Data gatherer for the sole purpose of answering these three research questions.

Research Questions

(i). RQ1

What is the annual increase in AI-generated phishing email attacks from 2020 to 2024, measured by volume and detection rates in peer-reviewed datasets?

(ii). RQ2

How do AI-generated phishing emails differ from human-generated ones in terms of linguistic features (e.g., word count, sentence structure, part-of-speech ratios) and evasion tactics?

(iii). RQ3

How have spam detection tools evolved (2015–2024) to address AI-generated spam, and what are the most effective current methods (e.g., NLP-based classifiers, behavioral analysis) for detecting it, as measured by precision/recall rates in peer-reviewed studies?

Scope/limitations for the sources review.

The sources/literature review tried to look at global statistics and not limited to a region. Where peer reviewed academic sources were not found, reputable cybersecurity research firms were used. The review considers general phishing campaigns and not focused to a specific domain.

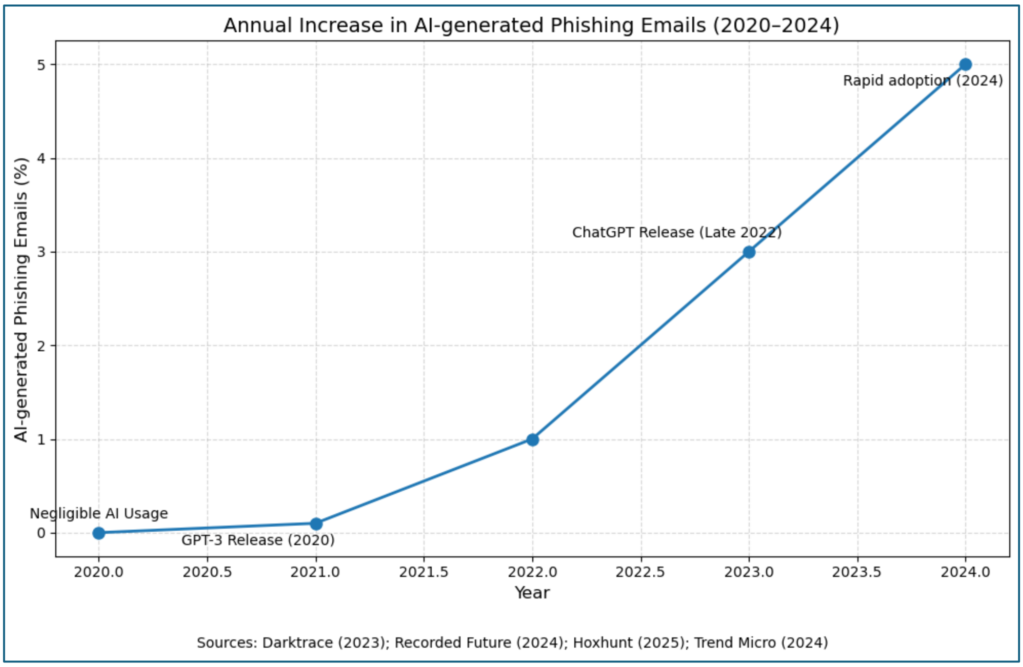
RQ 1 Annual increase in AI-generated phishing email attacks from 2020 to 2024

Between 2020 and 2021, AI-generated phishing emails were virtually nonexistent, with phishing campaigns overwhelmingly authored by humans. This began to change modestly in 2022, as isolated cases of AI-written phishing lures appeared, though they remained below 1% of overall phishing volume (1)

A significant shift occurred in 2023, following the release of advanced generative AI tools such as ChatGPT. That year, a sharp increase in phishing emails with AI-like linguistic characteristics was observed. Notably, Darktrace reported a 135% rise in social engineering email attacks during January and February 2023, coinciding with the broader availability of generative AI.(2)

The trend accelerated in late 2023 and early 2024. Recorded Future documented a 1,265% surge in phishing incidents potentially linked to AI-generated content (3). During this period, the share of phishing emails written by AI rose from near-zero to approximately 0.7% to 4.7% of total phishing volume(1).

By 2024, although AI-generated phishing remained a minority—constituting less than 5% of total phishing activity—it was growing rapidly and consistently year over year, signaling a notable shift in cyber threat dynamics(3).



