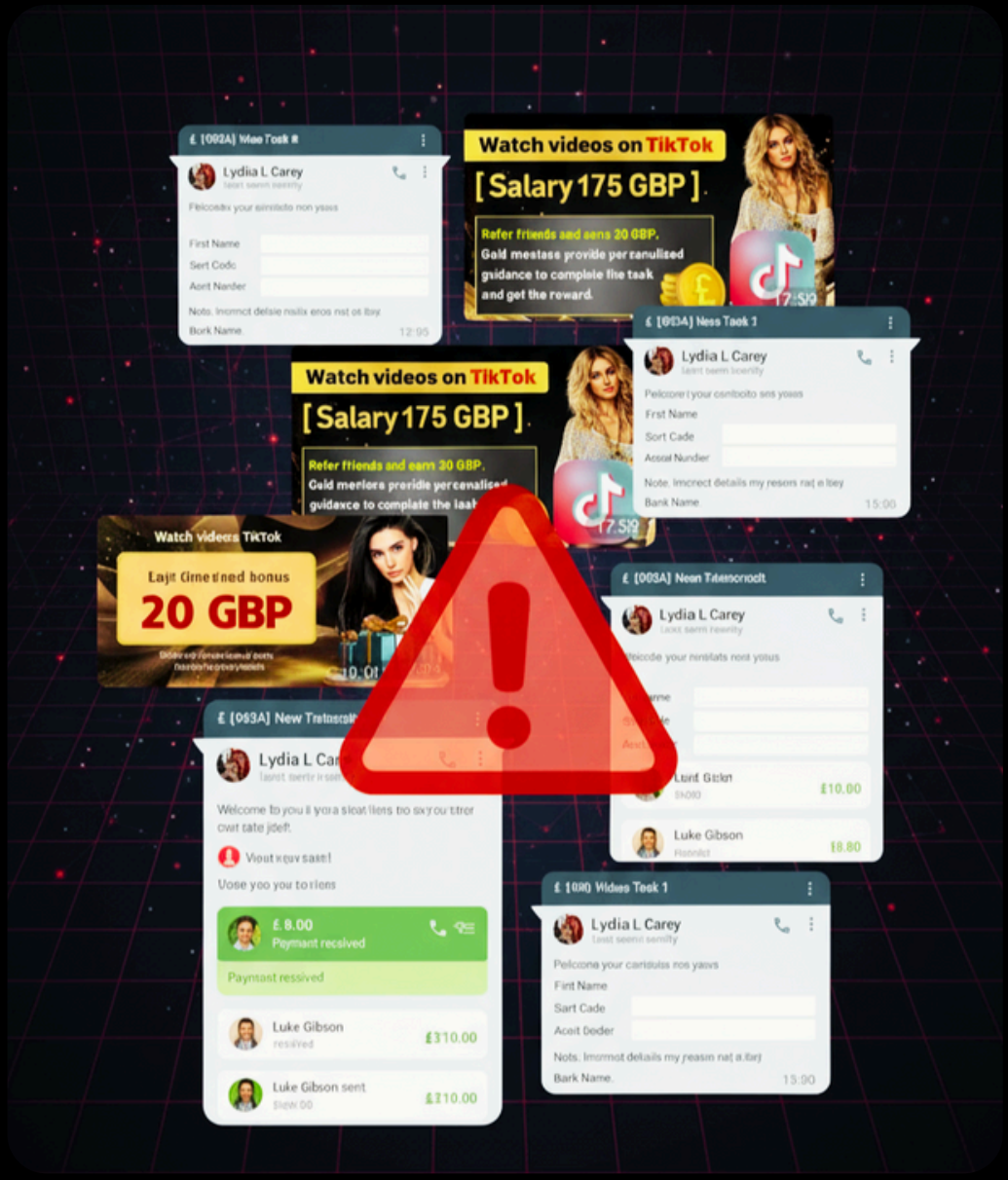
ARTIFACT SECURITY



Bitdefender Premium Security & Scam Copilot

Scam & Phishing Evaluation

Aug - Sep 2025

amtso
The cybersecurity industry's testing standard community
CONFIRMED COMPLIANT

AMTSO Tracking link can be found here.

Prepared by

Stefan Dumitrascu & Ana Maria Pricop

Table of Contents

Introduction

Stefan Dumitrascu

Chief Executive Officer

The Scam Spam "paradox"

Most users expect security products to block all "unwanted or bad things". However, security products have to deal with classifying unwanted messaging and treat them differently.

