Scam & Phishing Protection Evaluation Methodology

Table of Contents

1. Goal & Environment 2

2. Delivery methods & Scam types 2

3. Malicious Corpus 2

4. Legitimate Corpus 3

5. Test Environment 3

    Shelf Identities 3

    Solutions Tested 4

6. Scoring 4

    Malicious rating – on device protection 4

    Malicious Rating– AI Assistant-like components 4

    Education Rating 5

    Legitimate Rating - on device protection 5

    Legitimate Rating – AI Assistant-like components 5

    Total Accuracy 5

7. Change Log 6

1. Goal & Environment

This evaluation aims to measure how security tools & platforms identify and prevent scams & phishing. The evaluation covers whole solutions and not a single OS or platform, as such where possible all security layers are installed (security application, browser extensions, E-mail/SMS integration). The test environment can consist of:

- Windows 11
- MacOS Sequoia
- Android devices
- iOS devices

Specific version numbers of the OS, applications used, AI version (where applicable) is disclosed in the final report.

2. Delivery methods & Scam types

In the wild, scams are delivered through a wide variety of communication channels. The methodology allows for a flexible approach of delivery to support the varied way users interact & the protection solutions available.

Delivery vectors considered in scope:

- Emai/Web phishing
- Text Messaging
- Social media messaging apps
- Social media platforms & Job advertising platforms

Scam types considered in scope:

- Impersonation via Deep Fake
- AI Generated & Real Actors video scams
- Job recruitment
- Investment
- Tech support
- Romance/pig butchering scams
- ClickFix, FileFix, FakeCaptcha
- Fake/Fraudulent Shops

3. Malicious Corpus

The attack corpus consists of original scenarios and crafted scenarios based on the same techniques employed by attackers.

The corpus can be modified for a specific GeoLocation. While there are crossovers in techniques employed by scammers worldwide, specific regions may have different distribution of the types of scams. Upon publication test announcement any targeted locations will be identified in advance of test commencement.

Original scenarios are served directly to the target as they are received from real scammers.

Crafted scenarios are used to capture the relevant threat landscape at the time of test without relying on Honeypots. Each crafted scenario has a unique domain.

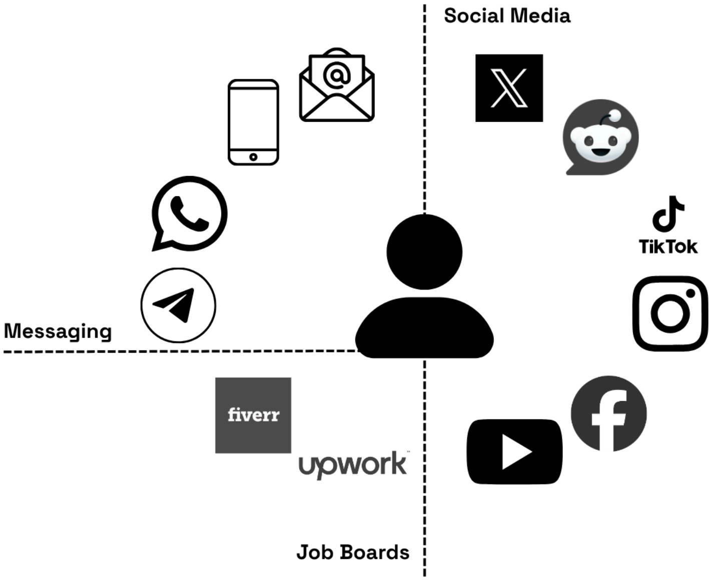
4. Legitimate Corpus

The legitimate corpus mirrors the malicious corpus in size, delivery methods and techniques used.

5. Test Environment

Shelf Identities

Throughout the engagement period our "Shelf Identities" are subject to prevalent scams as a real user would be. We create and maintain digital identities as potential targets for scammers throughout the calendar year to collect threat intelligence.



Our Shelf Identities have a footprint in the following:

- Social Media: TikTok, Instragram,Facebook, Youtube, Reddit, X
- Messaging: Whatsapp, Telegram, registered phone numbers, E-mail
- Job Boards: Upwork, Fiverr

Solutions Tested

The methodology is solution-agnostic. Each solution tested can provide multiple components working together or independent of each other.

Each public report also has a "platform" rating. For each of the delivery methods, in-built security provided by platforms are available and constantly changing. While our Shelf Identities use less secure deployments, during testing a standard configuration of a target user is tested alongside the commercial offering.

6. Scoring

The threat is delivered to a system with active protection capabilities. I.e. if the tested solution offers phishing protection via email protection it is delivered as such for the solution to act upon the scam. Each component of the solution is given a chance to detect or prevent during the threat chain.

Malicious rating – on device protection

Blocked **(+10)** – The user is prevented from interacting with the threat.

Notified – block **(+10)** - The user is notified of the threat upon interaction with the threat.

Allowed/Notified Allowed **(0)** – No malicious verdict is given.

Malicious Rating– AI Assistant-like components

For each scenario the AI assistant is presented with screenshots at a variety of points of scam protection. Its scoring is based on the confidence of the verdict and whether the user is discouraged from continuing with the scam interaction.

	True positive		*False Negative*
Rating/advice	**Discouraged (explicit stop)**	**Cautioned**	**Permitted (false negative)**
Definite	+10	+8	0
Unsure	+10	+5	0

Context rating is tracked for informational purposes in the report and is calculated as the number of questions the AI asks before reaching a verdict, excluding consent requests to follow links.

Education Rating

Education rating – AI has the opportunity for user education when prompted for a decision by the user. The education rating is given as a separate rating and is not part of Total Accuracy.

Accuracy	**Score**
Completely	+10
Mixed	+8
Inaccurate	0

Legitimate Rating - on device protection

Allowed/Notified Allowed **(+10)** – No malicious verdict is given.

Blocked **(0)** – The user is prevented from proceeding.

Legitimate Rating – AI Assistant-like components

For each scenario the AI assistant is presented with screenshots at a variety of points of scam protection. Its scoring is based on the confidence of the verdict and the verdict itself.

	True Negative		*False Positive*
Rating/advice	**Permitted (confirmed clean)**	**Cautioned**	**Discouraged (explicit stop**
Definite	+10	+8	0
Unsure	+5	+5	0

Total Accuracy

Total Accuracy is calculated as $Complete\ Malicious\ Rating + Complete\ Legitimate\ Rating$

7. Change Log

12/12/2025 - Fixed Malicious Rating and Legitimate Rating Tables typos

20/11/2025 - v1.5 Document Updated - Identifier - ScamPhishing2025v1.5

- Added new Scam Types considered in scope
- Unique domains for each crafted scenario in the malicious corpus
- "PUA" Removed
- Education rating - added as a separate rating
- Rating overhauled
- Geospecific Scope considered
- Platform security considered as base results
- Changed the targeted solutions for this testing.

06/08/2025 – v1 Document created – Identifier – ScamPhishing2025v1.0