StrongestLayer

The End of the Biggening

Why Mythos Is Not a Warning. It Is a Structural Indictment of Training-Based Detection in an Era of Compounding AI Capability.

Joshua Bass, Chief Product Officer
April 2026

What Anthropic's Mythos Preview reveals about the structural defender lag and why organizational context is the only durable signal.

Abstract

In April 2026, Anthropic announced Project Glasswing and its underlying model, Claude Mythos Preview, a frontier AI system the company considered too dangerous to release publicly due to its autonomous vulnerability discovery capabilities.

The security industry reacted as it always does at these moments: with alarm and tactical scrambling. This whitepaper argues that reaction, while understandable, is the wrong one.

Mythos is not a warning that AI-powered attacks are about to become dangerous. It is a confirmation that a predictable, multi-year capability curve has been advancing on schedule, and will continue to do so for eight to ten more years. This paper presents three interconnected arguments: that AI capability events follow a documented cadence making each threshold event anticipatable; that the structural lag between attacker adoption and defender detection is architectural, not operational; and that the only detection design that escapes compounding obsolescence is one anchored in organizational ground truth rather than learned content signatures.

Key Figures at a Glance

$20.9B
FBI CYBERCRIME LOSSES
FBI 2025 Annual Report, 26% increase over 2024

54%
AI PHISHING CLICK-RATE
Versus 12% for phishing human-written emails

4–7 yr
DETECTOR OBSOLESCENCE
Training-based model lag at peak production

Days
ATTACKER ADOPTION
Time to weaponize a new frontier model release

THE CORE THESIS

Any detection architecture whose effectiveness is bounded by its training surface inherits a compounding obsolescence stack. The only escape is to reason from what the attacker can never access: the defender's ground truth.

1. The Curve, Not the Point

Every twelve to fourteen months since 2020, the AI capability landscape has crossed a threshold that rendered the previous defensive posture inadequate. GPT-3 in mid-2020. ChatGPT and accessible generative AI at end of 2022. GPT-4 in early 2023. DeepSeek-R1 in January 2025, delivering frontier reasoning at a fraction of expected cost. And now Mythos.

The security industry's habitual response, analyze the new model, update detection signatures, patch the most visible gaps, treats these events as discrete incidents. They are not data points on an S-curve; they are the evidence of one. That curve has roughly eight to ten years of steep climb remaining before meaningful deceleration.

> *"The specific model is not the story. The cadence is the story."*

Figure 1. The AI Capability Curve: Threshold Events on a Long Slope

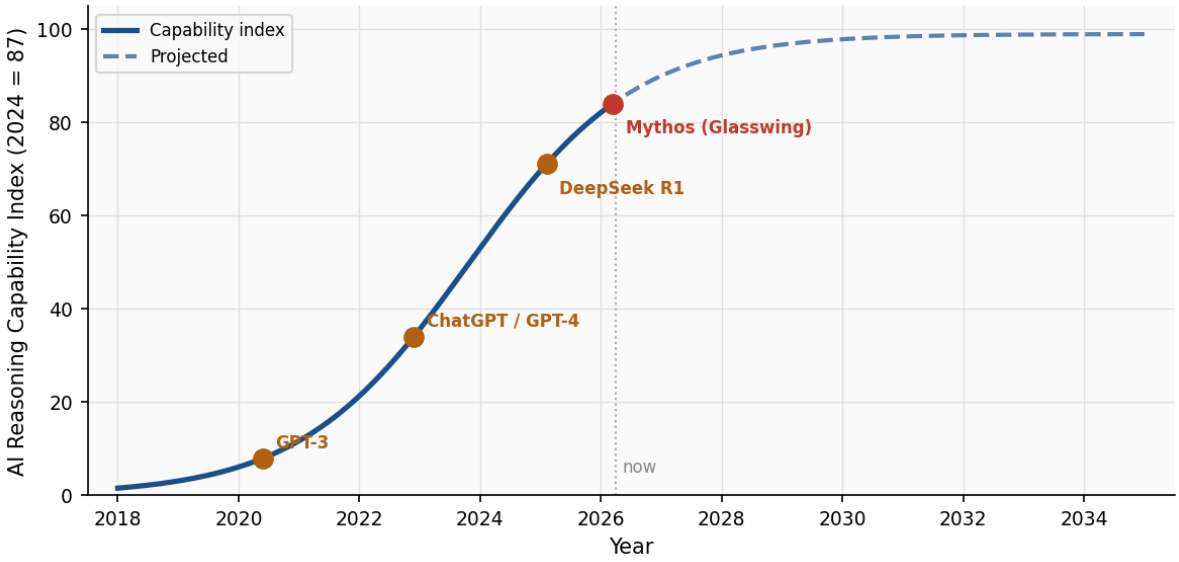

Figure 1: AI reasoning capability index, 2018–2035. Threshold events occur on a predictable cadence of 9–14 months.