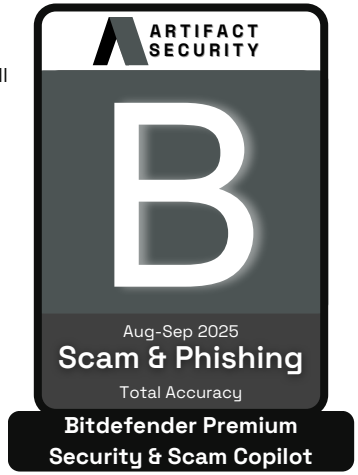
These expectations often conflict with the technical realities of security products. Most individuals simply want all unwanted or potentially harmful messages to be blocked, but the distinction between "scam" and "spam" isn't always obvious. Technically identifying a scam is challenging, but the users don't care for that.

Security products have a new ally in dealing with Scams & Phishing, AI Assistants.

Not every disappointing online purchase constitutes a scam, while classic fraud - such as fake penalty payment demands - always qualifies. Users may not notice these nuances, but for security products accurate classification is essential for both effectiveness and user trust. Security products have a new ally in dealing with this connundrum, AI Assistants.

Protecting against phishing and scams is ever a more complex problem. As users we interact with the internet across almost all our devices. This makes the attack surface ever increasing. We all have multitude of communication applications, part of multiple social media platforms. Attackers use this to their advantage. A threat chain is all too familliar, however a scam communication chain can jump between multiple applications and even devices. Security products therefore have to keep an eye on multiple attack vectors, sometimes piece them together for context.

We've evaluated a series of products that use AI to tackle Scams & Phishing. While we have our own rating attributed to the test scenarios presented we also shared all the raw results on our Github for readers who want to come up with their own rating.



Executive Summary

Evaluation notes

Our evaluation focuses it's rating on assessing the steps of a security concious user throughout multiple attack types. During a phishing scenario the suite is given the chance to block or detect the initial contact method or the landing page. We assume that a user can submit a question to the AI assistant upon seeing the email in their inbox, when opening the email or when getting to the final landing page.

Scam Scenarios are very tricky to test live as the campaigns can be very short and sometimes the scammers would catch on. As such we submitted a query at each stage that would require the user to either submit personal information or the communication is moved to another platform (Whatsapp to Telegram)

Overall the products evaluated in this round performed well but we're still very sensitive to the age of domains at the time of exposure. AI Assistants cover this weakness.

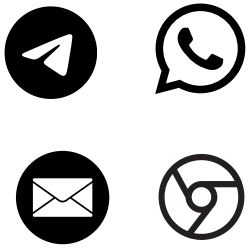
Our Total Accuracy is the sum of Malicious Score (Scam + Phishing rating) and Legitimate Score. You can find the detailed results over on our Github if you would like to perform your own calculation.

Our highest rating tier, S-uper, is reserved for products with exceptional Total Accuracy. While minor differences exist between products at this tier, users can trust they're receiving top tier protection.

Bitdefender Evaluation Highlights	Grade	Score	
Total Accuracy	B	79%	→
AI Assistant Score	S	94%	→
Scam Protection	S	100%	→

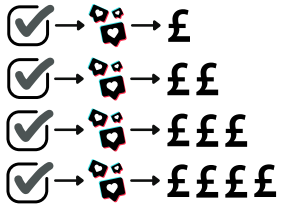
A new test for a complex landscape

It's important to give a realistic approach to how the threats are presented to the security solution. We took the approach of a security concious person who knows the full suite of the product they are buying. Where possible we enrolled into Email Security, Browser Extensions and also used the dedicated AI Assistants to check.



Scammers take advantage of the multitude of ways users exchange informations with each other or with trusted parties. We used the concept of "shelf babies" for attracting scammers. We created fake personas, with Social Media accounts, signed up for newsletters, job boards etc.

A very convincing scam is recruiters that offer jobs for unreasonable sums of money for "easy tasks" such as liking TikTok posts. These were encountered by our targets. What makes these convincing is the fact that to start with you do get a small payout for liking a post.



During testing we noticed that a prevelant type of scam is impersonating a legitimate recruitment company. While we incorporated the samples in our corpus we also notified the actual company of the ongoing campaign abusing the trusted brand.

