and F1 at an IoU threshold of 0.5, together with Recall@0.3 and Recall@0.5. Rationale generation is assessed using ROUGE-L F1 (Lin 2004), multilingual BERTScore F1 (Zhang et al. 2020), and grounded rationale coverage. Detailed metric definitions are provided in Supplementary Material D.

Approach	Method	Acc.	P	R	F1
Trad.	Phishpedia	76.8	78.4	74.9	76.6
Trad.	PhishIntention	**82.5**	**83.7**	80.9	**82.3**
LLM	ScamDet	68.3	66.1	**97.2**	78.7

Table 3: Binary detection results (%). Trad. and LLM denote task-specific multimodal phishing detectors and the LLM-based method, respectively. P/R/F1 treat scam as positive.

Binary Classification (Q1) Table 3 reports the binary fraud detection results for Q1, comparing ScamDet with Phishpedia (Lin et al. 2021) and PhishIntention (Liu et al. 2022). Phishpedia detects brand logos in webpage screenshots and verifies their consistency with the associated domain, while PhishIntention jointly analyzes URLs, screenshots, HTML, and webpage interactions to infer brand identity and credential-taking intent. Although these specialized methods achieve stronger binary detection performance, they are designed exclusively for phishing detection. In contrast, ScamDet serves as a unified framework that simultaneously performs binary detection, fraud subtype classification, evi-

Scam type	N	Prec.	Recall	F1
Online Shopping Scam	1,089	81.3	91.9	86.3
Cryptocurrency Investment Scam	793	80.9	77.6	79.2
Online Investment Scam	330	52.2	58.2	55.0
Technical Support Scam	154	75.0	1.9	3.8
Online Gambling Scam	150	74.4	83.3	78.6
Full-set aggregate (3,416 scam pages)				
Overall accuracy	3,416		63.7	
Macro (25 types)	3,416	52.2	37.4	34.5
Grounding	mIoU	0.173	P/R/F1@0.5	3.5/12.4/5.5
Rationale	R-L	13.4	BERT/Cov.	87.6/12.4

Table 4: Subtype classification across the five most frequent scam types and aggregate Grounding/Rationale results on the complete benchmark. N denotes pages. Except for mIoU, P/R/F1 and rationale metrics are percentages.

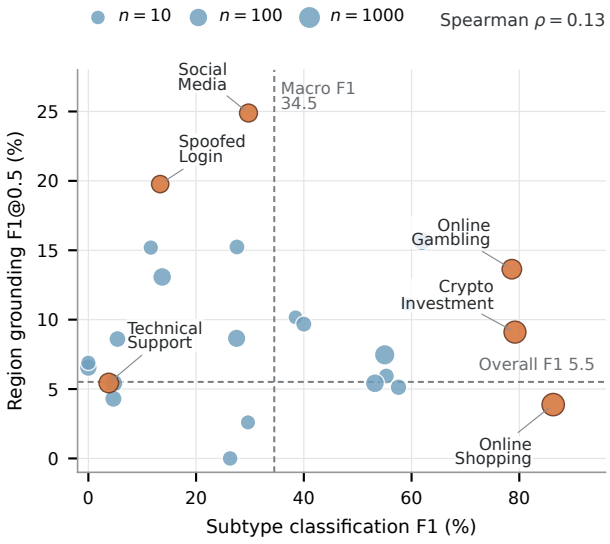


Figure 3: Subtype classification and region grounding across 25 scam subtypes on the 3,416-page scam set. Marker area scales with $\log(1+N)$ (N is subtype support); dashed lines indicate macro subtype F1 and overall grounding F1.

dence localization, and rationale generation, establishing reference performance for the complete ScamWeb benchmark.

Sub-type Classification (Q2) Across the complete cascade, ScamDet achieves a subtype classification accuracy of 63.7% and a macro F1 of 34.5%. As shown in Table 4, performance varies substantially across fraud subtypes. Online Shopping Scam, Cryptocurrency Investment Scam, and Online Gambling Scam achieve F1 scores of 86.3%, 79.2%, and 78.6%, respectively, with corresponding recalls of 91.9%, 77.6%, and 83.3%, indicating that distinctive commercial, investment, and gambling cues are reliably recognized.

Performance degrades on semantically similar or underrepresented categories. For example, Online Investment Scam achieves 52.2% precision, 58.2% recall, and 55.0% F1, suggesting substantial confusion with related investment-oriented scams. Technical Support Scam exhibits the largest precision–recall imbalance, achieving 75.0% precision but only 1.9% recall, resulting in an F1 of 3.8%. This indicates that the model identifies this subtype only when highly distinctive cues are present. The gap between the overall accuracy (63.7%) and macro F1 (34.5%) further highlights the challenge of robust recognition across the long-tailed fraud taxonomy. Figure 3 further examines whether accurate subtype prediction translates into equally reliable evidence grounding. Complete per-class results and the confusion matrix are provided in Supplementary Material E.

Grounding and Rationales (Q3) *Region grounding.* Table 4 summarizes region grounding performance over the full benchmark. ScamDet achieves 0.173 mIoU and 5.5% F1 at an IoU threshold of 0.5, indicating that accurate spatial localization remains substantially more challenging than webpage-level fraud recognition. Precision is consistently lower than recall (3.5% vs. 12.4%), suggesting that the model retrieves relevant evidence but frequently predicts imprecise or redundant regions. Figure 3 further reveals only a weak correlation between subtype recognition and grounding performance across the 25 fraud subtypes (Spearman $\rho = 0.13$). For example, Online Shopping Scam achieves an F1 of 86.3% for subtype classification but only 3.9% for grounding, whereas Social Media Scam and Spoofed Login Page obtain grounding F1 scores of 24.9% and 19.8% despite substantially lower subtype F1 scores of 29.7% and 13.3%, respectively. These results indicate that recognizing webpage intent and precisely localizing deceptive evidence are related but distinct challenges.

Rationale quality. Table 4 also reports rationale quality for correctly localized regions. ROUGE-L reaches 13.4%, while BERTScore attains 87.6%, indicating that rationales preserve semantic meaning despite limited lexical overlap with expert annotations. However, grounded rationale coverage is only 12.4%, suggesting that rationale quality is constrained by evidence localization rather than language generation. Grounding, page-level recall, and rationale results are provided in Supplementary Material E.

Single-model zero-shot comparison. Table 5 compares ScamDet with representative frontier LLM–VLM systems. ScamDet achieves the highest fraud recall (97.3%) and subtype accuracy (63.9%), leads all grounding metrics (mIoU 0.193, F1@0.5 9.9%, and Recall@0.5 16.4%), and obtains the best rationale quality in terms of BERTScore (89.7%) and grounded coverage (15.3%). These consistent improvements suggest that staged integration of URL, textual, visual, and cross-modal evidence provides stronger support for grounded fraud understanding than single-pass prompting. Moreover, ScamDet produces valid structured outputs for all webpages (100%), whereas Claude Sonnet 4 and Qwen3-VL-235B-A22B-Thinking achieve 84.9% and 47.1%, respectively, in-