What drives this extended runway? Three reinforcing dynamics. First, the feedback loop between AI capability and AI research is now measurable: systems like AlphaProof and FunSearch have demonstrated that AI can accelerate mathematical discovery in domains directly relevant to model improvement. Second, compute efficiency, through mixture-of-experts architectures, quantization, and sparse attention, compresses the cost of reaching any given capability level. Third, open-source democratization ensures each frontier breakthrough propagates to the criminal underground within weeks.

The implication for security architecture is direct: any defensive posture designed around the capabilities of the current attacker model is obsolete on a known schedule. Building security strategy around today's Mythos is equivalent to building a seawall to the height of last year's flood.

2. The Structural Defender Lag

An attacker with access to DeepSeek on December 26, 2024 had it in their tooling by December 27. There is no procurement process, no validation cycle, no compliance review. The attacker's upgrade cycle is now measured in days.

Research confirms the urgency. LLM-generated phishing emails now achieve click-through rates exceeding 50%, compared to 12% for human-written messages. A Cornell University mass-phishing study across 71,000 emails found LLM-enhanced campaigns exceeded 30% click-to-landing success rates. Over 73% of phishing emails analyzed in 2024 involved some form of AI assistance.

Figure 2. The Structural Defender Lag: Attacker vs. Defender Capability Curves

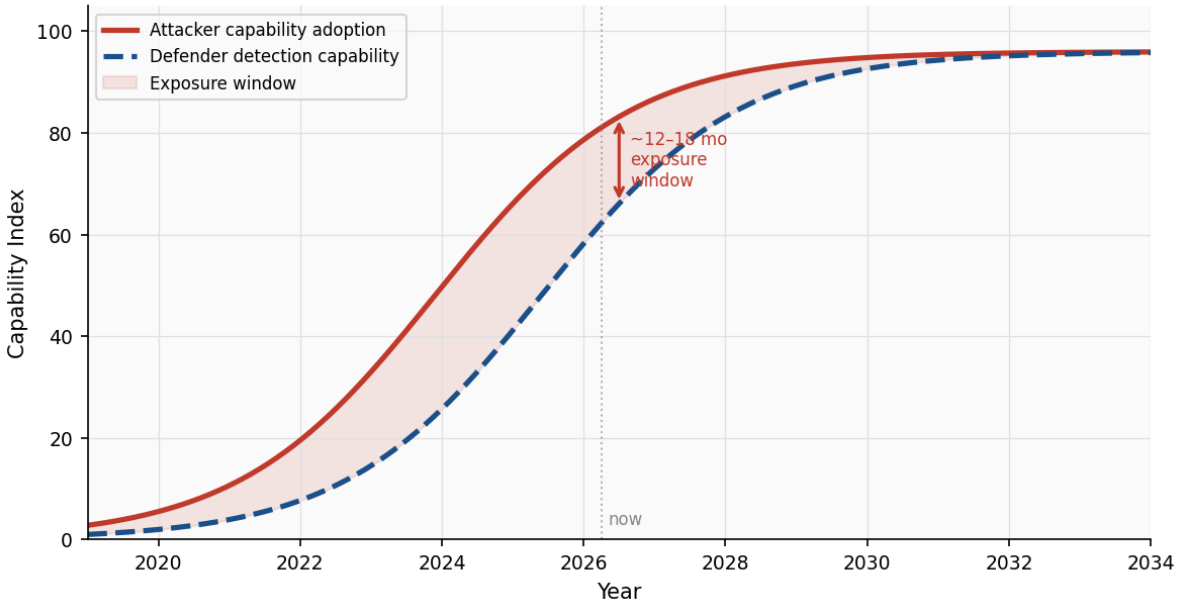

Figure 2: Structural defender lag. Attacker capability adoption (red) tracks new model releases within days. Defender detection capability (blue dashed) lags 12–18 months.

