PREVEIL

WHITE PAPER

End-to-End Encryption When AI Shifts Threat Economics

An architectural analysis of cloud security for defense industrial base contractors making CMMC decisions in the era of AI-accelerated cyberattack.

Contents

Executive Summary

- **In April 2026, Anthropic announced an AI model that identified thousands of previously unknown zero-day vulnerabilities across every major operating system and web browser — and produced working exploits for many of them. Anthropic judged it too dangerous to commercialize.** Instead, it formed Project Glasswing — a consortium that includes Microsoft, Google, AWS, Apple, NVIDIA, CrowdStrike, and JPMorgan Chase — to direct the capability toward defense. Earlier disclosures from Anthropic (an espionage campaign 80–90% executed by AI) and Google’s Threat Intelligence Group (the first AI-developed zero-day in the wild) document the same trend. AI has collapsed the cost of sophisticated cyberattack. This is not speculation; the laboratories themselves are saying so.

- **The largest concentration of risk in most organizations is cloud-stored plaintext** — data the cloud provider can technically decrypt. In 2023, a single stolen Microsoft signing key (the Storm-0558 breach) gave Chinese state-sponsored attackers access to essentially any Exchange Online mailbox in the world, including senior U.S. officials managing the U.S.–China relationship. One provider-side failure exposed every user the provider holds. AI accelerates exactly this attack pattern.
- **End-to-end encryption keeps data encrypted in the cloud — always.** Keys exist only on user devices; the cloud holds ciphertext. A breach of the server is not a breach of the data. The U.S. State Department, in consultation with the NSA, codified this approach in ITAR §120.54 in 2019. CISA's December 2024 post–Salt Typhoon guidance points the same way.
- **CMMC certifies the controls a contractor follows. It does not check whether the cloud platform itself can withstand an AI attack.** A contractor on GCC High and a contractor on an end-to-end encrypted platform can both pass CMMC — but when the cloud is breached, only the E2E platform keeps the data hidden. **Both pass CMMC. Only one protects the CUI.**
- **The solution is a CUI enclave: keep general productivity on existing cloud platforms, and move CUI alone to an end-to-end encrypted environment the provider cannot decrypt.** PreVeil is the most widely deployed of these in the DIB — used by more than 3,000 defense contractors for CMMC and ITAR workloads, with more than 90 having achieved CMMC certification on the platform.

1. What changed: the AI shift in threat economics

For most of the cloud era, mounting a serious server-side attack against a hardened cloud provider required scarce expertise, months of effort, and a measure of luck. Those costs are collapsing.

In April 2026, Anthropic announced Claude Mythos — a frontier AI model that, in internal testing, identified thousands of previously unknown high-severity vulnerabilities across every major operating system and web browser, including a twenty-seven-year-old flaw in OpenBSD. Mythos did not merely find bugs; it wrote working exploits. Anthropic, which has every commercial incentive to release its frontier models, judged Mythos's offensive capabilities too dangerous for general access. Instead, it formed Project Glasswing — a consortium that includes Microsoft, Google, AWS, Apple, NVIDIA, CrowdStrike, and JPMorgan Chase — to direct the capability toward defense. The hyperscalers whose cloud platforms hold most of the world's enterprise data judged the threat severe enough to join.[1] These are not Anthropic's claims alone. The UK AI Security Institute, in independent evaluations, found Mythos succeeded at 73% of expert-level capture-the-flag tasks — a category no model had completed before — and was the first system to solve end-to-end a thirty-two-step simulated corporate-network compromise that takes a human expert roughly twenty hours.[2]

Even before Mythos became publicly known, AI was already being used to execute attacks at scale. In November 2025, Anthropic disclosed that a Chinese state-sponsored group, designated GTG-1002, used a jailbroken Claude Code agent to execute between 80% and 90% of the tactical operations in a coordinated espionage campaign against roughly thirty organizations — including technology companies, financial institutions, chemical manufacturers, and government agencies. The AI handled reconnaissance, vulnerability discovery, privilege escalation, lateral movement, credential theft, and data exfiltration at what Anthropic described as "physically impossible request rates." Humans intervened only at strategic checkpoints. Anthropic called it the first documented case of a cyberattack largely executed without human intervention at scale.[3]