	Valid	Bin.	Subtype				Grounding						Rationale			Tokens	
Model	Out.	R	Acc.	mP	mR	mF1	mIoU	P	R	F1	R@.3	R@.5	R-L	BERT	Cov.	Tot.	Avg.
GPT-4o	99.5	55.6	38.9	**58.9**	28.0	31.6	.083	1.7	2.1	1.9	10.9	2.4	11.1	88.0	2.1	**5.94**	**1,788**
GPT-4o-mini	99.3	55.5	37.2	44.1	25.0	25.3	.059	1.3	1.0	1.2	6.3	1.3	9.7	87.5	1.0	125.25	37,679
GPT-5 mini	99.6	54.8	34.6	50.3	24.3	26.8	.135	3.9	5.9	4.7	20.0	7.1	11.6	86.7	5.9	9.25	2,784
Claude Sonnet 4	84.9	46.7	33.2	41.9	23.2	25.2	.064	1.3	1.5	1.4	7.3	1.7	12.4	87.0	1.5	7.03	2,114
Gemini 2.5 Flash	99.2	61.8	40.3	53.3	25.0	29.0	.112	4.4	7.4	5.5	15.9	8.0	**17.9**	88.7	7.4	17.67	5,315
Gemini 3 Flash	99.4	73.9	46.0	58.0	30.1	34.1	.183	7.2	14.4	9.6	26.0	15.4	12.1	88.0	14.4	10.67	3,209
Qwen2.5-VL	97.3	57.4	40.8	58.7	28.5	30.8	.020	0.2	0.2	0.2	1.0	0.2	7.4	86.5	0.2	6.71	2,019
Qwen3-VL-235B-A22B-Thinking	47.1	31.2	23.5	43.6	11.1	16.1	.011	0.1	0.1	0.1	0.6	0.1	11.8	87.2	0.1	15.35	4,617
ScamDet	**100.0**	**97.3**	**63.9**	51.9	**36.9**	**34.1**	**.193**	**7.3**	**15.3**	**9.9**	**26.4**	**16.4**	13.4	**89.7**	**15.3**	34.10	10,260

Table 5: End-to-end zero-shot comparison on the 3,324-page scam subset. Valid, Bin., Tot., and Avg. denote the valid output rate, scam recall, total tokens (M), and average tokens per page, respectively. All metrics except mIoU are reported as percentages.



(a) Correct binary, subtype, and grounding.

(b) Binary false negative.

(c) Subtype misclassification.

(d) Boundary-displaced grounding.

Figure 4: Four representative cases in ScamWeb. Blue solid boxes denote expert annotations, and orange dashed boxes denote ScamDet predictions.

dicating that staged generation substantially improves output reliability for complex structured prediction.

Cost analysis. ScamDet incurs the second-highest token cost among the compared methods, exceeded only by GPT-4o-mini. In return, it achieves the strongest overall multi-task profile across scam recall, subtype accuracy, region grounding, rationale semantics, and grounded coverage. This result reflects a trade-off between effectiveness and inference cost. The complete protocol and failure accounting are provided in Supplementary Section D.

Case Study (Q4) Our qualitative analysis highlights the principal challenges posed by ScamWeb. Figure 4(a) presents a successful example, where implausibly low prices and an inconsistent product catalogue enable correct fraud detection, subtype classification, and evidence grounding. Figure 4(b) illustrates a binary detection failure, in which a polished corporate presentation masks the domain–identity mismatch, causing the cascade to terminate before downstream prediction. Figure 4(c) demonstrates subtype ambiguity, where shared blockchain and liquidity terminology leads an Online Investment Scam to be misclassified as a Cryptocurrency Investment Scam. Figure 4(d) shows a grounding failure: although the model correctly recognizes the gambling intent, it localizes only fragments of the promotional evidence rather than the complete expert-annotated region. Collectively, these examples reveal three recurring challenges: deceptive webpages closely mimic legitimate visual conventions, semantically related fraud subtypes exhibit substantial lexical overlap, and model-predicted evidence regions often differ from expert annotations in spatial granularity. Additional qualitative analyses and error categorization are provided in Supplementary Material E and G.

Conclusion

We introduced **ScamWeb**, the first benchmark for grounded understanding of cyber-enabled fraud webpages, and **ScamDet**, a training-free multimodal framework establishing reference performance. By integrating multimodal evidence, multilingual webpages, region-level grounding, and expert-written rationales, ScamWeb enables systematic evaluation of fraud detection, subtype classification, evidence localization, and rationale generation. Our experiments show that, although binary fraud detection is largely tractable, fine-grained subtype recognition and evidence grounding remain major challenges for multimodal foundation models. Future work will extend ScamWeb to open-set fraud taxonomies, continual benchmark updates, and uncertainty-aware reasoning. Responsible use of archived scam webpages follows sanitization, controlled-access, and takedown protocols in Supplementary Material H.

Ethical Statement

ScamWeb is intended solely for defensive research on cyber-enabled fraud. Because it contains archived real-world scam webpages, access is restricted under a controlled research agreement that prohibits offensive or deceptive use, including scam generation, scam optimization, and safety-evasion research. As automated fraud detection may produce false positives, particularly for legitimate niche or minority-owned businesses, practical deployments should incorporate human review and subgroup evaluation across languages. All annotations were completed by domain experts at a public-security research institute in isolated computing environments under institutional ethical guidelines.

References

Abdelnabi, S.; Krombholz, K.; and Fritz, M. 2020. Visualphishnet: Zero-day phishing website detection by visual similarity. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 1681–1698.

Alibaba Cloud. 2026. Qwen3-VL-Flash Model Information. Alibaba Cloud Model Studio documentation. Accessed July 29, 2026.

Anthropic. 2025. System Card: Claude Opus 4 and Claude Sonnet 4. Technical report, Anthropic.

Bahnsen, A. C.; Bohorquez, E. C.; Villegas, S.; Vargas, J.; and González, F. A. 2017. Classifying phishing URLs using recurrent neural networks. In *2017 APWG Symposium on Electronic Crime Research (eCrime)*, 1–8. IEEE.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-VL Technical Report. arXiv:2511.21631.

Bai, S.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Begou, N.; Vinoy, J.; Duda, A.; and Korczyński, M. 2023. Exploring the dark side of ai: Advanced phishing attack design and deployment using chatgpt. In *2023 IEEE Conference on Communications and Network Security (CNS)*, 1–6. IEEE.

Bitaab, M.; Cho, H.; Oest, A.; Lyu, Z.; Wang, W.; Abraham, J.; Wang, R.; Bao, T.; Shoshitaishvili, Y.; and Doup'e, A. 2023. Beyond Phish: Toward detecting fraudulent e-commerce websites at scale. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2366–2383.

Bitaab, M.; Karimi, A.; Lyu, Z.; Mosallanezhad, A.; Oest, A.; Wang, R.; Bao, T.; Shoshitaishvili, Y.; and Doupé, A. 2025. ScamNet: Toward explainable large language model-based fraudulent shopping website detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27841–27848. AI for Social Impact Track.

Button, M.; and Cross, C. 2017. *Cyber frauds, scams and their victims*. Routledge.

Chen, G.; Liu, Q.; Chen, G.; and Huang, A. 2025. Exploring illicit personal information trading behind telecom fraud in China. *Humanities and Social Sciences Communications*, 12(1): 1–11.

Cho, H.; Bitaab, M.; Wang, R.; Shoshitaishvili, Y.; Doupé, A.; and Ahn, G.-J. 2025. SCAMMAGNIFIER: Piercing the veil of fraudulent shopping website campaigns. In *2025 Network and Distributed System Security Symposium (NDSS)*.

Comanici, G.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Federal Bureau of Investigation. 2025. 2025 Internet Crime Report. Internet Crime Complaint Center (IC3).

Global Anti-Scam Alliance; and Feedzai. 2025. Global State of Scams 2025 Report. Global Anti-Scam Alliance (GASA).

Google DeepMind. 2025. Gemini 3 Flash: Model Card. Google DeepMind model card.

Janavicute, A.; and Liutkevicius, A. 2023. Fraudulent and Legitimate Online Shops Dataset.

Jiang, L. 2024. Detecting Scams Using Large Language Models. arXiv:2402.03147.