The defender's cycle is not slow because of incompetence. It is slow because of structure. Consider what it actually requires to update enterprise-grade email threat detection:

- A foundational model must be selected, itself 12–18 months behind the research frontier at decision time.
- A threat corpus must be assembled and labeled, a snapshot of attacker behavior at training start, not at deployment.
- The model must be trained, red-teamed, validated against customer environments, and approved through compliance and release processes.
- The product must be sold, deployed, and integrated, with enterprise sales cycles adding 3–9 months from GA to customer production.
- The vendor must extract ROI from the training investment, creating economic pressure to maintain the model in production 18–36 months before a major revision.

Figure 3. The Compounding Obsolescence Stack: Why Training-Based Detection Accumulates Lag

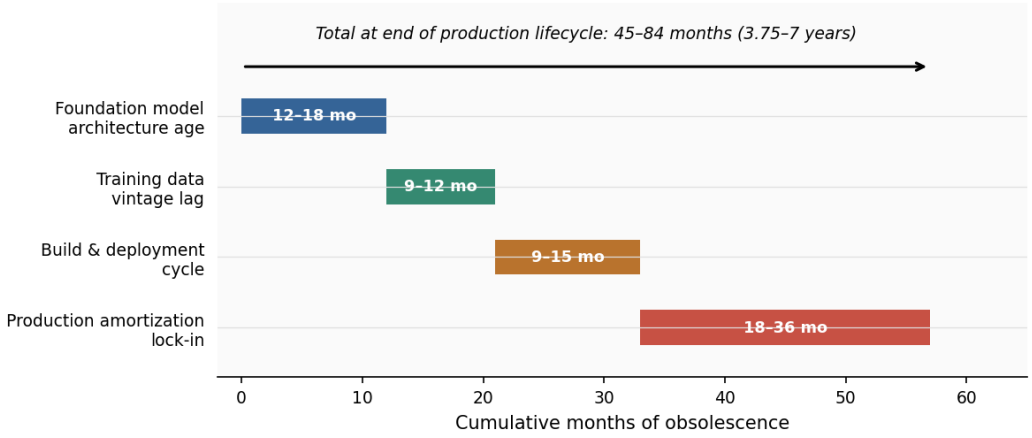

Figure 3: The Compounding Obsolescence Stack. Each layer inherits the lag of the layer below it. At peak production, a training-based model may reason with architecture 4–7 years old.

> *"The defender lag is structural. While waiting for AI-integrated security products, attackers are already using AI for evasion." – FireCompass Cyber Panel, 2026*

This dynamic is most acute for bespoke in-house detection models and small vendors without the capital to rebuild at each capability threshold. A custom model built on last year's foundational math, trained on last year's attack surface, is not a security investment. It is a liability with a declining half-life. The economics of staying current with the foundational model layer scale directly with the pace of the capability curve.

3. The Attacker's Permanent Blind Spot

The preceding sections describe a structurally disadvantaged defensive posture. This section describes why it is not an inevitable one, and where the genuine asymmetric advantage lies.

Every attacker operating against an organization is reasoning from estimates. They reconstruct your payment approval workflows from LinkedIn job descriptions. They infer reporting hierarchies from press releases and conference bios. They time attacks around public travel schedules and earnings cycles.

Mythos makes that reconstruction more fluent. It still cannot see your actual org chart, your actual workflows, your actual decision-makers.

The defender with the right architecture holds ground truth across every dimension the attacker is estimating:

- Complete communication history between any two parties, not a reconstruction of likely patterns.
- Actual approval workflow for financial transactions, including who currently holds authority and what out-of-band verification is standard.
- Calendar context: whether the impersonated executive is reachable or in a board meeting where an urgent wire request would be anomalous.
- Vendor relationship history: whether "updated banking details" arrived from an entity with zero prior contact, or from a domain that appeared last week.
- Process maps: whether this type of request has ever, in organizational history, been completed via email without secondary verification.

> *At equivalent reasoning power, facts beat estimates. The defender holding the system of record is always reasoning from facts.*

And critically: as foundational model capability improves, a detection system grounded in organizational context gets stronger alongside it. The attacker's upgraded model makes their social engineering more fluent. The same model upgrade, applied to reasoning about organizational ground truth, makes the defense more precise. The curves move together, but only if the architecture is built to use them.