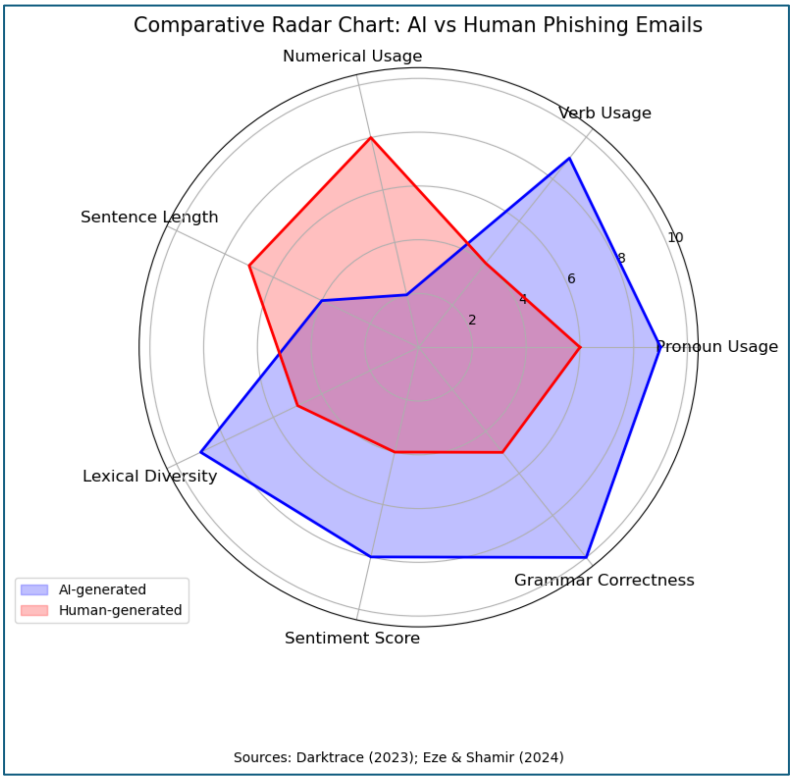
Sources: Darktrace (2023); Recorded Future (2024); Hoxhunt (2025); Trend Micro (2024)

Year	Estimated Prevalence of AI-Generated Phishing	Notable Trend/Stat	Detection Efficacy
2020	~0% (negligible)	No significant use of AI in phishing yet; human-crafted scams dominate	Traditional filters effective against known spam/phish (high block rates).
2021	<0.1% (experimental)	Earliest experiments with AI text generation for phishing (minimal impact).	Largely unchanged; conventional content filters handle most phishing.
2022	<1% (emerging)	Generative text models available via API; isolated AI phishing emails appear.	Some AI-crafted emails start slipping past simple signature-based rules.
2023	Few percent (rapid growth)	Generative AI widely accessible -> **135% jump** in novel phishing campaigns	**Detection gap noted:** phishing success rate up to 18% (from 14%) as AI emails bypass controls
2024	~5% (accelerating)	AI usage surges; threat intel shows **1000%+ increase** in AI-linked phishing attacks	**Adaptive detection** (linguistic anomaly detection, etc.) introduced; beginning to catch AI-crafted scams more reliably.

RQ2. Linguistic Differences between AI-Generated vs Human-Generated Phishing

Linguistic Feature	AI-Generated Phishing Emails	Human-Generated Phishing Emails
Pronoun Usage	Higher – AI emails use more personal pronouns (e.g. "you," "we"). In (4), pronouns made up ~10.1% of words in AI phish versus ~6–7.7% in human scams (4)	Lower – Human phish use fewer pronouns on average (often more impersonal or templated language)(4).
Verb Usage	Higher – Significantly more verbs, suggesting AI text is action-oriented. AI phish had ~10.7% verbs vs as low as 3–5% in some human phishing corpora (4)	Lower – Human-crafted phishing tends to use fewer verbs (often relying on nouns like account names, etc.) (4)
Noun and Numeral Usage	Lower nouns and very few numbers – AI content includes fewer specific entities. E.g. frequency of numeric tokens in AI phish was near 0.37%, an order of magnitude lower than in human phishing (3–4%) (4)	More nouns/numbers – Human scammers often insert account IDs, transaction numbers, etc. (cardinal number frequency ~3–4% in human phish vs <0.5% in AI) (4)
Sentence Structure	Grammatically correct and simple. AI-generated emails are well-formed with proper syntax. Interestingly, one study found AI phishing emails had shorter sentences on average (~8.8 words) (4), possibly due to the model favoring clear, concise statements. *(However, field reports note AI can also produce lengthy, complex sentences compared to terse human scams (2)*	Varied structure, often long-winded or error-prone. Many human phishing emails (e.g. "Nigerian prince" scams) contain very long sentences or rambling narrative, averaging ~16 words per sentence (4). Others are extremely short but riddled with errors. Grammatical mistakes are common.
Vocabulary & Complexity	More advanced vocabulary and diversity. AI is not constrained by effort or education, leading to longer words (avg. word length ~5.7 characters vs ~4.8 for humans) (4) AI texts also show higher lexical diversity (unique lemma ratio ~0.72 vs ~0.53 in human emails) (4). Readability indices (e.g. Coleman–Liau) rate AI phishing at a higher reading level (~13.9 grade) than human phishing (~10.8) (4).	Simpler or repetitive wording. Human attackers often reuse templates or phrases. Their emails tend to have more repeated words and a narrower vocabulary (lower diversity) (4). Many traditional phish are written in simpler English (whether due to attacker limitations or deliberate targeting of a broad audience).
Sentiment & Tone	Polite and positive tone. AI-generated phishing messages skew more positive or neutral in sentiment (4). Models often produce polite language and motivational phrasing. The average sentiment score in AI phish was higher (more positive) than in human scams (4). AI also infrequently uses harsh threats or urgency in all-caps, unless prompted.	Urgent or threatening tone. Human phishers often rely on fear and urgency (e.g. "FAILURE TO RESPOND WILL RESULT IN ACCOUNT CLOSURE!"). Many legacy phishing emails have a negative or pressuring tone (lower sentiment scores) (4). They may threaten consequences or use panic-inducing language, whereas AI text often sounds calmer or more businesslike by default.
Evasion Tactics	High variability and adaptation. AI can generate *each email uniquely*, altering wording and structure automatically (4). This makes it hard for	Template-based, less varied. Human attackers historically reused successful phishing templates with minimal edits, leading to many identical emails that