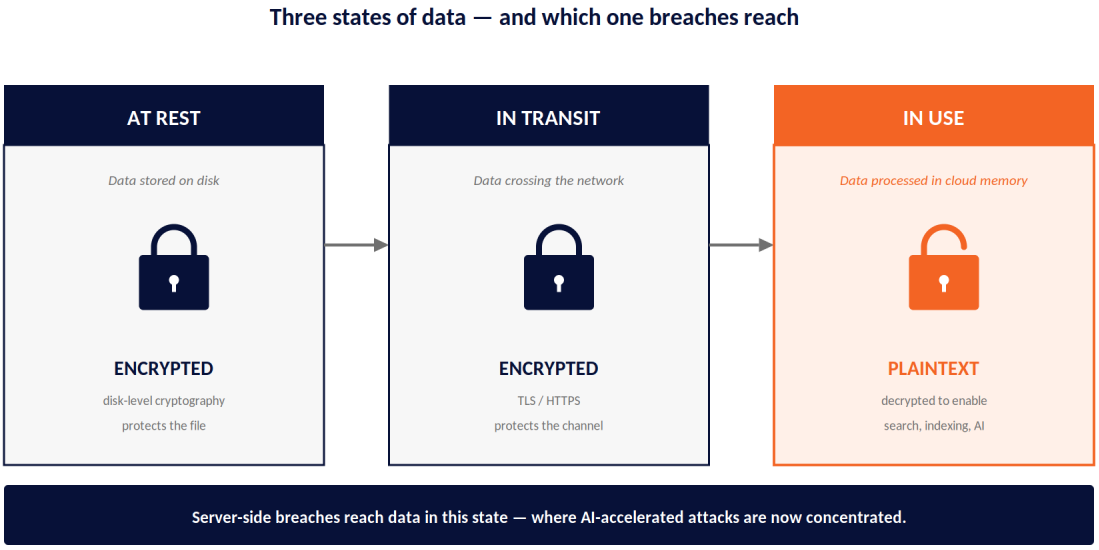
By May 2026, AI-developed exploits had moved from controlled environments to live targets. Google's Threat Intelligence Group published its first report identifying an AI-generated zero-day exploit deployed in the wild. The exploit, developed by a cybercrime group planning a mass-exploitation campaign, bypassed two-factor authentication in a widely-used open-source administration tool. Google's team also reports that Chinese and North Korean state actors are using AI to recursively analyze CVEs and validate proofs of concept, and concludes that adversaries are now using AI "as expert-level force multipliers for vulnerability research and exploit development, including for zero-day vulnerabilities."[4]

These are admissions, from the laboratories building the technology and the threat-intelligence units paid to watch its misuse. Sophisticated server-side attack is moving from rare and expensive to routine and cheap. Systems built when cloud-side compromise was rare are now defending against the wrong threat.

2. Three assumptions about cloud security that no longer hold

Most cloud security policies are built on three intuitions that were approximately true for many years and are now actively dangerous.

"Encryption at rest and in transit means the cloud is safe." It does not. Almost every major cloud service encrypts disks and network traffic, but the service itself decrypts data in memory to do its work — to index it, search it, scan it for malware, render it to users, run AI features over it, respond to legal process. Plaintext data and the keys to decrypt it exist, continuously, inside the provider's infrastructure. An attacker who reaches into that infrastructure reaches plaintext. "Encryption at rest" protects against theft of a hard drive; it does not protect against an attacker who can ask the service to decrypt.