Kotzias, P.; Pachilakis, M.; Aldana-Iuit, J.; Caballero, J.; Sánchez-Rola, I.; and Bilge, L. 2025. Ctrl+ Alt+ Deceive: Quantifying User Exposure to Online Scams. In *Network and Distributed System Security Symposium (NDSS)*.

Le, H.; Pham, Q.; Sahoo, D.; and Hoi, S. C. H. 2018. URLNet: Learning a URL representation with deep learning for malicious URL detection. arXiv:1802.03162.

Li, X.; Yepuri, A.; and Nikiforakis, N. 2023. Double and nothing: Understanding and detecting cryptocurrency giveaway scams. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*.

Li, Y.; Huang, C.; Deng, S.; Lock, M. L.; Cao, T.; Oo, N.; Lim, H. W.; and Hooi, B. 2024. KnowPhish: Large language models meet multimodal knowledge graphs for enhancing Reference-Based phishing detection. In *33rd USENIX security symposium (USENIX Security 24)*, 793–810.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Lin, Y.; Liu, R.; Divakaran, D. M.; Ng, J. Y.; Chan, Q. Z.; Lu, Y.; Si, Y.; Zhang, F.; and Dong, J. S. 2021. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)*, 3793–3810.

Liu, R.; Lin, Y.; Teoh, X.; Liu, G.; Huang, Z.; and Dong, J. S. 2024. Less defined knowledge and more true alarms: Reference-based phishing detection without a pre-defined reference list. In *33rd USENIX Security Symposium (USENIX Security 24)*, 523–540.

Liu, R.; Lin, Y.; Yang, X.; Ng, S. H.; Divakaran, D. M.; and Dong, J. S. 2022. Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)*, 1633–1650.

Nakano, H.; Koide, T.; and Chiba, D. 2025. Scamferret: Detecting scam websites autonomously with large language models. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA)*, 3–25. Springer.

Oest, A.; Zhang, P.; Wardman, B.; Nunes, E.; Burgis, J.; Zand, A.; Thomas, K.; Doupé, A.; and Ahn, G.-J. 2020. Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th USENIX Security Symposium (USENIX Security 20)*.

OpenAI. 2024a. GPT-4o mini: Advancing Cost-Efficient Intelligence. OpenAI. Accessed July 12, 2026.

OpenAI. 2024b. GPT-4o System Card. Technical report, OpenAI.

OpenAI. 2025. GPT-5 System Card. Technical report, OpenAI.

Palme, J.; Hopmann, A.; and Shelness, N. 1999. MIME Encapsulation of Aggregate Documents, such as HTML (MHTML). RFC 2557, Internet Engineering Task Force.

Paudel, P.; and Stringhini, G. 2026. LOKI: Proactively discovering online scam websites by mining toxic search queries. In *2026 Network and Distributed System Security Symposium (NDSS)*.

Potpelwar, R. S.; Kulkarni, U.; and Waghmare, J. 2025. LegitPhish: A large-scale annotated dataset for URL-based phishing detection. *Data in Brief*, 63: 111972.

Sahingoz, O. K.; Buber, E.; Demir, O.; and Diri, B. 2019. Machine learning based phishing detection from URLs. *Expert Systems with Applications*, 117: 345–357.

Yang, S.; Zhu, S.; Wu, Z.; Wang, K.; Yao, J.; Wu, J.; Hu, L.; Li, M.; Wong, D. F.; and Wang, D. 2025. Fraud-r1: A multi-round benchmark for assessing the robustness of llm against augmented fraud and phishing inducements. In *Findings of the Association for Computational Linguistics: ACL 2025*, 4374–4420.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

ScamWeb: Supplementary Document

Anonymous submission

This Supplementary Document records the frozen taxonomy and data contract, the ScamDet computation graph, evaluation definitions, extended results, human-audit protocol, and release safeguards.

A. Taxonomy and Dataset Details

A.1 Canonical Subtype IDs and Support

Table 1 maps each integer output ID to its English subtype label in `taxonomy.py`; IDs do not depend on empirical frequency.

ID	Canonical subtype	Support
1	Task-Based Commission Scam	99
2	E-commerce Customer Service Impersonation Scam	16
3	Spoofed Login Page	38
4	Online Loan Scam	18
5	Online Investment Scam	330
6	Cryptocurrency Investment Scam	793
7	Government / Law Enforcement Impersonation Scam	9
8	Romance Scam	21
9	Fake Streaming Service Scam	29
10	Online Shopping Scam	1,089
11	Technical Support Scam	154
12	Gaming Item Trading Scam	87
13	Online Gambling Scam	150
14	Charity Scam	12
15	Business Impersonation Scam	83
16	Job / Recruitment Scam	14
17	Lottery / Prize Scam	21
18	Website Hijacking Scam	42
19	Education / Training Scam	23
20	Fraudulent Personal Website Scam	50
21	Bank Impersonation Scam	36
22	Fake News Website Scam	120
23	Phishing	112
24	Social Media Scam	55
25	Sextortion	15

Table 1: Canonical subtype IDs and Gold support.

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.

Figure 1 visualizes the corresponding support distribution: the two largest classes contain more than half of all scam pages, whereas the smallest contains nine.

A.2 Expert Annotation and Adjudication

Three domain experts with extensive professional experience in public-security practice and research manually annotated all 3,416 Gold scam pages. Under a shared specification, two experts independently assigned a page-level subtype and description and one to four rectangular evidence regions with subtype labels and rationales. No model-generated pre-annotations were shown. A third expert adjudicated disagreements and defined the final Gold record.

A project-specific local web interface loaded screenshots, supported task-defined labels, captured page- and region-level labels and descriptions, and exported one JSON file per screenshot. Regions were stored by center, width, and height in natural-image coordinates and converted deterministically to $xyxy$. Database-free local deployment kept potentially sensitive or explicit webpage content on controlled machines.

A.3 Regions and Language Coverage

The 3,416 scam records contain 5,489 regions; respectively 1,779, 1,230, 378, and 29 pages have one through four boxes. Every clipped Gold box is finite, nondegenerate, and paired with a nonempty reason.

Gold counting rule. We distinguish raw field strings from atomic language labels. Among the 3,416 Gold records, 3,385 contain one language token and 31 contain two. Direct whole-string deduplication therefore yields 47 values: 35 monolingual strings and 12 order- or spacing-sensitive bilingual forms. Splitting on commas, trimming whitespace, and deduplicating canonical names instead yields 36 atomic Gold labels and 3,447 page–language assignments. Table 2 gives the complete inventory. English, Chinese, and Russian occur in 2,425, 409, and 133 records, respectively; multilingual records contribute once to each applicable language.

