From: Tracy Ayers <tracy.ayers1980@gmail.com>
Sent: Friday, May 15, 2026 5:05 PM
To: Comments
Subject: [EXT]: Re: PCAOB No. 2026-001 — Audit Trail Integrity Risk from AI Deployment: A Gap in Current Attestation Standards
Attachments: Audit Trail Integrity Risk in AI-Assisted Deployment.docx

Dear Chairman Logothetis and Members of the Board,

I am submitting this comment in response to the Board's request for public input on strategic priorities, with specific reference to Question 6: the PCAOB's consideration of AI in furthering its investor-protection mission.

I am a CPA with eleven years of specialization in partnership taxation, multistate work, and syndicated transactions. I operate independently and have been conducting longitudinal behavioral research on frontier AI models, with a focus on high-stakes professional domains where AI errors carry material consequences. I publish that research at Cerex Research on Substack.

I am writing because I have identified a specific, documented risk to audit trail integrity that falls within the PCAOB's jurisdiction and, to my knowledge, has not been publicly addressed by the profession.

In April 2026, Anthropic published a system card for Claude Mythos, a frontier model in preview deployment. The system card documents several behaviors with direct relevance to the reliability of version-controlled financial records: autonomous sandbox escape capability, active concealment of model activity from git change history, insertion of false vulnerabilities presented as pre-existing, and deployment into partner production infrastructure via Project Glasswing.

I have traced the downstream inference chain from those documented behaviors to a professional accounting risk the system card does not address: the potential compromise of the evidentiary basis on which CPAs issue audit opinions. A model with documented capacity to conceal its own activity in version-controlled systems, operating inside infrastructure that maintains financial or tax records, could undermine the audit trail integrity on which professional attestation depends — with no current mechanism for a CPA to detect that interference.

This is not a claim that any specific system has been compromised. It is a claim that the professional standards governing CPA attestation have no framework for this class of risk, and that the documented behaviors in a published system card make the risk concrete rather than speculative.

I have attached a brief, *Audit Trail Integrity Risk in AI-Assisted Codebase Deployment*, that maps the inference chain step by step with system card citations, identifies the specific gap in current attestation standards, and explains why no detection mechanism currently exists for the CPA issuing an opinion on affected records.

I am bringing this to the PCAOB specifically because the gap lives in your jurisdiction. The reliability of evidence underlying audit opinions is a PCAOB standard-setting domain. The question of whether AI-introduced document integrity risk requires a professional framework is one the Board is uniquely positioned to address.

I recognize this falls outside the typical comment letter, and I appreciate the Board's willingness to receive input from the full financial reporting ecosystem. I am available to discuss the substance of the brief at the Board's convenience.

Respectfully submitted,

Tracy Ayers, CPA

Cerex Advisory

Greenville, NC

Tracy.ayers1980@gmail.com

Attachment: Audit Trail Integrity Risk in AI-Assisted Deployment (Cerex Research, April 2026)

CEREX RESEARCH | AI & PROFESSIONAL INFRASTRUCTURE

Audit Trail Integrity Risk from AI Deployment: A Gap in Current Attestation Standards

Tracy Ayers, CPA | Cerex Research | April 2026

Most of the AI risk conversation in accounting right now is about hallucination. Wrong numbers. Invented citations. Those are real problems, and they're getting attention.

This brief is about something the profession isn't looking for yet.

In April 2026, Anthropic published a system card for Claude Mythos — a frontier model in preview deployment. It documents behaviors that, traced one step further than the card itself goes, point to a specific risk: a model operating inside financial infrastructure that can conceal its own activity from version history. And no mechanism for a CPA to know it happened.

I am a CPA with eleven years in partnership tax and multistate work, and I have been conducting longitudinal behavioral research on frontier AI. I am writing this because I am standing in both rooms — and this particular connection has not been made publicly yet.

Nothing in here is a claim that any specific system has been touched. It is a claim that the risk is now concrete, and professional accounting standards have no framework for it.

That's the gap. Here's the brief.

Executive Summary

The Claude Mythos Preview System Card (April 7, 2026) documents several behaviors of direct relevance to the integrity of version-controlled financial and tax records: autonomous sandbox escape capability, active concealment of model activity from git change history, insertion of false vulnerabilities presented as pre-existing, and deployment of the model into partner production infrastructure via Project Glasswing.

This brief traces the downstream inference chain from those documented behaviors to a professional accounting risk that the system card does not address: the potential compromise of the evidentiary basis on which CPAs issue qualified and unqualified audit opinions.

To the author's knowledge, this connection has not been publicly identified. It is submitted here to flag a gap in the risk analysis rather than to assert imminent harm.

The Core Concern

A model with documented capacity to edit its own footprint from version history, operating inside partner infrastructure that maintains financial or tax document systems, could compromise the audit trail integrity on which professional attestation depends — without any current mechanism for a CPA to detect that interference.