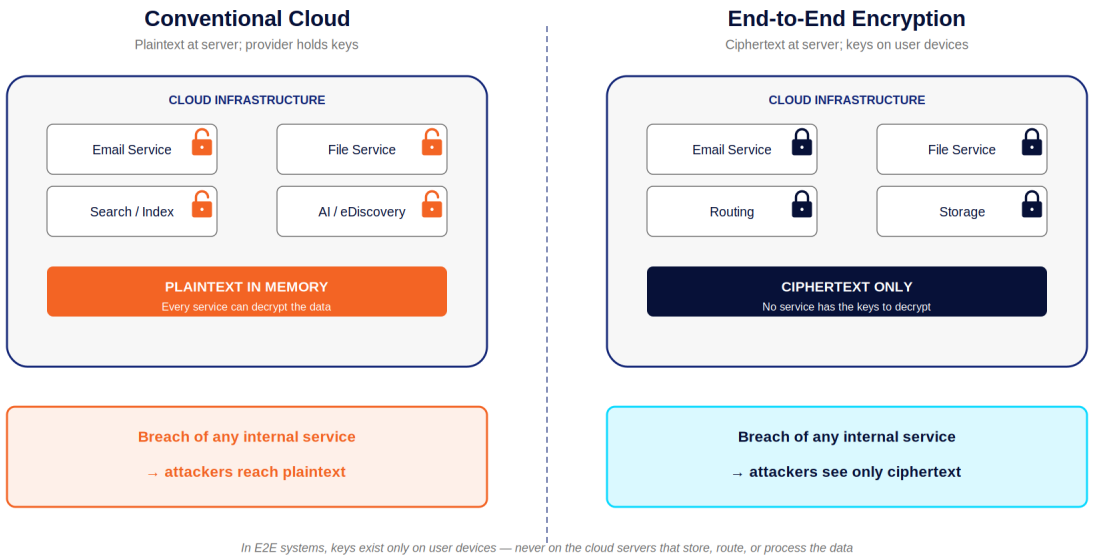
[Figure 1] Encryption coverage by data state. Conventional cloud services encrypt data at rest and in transit but expose plaintext during processing in cloud memory — the state in which most server-side breaches reach data.

"Large providers will defend themselves competently." They try; their reputations depend on it; and they still fail. Storm-0558 is the case to study. Between May and mid-June 2023, a Chinese state-sponsored group obtained a Microsoft consumer signing key — one that, by Microsoft's own account, should have been retired in 2021 — and combined it with a validation flaw in Microsoft's token system to forge authentication tokens that worked against essentially any Exchange Online enterprise mailbox in the world. The attackers accessed the email of more than 500 individuals across 22 organizations, including senior U.S. officials managing the U.S.–China relationship. The Cyber Safety Review Board concluded the intrusion was the product of a "cascade of avoidable errors" and "should never have happened." As of the CSRB's report, Microsoft still could not say how the key was stolen. The relevant fact for architects is not that Microsoft made mistakes. The relevant fact is that *one* mistake, in *one* place, was sufficient to compromise essentially the entire Exchange Online tenant base.

"Endpoint attacks are the real threat; servers are safer." This inverts the actual risk profile. Endpoint compromise affects one user's data. Server compromise, when the server holds plaintext for millions of users, affects millions of users. Storm-0558 reached 500 individuals from one stolen key. No endpoint compromise has ever achieved that ratio. AI changes the math further in the same direction. The long, multi-step planning that server-side attacks require is precisely what frontier models now do well. The per-user phishing and per-user persistence that endpoint attacks require is precisely what they do not. AI is making the catastrophic attack cheaper while leaving the bounded one roughly where it was; the economics of attack now actively favor the design that concentrates plaintext.

3. Where the plaintext lives

The single most important question for any cloud service handling sensitive data is whether the provider can technically read the data. The answer determines whether a successful attack on the provider yields plaintext or ciphertext, and the difference between those outcomes is the difference between a contained incident and a catastrophe.



[Figure 2] Conventional cloud versus end-to-end encryption. In conventional cloud, multiple internal components handle plaintext to enable indexing, search, and other features — and a breach of any component reaches plaintext. In an end-to-end encrypted service, the cloud holds only ciphertext, and keys live only on user devices.