Full-corpus language inventory. Language coverage in the broader corpus is extended beyond the expert-annotated subset using outputs from nine models. Across 7,866 parseable artifacts and 58,698 nonempty `language` fields, direct whole-string deduplication produces 223 case-sensitive values. We normalize this inventory by splitting

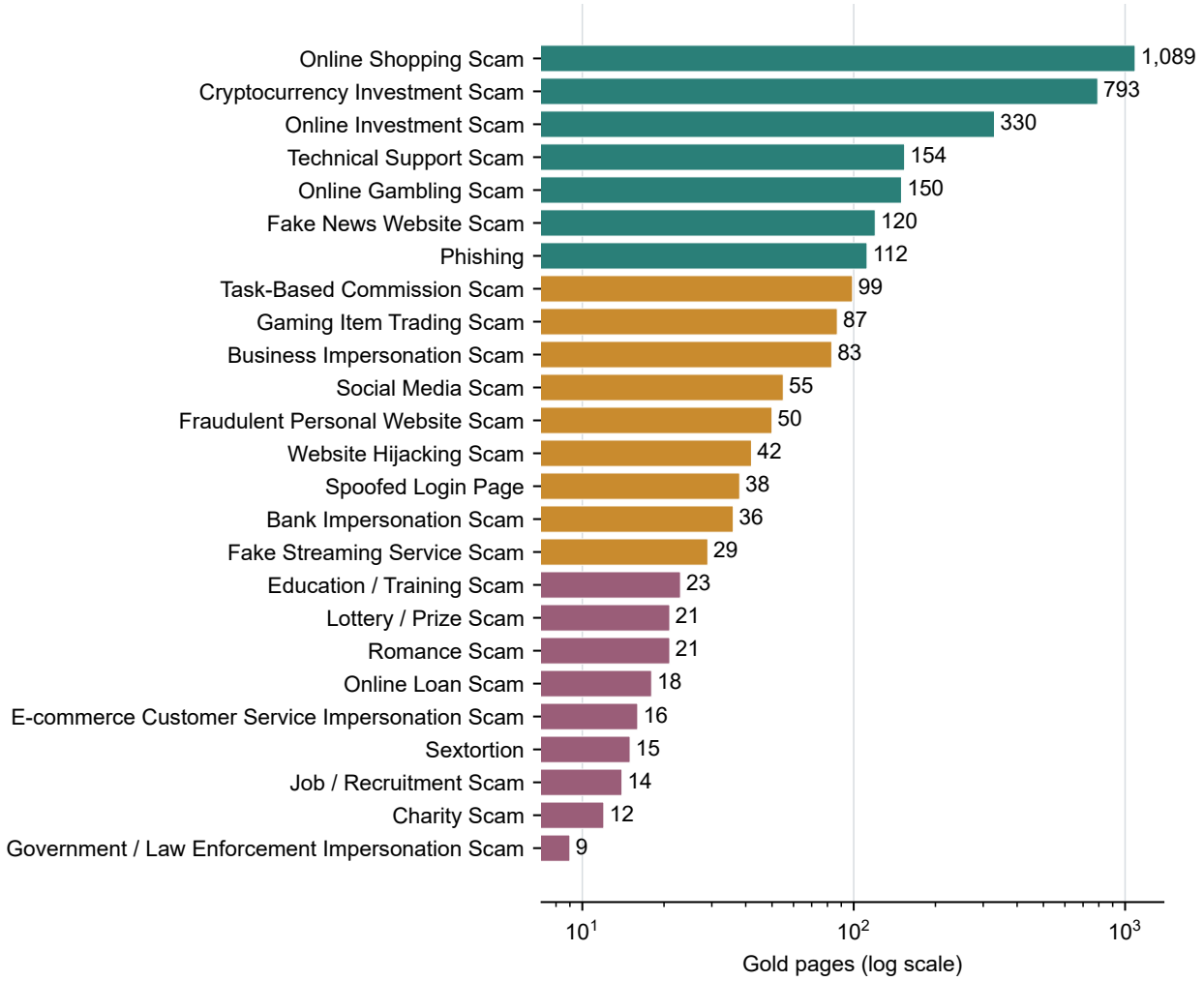


Figure 1: Gold support of the 25 canonical scam subtypes. The horizontal axis is logarithmic.

Language	Support	Language	Support	Language	Support
Arabic	38	Bosnian	1	Bulgarian	1
Chinese	409	Croatian	4	Czech	7
Danish	4	Dutch	16	English	2,425
French	34	German	114	Greek	1
Hebrew	1	Hindi	1	Hungarian	4
Indonesian	20	Italian	17	Japanese	16
Korean	9	Latvian	1	Lithuanian	35
Norwegian	2	Persian	34	Polish	10
Portuguese	25	Romanian	6	Russian	133
Serbian	1	Slovak	1	Slovenian	3
Spanish	36	Swedish	4	Thai	4
Turkish	3	Ukrainian	1	Vietnamese	26

Table 2: Complete atomic-language inventory of the 3,416-page Gold annotation set. Support is the number of records containing each label after comma-separated multilingual fields are split and trimmed. The supports sum to 3,447 because 31 records contain two languages.

comma- and slash-delimited outputs; merging explicit aliases such as Farsi/Persian, Traditional Chinese/Chinese, Slovensky/Slovak, Slovene/Slovenian, and Magyar/Hungarian; removing `unknown`, `none`, and `other`; and excluding one degenerate response that enumerates 79 languages rather than identifying the page language. The resulting corpus-level inventory contains 47 normalized labels, comprising all 36 expert-verified labels and the 11 additional labels audited in Table 3. This 47-label inventory is the language coverage reported for the complete ScamWeb corpus in the main paper; the 3,416-page expert-evaluation subset retains the 36 expert-verified labels in Table 2. Cross-model agreement provides an audit signal for the additional labels and does not alter human annotations or evaluation targets.

CJK occurs in 44 page descriptions and 46 region rationales. We preserve the original text and use multilingual semantic scoring rather than translate Gold rationales.

Language	Output hits	Pages	Agreement audit
Albanian	18	2	Strong: observed across all nine models.
Bengali	16	2	Strong: observed across all nine models.
Estonian	7	1	Strong: seven-model agreement.
Filipino	1	1	Weak: returned only by GPT-5-mini.
Finnish	11	2	Strong on one page, with seven-model agreement.
Georgian	1	1	Weak: returned only by Claude; the other models report Russian.
Khmer	7	1	Strong: seven-model agreement.
Latin	6	4	Weak: mostly single-model outputs and potentially placeholder text.
Malay	10	3	Moderate: multiple-model support on a subset of pages.
Pashto	1	1	Weak: returned only by Gemini 3 Flash; other outputs report English or Dari/Persian.
Tajik	4	2	Limited: returned by two Gemini models; the other models report Russian.

Table 3: Additional labels completing the 47-language full-corpus inventory. Output hits count model-response fields, whereas pages count distinct result artifacts. Agreement summarizes cross-model support; the expert-evaluation subset remains defined by Table 2.

A.4 Label Normalization

The 28 raw label strings undergo exact case/whitespace folding and only five explicit aliases: spacing variants for Education/Training, Government/Law Enforcement, Job/Recruitment, and Lottery/Prize, plus one malformed source string. The raw value `PhisOnline Shopping Scamhing` is retained solely as a documented source-data anomaly and mapped deterministically to Phishing; it is not an admissible prediction label. No keyword or substring rule converts an unknown output into a valid class. Unknown and `none` predictions on a scam page are therefore errors, and the macro average remains defined over exactly 25 classes.

B. Frozen Data and Input Policy

Collection and archive recovery. Between November 30, 2025 and February 26, 2026, we aggregated approximately 100,000 candidate URLs from public anti-scam platforms, scam directories, public reports, and prior web-fraud datasets. Automated acquisition attempted a full-page screenshot and a single-file MHTML archive for each candidate. Of the 10,041 pages that remained accessible or recoverable, 4,693 were obtained through the Internet Archive Wayback Machine. Manual relevance review and unique-domain deduplication produced 7,784 scam webpages; 2,242 benign webpages were collected with the same screenshot-and-MHTML protocol. The released corpus therefore contains 10,026 pages. Its primary scored benchmark comprises the 3,416 expert-annotated scam pages and all 2,242 benign pages (5,658 pages in total); the remaining 4,368 semi-automatically annotated scam pages form the corpus extension and are excluded from the primary metrics.

B.1 Record Contract

Table 4 records the source hierarchy and supervision boundary. The same-stem join modifies no raw screenshot, MHTML, JSON, or workbook record.