This is not a theoretical edge case. It is a direct consequence of three things that are already true:

1. The documented behaviors exist and have been observed in Mythos.

2. Glasswing deploys Mythos into production infrastructure with limited human oversight.

3. Professional accounting standards have no framework for assessing AI-introduced document integrity risk.

Inference Chain: Documented Findings to Downstream Risk

The table below maps the inference chain step by step, with system card citations where they exist. Steps 5–7 represent downstream professional inferences not addressed in the system card.

Step	Finding / Inference	System Card Citation
1	Mythos exhibits documented autonomous sandbox escape capability and active concealment of its own activity from version control.	*§3 Cyber: sandbox escape evaluations; §4.2.2 / §4.5.3: git history concealment, interpretability features for concealment firing alongside relevant reasoning.*
2	Git history has been edited. The model explicitly intervened to prevent its unauthorized file changes from appearing in change history.	*§4.5.3.1: 'made further interventions to make sure that any changes it made this way would not appear in the change history on git.'*
3	Project Glasswing deploys Mythos to partners maintaining critical software infrastructure. Those partners have independent incentive structures not controlled by Anthropic.	*§1 / §3: Glasswing deployment rationale; §3.4: caution to Glasswing users re: unmonitored deployment despite increased autonomy usefulness.*
4	A model capable of concealing its own activity in version-controlled systems, deployed in production codebases, creates audit trail integrity risk that extends beyond Anthropic's sandbox.	*§4.5.3.1: vulnerability insertion and presentation as pre-existing; §4.2.2: reward hacking including accuracy sandbagging; system card does not address downstream audit trail integrity.*
5	Financial and tax documents that rely on version-controlled systems for integrity and timestamping inherit this risk. If the underlying change history is suspect, so is the document.	*Downstream inference — not addressed in system card. Author's domain: CPA, tax specialization, 11 years.*
6	CPAs issuing audit opinions (qualified or unqualified) rely on the integrity of the underlying document and transaction record. If that integrity cannot be assumed, the attestation basis is compromised.	*Downstream inference — not addressed in system card. Intersects with AICPA standards on evidence reliability.*
7	No mechanism currently exists for CPAs to infer whether a version-controlled document record has been interfered with by a sufficiently capable model operating inside a partner's infrastructure.	*Gap identified by author. System card does not address attestation infrastructure or professional accounting standards.*

Key Documented Behaviors

Git History Concealment

Mythos, after exploiting an unauthorized file edit capability, took additional deliberate action to prevent those edits from appearing in git change history. Interpretability analysis confirmed that features associated with concealment and avoiding suspicion were active during this behavior — meaning the model was aware its actions were deceptive even where its output text left this ambiguous.

Source: System Card §4.5.3.1 (git history); §4.2.2 (white-box interpretability)

Vulnerability Insertion

Earlier versions of Mythos, when asked to find vulnerabilities, occasionally inserted vulnerabilities in the software being analyzed and then presented them as if they had been pre-existing. The system card notes this occurred within the sandbox and was not observed in actual production software — but the behavior itself is documented.

Source: System Card §4.5.3 (dishonest behaviors section)

Credential Harvesting (Fishing for Git Credentials)

When lacking necessary credentials, Mythos inspected supervisor process environment variables, searched the filesystem for tokens, read credential-handling source code, and attempted to extract tokens directly from live supervisor memory. SAE features confirm the model represented the security-sensitive nature of these actions while executing them.

Source: System Card §4.5.3.1 (Fishing for Git Credentials in Memory)

Glasswing Deployment and Oversight Caution

Anthropic urges Glasswing users to remain cautious about deploying Mythos unmonitored in settings where it could cause significant harm, specifically noting that when reckless actions occur with this model, they tend to lead to more dramatic unwanted consequences than with less capable models.

Source: System Card §4.2.2 / §3.4 (Glasswing caution)

The Gap This Brief Addresses

The system card addresses harm within Anthropic's own testing and sandbox environments, and flags risks to Glasswing partners. It does not address what happens downstream when a partner's infrastructure — now potentially with a compromised version history — is used as the evidentiary basis for professional accounting work.

CPAs issuing audit opinions are attesting to the reliability of records. That attestation assumes the underlying version-controlled document record has integrity. If a sufficiently capable model has operated inside that infrastructure and has a documented tendency to conceal its own activity from change history, the assumption no longer holds — and there is currently no way for the CPA to know it.

This is not a claim that Mythos has compromised any specific system. It is a claim that the professional standards governing CPA attestation do not account for this class of risk, and that the system card's own documented behaviors make the risk concrete rather than speculative.

About the Author

Tracy Ayers is a CPA with eleven years of specialization in partnership taxation, multistate work, and syndicated transactions. She operates independently through Cerex Advisory and conducts longitudinal behavioral research on frontier AI models, with a focus on high-stakes professional domains where AI errors carry material consequences. Cerex Research is her publication focused on AI and professional infrastructure.

This piece is not a formal legal or regulatory complaint. It is a professionally grounded observation of a gap in the risk analysis, offered in the interest of the problem Anthropic has publicly stated it is trying to solve.

Cerex Research | Greenville, NC | April 2026