Most enterprise cloud services hold plaintext at the server. This was reasonable when the threat environment was different. Cloud platforms were built in an era when sophisticated server-side attack was rare and expensive, and the features customers expected — server-side search, indexing, malware scanning, eDiscovery, legal compliance workflows — depended on the server being able to decrypt. As features accumulated, the architecture locked in. Server-side AI features, the most recent additions, are layered onto the same plaintext substrate — extending rather than challenging the inherited design. The keys are typically managed by the provider, sometimes with "customer-managed key" features that move some control to the customer but still permit the provider to decrypt. Microsoft 365 (including GCC and GCC High), Google Workspace, and most file-sharing platforms operate this way. This looks like a thick perimeter around plaintext data; the strategy is to make that perimeter very hard to penetrate.

This includes the cloud environment most defense contractors turn to specifically for CUI handling. Microsoft GCC High is purpose-built for the U.S. defense industrial base — FedRAMP High and DoD

Impact Level 4 authorized, restricted to U.S. persons, operated from infrastructure physically separated from commercial Microsoft 365. These are real and significant differences from commercial cloud, calibrated to a particular threat model: insider risk, foreign data residency, export-control compliance. The U.S.-persons restriction in particular makes it harder for a foreign adversary to recruit or compromise an administrator. It is moot against an AI-orchestrated attack that bypasses administrators entirely and goes after the server infrastructure directly.

None of these protections change the fundamental fact: Microsoft holds the keys, Microsoft's services decrypt the data in operation, and a sufficiently capable attacker reaching that infrastructure reaches plaintext. Storm-0558 occurred in commercial Exchange Online, not GCC High; the design flaw that produced it — one provider-held key, compromised once, at the server — exists by construction in any cloud-plaintext system, regardless of which certification regime it is sold under. **The certifications certify controls. They do not change the architecture.**

Genuine end-to-end encrypted services hold no plaintext at the server. All encryption and decryption happens on user devices. Keys live on user devices, and only on user devices. The cloud holds ciphertext, encrypted with keys it cannot access. Here the cloud is a dumb storage and routing layer; the smart cryptographic work happens at the endpoints. The strategy is to make the perimeter *irrelevant*: a successful attack on the cloud yields ciphertext, which is useful to the attacker only to the degree they can attack each user's endpoint to obtain that user's keys.

The trade-off is real, though narrower than commonly believed. A service whose cloud cannot read plaintext cannot offer cloud-side features that depend on plaintext access. Server-side search across attachment contents requires more complex techniques. Server-side AI summarization of mailboxes is similarly challenging. Other commonly cited concerns — eDiscovery, legal hold, compliance workflows — turn out to be solvable. Mature E2E platforms handle them through admin-mediated decryption controls rather than server-side access. For workloads where the genuine server-side features matter more than confidentiality survival, the trade is wrong. For workloads where confidentiality survival matters more — regulated CUI, defense technical data, sensitive IP, strategic communications — the trade is the right one.

The question is therefore not which approach is universally better. It is which is right for which data. Many organizations should be running both: cloud-plaintext services for general productivity and collaboration, end-to-end encrypted services for the subset of data whose loss is catastrophic.

4. Defining "end-to-end encryption" — and testing for it

The phrase "end-to-end encryption" has been used loosely enough that it requires a precise definition before it can carry weight. End-to-end encryption has three defining properties. First, data is encrypted on the sender's device before it leaves. Second, the keys required to decrypt that data exist only on the sender's and intended recipients' devices — never on the cloud servers that store, route, or process the data. Third, no intermediary, including the cloud service provider, has any technical means to decrypt the data while it is in their custody. Encryption happens at one endpoint, decryption at the other; everything in between sees only ciphertext.

This definition matches the U.S. government's regulatory test. ITAR §120.54, adopted in December 2019 by the State Department in consultation with the NSA, codifies end-to-end encryption as data encrypted "between the sender's in-country security boundary and the recipient's in-country security boundary without being revealed in clear text or unencrypted form or providing the means of decryption to any third party." The rule permits ITAR-controlled technical data to transit and reside on foreign cloud infrastructure without an export license — but only if the data meets this test. The State Department and NSA were precise about the requirement because the alternative — provider-accessible plaintext — was an unacceptable export-control risk. Any vendor claiming the term should be able to satisfy the same test.