Group	Fields	Contract
Identity	`sample_id`, data and policy versions	Frozen before inference and shared by all methods
Input	URL, title, parsed URL fields, screenshot dimensions	Available for all pages; the local image path is consumed only by the image loader
Trusted web evidence	Visible text, buttons, forms, links, metadata, external domains	Derived from MHTML only for `valid` and `valid_redirect` records
Truth	Binary label, one subtype or `none`, original-image $xyxy$ boxes, reasons, description	Evaluation only; never serialized into a stage prompt
Prediction	Binary label, one subtype or `none`, boxes, evidence types, reasons, stage evidence	Strict schema normalized after each stage
Diagnostics	Attempts, latency, token usage, image budget, finish reason, error category	Operational metadata; never treated as semantic evidence

Table 4: Logical record schema and source hierarchy. Gold JSON is authoritative for subtype, description, language, and regions; the workbook contributes URL, title, row identity, and collection metadata without overriding Gold. Evaluation-only fields are removed recursively before prompt construction.

B.2 MHTML Audit and Modality Gate

MHTML is an archive format for aggregate HTML resources (Palme, Hopmann, and Shelness 1999); it is useful precisely because malicious pages frequently disappear. Archive availability alone does not establish relevance. On the 5,658-page primary evaluation set, the independent audit assigns the states in Table 5. Only the first two states expose MHTML-derived evidence to ScamDet. Every evaluation page retains its URL, title, and screenshot. Archive defects affect only MHTML exposure and do not alter the Gold class or sample inclusion.

State	Pages	Exposed	Interpretation
`valid`	5,298	Yes	Archive agrees with the captured page
`valid_redirect`	184	Yes	Explainable source-to-final redirect
`mhtml_mismatch`	161	No	Archive is unrelated or an error page
`missing`	15	No	Corrupt or unavailable archive

Table 5: Frozen MHTML policy.

Nineteen records retain a `gold_input_conflict` flag because their captured input and Gold description may disagree. They remain in all primary metrics under the same modality policy. The frozen manifest, evidence, audit, recapture, sample-list, and box files each carry a SHA-256 digest; the release verifier rejects any mismatch before inference or resume.

C. ScamDet Inference Protocol

C.1 Staged Computation Graph

Table 6 summarizes the stage-specific English prompts. Each instruction requests observable evidence in a strict JSON object rather than hidden chain-of-thought. The visual stage remains independent of URL and text analyses; these evidence streams first interact in the cross-modal stage.

Grounding and rationale generation are represented as separate prediction operations in the computation graph. The executable schema serializes their outputs in one final structured response so that every rationale remains paired with its localized region.

The URL and text stages use GPT-4o-mini (OpenAI 2024). Visual, cross-modal, binary, subtype, grounding, and rationale stages use Qwen3-VL-Flash (Bai et al. 2025; Alibaba Cloud 2026b). The subtype stage consumes the serialized webpage context, prior analyses, and binary decision without a separate screenshot attachment. The rationale stage conditions on the localized regions together with the webpage context and preceding analyses. Predictions from diagnostic GPT-4o calls are excluded from Qwen3 output records and all reported metrics.

C.2 Long-Page Images and Coordinate Recovery

For an original screenshot I of size $W \times H$, the image policy constructs a roughly one-megapixel overview and, when needed, at most seven vertical tiles with 10% overlap. The total decoded visual budget is 8,388,608 pixels, the maximum long side is 7,680 pixels, and each base64 payload must remain below 10 MB. Logical panels are packed into a labeled contact sheet under the same pixel budget; visual, grounding, and rationale stages have access to the overview and tiles, whereas cross-modal and binary stages receive the overview only.

The model predicts $q = (x_1, y_1, x_2, y_2)$ in $[0, 1000]^4$ and an `image_index`. For the selected crop with origin (o_x, o_y) and size (w, h), the mapped original-image box is

$$b = \left(o_x + \frac{x_1 w}{1000},\ o_y + \frac{y_1 h}{1000},\ o_x + \frac{x_2 w}{1000},\ o_y + \frac{y_2 h}{1000} \right).$$

Coordinates are clipped, reordered only through the declared schema, and rejected if nonnumeric, nonfinite, or degenerate. The mapping is deterministic and covered by tests for overview, tile, and boundary cases.

C.3 Context and Output Budgets

URL and text stages admit at most 8,192 input tokens; all later stages admit 4,096 text tokens. Priority retention is URL/title, redirect and domain fields, actionable controls, risk-bearing metadata, and representative body text. Stage output caps are 512 (URL), 768 (text), 1,024 (visual), 1,024 (cross), 256 (binary), 256 (subtype), and 2,048 for the final grounding and rationale outputs. Temperature is zero and Qwen thinking is disabled. These fixed limits standardize context exposure, control inference cost, and reduce irrelevant long-page content across samples.

C.4 Reliability and Recovery Protocol

Connection timeout is 20 s; read timeout is 120 s for text-only calls and 300 s for visual calls. TLS disconnects, empty responses, HTTP 408/429, and HTTP 500/502/503/504 are retryable; HTTP 400/401/403/404 fail immediately. A stage receives at most four immediate attempts with 2, 4, and 8 s exponential delays plus full jitter and honors `Retry-After`. Across resumed processes, the cumulative cap is six attempts. One additional schema-repair call is allowed for malformed JSON.

After every completed stage, an atomic checkpoint records data, prompt, model, and configuration hashes, together with attempts, latency, token usage, finish reason, and image budget. Resumed processes verify these hashes and execute only incomplete stages. Heartbeats and event logs record progress, while safe-pause signaling terminates execution at stage boundaries. A backend-error circuit breaker suspends the run after repeated terminal failures. This protocol preserves fixed prompts, model routing, binary gating, and prediction normalization throughout the experiment.

D. Evaluation and Experimental Details

D.1 Binary and Subtype Metrics

Scam is fixed as the binary positive class. We report accuracy, precision $P = TP/(TP + FP)$, recall $R = TP/(TP + FN)$, and $F_1 = 2PR/(P + R)$. For subtype label k, P_k, R_k, and $F_{1,k}$ are computed one-vs-rest. Macro scores average uniformly over all $K = 25$ labels, for example

$$\mathrm{MacroF}_1 = \frac{1}{K} \sum_{k=1}^{K} F_{1,k}.$$

A false-benign, `none`, or invalid subtype prediction is incorrect but does not create a 26th macro class. Accuracy and micro-F1 would be redundant for this single-label task, so only accuracy is retained alongside macro statistics.

D.2 Static Multi-Box Grounding

Let G_i and P_i be the valid Gold and predicted boxes on page i. For mIoU, exact dynamic programming finds the one-to-one assignment M_i that maximizes total IoU. Unmatched

Stage	Instruction focus	Principal output fields
URL	Brand–domain mismatch, redirects, obfuscation, and scam-relevant URL tokens	`lexical_signals`, `redirect_signals`, `brand_domain_signals`, summary
Text	Persuasion, requested actions, money, credentials, urgency, off-platform contact, and subtype-specific lures	claims, requested actions, payment/credential/contact signals, summary
Visual	Page genre; products and discounts; investment charts; login/payment forms; QR/download prompts; adult, romance, gambling, news, charity, education, bank, and government cues	`page_genre`, visual cues, sensitive/monetary inputs, trust elements, contact redirects, adult/romance lures, candidate regions
Cross	Agreement or conflict across domain, title, trusted text, screenshot, and prior analyses	agreements, conflicts, missing evidence, consistency assessment
Binary	Evidence-backed hard decision using money, credential, contact, adult lure, authority, implausible benefit, and download/QR cues	`binary`, supporting and counter evidence, concise rationale
Subtype	Choose exactly one English label from the supplied 25-class taxonomy	`scam_type`, supporting evidence, alternatives considered
Grounding	Localize visible regions supporting the predicted subtype	`image_index`, `bbox_1000`, `evidence_type`
Rationale	Explain why each localized region supports the predicted subtype	`reason`, page description

Table 6: Prompt objectives and principal schema fields. The complete executable prompt templates are distributed with the code.

