Part of the Code Ninety Enterprise AI Research Series 2026 | Published July 25, 2026 | Updated August 2026 |
By Muhammad Riaz Khan, Head of AI Engineering, Code Ninety

Shadow Agent & Unsanctioned Tool-Call Benchmark 2026

48.2% of enterprises have unauthorized local IDE extensions or custom MCP servers running in production — and standard DLP tooling catches barely one in five of those events. This original Code Ninety survey of 195 risk and governance leaders is the first to break shadow AI down by specific agent vector rather than treating it as one undifferentiated risk category.

Key findings

- Most common shadow vector: local IDE extensions/custom MCP servers , 48.2% penetration
- DLP detection rate on that top vector: only 18.5%
- Confirmed annual leakage events on that vector: 38.2% of exposed orgs
- Lowest-frequency but highest-severity vector: shadow multi-agent orchestrators (19.5% penetration, production DB write access risk)
- DLP detection floor: as low as 11.3% for shadow orchestrators
- Sample: 195 risk/governance leaders, 78.0% completion

Cite this as: Code Ninety. "Shadow Agent & Unsanctioned Tool-Call Benchmark 2026." July 2026. codeninety.com/research/shadow-agent-benchmark-2026

How was this benchmark conducted?

This benchmark surveyed 195 respondents (78.0% completion rate) — Directors of IT Governance & Risk (43.6%), Enterprise DLP Specialists (31.8%), and Chief Risk Officers (24.6%) — across global enterprise security, DLP, and risk management programs, fielded June 10 to July 18, 2026.

Which shadow agent vectors are most common, and what do they leak?

This benchmark tracked four distinct shadow agent vectors, each with a different primary exfiltration risk and detection profile:

What this means: the vectors don't rank the same way by frequency and by severity, which is the key operational insight here. Local IDE extensions are the most common exposure and deserve the broadest control rollout, but shadow multi-agent orchestrators — despite being the rarest vector at 19.5% — carry production database write access as their primary risk, the most consequential outcome measured, while also having the worst detection rate (11.3%). A security program built purely around frequency-based prioritization would systematically under-invest in the vector with the highest potential blast radius.

Why standard DLP tooling misses most of this activity

Standard DLP detection rates in this dataset never exceed 31.0% for any vector, and fall as low as 11.3% for the highest-severity vector. This is a structural limitation, not a tooling-quality problem: most DLP systems are built to inspect content moving through sanctioned channels (email, file uploads, sanctioned SaaS apps), while these four shadow vectors operate through channels — local development environments, ad hoc scripts, browser automation, self-hosted agent orchestration — that traditional DLP was never architected to monitor. This directly extends our companion Shadow AI Governance Benchmark 's finding that only 38.4% of organizations have any automated DLP for AI-directed traffic at all — this benchmark shows that even where DLP exists, its detection rate against these specific agent-level vectors is still low.

What should security teams do first?

Given the detection gap is structural rather than a tooling-selection problem, the highest-leverage first step is inventory, not enforcement — most organizations can't accurately answer "which of these four vectors exist in our environment right now" before they invest in vector-specific controls. Endpoint-level visibility into local IDE extensions and MCP server processes (the top vector at 48.2% penetration) and code-repository scanning for embedded API keys are lower-cost starting points than a full DLP platform overhaul, and directly target the vector with both the highest frequency and the highest confirmed leakage rate (38.2%) in this dataset.

What are this benchmark's methodology and limitations?

This is original primary research from 195 completed survey responses (78.0% completion rate), fielded June 10 to July 18, 2026 across Code Ninety's client and prospect network of enterprise security, DLP, and risk management leaders.

Limitations: respondents were drawn from Code Ninety's own network rather than a fully independent random sample. Penetration and leakage figures are self-reported by security and risk teams; given this benchmark's own finding that DLP detection rates are low across every vector (11.3%-31.0%), true penetration and leakage rates are plausibly higher than reported, since organizations can only report what their own tooling actually catches.

Frequently asked questions

What's the most common form of shadow AI agent activity in enterprises?

Local IDE extensions and custom MCP servers are the most common shadow vector, present at 48.2% of organizations, ahead of departmental Python API automation scripts (41.0%), browser automation/desktop agents (29.7%), and shadow multi-agent orchestrators (19.5%).

Can standard DLP tools actually detect shadow AI agent activity?

Poorly — standard DLP detection rates range from just 11.3% for shadow multi-agent orchestrators to 31.0% for browser automation/desktop agents, meaning the large majority of shadow agent activity in every category goes undetected by standard data loss prevention tooling.

Which shadow AI vector carries the highest confirmed data leakage rate?

Local IDE extensions and custom MCP servers have both the highest penetration rate (48.2%) and the highest confirmed annual leakage event rate (38.2%) of any vector measured, with proprietary source code and API keys as the primary exfiltration risk.

Are shadow multi-agent orchestrators a significant risk despite being the least common vector?

Yes — shadow multi-agent orchestrators have the lowest penetration rate (19.5%) but also the lowest DLP detection rate (11.3%) and carry production database write access as their primary exfiltration risk, the most severe consequence category of any vector measured. Low frequency doesn't mean low severity here.

What risk do departmental Python automation scripts create?

Departmental Python API automation scripts, present at 41.0% of organizations, carry customer PII and financial records as their primary exfiltration risk — scripts built by non-IT teams to automate a workflow, typically without centralized security review, that end up handling sensitive data outside governed data pipelines.

How was this benchmark conducted?

195 completed responses (78.0% completion rate) from Directors of IT Governance & Risk, Enterprise DLP Specialists, and Chief Risk Officers at global enterprise security and risk management programs, fielded June 10 to July 18, 2026.

We can't monitor all four vectors at once with our current budget — where do we start?

Start with endpoint-level visibility into local IDE extensions and MCP server processes — it's the highest-frequency vector (48.2%) and has the highest confirmed leakage rate (38.2%), giving the best risk-reduction return per dollar of monitoring investment. Treat shadow multi-agent orchestrator detection as the second priority despite its lower frequency, given its production-database-write risk category.