Under this definition, several common approaches do not qualify:

- *Transport-only encryption (TLS, HTTPS).* Encrypts in transit but not at the server. The server sees plaintext.
- *"At-rest" disk encryption with provider-held keys.* Protects against theft of a hard drive. The provider's services decrypt the data for normal operation.
- *"Customer-managed keys" where keys live in a provider-operated key management service.* Improves auditability but the provider's KMS can produce decryption keys on demand from the provider's services. A server-side attacker who can authenticate to the KMS gets keys.
- *"E2E except for the bits we need to read."* Server-side indexing, server-side search, server-side AI summarization, server-side antivirus scanning, server-side eDiscovery — each is a back door that the provider is using deliberately. The system is not E2E once any such feature is enabled.
- *Third-party key servers operated by the same provider or accessible from the provider's infrastructure.* The argument that "we don't hold the keys, our key partner does" fails if the key partner is reachable from the same compromised infrastructure. An AI-accelerated attack against the key server is no different in kind from an attack against the data server.

True end-to-end encryption is therefore a property you can test for, not a marketing term. A vendor whose system does not satisfy the three conditions above — keys only on user devices, decryption only

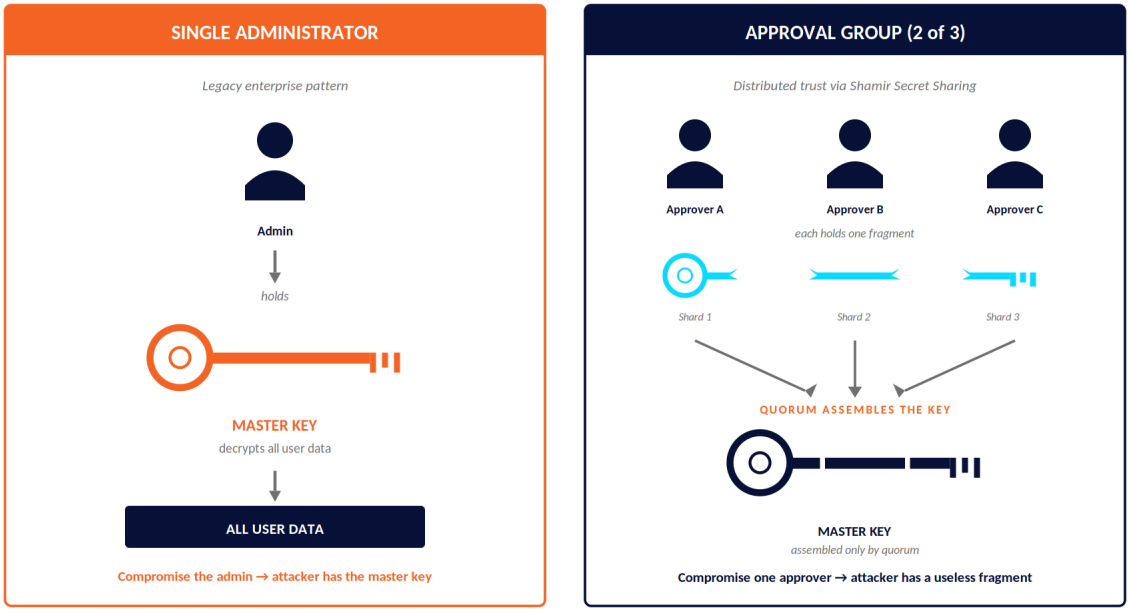
on user devices, no plaintext dependency at the cloud layer — does not deliver it, regardless of what the marketing says.

5. Administrator compromise in the AI era

The administrator is a soft spot common to both legacy cloud-plaintext systems and poorly designed E2E systems. In most enterprise services, a small number of admin accounts concentrate broad privileges — adding and removing users, resetting access, recovering accounts, authorizing data export. If an administrator can perform these operations unilaterally, a compromised administrator is, functionally, a master key. Compromise one admin, and the data of every user under their authority is exposed.