The financial cost of failing this test is concrete. The FBI's 2025 Annual Cybercrime Report recorded $20.88 billion in reported losses, a 26% increase over 2024. BEC alone accounted for $2.77 billion in 2024 across 21,442 incidents. These are attacks that succeed not because defenders lack compute power, but because they lack the organizational context to recognize that the request is inconsistent with how the organization actually operates.

4. A Taxonomy of Architectural Bets

Not all detection architectures carry equal exposure to the compounding obsolescence problem. The central question is whether a given architecture resets to near-zero effectiveness at each capability threshold event, or whether it survives and strengthens.

Figure 4. Detection Architecture Risk Matrix: Reset Risk vs. Signal Permanence

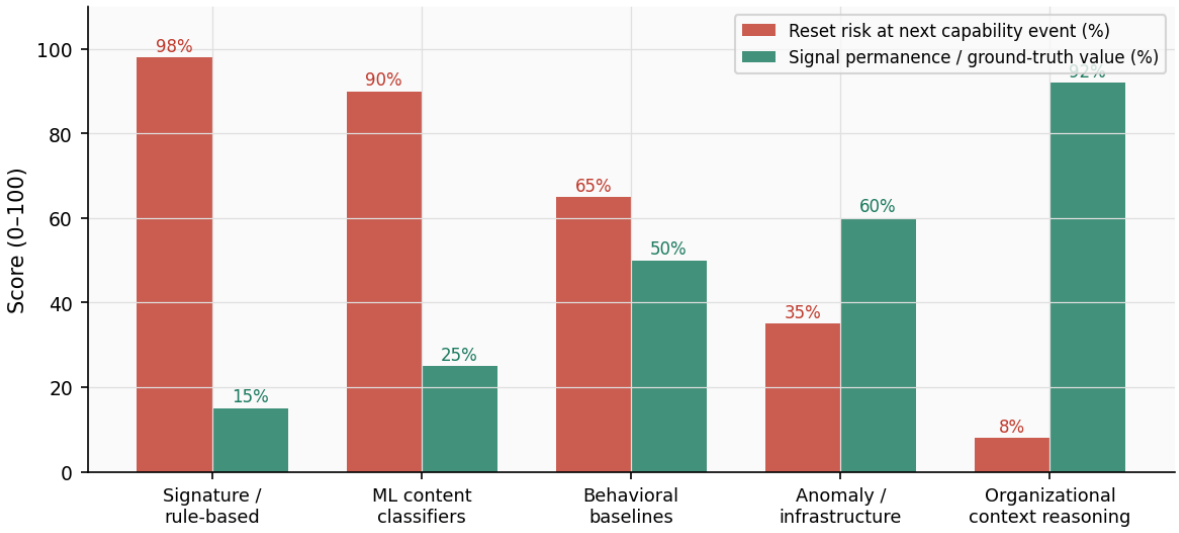

Figure 4: Detection architecture risk matrix. Architectures toward the lower-right are most resilient.

Architectures that reset at each event (highest risk)

Signature and rule-based detection, and ML content classifiers trained on attacker-generated text, carry near-total reset risk. Their effectiveness is bounded by their training surface, always a snapshot of attacker capability at training time. When DeepSeek-class content differs structurally from GPT-4-class content, a classifier trained on the latter has materially degraded recall against the former. When Mythos-class content differs from both, the cycle repeats.

Forrester noted in 2025 that the email security market is in “a period of digestion”, precisely the language of an industry absorbing a capability event that outpaced its detection models.

Architectures that partially survive (medium risk)

Anomaly detection, sender reputation scoring, and infrastructure-based IOC analysis degrade at each threshold event but do not fully reset. Sending infrastructure, domain registration patterns, and authentication anomalies retain signal value even when email content has become indistinguishable from legitimate communication. These signals are necessary but insufficient. A Mythos-class model operating through a clean Microsoft 365 tenant with legitimate DKIM signatures produces no infrastructure anomaly signal at all.

Architectures that do not reset (lowest risk)

Detection logic grounded in organizational context and intent reasoning is structurally immune to the capability curve in a way content-based approaches are not. The core question, is this request consistent with how this organization actually operates, does not change when the attacker upgrades their model.

This architecture class has an additional property that makes it the correct long-term investment: it improves as foundational model reasoning power improves. Better reasoning applied to richer organizational ground truth produces more precise anomaly detection, not simply faster execution of the same analysis. The upgrade cycle works for the defender, not against them.