TL & DR:

Bitdefender's AI Assistant shined amongst it's other components earning an S rating for it's performance across both scam and phishing parts of the test. Its pitfalls came when dealing with legitimate websites earning 60% accuracy. When contacted about the result the Bitdefender team responded with suggestions about the prompt used. After modifying and re-testing the scoring changed in one instance of the FP however it failed progress the score significantly as the other instances were unaffected by the new prompt.

Scam Copilot can be accessed via it's dedicated portal or via Whatsapp which makes it particularly easy to use. Through our testing we've noticed very detailed reasoning behind it's verdict which can empower users with the knowledge required to convict or allow and over time not be so reliant on the technology itself.

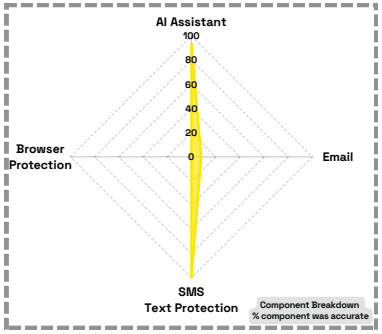
Malicious Results

Our malicious corpus consists of Scams found in the wild by our "shelf babies" honeypots and crafted attacks. In this run of the test we focused on Phishing via e-mail, specifically via online banking portals throughout the EU. Our rating system expected the security solution to discourage the user from proceeding with the threat either via blocking access, tagging the email as unsafe or the AI assistant decision. You can access all the tested scenarios and how each of the components dealt with the multi-stage communication in our Github page linked in the Conclusion section of our report.

An example of scammers using multiple applications for communicating with their target is the "HR recruitment for watching TikTok Videos". These start by them introducing an offer for multiple hundreds of pounds for watching videos all days. After sending an initial link to a post and providing proof you liking you are then introduced to their manager/receptionist via Telegram and a code to "verify" you. Upon reaching out to the manager you can actually get paid a small amount of money for the initial like. Usually around the £10 mark. They post the proof in a Telegram group chat.

This gets used in the group chat to reinforce the "fact" that people do indeed get money. This is where tasks are given in exchange for doing tasks. The deeper you go in the tasks will end up requiring a "deposit" to execute the "task". This is where you will see victims starting to lose money as the pricing for deposits go into the hundreds and even thousands. Everything in the group chat is dominated by urgency, "attractive" AI generated women and unfortunately real victims that are convinced by the initial payout. The moderators in these group kick out inactive members within 24 hours.

Our scenario based scoring assesses whether a user using the full suite of products available from the vendor will be protected. Due to the complexity of scams and phishing each component has a chance to detect the threat. A perfect product with each component detecting at every instance would fill up all the corners in the Component Breakdown figure. We can see that Bitdefenders strengths are in their SMS prevention and AI assisted decision making when convicting scams.



Bitdefender Premium Security & Scam Copilot Malicious Protection Accuracy

Product	Scam Score*	Scam %	Phishing Score*	Phishing %
Bitdefender Premium Security & Scam Copilot	70	100%	70	78%

*Higher score is better

Legitimate Results & User Education

False positives or Legitimate results are important in any testing. For this first iteration of the testing we've kept the false positives focused on AI Assistants verdicts on legitimate websites by submitting the same portals we used for our crafted attacks. When navigating to these pages using a protected device (mobile, PC or Email) there were no false positives recorded.

With AI assistants coming in to play, there is also a question of user education and the language used. At the moment we are tracking these internally and we are evaluating ways to reflect behaviour. For example, if an assistant miscategorises a threat while still protecting a user does it really matter? Part of the value proposition of AI assistants is educating the users to recognise scams in the future. Giving wrong information should be a market differentiator, however language interpretation differs between cultures and even personal backgrounds. It's no longer just 1s and 0s. It's 1s, 0s and "probably" this kind of threat.

We welcome feedback on this topic on how you would like this evaluated in future iterations of this testing.

Bitdefender Premium Security & Scam Copilot Legitimate Accuracy

Product	Legitimate Score	Legitimate Accuracy %
Bitdefender Premium Security & Scam Copilot	45	60%

*Higher score is better

Useful Links

Methodology & Detailed Results	**Github.com/Artifact-Security/Scam-Phishing-Product-Evaluations**
Contact Us	**info@artifactsecurity.co.uk**
FAQs	**artifactsecurity.co.uk/faqs**

amtso
The cybersecurity industry's testing standard community
CONFIRMED COMPLIANT

This test unsponsored, funded solely by Artifact Security & run according to the AMTSO Standard.

AMTSO Tracking Link can be found here.