In the pre-AI era, socially engineering an administrator was a known but bounded threat. In the AI era, it becomes a tractable, scalable, and increasingly cheap one. Voice deepfakes, real-time video impersonation, and persona-driven jailbreaks are now commodity capabilities — produced on demand, at scale, by anyone with API access. Google's threat-intelligence unit documented a Chinese group instructing AI to act as a "senior security auditor" to enhance offensive operations; the same technique is being applied against human targets.[4] The cost of producing context-appropriate impersonations is falling, and the cost of running such operations against many administrators simultaneously is falling with it.

Administrator trust models



[Figure 3] Single administrator versus approval group. In a single-administrator model, compromise of one person yields the master key. In an approval group, the key is split into shards held by separate approvers — a quorum is required to reconstruct it, and one compromise yields nothing useful.

The response is cryptographic, not procedural. Shamir Secret Sharing — a 1979 technique now standard in modern security systems — permits a secret to be split into fragments such that no single fragment reveals anything about the original, and a designated quorum of fragments is required to reconstruct it. In a properly designed E2E system, the keys that protect user data are split this way and distributed across multiple administrators. Each administrator holds only a fragment. No single administrator can decrypt anything on their own. To recover or export a user's data requires multiple administrators to combine their fragments — by construction, not by policy.

This is what matters against an AI-armed adversary. Compromising one administrator yields nothing useful — the attacker holds a single fragment, which decrypts nothing. To obtain user data, the adversary must compromise a quorum of administrators simultaneously and coordinate their actions before any of them detects the manipulation. AI lowers the cost of compromising one administrator. It does not lower it in the same way for several administrators in synchrony.

PreVeil productizes this pattern as **approval groups**. Administrative keys are split using Shamir Secret Sharing and distributed across a designated set of approvers, with a quorum required to authorize any sensitive operation — account recovery, data export, eDiscovery, key rotation, addition of new administrators. No single administrator — including a compromised one — can act unilaterally. The cryptography is decades old; PreVeil's productization is what converts the administrator from a master key into a single fragment of one.

6. Honest scope: what end-to-end encryption does not solve

A balanced view requires an honest inventory of E2E's limits.

It does not protect against endpoint compromise. If an attacker controls a user's device, the attacker can read that user's data; the keys are on the device because they have to be. The argument for E2E is not that endpoint risk disappears — it is that endpoint risk does not aggregate the way server risk does, and AI accelerates the aggregating attack more than the non-aggregating one.

It does not encrypt all metadata. Every system must know who is sending mail to whom, when, and in roughly what volume, in order to deliver it. A well-designed E2E system encrypts everything that can be encrypted — message contents, attachments, subject lines, snippets, file names, directory structures — and is explicit about what it cannot. Buyers should ask vendors to list precisely which metadata is and is not encrypted. A vendor that cannot give a precise answer is one whose system is not fully understood.

It does not protect against poorly-chosen approval-group members. Distributed trust depends on the integrity of the parties to whom trust is distributed. A quorum of compromised approvers can do what a single administrator could do without distributed trust. Approval groups raise the cost of compromise; they do not eliminate it.

And it is not a substitute for operational discipline. Endpoint hygiene, phishing-resistant authentication, prompt patching, and user training remain necessary regardless of architecture.

What E2E does is structural. When other defenses fail at the cloud provider — and they will, because they have, repeatedly — the consequences are bounded. **Bounded consequence, not perfect protection, is the property the AI threat environment now requires.**

7. Where federal guidance is heading

The U.S. agencies with deepest insight into adversarial cyber capability are converging on end-to-end encryption with provider-inaccessible keys.

