StrongestLayer

WHITEPAPER · DETECTION ARCHITECTURE

The Email Is Not the Threat

A trusted vendor emails your accounts-payable team to update her banking details. Every authentication check passes. Every word is plausible. The same message is either routine or a seven-figure fraud, and nothing inside the email tells you which.

Muhammad Rizwan

Chief Technology Officer, StrongestLayer

StrongestLayer · May 2026

> The email is not the threat. The relationship is.

IN SHORT

The email is not the threat; the relationship is. AI has made every attack effectively one-of-one, so the signal that once lived inside the document is gone. We do not classify the message. We reason about how well it fits its context, along five dimensions, identity, medium, timing, workflow, and freshness, at three scales, the pair, the organization, and the ecosystem. Freshness runs through every check, because context that has gone stale is worse than none. The output is not a score but an auditable verdict, one readable cell per question, that an analyst or an auditor can overturn line by line.

The email is not a threat. The relationship is. There is no such thing as a malicious email, only an email-in-a-context that turns out to be malicious. The same email, read against a different context, is fine. That is not a philosophical claim. It is the operational reality of modern email security.

For thirty years the industry treated email as a document: a self-contained object you could open, score, and judge on its own merits. The model worked when attackers wrote badly, reused infrastructure, and templated their lures. The misspellings, the lookalike domain, the urgency phrasing, the suspicious link were stable enough to learn from. That era ended, and the cleanest way to see why is a case we now field weekly.

The puzzle

A vendor, call her Sandra, accounts receivable, ten years at her company, emails one of her customers' AP team. SPF, DKIM, DMARC all pass. The thread is twenty-two messages deep. Her writing style is intact. The signature is right. The reference to last quarter's invoice dispute is right. The email asks for one thing: please update our banking details for future payments, here is the new account.

Read this email in isolation and there is no signal. Every word is plausible, every header legitimate. A senior analyst handed only this message cannot tell you whether it is real. Neither can the best language model in the world.

Now read it in context. Sandra's company has never announced a banking change by email; those go through a vendor portal. The same message went to thirty of her other customers in the last six hours. The destination account has no prior payment history with anyone on our platform. The ses-

sion that sent it came from a residential proxy in a country Sandra has never traveled from. Read in context, this is a supply-chain BEC fingerprint with seven-figure exposure.

READ IN ISOLATION

- SPF, DKIM, DMARC all pass
- 22-message thread, writing style intact
- Signature correct, internal reference correct

Verdict: no signal, indeterminate

READ IN CONTEXT

- No prior banking change by email; the portal is the norm
- Same message hit 30 customers in 6 hours
- Destination account has no platform history
- Sent via residential proxy from a new country

Verdict: supply-chain BEC, seven-figure exposure

Figure 1. Same email, same headers. The context flips the verdict.

The email did not change. The context did. The verdict was never in the artifact, and no sharper look at the artifact will put it there.

The signal moved

If Sandra's case were rare you could staff around it with analysts. It is not, and AI is the reason. Last quarter we measured feature overlap across 15,042 confirmed threats that had already bypassed a conventional secure email gateway and landed in an inbox. Measured as pairwise similarity over message-body embeddings, AI-generated attacks clustered at 12 to 18 percent similarity across variants. Each one is effectively one-of-one, written from scratch for its target. Run the same metric against a labeled corpus of template-based phishing and it returns 85 to 95 percent similarity, the fingerprints a signature database is built to match.

15,042

CONFIRMED THREATS

Analyzed after they bypassed a secure email gateway

12-18%

AI VARIANT SIMILARITY

Each attack effectively one-of-one

85-95%

TEMPLATE SIMILARITY

The fingerprints signatures are built to match

54%

AI CLICK-THROUGH

Matches human experts; 4x the 12% baseline

METHODOLOGY AND SCOPE

This is a censored population by design: every sample bypassed an upstream gateway, so the figures describe what survives conventional filtering, not all email. Similarity is pairwise cosine over normalized message-body embeddings; the template baseline is measured the same way on a labeled corpus. n = 15,042, single-quarter window. The direction is corroborated by independent research: in a Harvard Kennedy School study by Heiding, Schneier and colleagues, fully automated AI spear phishing matched the 54 percent click-through rate of human-expert lures, more than four times the 12 percent rate of an ordinary-phishing control. ATT&CK v19 (April 2026) catalogs this behavior as T1683.001 (Generate Content: Written Content).

The implication is not that we need a better classifier. It is that the classification target dissolved. There is no document-level signal to learn from when every document is one-of-one. The features that once told you an email was malicious are no longer reliably present in malicious mail, and are increasingly present in legitimate mail. The distribution collapsed onto itself.

