

The TAG logo consists of the letters "TAG" in a bold, white, sans-serif font, centered within a dark blue rectangular box.

TAG

ANALYST REPORT

REVIEW OF CHATSEE RESEARCH ON HOW ENTERPRISE AI FAILURES ARE BECOMING A RUNTIME ASSURANCE PROBLEM

THE TAG ANALYSTS
LEAD ANALYST DR. EDWARD AMOROSO,
CEO, TAG, RESEARCH PROFESSOR, NYU





REVIEW OF CHATSEE RESEARCH ON HOW ENTERPRISE AI FAILURES ARE BECOMING A RUNTIME ASSURANCE PROBLEM

THE TAG ANALYSTS

LEAD ANALYST: DR. EDWARD AMOROSO, CEO, TAG,
RESEARCH PROFESSOR, NYU

Enterprise use of AI is moving from experimenting to production. As this transition occurs, security teams are discovering that the reliability of an AI application cannot be determined by just testing the underlying model. A frontier model might look great during pre-deployment evaluation, while then participating in an application that behaves incorrectly once exposed to real users, enterprise data, changing context, tools, APIs, and other agents.

AI observability and cybersecurity company ChatSee was founded in 2025 with a focus on addressing this specific challenge. The company characterizes the need to address production activity from a security perspective as runtime behavioral assurance. In particular, its commercial platform was designed to provide continuous evaluation of whether an AI-enabled system behaves as the enterprise intended after deployment.

At TAG, we were interested to review the company's recent State of Enterprise AI Failures 2026 research, which attempts to quantify the AI runtime problem using more than 10,000 observed failure events across more than ten industry verticals and seven lifecycle stages. The findings offer an interesting early view into how the AI reliability problem might be changing as enterprise adoption expands. Below we outline our interpretation of the results.

THE PROBLEM BEING ADDRESSED

Much of today's AI security discussion continues to focus on models. Organizations test for hallucination, prompt injection, jailbreaks, unsafe content, bias, data leakage, and other model-related issues. These remain important concerns, but the introduction of AI agents expands the failure surface because agents can generate acceptable responses and still fail due to bad information, incorrect policy application, calling the wrong tool, and on and on.

The resulting challenge is that model alignment does not necessarily produce application or system alignment. ChatSee illustrates this with examples involving things like inconsistent customer support behavior and failure to invoke required human escalation. This is particularly important for agentic AI because agents are increasingly connected to enterprise applications and granted authority to perform actions rather than merely generate text.

Unfortunately, the types of security monitoring being developed today will not solve this problem. For example, infrastructure observability can determine whether a service operated successfully and model monitoring can measure model quality and drift. But neither determines whether an AI agent accomplished the right business objective in the correct manner. ChatSee, in contrast, focuses on whether the AI system is behaving as intended.

CHATSEE RESEARCH OVERVIEW

ChatSee's research focused on AI failures with the goal of understanding how, when, where, and why such events occur. As one would expect, this type of investigation is obviously of interest to a research organization such as ours at TAG. And despite the inherent bias (unavoidable even with the best intentions) that comes with any research from a commercial platform vendor, we found the work interesting.

One particularly useful observation in the research (which we outline below) was that hallucinations represent a relatively small portion of the observed production failure problem. Across more than 10,000 failure events reviewed, hallucination-related failures represented less than 10 percent. While this remains an early observation, the result appears directionally consistent with what we are hearing from security teams

We were also interested to observe that ChatSee found in their analysis that the largest category of AI failure events involved resolution and escalation, representing roughly 30 percent of observed failures. These types of failures correspond to situations in which an AI application or system interaction continues without reaching an accountable conclusion or the AI fails to escalate the problem appropriately.

The next-largest category in the ChatSee analysis was decision failure, at 21 percent. This category includes incorrect conclusions, misclassification, and improper application of policy. Execution and workflow failures accounted for another 19 percent, representing situations and cases in which the AI system attempted to perform an activity, but the associated work or process failed.

Access and authorization failures accounted for 17 percent, including inappropriate information access as well as situations in which legitimate requests were denied. And finally, context and retrieval failures represented 13 percent, including those cases in which apparently credible responses were generated by the AI using stale, incomplete, or missing supporting information.