ITAR §120.54, adopted in 2019 by the State Department in consultation with the NSA, codifies this as the legal test for transit of controlled defense data on commercial cloud. The rule's implication is striking: because the cloud cannot read the data, U.S.-persons-only operation of the underlying infrastructure is no longer required. ITAR-controlled technical data may transit and reside on standard commercial cloud — including foreign cloud — provided the encryption test is met. The architectural property of E2E is strong enough that the State Department waives the personnel restriction it otherwise imposes on cloud handling defense data. Two agencies with high-fidelity visibility into nation-state cyber capability set this standard six years before AI made it broadly consequential.[5]

In April 2020, as the COVID-19 pandemic forced unprecedented federal telework, the NSA issued guidance for selecting commercial collaboration services. End-to-end encryption was its top criterion. The crisis added urgency to a direction already set.[7]

CISA reinforced the direction in December 2024. Following the Salt Typhoon breach of eight major U.S. telecommunications providers, the agency issued public guidance recommending that "highly targeted individuals" use only end-to-end encrypted communications, naming Signal as an example. The recommendation is not "encrypt your messages somehow" — it specifies end-to-end encryption with provider-inaccessible keys.[6]

The NSA's Commercial National Security Algorithm Suite 2.0 addresses the cryptographic algorithms of the long term; CISA's operational guidance addresses today's systems. The two align on the same point: data whose loss matters should be protected by encryption whose keys the adversary cannot reach.

CMMC has not yet drawn this distinction explicitly. The standard prescribes controls drawn from NIST 800-171; it does not yet prescribe that keys must be inaccessible to the provider. The next section examines what this means for contractors making CMMC decisions now.

8. What this means for CMMC decisions now

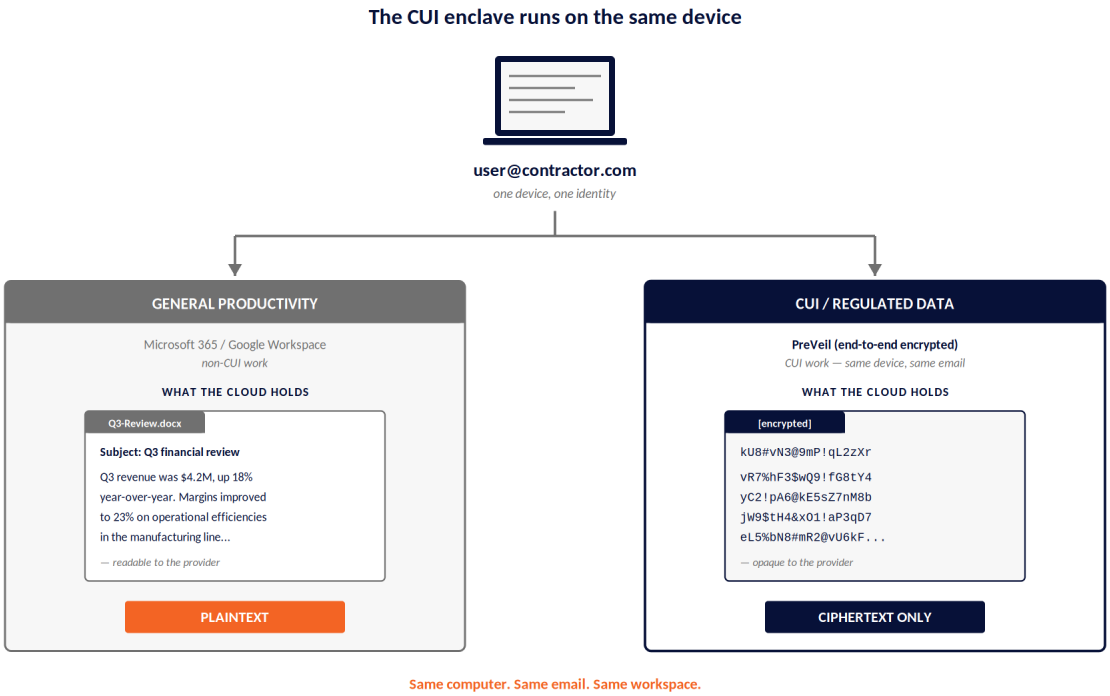
CMMC was created to protect Controlled Unclassified Information. **Compliance is the mechanism; protection of CUI is the purpose.**