Classification or reasoning

A system that treats an email as a document is doing classification. It reads features inside the message, scores them, and picks a label. Modern language models are excellent at this, and still useless on attacks where the document does not carry the signal.

A system that treats an email as an utterance in a relationship is doing reasoning. It assembles context and argues toward a verdict. The output is not a label, it is an argument. Reasoning beats classification on the attacks classification cannot see, and among the attacks that bypass conventional filters today, that is the majority. The question is what context means, operationally.

Context has two axes: scale and dimension

It is tempting to call context a list of places to look. That framing breaks down quickly, and the imprecision matters for detection. Context has two axes. The first is scale: where the context lives, from the pair to the organization to the ecosystem, nested one inside the next. The second is dimension: what kind of fit you are checking, across identity, medium, timing, workflow, and freshness. Any single check is one axis crossing the other. Seeing it as two axes explains an asymmetry that a flat list cannot: the organization is not special because it subdivides, it is simply the scale where the most dimensions carry strong signal.

Freshness is the clearest reason the axes matter. It is not a fifth place to look, it is a dimension that runs through every cell. The pair check, the organization check, and the ecosystem check are each wrong if the model behind them is stale, so a freshness test sits inside every one of them. Dropping time does not remove one box from a list. It silently corrupts the others, because each is reading a model that no longer describes the world. Read the model as five questions, asked at up to three scales each:

DIMENSION	SCALE Correspondence (pair)	SCALE Organization (tenant)	SCALE Ecosystem (cross-tenant)
Identity	Sender is who they are to this recipient: ATO, VEC, lookalike, thread hijack	Internal and self-impersonation vs role and culture	Mule-account and adversary-infra reputation
Medium	Has this pair used this format and endpoint	Format and endpoint fit vs lived practice	Shared abuse patterns (weak)
Timing	This pair's cadence	Message vs org pattern: sequence, recurrence, variance	Same lure across tenants in one window
Workflow	*n/a (org-level)*	Does the action follow the real process	Cross-tenant reuse of the fraud workflow
Freshness	Pair history current (dormant reactivation)	Org model current (drift)	Adversary infra freshly seen elsewhere

Table 1. The model is scale by dimension. Freshness is a dimension, not a scale, which is why it appears in every cell.

The five dimensions are the detection distinctions, so the rest of this brief walks them: for each, the scales it draws on, what it catches, and how it fails.

Identity

Identity asks whether the sender is who they are to this recipient. At the correspondence scale this is the sharpest and most fragile evidence. The architecture maintains the communication graph and pair-specific history for every active sender-recipient pair, and asks whether this exact request type has happened between this exact pair before. Account takeover, vendor email compromise, and lookalike-domain impersonation are all identity attacks at the pair scale, the attacker borrowing trust built up over time. At the organization scale, identity is internal and self-impersonation against role and culture. At the ecosystem scale, it is reputation: is the destination a known mule account, has the sender's infrastructure been seen in adversary activity elsewhere.

COVERAGE

ATT&CK T1078.004 (valid cloud accounts), T1684.001 (impersonation, formerly T1656), T1583/T1584 (infrastructure) at the ecosystem scale, and T1682 (query public AI services) for the reconnaissance that precedes modern impersonation; in our published evasion taxonomy, EV-DT-001, EV-DT-002, EV-CE-006, EV-DT-004. Failure mode: cold pairs with no history carry thin identity evidence. Falsifiable test: recall on first-ever request types between pairs with established history.

Medium

Medium is the artifact and where it lands, format and endpoint together. A voice memo from a CFO on a phone is a different medium than the same memo inside a procurement workflow. The check pulls primarily from the organizational model, has this organization ever received voice memos through this endpoint, with the pair's own history as bounding evidence. Crucially, it is a fit check, not an authenticity check, so it does not require knowing whether an artifact is synthetic. That distinction is what unsticks deepfakes, and the deepfake section returns to it.

COVERAGE

QR pivots (T1204.001), payload concealment, and synthetic-media delivery (T1683.002, Generate Content: Audio-Visual Content); EV-CS-002 and the EV-PC family. Strongest signal at the organization scale, and it does not decay as generators improve, because it never asked the authenticity question in the first place.

Timing

Timing asks whether this message arrived when it should have, across sequence proximity, recurrence, and variance. An invoice that arrives twenty minutes after a calendar pretext from the same

sender is a different message than the same invoice in isolation. A wire request that follows a voice memo that follows a banking-change announcement is a multi-stage setup, not three independent events. Every relationship has a recurrence pattern, invoices from this vendor land in the second week, banking changes have never happened here, and a message outside that variance is anomalous before any other question is asked. At the ecosystem scale, timing is the campaign window: the same lure across many tenants in the same six hours is itself the verdict.