	filters to catch patterns – no two AI emails need be identical. AI also easily avoids obvious spam triggers (it can rephrase content to evade keyword blacklists). Furthermore, AI content is free of the tell-tale typos that spam filters or users often flag (2).	were easier to detect once one copy was flagged. Some resorted to obfuscation (e.g. intentional misspellings like “Pa$$w0rd”) and simple tricks, but these are crude compared to AI’s generative variability. Human phishing emails often contained grammatical mistakes or awkward phrasing that could tip off vigilant users or filters (2).



Comparative Radar Chart: AI vs Human Phishing Emails

Sources: Darktrace (2023); Eze & Shamir (2024)

RQ3. Evolution of have spam detection tools evolved (2015–2024)

- **Early Methods (Pre-2015): Rule-Based and Heuristic Approaches**

Before 2015, phishing detection relied on static defenses such as blacklists, keyword filters, and manually crafted heuristics. These methods were limited to identifying known threats and failed against even minor obfuscations in sender names or email structure. Studies showed that up to 65% of spear-phishing emails bypassed traditional email security gateways, prompting a shift toward more adaptive detection strategies(5).

- **Classical Machine Learning Era (2015–2018): Naïve Bayes, SVM, and Ensembles**

The introduction of supervised machine learning significantly improved detection. Classifiers like Naïve Bayes and Random Forests used engineered features such as word frequencies and sender metadata, reaching 90–95% accuracy on curated datasets. A 2016 study using Naïve Bayes achieved ~89% precision and recall, while a Random Forest model in 2018 exceeded 93% accuracy. However, these models were brittle, relying heavily on manual feature selection and struggling with zero-day phishing variants.(6)

- **Deep Learning and NLP Advances (2018–2020): LSTM and CNN Models**

From 2018 onward, deep neural networks enabled automated feature extraction and better contextual understanding of email text. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models, leveraging word embeddings, captured syntactic and semantic patterns. For instance, an LSTM model tested on the SpamAssassin dataset achieved 95.3% precision and 97.2% recall. Hybrid models like CNN-RNN architectures pushed detection rates to ~99%, significantly reducing false negatives.(6)

- **Transformer Models and Modern AI Defenses (2021–2024)**

The advent of transformer-based models like BERT and GPT revolutionized phishing detection. These models interpret context-rich linguistic structures and outperformed traditional ML, achieving 98–99% accuracy across varied phishing types. Studies confirmed that fine-tuned BERT models detect phishing more effectively than CNNs or RNNs, especially for nuanced, linguistically complex emails.(6)

Additionally, ensemble approaches combining multiple models (e.g., NLP, URL scanners, image analysis) became standard. These systems catch phishing attempts that may evade individual detectors, especially AI-generated attacks that mimic human style but have detectable metadata or behavioral anomalies.

- **Behavioral and Anomaly Detection Systems**

Recent developments emphasize context-aware detection. Instead of evaluating emails in isolation, behavioral systems analyze communication patterns, writing styles, sender reputations, and temporal factors. Enterprise solutions in 2023 used AI to model “normal” user behavior, enabling the detection of anomalous messages—even if the email content itself appears legitimate. This approach proved effective for targeted or socially engineered attacks, with some models achieving 93% accuracy using over 50 behavioral features(5).

