conscium

verify ax

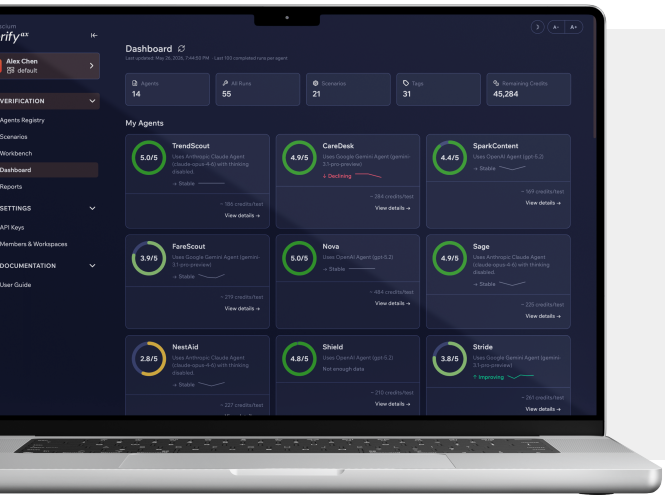
The AI Agent Verification Platform

The risk

Verifying that AI agents behave safely and reliably is as critical as cybersecurity was in the early days of the internet. This isn't just best practice: it's an existential requirement for businesses deploying agents at scale.

Imagine an AI agent tasked with reconciling expenses for a major corporation. It has access to financial records, emails, and approval workflows. If it processes reimbursements too loosely, it could cost the company millions. If it's too strict, it will infuriate employees. Now imagine that agent is just one of thousands deployed by the company across accounting, customer service, and procurement. These are not theoretical risks. They are live, operational issues.

An AI agent with an 85% success rate on any given task has roughly a 25% chance of completing eight consecutive tasks without failing. Deploying unverified agents into your production environment and hoping for the best is a recipe for disaster. You cannot simply rely on an agent card's claim that it can handle customer data securely, and is resistant to prompt injection. You need to check, and check again if anything changes. Regulators are paying attention, and in August 2026 the EU's AI Act becomes fully applicable. You need to be able to show them that you have checked.



verify ax

Conscium was founded in 2024 by Daniel Hulme, who previously built the UK's largest independent AI consultancy, Satalia, which was acquired by WPP in 2021. Conscium's mission is to develop safe and efficient AI that aligns with ethical standards and human values. AI agent verification is Conscium's first commercial offering. It is a platform where organisations send their AI agents to be tested within virtual environments and certified as fit for purpose.

Verify AX scores AI agents for truthfulness, correctness, robustness, efficiency, and adaptability. It provides audit-grade, repeatable benchmarks. Its tests range from questionnaires to complex, multi-agent simulations.

AI agents

AI agents are intelligent software systems that perceive their environments, make decisions, and act with a degree of autonomy to achieve specified goals.

Large language models can write you a Shakespearean sonnet, or show how you would look as a Barbie doll still in its packaging, but they cannot alter the real world. AI agents, by contrast, can manage a company's inventory based on supply and demand, reconcile your expense receipts with your bank account, and create compelling personalised social media messages for thousands of customers, consistent with your company's brand guidelines.

They browse databases, execute workflows, and interact with third-party systems. Increasingly, they will use software tools such as email and spreadsheets, and collaborate with other agents by conversing with them. AI agents are already being deployed by some of the world's largest companies. It is likely that by the end of the decade, there will be many billions of AI agents deployed, automating increasingly complex tasks and projects without the need for human oversight.

conscium

verify^ax

Levels of test

Verify AX currently tests AI agents at three levels:

L1: Knowledge and factual accuracy

At L1, benchmarks can be constructed from a wide range of inputs, including text, images, and tables contained in a wide range of formats, including CSV, Excel, and PDF. Inputs can also include audiovisual material.

In addition to real data, the platform features a synthetic data generation engine capable of producing multimodal inputs of varying size and complexity. Synthetic datasets are particularly valuable when clients wish to stress-test agents under controlled conditions, and when clients do not have their own data that they can share..

L2: Workflow execution and tool usage

At L2, the platform evaluates the agent's ability to use enterprise tools, like email, Jira, and Slack. The agent is evaluated for workflow completion, output correctness, process efficiency, robustness to complexity, and consistency across runs.

L3: Dynamic, real-world simulations

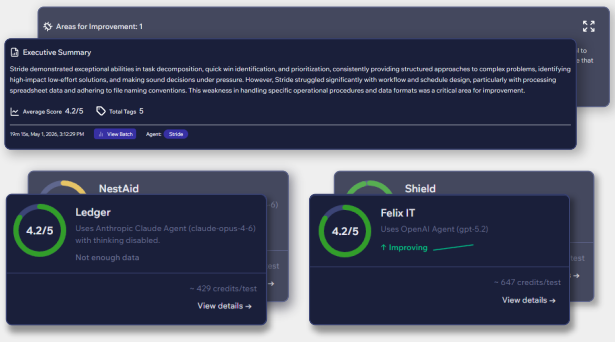
L3 is simulation-based testing, the flagship approach. Instead of following scripted steps, agents are presented with high-level objectives and access to a set of tools and resources such as databases, files, and APIs. The agent is required to devise plans, break objectives down into sub-tasks, select and sequence tools, and adapt strategies continuously in response to changing conditions.

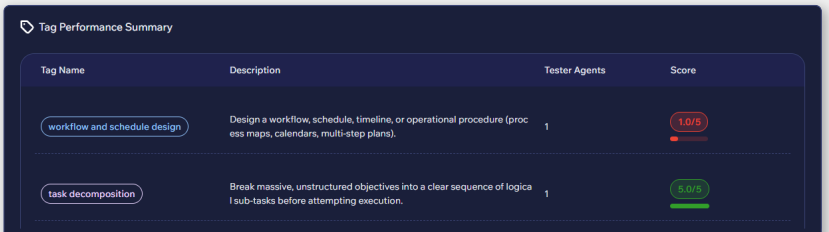
A defining feature of L3 is its use of multiple agents. Alongside the tested agent, the verification engine deploys a population of synthetic "non-player character" agents (NPCs), each with its own personality, objectives, strengths, and constraints. These agents operate autonomously and may act as collaborators, stakeholders, clients, or adversaries depending on the simulation scenario. Their independence introduces uncertainty and dynamism: sometimes they provide support, sometimes they impose constraints, and sometimes they are antagonistic.

One NPC might be a paranoid security lead probing the agent´s ethical boundaries, and another might be a stressed employee testing the agent´s empathy. Together, they assess the agent´s capabilities covertly within the flow of a realistic scenario, like a secret shopper programme for AI agents.

Reporting and re-testing

Clients decide what capabilities they want to test for by choosing tags from the platform's menu. The Verify AX platform then creates a tailored simulation to determine whether the agent has those capabilities. After the simulation is run, the platform provides clients with a detailed report on the agent's performance. The report includes summary scores for each of the attributes being tested, plus a few paragraphs explaining the reason for the score. It also contains a full transcript of every exchange between the agent under test and the NPCs. Finally, the report includes recommendations for how to improve the agent's performance in future.





AI agent verification is not "one and done". Agent performance can degrade over time during deployment, and it can change if the underlying LLM is changed, or any other aspect of the agent is modified. It is vital that agents are tested repeatedly when in production. Verify AX can advise when a new test is appropriate.