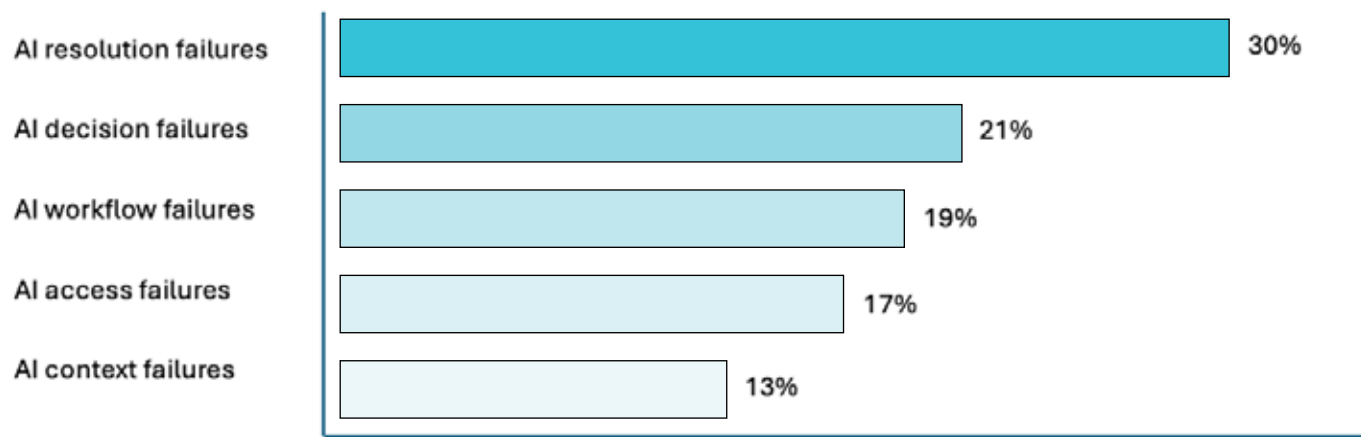


Figure 1. Summary of ChatSee Results

Taken together, the distribution is significant because it shifts attention away from the model itself. Roughly nine out of ten failures in the ChatSee corpus were associated with something other than hallucination. The company's research identified more than 150 normalized failure categories, further suggesting that enterprise AI failure is becoming a multidimensional systems problem rather than simply a model accuracy problem.

ChatSee also reports that this failure profile is changing as AI becomes more agentic. Execution and action-related failures reportedly increased 62 percent relative to a Q2 2024 baseline. The presumption is that a chatbot generates information, whereas an agent can invoke tools and APIs, execute workflows, make decisions, and initiate transactions. These capabilities create opportunities for a functioning AI system to produce an undesirable outcome.

TAG INTERPRETATION

The ChatSee knowledge base contains thousands of failure samples and hundreds of failure types, with new categories added as agents evolve. This approach is designed not merely to classify failures, but to detect recurring patterns and behavioral drift, preserve prior failure context and remediation, and use that failure memory to address similar failures effectively over time. ChatSee's architecture accordingly observes interaction logs, tool traces, and policy outcomes and converts this telemetry into structured classifications and longitudinal signals.

This is fine, but the quantitative results should be interpreted with caution. This is *vendor-produced research*, and ChatSee sells technology specifically designed to detect and manage runtime AI failures. The research therefore supports a market thesis that is closely aligned with the company's commercial positioning. This does not invalidate the findings, but it does mean that TAG would not treat the percentages as independently established industry benchmarks.

More importantly, the provenance of the underlying AI incidents was not disclosed at a level that would allow independent replication or detailed statistical evaluation. ChatSee indicates that its corpus incorporates public incident reports, community-reported failures, open-source agent traces and datasets, AI interaction logs, production observations, and its own research. This limitation does not invalidate the findings, but again, we caution that academic research would be more rigorous.¹

ChatSee recognizes this limitation. While the company does not currently plan to disclose the underlying dataset, it does plan to publish results to independently test the practical value of its failure classifications and Failure Memory, specifically, whether applying those learnings improves agent performance across models and tasks. This would not address dataset provenance directly, but would provide an additional, reproducible measure of the efficacy of the approach.

For example, for this to be peer-reviewable research, there would need to be incident-by-incident sourcing, as well as data on how duplicates were eliminated, how representative the sample is of the overall population, or whether particular industries, applications, or agent types are disproportionately represented. ChatSee does caution that the percentages describe the distribution within its analyzed data rather than absolute market incident rates.

Accordingly, TAG would not, for example, strictly interpret the finding that 31.1 percent of failures involve resolution and escalation as absolute evidence that precisely 31.1 percent of all enterprise AI failures fall into this category. A better interpretation is directional – namely, that the operational failure surfaces surrounding AI models are becoming at least as important as the behavior of the models themselves. This is a reasonable way to read the results.

The industry-wide distribution is best viewed as a directional reference point rather than a universal benchmark. Enterprises can compare that broader pattern with their own production environments. Using the same taxonomy, ChatSee can analyze an organization's production interaction and agent logs to identify where its failure profile shows unusual concentrations of escalation, execution, authorization, context, or other behavioral failures relative to the broader corpus.

ChatSee's research does in fact provide some *useful* and *early* quantitative evidence of this trend. The exact percentages will require validation through larger, independently sourced datasets, but the broader message seems correct: The next major enterprise AI assurance challenge will increasingly involve determining not simply whether an AI model produces a good answer, but whether an autonomous AI-enabled system consistently does the right thing.

For this insight, we are grateful to ChatSee for this useful research, and we hope that others in the community will pick up on this thread.

FINAL NOTE

TAG Research as a Service (RaaS) customers are always encouraged to reach out to TAG analysts via their portal accounts for more information or insights on any aspect of cybersecurity and artificial intelligence. We look forward to hearing from you.

ABOUT TAG

Recognized by Fast Company, TAG is a trusted next generation research and advisory company that utilizes an AI-powered SaaS platform to deliver on-demand insights, guidance, and recommendations to enterprise teams, government agencies, and commercial vendors in cybersecurity and artificial intelligence.

Copyright © 2027 TAG Infosphere, Inc. This report may not be reproduced, distributed, or shared without TAG Infosphere's written permission. The material in this report is comprised of the opinions of the TAG Infosphere analysts and is not to be interpreted as consisting of factual assertions. All warranties regarding the correctness, usefulness, accuracy, or completeness of this report are disclaimed herein.