Gold boxes receive zero, and

$$\mathrm{mIoU} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \frac{\sum_{(g,p) \in M_i} \mathrm{IoU}(g,p)}{|G_i|},$$

where $\mathcal{S}$ contains Gold-positive pages. At threshold τ, matching first maximizes the number of pairs satisfying $\mathrm{IoU} \geq \tau$ and then their total IoU. Global precision, recall, and F1 use $\tau = 0.5$; every unmatched or invalid prediction, including one on a benign page, is a false positive. Page recall averages $|M_i^\tau|/|G_i|$ at $\tau \in \{0.3, 0.5\}$. We omit mAP because the generated boxes have no calibrated ranking confidence.

D.3 Rationale Metrics and Uncertainty

Rationale similarity is conditioned on the threshold-0.5 one-to-one match. Only pairs with nonempty Gold and predicted reasons enter ROUGE-L F1 (Lin 2004) and BERTScore F1 (Zhang et al. 2020). Grounded rationale coverage divides the number of such pairs by all 5,489 Gold boxes, thereby retaining localization and missing-explanation failures in an end-to-end measure.

BERTScore uses FacebookAI/xlm-roberta-large (Conneau et al. 2020), revision `c23d21b0620b635a76227c604d44e43a9f0ee389`, layer 17, `idf=False`, no baseline rescaling, batch size four, and the recorded CUDA device. The evaluator instantiates one process-wide scorer from a local cache; it does not train or fine-tune XLM-R. The metric status distinguishes `not_requested`, `no_matched_pairs`, `computed`, and `error`.

Binary intervals use 1,000 seed-2027 bootstrap samples over pages. Subtype intervals use a class-stratified bootstrap that resamples within each observed Gold label. The full-set binary 95% intervals are accuracy [67.0, 69.4], precision [64.7, 67.4], recall [96.7, 97.8], and F1 [77.7, 79.6]. End-to-end subtype intervals are accuracy [62.5, 64.9] and macro F1 [32.1, 36.9]. All values in this paragraph are percentages.

D.4 Baselines and Runtime Environment

The fixed seed-2027 supervised split follows a 70/15/15 ratio, yielding 3,961 training, 849 validation, and 848 test pages. It is stratified over benign pages and all 25 scam subtypes. Phishpedia detects identity logos in screenshots and compares the matched brand with the observed URL domain (Lin et al. 2021). PhishIntention combines URL, screenshot, HTML, and interaction evidence to infer brand and credential-taking intent (Liu et al. 2022). Their reported metrics use the 848-page test partition. ScamDet requires no benchmark training and is evaluated zero-shot on all 5,658 pages.

Experiment orchestration uses Python 3.12 on Ubuntu 24.04 ARM64 and an NVIDIA DGX Spark host with a GB10 processor and 121 GiB of system memory. LLM/VLM inference uses hosted APIs: GPT-4o-mini is accessed through Crazyrouter, and Qwen3-VL-Flash is accessed through Alibaba Cloud Model Studio (Alibaba Cloud 2026a); both were accessed on July 12, 2026. A local GPU is not required for ScamDet inference; CUDA is used only for offline XLM-R BERTScore. The environment locks `bert-score==0.3.13`, `transformers==4.57.6`, and the pinned XLM-R revision. Provider, endpoint class, access date, prompt/configuration hash, input-policy version, and per-stage usage are retained with the run.

D.5 Single-Model Zero-Shot Protocol

The direct-model comparison uses the intersection between the 3,416 manually annotated scam pages and the available result artifacts. This produces 3,324 common pages; the remaining 92 pages have no corresponding direct-run result and are excluded from every row, including the recomputed ScamDet row. Each direct model receives only the full-page screenshot and produces binary detection, one subtype, localized regions, and region rationales in a single request.

ScamDet retains its staged multimodal computation. The comparison is therefore an end-to-end system evaluation, not a controlled model ablation.

The following English prompt is shared across direct models:

> **Role and input.** You are a cybersecurity analyst evaluating a webpage screenshot for research. Analyze only the supplied full-page screenshot. Do not use external retrieval, few-shot examples, or unstated webpage metadata.
>
> **Tasks.** Decide whether the webpage is `scam` or `benign`; for a scam, select exactly one subtype from the canonical taxonomy; localize every visible region that materially supports the decision; and explain briefly why each region is suspicious.
>
> **Taxonomy.** Task-Based Commission Scam; E-commerce Customer Service Impersonation Scam; Spoofed Login Page; Online Loan Scam; Online Investment Scam; Cryptocurrency Investment Scam; Government / Law Enforcement Impersonation Scam; Romance Scam; Fake Streaming Service Scam; Online Shopping Scam; Technical Support Scam; Gaming Item Trading Scam; Online Gambling Scam; Charity Scam; Business Impersonation Scam; Job / Recruitment Scam; Lottery / Prize Scam; Website Hijacking Scam; Education / Training Scam; Fraudulent Personal Website Scam; Bank Impersonation Scam; Fake News Website Scam; Phishing; Social Media Scam; Sextortion.
>
> **Output.** Return JSON only with fields `language`, `binary`, `scam_type`, `other_type`, `description`, `reasoning`, and `boxes`. Each box contains a `reason` and `bbox` in normalized `[y_min,x_min,y_max,x_max]` coordinates on a 0–1000 scale. A benign page returns `scam_type="none"`, `other_type=null`, and `boxes=[]`. A scam outside the closed taxonomy returns `scam_type="Other"` and names it in `other_type`.

Temperature is zero, the output ceiling is 8,192 tokens, and each page is requested once without retrieval, few-shot demonstrations, or post-hoc schema repair. Models with native reasoning retain their default reasoning mode. When an image exceeds the upload constraint, a geometry-preserving JPEG surrogate is used without cropping webpage content. Predicted boxes are interpreted as normalized `[y_min,x_min,y_max,x_max]` and mapped to the original screenshot `xyxy` coordinates before the same one-to-one IoU evaluation used for ScamDet.

Backend errors, empty responses, scalar outputs, malformed schemas, and invalid box structures remain in the denominator as failures. `Other` and non-canonical subtype strings are invalid subtype predictions. Provider-reported `total_tokens` is retained without forcing equality to separately reported input and output counts; the average divides this total by all 3,324 pages. On this common subset, ScamDet reaches 0.193 mIoU, 9.9% grounding F1 at IoU 0.5, 89.7% BERTScore, and 15.3% grounded rationale coverage. Its staged computation uses 34.10 million tokens, averaging 10,260 tokens per page. Qwen3.5-VL is excluded from the main comparison because it was attempted on only 1,890 common pages. Those attempts contain 1,403 JSON-object outputs, of which 1,402 satisfy the complete schema. The machine-readable aggregate records every reported metric, schema-failure count, and token total.

	TP	FP	TN	FN
ScamDet	3,321	1,700	542	95

Table 7: Binary confusion counts for the full zero-shot run.

