Cisco
The bridge to possible

How large language models enhance Cisco Secure Email Threat Defense

Email-based attacks remain a significant and costly threat worldwide. Business Email Compromise (BEC) is a sub-type of email-based attack where an attacker runs a sophisticated scam against a business or individual for financial gain. As per the US Federal Bureau of Investigation (FBI), BEC is one of the fastest-growing and most financially damaging internet-enabled crimes. The FBI's data shows that BEC-related losses increased from $360 million in 2016[1] to $2.9 billion in 2023.[2]

As email-based attacks have increased in sophistication, defensive controls and detection techniques have had to evolve. Modern email security products make heavy use of artificial intelligence (AI), including Generative AI (GenAI) technologies such as Large Language Models (LLMs).

Cisco Secure Email Threat Defense is a modern high-performance system that detects email-borne threats, including BEC.[3] Email Threat Defense functions by assigning a maliciousness grade to each email it processes. It also declares its confidence in the maliciousness grade as medium or high. Emails are only declared as BEC emails and blocked if a high maliciousness grade for an email also comes with high confidence.[4]

Recently, Email Threat Defense's designers asked themselves: Could some of the medium-confidence emails with a high maliciousness grade be processed using AI techniques so that some of these emails are (correctly) classified as BEC emails? LLMs can, in fact, enhance Email Threat Defense's performance on BEC attacks.

How Email Threat Defense works on BEC emails

Email Threat Defense works via specialized detectors. Each detector produces a maliciousness grade for an email under inspection. These grades are aggregated to create a cumulative maliciousness grade for the email. Email Threat Defense marks the email for potential blocking if the cumulative grade exceeds a preset threshold on an email.

Currently, there are 90 or so detectors in Email Threat Defense. These detectors are generally powered by AI techniques.

Example detectors include:

- An urgency detector detects if the sender of an email is asking the recipient for a quick reply or similar action.
- A call-to-action detector detects if the sender of an email is asking the recipient to take an action, such as opening an attachment, clicking a URL, or disclosing sensitive data.
- A credential request detector that detects if the sender of an email is asking the recipient to provide privileged information such as login-id and password.
- An email address masquerade detector detects whether the sender is masking their actual email address behind an email address familiar to the recipient.
- A Unicode masquerade detector that detects the use of lookalike symbols (homoglyphs). Here, the sender tries to evade email security systems by using symbols from a language different from the one the recipient typically uses.
- A communication frequency detector detects rare communications between the sender and the recipient and between the sender's and recipient's organizations. Frequent communication between a sender and a recipient can indicate a safe email, whereas rare communication can be a sign of maliciousness.

Based on the combination of detectors that produced the cumulative maliciousness grade for an email, the grade is deemed either medium or high confidence. Typically, at least one of the detectors must give an email a high maliciousness grade for the cumulative grade to be marked "high confidence."

Email Threat Defense blocks an email only if both the cumulative maliciousness grade and confidence in that grade are high.

The LLM idea in brief

The Email Threat Defense team considered a performance enhancement that involves routing high-grade and medium-confidence emails through a customized LLM. The goal is to get the LLM to pass a malicious/safe decision on these emails.

A customized LLM prompt is created for every high-grade, medium-confidence email being considered. This prompt includes:

- The email header, subject, and body.
- Intermediate information produced by the detectors.
- An instruction to the LLM to produce a verdict on the email (i.e., a directive to classify the email).

When an appropriate LLM is chosen and an adequately engineered prompt is sent, the LLM sends back a short verdict, such as "malicious" or "safe," per email (see Figure 1).

Email Threat Defense's designers have experimented with the idea mentioned above and found that for the specific input considered – high-grade medium-confidence emails – some LLMs can accurately identify BEC emails. The designers reached this conclusion after independently inspecting and labeling a sample of high-grade, medium-confidence emails and comparing their results with those from their chosen LLM. Email Threat Defense's designers found that the LLM's verdict matched the verdict from their independent inspection in a high percentage of cases.

The Email Threat Defense team is currently deploying LLM-based classification of high-grade medium-confidence email to its cloud-based infrastructure. With this enhancement, emails that Email Threat Defense refrained from convicting as malicious to keep false positives in check can be safely blocked. The result is a safer email environment for Email Threat Defense customers.

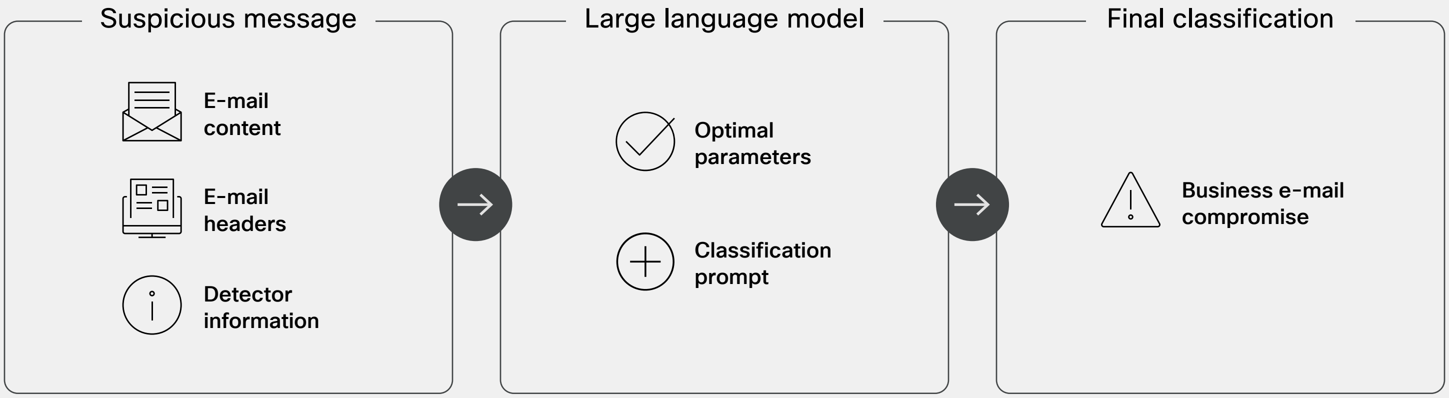

Figure 1. Using an LLM to improve business email compromise classification.

The bridge to possible

In closing

Cisco has been working on email security systems for over two decades. In recent years, Cisco has invested heavily in "AI for Security."[5] Cisco uses AI to assist security administrators, augment human ability to detect incoming threats and automate mundane and repetitive security tasks.

Deploying LLMs in the BEC detection pipeline is an example of using AI to augment human ability. Here, AI improves BEC detection at scale. This use of LLMs also demonstrates Cisco's continued commitment to innovation in security with AI.

Explore the ways that AI in Email Threat Defense protects your email environment from advanced threats like BEC.
Start a free trial today.

[1] Federal Bureau of Investigation, **Business Email Compromise and Real Estate Wire Fraud**, 2022.

[2] Federal Bureau of Investigation, Internet Crime Complaint Center, **Internet Crime Report**, 2023

[3] Brabec et al., **A Modular and Adaptive System for Business Email Compromise Detection**, August 21, 2023, arxiv.org.

[4] The "maliciousness grade" is a system-internal measure. It is not managed by security administrators.

[5] **Cisco's Pioneering Identity Intelligence Defends Against Most Persistent Cyber Threat**, February 6, 2024