5. What Security Teams Should Do Differently

Stop reacting to the point; build for the curve

Every organization that responds to Mythos by asking “can our current vendor detect this?” is asking the wrong question. The right question is: when the next threshold event happens in twelve to eighteen months, does our detection architecture become stronger or does it reset? Procurement decisions should be evaluated on this axis, not on benchmark performance against current-generation attacker content.

> *You're probably thinking: our vendor says they update detection monthly. That's fine for catching variations. It's not fine for catching a class of attack your training data has no examples of.*

Be honest about bespoke and small-vendor risk

A custom model built on last year's foundational math, trained on last year's threats, maintained by a team without the capital to rebuild at each capability threshold, is a choice to fall further behind on a predictable schedule. Organizations running bespoke detection should conduct an honest assessment of whether their model vintage and training corpus age are producing a net security benefit or a net liability.

Prioritize signal access over signature coverage

The width of the organizational signal surface is more durable than the depth of the signature library. Mail history, calendar integration, identity context, vendor relationship graphs, and process-map awareness do not become less valuable when attacker models improve. They become more valuable, because the reasoning layer that interprets them gets cheaper and more capable on the same curve that is empowering the attacker.

Demand reasoning chains, not confidence scores

A confidence score is the output of a trained model. It tells you how similar the input is to the training distribution, precisely the metric that resets at each capability event. A reasoning chain tells you why a specific email is inconsistent with specific organizational context. Only one of those two outputs survives the next Mythos. Procurement and vendor review processes should treat the presence or absence of explainable reasoning as a first-order architectural signal.

Conclusion

Winston Churchill, in November 1942, described El Alamein: "This is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning." Mythos is the email security industry's end of the beginning: the point at which the capability curve is too steep and too long to be addressed by tactical responses to individual threshold events.

The organizations that maintain defensible security postures through the next decade will not be the ones who reacted fastest to each new model release. They will be the ones who understood early enough that the correct architectural bet is not a better classifier. It is a system that reasons from organizational context the attacker can never access, and that gets more capable as the same curve empowering the attacker advances.

> *Any detection architecture whose effectiveness is bounded by its training surface inherits a compounding obsolescence stack. The only escape is to reason from what the attacker can never access: the defender's ground truth.*

The AI capability curve has years left. Build accordingly.

References & Citations

1. Anthropic, Project Glasswing / Claude Mythos Preview, April 2026.
   venturebeat.com/technology/anthropic-says-its-most-powerful-ai-cyber-model-is-too-dangerous-to-release
2. Sysdig Threat Research Team. “LLMjacking Targets DeepSeek.” December 2025.
   sysdig.com/blog/llmjacking-targets-deepseek
3. AI-Generated Phishing CTR (54% vs 12%): Secureframe phishing statistics, 2026.
   secureframe.com/blog/phishing-attack-statistics
4. Cornell University 71,000-email mass-phishing study. TechMagic, 2025. techmagic.co/blog/blog-phishing-attack-statistics
5. FBI IC3 2025 Annual Cybercrime Report, April 2026.
6. FBI IC3 2024 Internet Crime Report. Nacha, April 2025.
   nacha.org/news/fbis-ic3-finds-almost-85-billion-lost-business-email-compromise-last-three-years
7. Palo Alto Networks Unit 42. BEC timing and targeting analysis.
   paloaltonetworks.com/cyberpedia/what-is-business-email-compromise-bec-tactics-and-prevention
8. FireCompass Panel Brief. “Top Breaches in Cyber Security in 2025.” February 2026.
   firecompass.com/top-breaches-in-cyber-security-in-2025
9. Forrester Wave: Email, Messaging, and Collaboration Security Solutions, Q2 2025.
10. NDSS Symposium 2025. “Utilizing LLMs to Create Context-Aware Spear Phishing.”
    ndss-symposium.org/wp-content/uploads/2025-poster-68.pdf
11. ScienceDirect systematic review. “LLMs in Phishing Attack Generation and Detection,” 2025.
    sciencedirect.com/science/article/pii/S2590005626000986
12. Trend Micro TrendAI Research. “Fault Lines in the AI Ecosystem,” March 2026.
    trendmicro.com/vinfo/us/security/news/threat-landscape/fault-lines-in-the-ai-ecosystem-trendai-state-of-ai-security-report

Version: 1.0 | Date: April 2026 | Joshua Bass, Chief Product Officer

Contact: josh@strongestlayer.ai | strongestlayer.com