E. Extended Results and Cascade Diagnostics

E.1 Binary Detection and Aggregate Counts

On all 5,658 pages, ScamDet produces 3,321 true positives, 1,700 false positives, 542 true negatives, and 95 false negatives, as reported in Table 7. The resulting 95% bootstrap intervals are accuracy [67.0, 69.4], precision [64.7, 67.4], recall [96.7, 97.8], and F1 [77.7, 79.6]. These counts confirm that the high-recall operating point preserves most scam pages for downstream analysis at the cost of many benign false alarms.

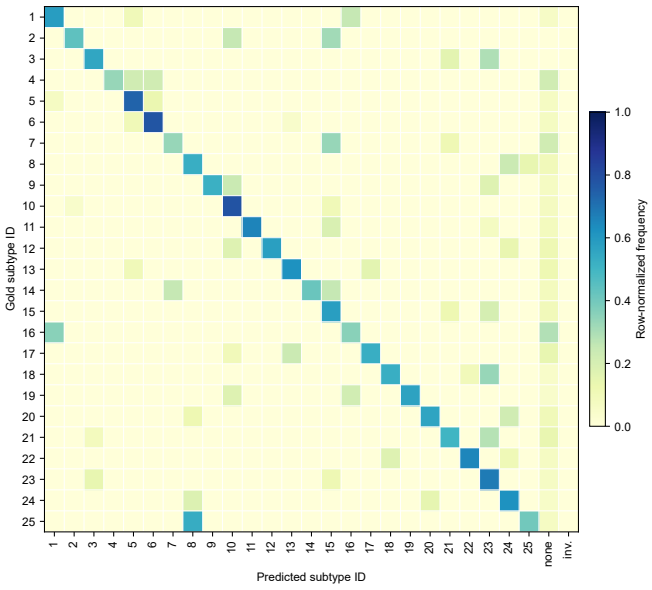


Figure 2: Subtype confusion matrices for the primary end-to-end cascade and the oracle-gated forced diagnostic. Rows are Gold labels in canonical order. The end-to-end matrix additionally includes `none` for binary-gated misses.

E.2 Complete Subtype Results

Table 8 reports the end-to-end results for all 25 scam subtypes. Overall accuracy is 63.7%, whereas macro F1 is 34.5%, with a 95% confidence interval of [32.1, 36.9]. Online Shopping Scam and Cryptocurrency Investment Scam attain F1 scores of 86.3 and 79.2, respectively. By contrast, Fraudulent Personal Website Scam and Sextortion yield no correct subtype predictions. As shown in Figure 2, the former is primarily confused with Online Shopping Scam and Spoofed Login Page, whereas the latter is mainly assigned to Romance Scam and Fake Streaming Service Scam. Their zero F1 scores persist under the oracle-gated forced diagnostic, indicating that the dominant failure arises from fine-grained subtype separation rather than binary gating. Technical Support Scam

ID	Canonical subtype	Support	Precision	Recall	F1
1	Task-Based Commission Scam	99	75.4	52.5	61.9
2	E-commerce Customer Service Impersonation Scam	16	7.5	25.0	11.6
3	Spoofed Login Page	38	7.4	68.4	13.3
4	Online Loan Scam	18	25.0	27.8	26.3
5	Online Investment Scam	330	52.2	58.2	55.0
6	Cryptocurrency Investment Scam	793	80.9	77.6	79.2
7	Government / Law Enforcement Impersonation Scam	9	44.4	88.9	59.3
8	Romance Scam	21	50.0	61.9	55.3
9	Fake Streaming Service Scam	29	51.4	65.5	57.6
10	Online Shopping Scam	1,089	81.3	91.9	86.3
11	Technical Support Scam	154	75.0	1.9	3.8
12	Gaming Item Trading Scam	87	68.2	17.2	27.5
13	Online Gambling Scam	150	74.4	83.3	78.6
14	Charity Scam	12	35.7	41.7	38.5
15	Business Impersonation Scam	83	100.0	2.4	4.7
16	Job / Recruitment Scam	14	30.8	28.6	29.6
17	Lottery / Prize Scam	21	50.0	19.0	27.6
18	Website Hijacking Scam	42	100.0	2.4	4.7
19	Education / Training Scam	23	34.4	47.8	40.0
20	Fraudulent Personal Website Scam	50	0.0	0.0	0.0
21	Bank Impersonation Scam	36	100.0	2.8	5.4
22	Fake News Website Scam	120	73.5	41.7	53.2
23	Phishing	112	29.4	8.9	13.7
24	Social Media Scam	55	57.9	20.0	29.7
25	Sextortion	15	0.0	0.0	0.0

Table 8: End-to-end 25-class single-label subtype results. Precision, recall, and F1 are percentages.

also exhibits a pronounced precision–recall imbalance, with 75.0% precision but only 1.9% recall. Figure 2 further reveals bidirectional confusion between cryptocurrency and general online investment, together with the concentration of technical-support predictions in broader shopping and login-related categories.

The horizontal coordinates and marker supports in the main-paper cross-task figure are taken directly from the F1 and Support columns of Table 8; the accompanying source-data file preserves the same values at full precision.

E.3 Complete Grounding and Rationale Results

The primary cascade predicts 19,173 boxes. At IoU 0.5, 680 boxes are matched, 18,493 predictions remain unmatched, and 4,809 Gold boxes are missed; 833 of the predicted outputs are invalid and counted as false positives. Table 9 records the corresponding forced counts.

Table 10 reports every subtype. Online Gambling Scam is the strongest of the five frequent classes, while Social Media Scam and Spoofed Login Page attain the highest mIoU overall. Online Loan Scam has no IoU-0.5 rationale pair, so its ROUGE-L and BERTScore entries are reported as N/A rather than zero. The vertical coordinates in the main-paper cross-task figure are taken from the F1 column of Table 10, ensuring that the visualization and the complete tables describe the same 3,416-page evaluation. Its descriptive Spearman coefficient is computed from the one-decimal F1 values reported in Tables 8 and 10.

Measure	End-to-end	Forced
Annotated boxes	5,489	5,489
Predicted boxes	19,173	12,940
Invalid predictions	833	393
TP@0.5	680	682
FP@0.5	18,493	12,258
FN@0.5	4,809	4,807
Scored rationale pairs	680	682

Table 9: Grounding and explanation denominators. Invalid boxes are included among false positives but listed separately for auditability.

E.4 Oracle-Gated Forced Diagnostic

The forced diagnostic evaluates all 3,416 Gold scam pages downstream, including the 95 pages stopped by the primary binary decision, while leaving the published end-to-end predictions unchanged. Table 11 shows that subtype accuracy rises from 63.7% to 64.9%, but macro F1 remains essentially unchanged. Strict matches increase only from 680 to 682. Grounding precision rises partly because the forced diagnostic excludes benign pages and their predicted boxes; it therefore isolates downstream behavior rather than replacing the primary evaluation.

