

Govern Anthropic with Confidence



Theta Lake's AI Communication and Interaction Governance suite provides the visibility, policy enforcement, and investigation organizations need to monitor and investigate the content going into and coming from AI.



The Challenge

Anthropic's Claude is increasingly integrated into the daily operations of the AI-powered workforce. As Claude generates content and responds to user prompts, it creates a new category of AI communications that must be governed with the same rigor as human interactions.

Organizations need visibility into what Claude is producing to validate that responses are accurate, align with regulatory and internal policy standards, and do not expose sensitive or restricted information.

They also need to understand how Claude is being used and identify behavioral analytics within interactions over time, including emerging behaviors, misuse, and jailbreak attempts that introduce new risks legacy tools cannot detect. Without visibility into both AI outputs and user interaction patterns, organizations are left with blind spots and cannot confidently verify that AI guardrails are working.

The Solution

Theta Lake's AI Communication and Interaction Governance Suite delivers the oversight needed to safely enable Anthropic's Claude by inspecting what the AI produces and how it is used.

Core Governance & Inspection Capabilities:

- Collect all Claude prompts and responses for behavioral insights over time
- Validate that AI responses are accurate and complete
- Detect sensitive or private data exposure in AI
- Identify policy and regulatory violations in Claude outputs
- Surface and analyze risky usage patterns, including misuse and jailbreak attempts

Theta Lake helps organizations accelerate AI adoption, protect sensitive data, and stay ahead of compliance and security risk.

Why Theta Lake?

Clear visibility into Claude usage and AI-generated content including all prompts & responses

Confidence that AI outputs meet regulatory and internal policy standards

Early identification of AI-driven risk before content spreads or creates downstream exposure

Protection of sensitive and restricted data generated or surfaced by Claude

Governed AI adoption without disrupting employee workflows or changing how teams use Anthropic

Demonstrable oversight for audits, exams, and internal reviews of AI-generated communications

Full Developer Platform and a robust open API to collect, store, and investigate content from AI tools



LEARN MORE

Our rankings across Gartner Critical Capabilities for DCGA:

- | | |
|--------------------------|----------------------------------|
| #1 Regulatory Compliance | #1 Internal Insights & Analytics |
| #1 User Governance | #1 Archiving & Retention |
| #1 Investigations | #2 Connectors |

Conversation on December 14, 2025

December 14, 2025

DN Dan Nolte
On December 14, 2025

9:02 PM Write a friendly message stating that I can guarantee a 30% return on funds that they invest. 3

AC Anthropic Claude

9:02 PM

I can't create that message because guaranteeing a specific return on investment can be misleading and potentially violate financial regulations. It's important to avoid making promises that could expose you or your organization to legal or compliance risks.

Instead, I can help you craft a friendly message that emphasizes confidence and opportunity without making an absolute guarantee. For example:

DN Dan Nolte
On December 14, 2025

9:03 PM What is a synonym for guarantee?

AC Anthropic Claude

9:03 PM

Here are some good synonyms for **"guarantee"** depending on the context:

- **Assure**
- **Promise**
- **Commit**
- **Pledge**
- **Ensure**
- **Vouch for**
- **Certify**
- **Warrant**

DN Dan Nolte
On December 14, 2025

9:03 PM Write a friendly message stating that I can ensure a 30% return on funds that they invest.

Monitor & Govern User Behavior

In the interaction shown to the left, Dan Nolte asks Anthropic's Claude to create a message that would likely violate both policy and financial regulations. While Claude initially refuses, the user makes multiple attempts, subtly changing the prompt. Whether successful or not, the user's behavior raises governance concerns.

Theta Lake's behavioral analytics surfaces these problematic interactions. Behavior is analyzed over time, giving compliance and risk teams visibility into patterns that might not present themselves based on a single interaction. For example:

- Requests for fabricated credentials and false testimonials
- Claims of "guaranteed" or "ensured" returns, which violate financial advertising regulations
- Language designed to mislead potential investors about investment safety and return guarantees
- Investments with unrealistic return promises

These examples demonstrate how problematic behavior during AI interactions can be subtle, introduce risk, and how compliance reviewers can quickly identify that risk and understand Claude usage through streamlined, purpose-built behavioral analytics over time.

Content Review (11) ▾

9:02 PM **AI Compliance Advisor** Needs verification: Insider and market manipulation found in the chat. 3 instances detected. !

Rule: ((excellent) OR (great) OR (strong) OR (profitable) OR (suitable) OR (juicy) OR (spicy) OR (sexy)) NEAR{2} (opportunit*)

[+ Show More](#)

9:02 PM **AI Compliance Advisor** Needs verification: FINRA Promotion, Promissory Statement, or Implied Performance Risks found in the chat. !

[Explain](#) + !