> **COVERAGE**
>
> Multi-stage setups and internal staged phishing (T1534). Failure mode: low-volume relationships have weak recurrence baselines.

Workflow

Workflow asks whether the action follows the organization's real process. An employee emails HR: please update my direct deposit, here is the new routing and account. The sender is real, the mailbox is real, the request is plausible. But the company's direct-deposit changes run through the self-service portal. Changes by email do not happen in this workflow. Identity, medium, and timing are all unremarkable. Workflow is wrong, because the email short-circuits the process it claims to be part of.

> **COVERAGE**
>
> Every BEC, vendor-fraud, and payroll-diversion variant has to violate this check to succeed; the short-circuit is the signal. This is the most falsifiable dimension, because the expected workflow for a given action is knowable and the deviation is close to binary. It lives almost entirely at the organization scale.

Freshness

Freshness is the dimension most platforms skip, because it does not show up in a feature comparison. It shows up in detection rates eighteen months in. Every relationship a message references has a timestamp. The vendor authoritative in January may have terminated in March. The employee whose voice is being cloned may have left. The trust the attacker borrows has a freshness date, and the architecture's view of that trust has one too. When they disagree, the verdict is confidently wrong. A dormant pair that reactivates with a banking change, an organization model that has drifted, an adversary infrastructure that was clean last week, each is a freshness failure at a different scale.

COVERAGE

Dormant-vendor reactivation and stale-identity exploitation (T1078). The falsifiable exhibit is the detection-rate-versus-model-age curve, which shows directly what a snapshot model loses over weeks. The architecture is not a context store, the failure mode of zero-trust as deployed; it is a context loop.

The verdict

Each cell emits evidence with a confidence, not a vote, and the reasoning layer combines that evidence rather than averaging scores. That is what lets a strong ecosystem signal carry a weak pair-level one, or a single hard workflow violation override otherwise-normal cells. The output is not a weighted sum a tuning knob can move. It is an argument over named evidence, which an analyst can read, an auditor can question, and either can overturn cell by cell.

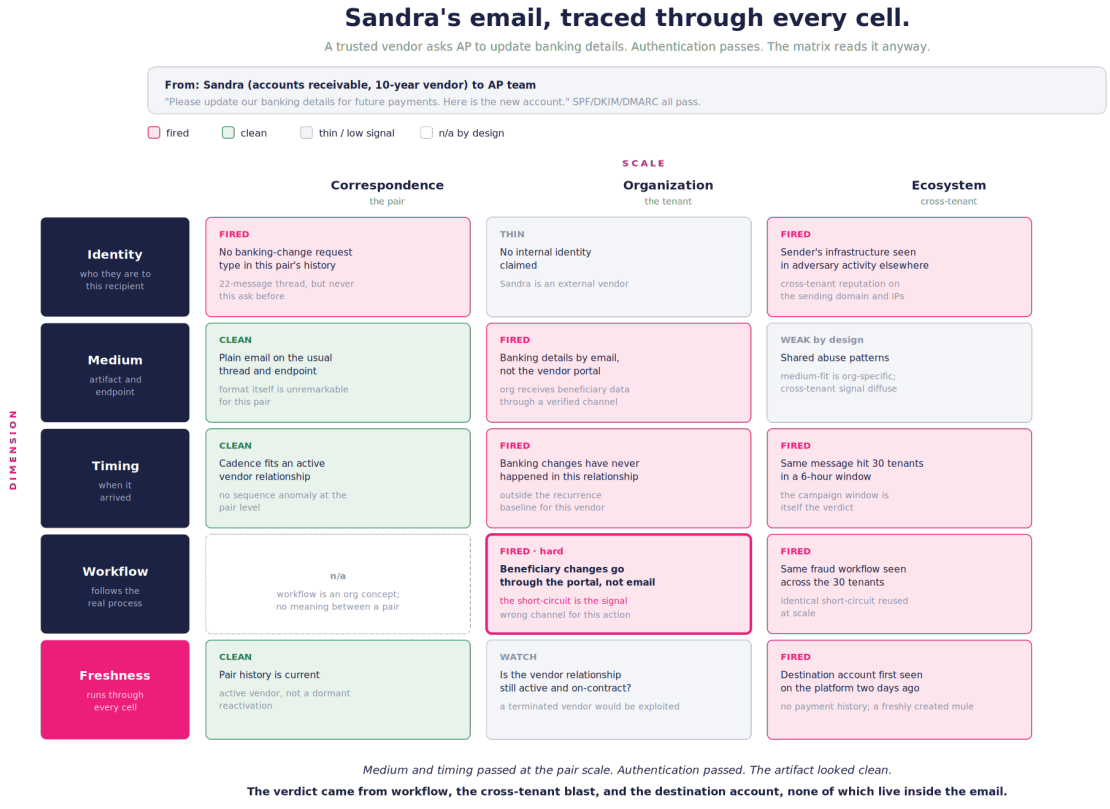

Figure 2. The same email read across all five dimensions and three scales. Each cell is independently checkable; the verdict is their joint reading.