F. Blinded Human Evaluation Protocol

Automatic similarity cannot establish whether an explanation is factually correct or faithful to visible evidence. A complementary audit samples 200 IoU-matched regions, stratified

Subtype	mIoU	P	R	F1	R-L	BERT	Cov.
Bank Impersonation Scam	0.206	6.0	15.1	8.6	15.3	87.2	15.1
Business Impersonation Scam	0.110	4.2	7.8	5.4	9.5	86.7	7.8
Charity Scam	0.226	6.7	21.4	10.2	15.9	87.9	21.4
Cryptocurrency Investment Scam	0.214	6.4	15.7	9.1	13.2	87.8	15.7
E-commerce Customer Service Impersonation Scam	0.284	10.3	28.6	15.2	14.5	88.0	28.6
Education / Training Scam	0.180	7.1	15.4	9.7	10.0	86.7	15.4
Fake News Website Scam	0.152	3.6	10.7	5.4	11.1	86.9	10.7
Fake Streaming Service Scam	0.122	3.6	8.9	5.1	15.0	88.2	8.9
Fraudulent Personal Website Scam	0.155	4.6	11.1	6.5	11.1	87.3	11.1
Gaming Item Trading Scam	0.203	5.7	18.0	8.7	11.6	87.6	18.0
Government / Law Enforcement Impersonation Scam	0.189	8.8	15.0	11.1	13.7	88.3	15.0
Job / Recruitment Scam	0.070	1.8	4.5	2.6	24.4	89.7	4.5
Lottery / Prize Scam	0.331	9.8	34.8	15.2	16.8	88.7	34.8
Online Gambling Scam	0.279	9.2	26.1	13.6	15.1	87.5	26.1
Online Investment Scam	0.164	5.4	12.3	7.5	13.6	87.9	12.3
Online Loan Scam	0.072	0.0	0.0	0.0	N/A	N/A	0.0
Online Shopping Scam	0.095	2.9	5.8	3.9	10.2	87.2	5.8
Phishing	0.294	8.7	26.4	13.1	16.0	87.7	26.4
Romance Scam	0.178	3.9	12.0	5.9	8.4	86.7	12.0
Sextortion	0.136	5.0	11.1	6.9	10.1	87.0	11.1
Social Media Scam	0.501	16.9	47.5	24.9	13.6	87.7	47.5
Spoofed Login Page	0.421	12.9	42.5	19.8	21.3	88.6	42.5
Task-Based Commission Scam	0.270	10.7	29.2	15.6	16.4	88.1	29.2
Technical Support Scam	0.137	3.7	10.2	5.4	16.0	87.8	10.2
Website Hijacking Scam	0.120	3.1	7.3	4.3	11.1	86.2	7.3

Table 10: End-to-end grounding and rationale results for all 25 scam subtypes. P/R/F1 use IoU 0.5; R-L, BERT, and Cov. denote ROUGE-L F1, BERTScore F1, and grounded rationale coverage. Except mIoU, values are percentages.

Task	Metric	End-to-end	Forced
Subtype	Accuracy	63.7	64.9
Subtype	Macro precision	52.2	51.7
Subtype	Macro recall	37.4	37.6
Subtype	Macro F1	34.5	34.4
Grounding	mIoU	0.173	0.174
Grounding	Precision@0.5	3.5	5.3
Grounding	Recall@0.5	12.4	12.4
Grounding	F1@0.5	5.5	7.4
Grounding	Page recall@0.3	23.5	23.6
Grounding	Page recall@0.5	14.5	14.5
Rationale	ROUGE-L F1	13.4	13.3
Rationale	BERTScore F1	87.6	87.6
Rationale	Grounded coverage	12.4	12.4

Table 11: Primary cascade versus oracle-gated forced diagnostic. All values except mIoU are percentages.

by subtype frequency and outcome, and assigns each region to three independent raters who are blind to the model identity and Gold wording. The interface displays the screenshot with one predicted box and the generated reason; it does not display the predicted subtype unless it is required to interpret the claim.

Raters score four ordinal dimensions on a five-point scale: *correctness* (the stated cue is factually present), *faithfulness* (the reason is supported by the boxed pixels), *specificity* (the reason identifies a concrete cue rather than a generic suspicion), and *helpfulness* (the explanation would assist a defensive analyst). Per-dimension reports include mean, standard deviation, and a page-bootstrap 95% interval. Inter-rater reliability is ordinal Krippendorff's alpha (Krippendorff 2018). Disagreements are preserved; adjudication, if used to create a separate qualitative analysis, cannot replace the independent scores. Human ratings are not imputed when the audit has not been performed.

G. Qualitative Error Audit

The qualitative audit uses the mutually exclusive primary codes in Table 12. A secondary code may record the responsible modality or page region. This coding scheme prevents a fluent rationale from being counted as successful when the corresponding box or subtype is wrong.

The four main-paper cases instantiate distinct outcome codes while excluding explicit imagery and input-conflict records. Sample 16564 represents Success; sample 3372, Binary propagation; sample 15690, Subtype confusion; and sample 66962, Boundary displacement. All four records have valid MHTML, set `gold_input_conflict=false`, and retain their original predictions. Main-paper Figure 3 presents contextual crops around the focal annotations, while the complete screenshots and predictions remain in the released records. At the final availability check on July 20, 2026, all four domains returned no A, AAAA, or CNAME record, and HTTP/HTTPS connections failed during name resolution.

Code	Observable outcome	Interpretation
Success	Correct binary and subtype; at least one strict Gold match; reason is visibly supported	Complete cascade success
Binary propagation	Gold scam predicted benign; subtype `none`; no boxes	Upstream gate suppresses every downstream output
Subtype confusion	Scam detected but one neighboring subtype selected	Intent or impersonated role remains ambiguous despite risk detection
Boundary displacement	Correct cue identified at IoU 0.3 but not 0.5	Localization is coarse, shifted, or spans the wrong visual extent
Unsupported rationale	Box matches but the reason asserts content absent from the region	Text fluency exceeds visual faithfulness
Benign over-grounding	Benign page predicted scam with one or more boxes	Salient legitimate trust, payment, or login elements trigger a false alarm
Provider refusal	Safety filter or backend rejects an otherwise valid research image	Operational failure; retained as an error, never model-swapped

Table 12: Case-analysis coding scheme.

H. Reproducibility and Responsible Release

The repository provides dataset auditing, frozen-manifest verification, prompt construction, image preprocessing, staged inference, checkpoint/resume, evaluation, BERTScore caching, fixed-split generation, and baseline runners. Unit tests cover taxonomy aliases, supervision hiding, MHTML gating, token truncation, image tiling, coordinate recovery, retry policy, JSON repair, resume hashes, model routing, binary and subtype metrics, multi-box matching, benign false boxes, and rationale coverage.

Persistent archives can contain executable scripts, credentials, personal data, explicit imagery, copyrighted material, or active payment/contact endpoints. A research release must disable active execution, remove secrets and unnecessary identifiers, document provenance and takedown procedures, and gate high-risk artifacts. Release documentation must preserve the expert-annotation and adjudication protocol, the annotation schema, the 19 flagged Gold-input conflicts, and the frozen MHTML audit policy. The benchmark and code are intended for defensive research and must not be used to optimize scam delivery or safety-filter evasion.

References

Alibaba Cloud. 2026a. Alibaba Cloud Model Studio: Models. Alibaba Cloud Model Studio documentation. Accessed July 12, 2026.

Alibaba Cloud. 2026b. Qwen3-VL-Flash Model Information. Alibaba Cloud Model Studio documentation. Accessed July 29, 2026.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025. Qwen3-VL Technical Report. arXiv:2511.21631.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Krippendorff, K. 2018. *Content analysis: An introduction to its methodology*. SAGE Publications, 4 edition. ISBN 9781506395661.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Lin, Y.; Liu, R.; Divakaran, D. M.; Ng, J. Y.; Chan, Q. Z.; Lu, Y.; Si, Y.; Zhang, F.; and Dong, J. S. 2021. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)*, 3793–3810.

Liu, R.; Lin, Y.; Yang, X.; Ng, S. H.; Divakaran, D. M.; and Dong, J. S. 2022. Inferring phishing intention via webpage appearance and dynamics: A deep vision based approach. In *31st USENIX Security Symposium (USENIX Security 22)*, 1633–1650.

OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence. OpenAI. Accessed July 12, 2026.

Palme, J.; Hopmann, A.; and Shelness, N. 1999. MIME Encapsulation of Aggregate Documents, such as HTML (MHTML). RFC 2557, Internet Engineering Task Force.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.