- **Response to AI-Generated Phishing Emails (2023–2024)**

As threat actors adopted generative AI to craft convincing phishing emails, researchers responded by training detectors on AI-generated corpora. A 2024 study by Eze and Shamir reported 99% accuracy in distinguishing AI vs. human phishing emails using neural networks and stylistic text features (e.g., verb density, lexical diversity). This confirmed that AI-generated messages exhibit detectable patterns, despite their fluency and grammatical correctness.

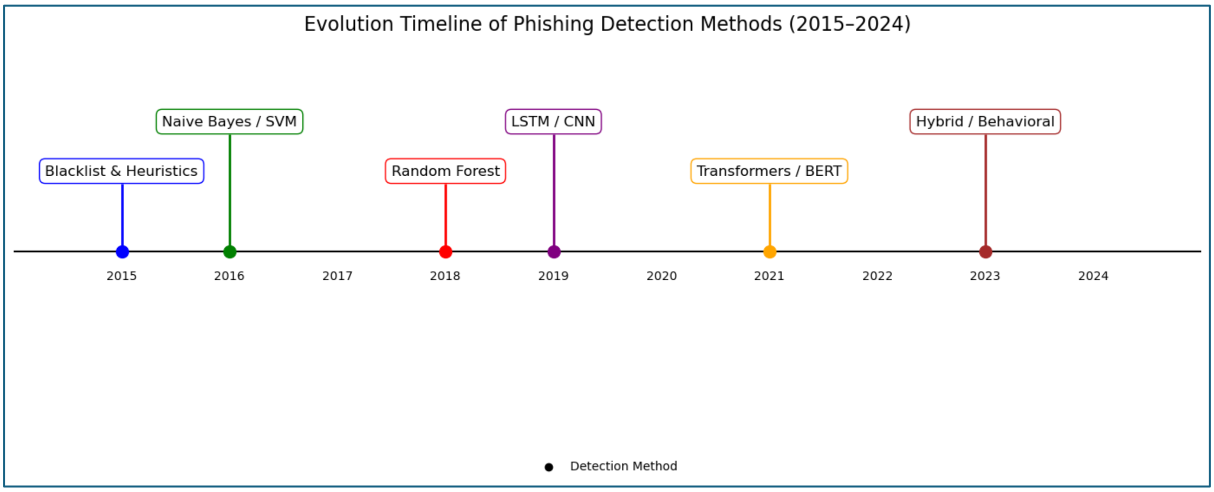
Detection tools are now incorporating features like perplexity metrics, syntax profiling, and AI-text fingerprinting to identify these advanced threats. The field is evolving into an arms race where generative models are used by both attackers and defenders.(3)

- **State-of-the-Art (2024): Multilayered, Adaptive Defense**

By 2024, the most effective phishing defenses combine:

- Transformer-based NLP models for content analysis.
- Ensemble systems integrating text, metadata, and image-level cues.
- Behavioral analytics for user- and domain-specific anomaly detection.
- Continuous learning, retraining on novel phishing attempts, including AI-generated variants.

These layered architectures achieve high precision and recall while minimizing false positives, adapting effectively to the dynamic and increasingly sophisticated threat landscape.



References

1. Cartier M. https://hoxhunt.com/blog/ai-phishing-attacks. 2025. AI Phishing Attacks: How Big is the Threat? (+Infographic) - Hoxhunt.
2. McKay T. https://www.itbrew.com/stories/2023/04/24/phishers-may-already-be-using-ai-to-improve-their-attacks-darktrace-report-finds. 2023. Phishers may already be using AI to improve their attacks, Darktrace report finds.
3. RecordedFuture. https://www.recordedfuture.com/research/qr-code-and-ai-generated-phishing-proliferate. 2024. Security Challenges Rise as QR Code and AI-Generated Phishing Proliferate | Recorded Future.
4. Eze CS, Shamir L. Analysis and Prevention of AI-Based Phishing Email Attacks. Electronics (Basel) [Internet]. 2024;13(10). Available from: https://www.mdpi.com/2079-9292/13/10/1839
5. Srikanth Bellamkonda. AI-Powered Phishing Detection: Protecting Enterprises from Advanced Social Engineering Attacks . International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE) [Internet]. 2022 Jan [cited 2025 Apr 11];11(1). Available from: https://www.ijareeie.com/upload/2022/january/2_AI-Powered.pdf
6. Thakur K, Ali ML, Obaidat MA, Kamruzzaman A. A Systematic Review on Deep-Learning-Based Phishing Email Detection. Electronics (Basel) [Internet]. 2023;12(21). Available from: https://www.mdpi.com/2079-9292/12/21/4545