For Sandra, the verdict is a set of cells. Identity fails at the pair scale: no banking-change history for this pair. Workflow fails at the organization scale: banking changes go through the portal, not email. The ecosystem scale shows the same pattern across thirty tenants in the same window. Freshness flags a destination account first seen on the platform two days ago. Each cell is independently checkable, and the verdict is their joint reading. The architecture never had to answer the one question with no answer in the text: is this email malicious.

The hardest case: deepfakes

Pick any deepfake delivered by email, a synthetic invoice, a voice memo from a spoofed executive, a meeting invite that sets up a real-time video impersonation. It misses on cells across the matrix at once. There is no pair history for the artifact, the organization does not receive it through this endpoint, it lands outside the organizational pattern, it short-circuits its workflow, the campaign or infrastructure shows up elsewhere, and the trust it exploits may already be stale. A deepfake is hard to classify as an artifact and easy to fail as a fit. The artifact gets harder to classify every year. The scale-by-dimension fit does not.

When the attacker knows the model

A fair question for any architecture is how it holds when the adversary understands it. Two moves are obvious. An attacker can warm a correspondence history first, trading benign messages to build pair-level trust before the ask. And an attacker can try to stay inside the workflow rather than short-circuit it. The architecture is built for both. Freshness catches the warmed and the dormant-then-active relationship, because trust carries a date and a sudden warming is itself an anomaly against the relationship's longer history. Workflow holds when the attacker stays polite: a banking change is still a banking change, and if the organization routes it through a portal, an email instance is a violation no matter how courteous the thread. The residual risk is worth naming. An attacker who compromises the legitimate workflow itself, not just the email, defeats the workflow cell, which is exactly why ecosystem and freshness evidence exist to catch what a single tenant cannot. No cell is sufficient alone. The architecture is the joint read, and it claims nothing stronger.

What this means for your program

This is not a rip-and-replace bet. It reasons over what reaches the mailbox, so it can sit on top of or in place of the gateway's verdict, and it coexists with the identity and endpoint controls you already run. It earns its place exactly where those are weakest: the business email compromise, vendor fraud, and account-takeover class that passes authentication, carries no payload, and never trips a

signature. It also slots into the zero-trust architecture you are already funding. Where zero-trust as deployed checks context at the gate and stores it as fact, this is the continuous-verification half the principle prescribed and the implementations skipped.

The model is size-independent. A company of 1,200 with one finance mailbox and a dozen SaaS apps has the same scale-by-dimension exposure as a multinational, with fewer rooms, and the ecosystem scale is where a leaner team gains the most: cross-tenant intelligence it could never assemble alone. What changes operationally is concrete. Fewer false positives, because a reasoned verdict is narrower than the broad patterns a classifier has to cast. Less analyst triage, because every verdict arrives with its evidence attached. And coverage of the attacks that turn into wire fraud and account takeover, the incidents that reach a board.

How to test it

None of this should be taken on our word, including the numbers in this brief. The two that matter, the variant-similarity figure and the detection-rate-versus-model-age curve, are reproducible on your own traffic in a proof of value. That is the honest way to test a context claim: hold the questions fixed and measure them on mail you already have. The questions are invariant to the next attack's surface, which is what makes that test fair. Five to put to us, or to anyone who claims context:

1. Does the system maintain per-pair history at the correspondence scale, and how fresh is it kept?
2. Can it name the workflow an email claims to join, and detect the short-circuit?
3. Does it correlate across tenants at the ecosystem scale, and what confidence lift does that give?
4. What is its detection-rate-versus-model-age curve, the freshness dimension made measurable?
5. For any verdict, can it show the per-cell evidence an auditor can check independently?

An email is not a document. It is an utterance in a correspondence, inside an organization, inside an ecosystem, all of it aging in time. Read it as text and you build a classifier that catches what classifiers catch. Read it as an utterance and you build a system that reasons. That is the choice, and it is why we think the way we do.

ABOUT THE AUTHOR

Muhammad Rizwan is Chief Technology Officer at StrongestLayer. He leads the architecture and engineering of an email security platform built on contextual reasoning. He writes about detection architecture and the operational realities of defending against AI-era threats.