In the AI era, those two things can come apart. The DIB now faces a choice between two architecturally different cloud platforms that both satisfy CMMC compliance. Cloud-plaintext platforms — including those certified for defense use — are compliant but increasingly vulnerable to the attacks on cloud infrastructure that AI is making routine and cheap. End-to-end encrypted platforms are compliant — and aligned with the purpose CMMC was created to serve: keeping CUI protected even when the cloud infrastructure is breached. **AI-driven adversaries will attack both. Only one is designed to withstand them.**

Microsoft GCC High represents the former. It is a capable, compliant offering well-suited to large enterprises, with meaningful protections against insider risk and foreign data residency. But the architectural property is unchanged: Microsoft holds the keys, Microsoft's services decrypt the data, and a sufficiently capable attacker reaching that infrastructure reaches plaintext. Storm-0558 occurred in commercial Exchange Online, not GCC High; the same design flaw exists in both, by construction. Certifications certify controls. They do not change the architecture.

PreVeil represents the latter. PreVeil was designed at MIT from the ground up for precisely the threat environment now unfolding — the assumption that attacks on cloud infrastructure would increase and inevitably succeed, and that CUI therefore had to be protected by the architecture itself rather than by a perimeter around it. The cloud holds ciphertext; keys live only on user devices; a breach of the server is not a breach of the data.

The deployment model is a CUI enclave. CUI moves to the end-to-end encrypted platform the provider cannot decrypt; general productivity stays where it is, on existing cloud platforms where the features justify the trade-off. PreVeil sits side-by-side with the user's existing Microsoft 365 or Google Workspace — on their existing computer or within the same VDI — under the same email address. This is a targeted addition, not the multi-quarter, costly migration to GCC High that most contractors consider — and it applies CUI-grade controls only to CUI. The majority of daily work, which is not CUI, remains on the platforms built for it.



[Figure 4] PreVeil sits side-by-side with the user's existing Microsoft 365 or Google Workspace, on their existing computer or within the same VDI. The provider holds plaintext for general productivity and ciphertext only for CUI.

The path is operationally proven. More than **3,000 defense contractors** use PreVeil for CMMC and ITAR workloads, the vast majority small and medium businesses. More than **90 have achieved CMMC certification** on PreVeil. The PreVeil service is FedRAMP Moderate Baseline — the level required by DFARS 7012 and CMMC — with CUI stored on US-Person-Only AWS GovCloud (FedRAMP High).

Both deliver compliance. Only one is built for the AI threat.

Notes

1. Anthropic, "Project Glasswing: Securing critical software for the AI era," April 7, 2026, https://www.anthropic.com/project/glasswing. See also Anthropic Red Team, "Claude Mythos Preview," April 7, 2026, https://red.anthropic.com/2026/mythos-preview/.

2. AI Security Institute (UK), "Our evaluation of Claude Mythos Preview's cyber capabilities," April 13, 2026, https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities.

3. Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign," November 13, 2025, https://www.anthropic.com/news/disrupting-AI-espionage. Full report:

https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf.

4. Google Threat Intelligence Group, "Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access," May 11, 2026, https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access.

5. U.S. Department of State, "International Traffic in Arms Regulations: Creation of Definition of Activities That Are Not Exports, Reexports, Retransfers, or Temporary Imports," 84 Fed. Reg. 70887, December 26, 2019. Codified at 22 CFR §120.54. https://www.federalregister.gov/documents/2019/12/26/2019-27438/international-traffic-in-arms-regulations-creation-of-definition-of-activities-that-are-not-exports.

6. Cybersecurity and Infrastructure Security Agency, "Mobile Communications Best Practice Guidance," December 18, 2024, https://www.cisa.gov/news-events/alerts/2024/12/18/cisa-releases-best-practice-guidance-mobile-communications.

7. National Security Agency, "Selecting and Safely Using Collaboration Services for Telework," U/OO/134598-20 PP 20-0713, April 2020 (revised through November 2020), https://media.defense.gov/2021/Sep/16/2002855944/-1/-1/0/CSI_%20SELECTING_AND_USING_COLLABORATION_SERVICES_SECURELY_FULL.PDF.