# 76%

## of tool calls in AI agents have **zero safeguards**

Reality Check — A static analysis benchmark across 19 open-source AI agent codebases

16 Python repos · 3 TypeScript codebases · 16,500+ source files scanned

| | |
|---|---|
| **Scanner** | diplomat-agent v0.2.0 (Python) · diplomat-agent-ts v0.1.0 (TypeScript) |
| **Date** | May 15, 2026 |
| **License** | Apache 2.0 |
| **Method** | Static AST analysis only — zero runtime dependencies (Python) |
| **Contact** | contact@diplomat.run · diplomat.run |

AT A GLANCE

## Benchmark summary

| 76% | 83% | 77% |
|---|---|---|
| Python — unguarded | TypeScript — unguarded | Combined — unguarded |
| 14 valid repos, 7,029 tool calls | 4 scopes, 669 tool calls | 19 codebases, 16,500+ files |

| Ecosystem | Repos | Files | Tool calls w/ side effects | Unguarded |
|---|---|---|---|---|
| Python | 16 (14 valid) | ~5,100 .py | **7,029** | **5,344 (76%)** |
| TypeScript | 3 codebases, 4 scopes | 11,379 .ts/.tsx | 669 | 553 (83%) |
| **Combined** | **19 codebases** | **~16,500** | **7,698** | **5,897 (77%)** |

The 76% figure is the signature finding — Python is the more mature ecosystem with the larger dataset. The TypeScript benchmark confirms the same pattern in a second ecosystem: agent frameworks and applications ship without intra-functional guards, by convention more than by oversight.

READ THIS IF YOU ARE…

| | |
|---|---|
| **CISO / Head of Security** | A defensible inventory of agent side-effect surface in 2026. Use it to frame the agent-governance conversation with engineering before an incident frames it for you. |
| **Platform / Infra Lead** | A reproducible scan recipe (pip install, npm install, run on your repos) and a pre-deploy checklist mapping side-effect categories to runtime gates. |
| **Founder shipping an agent** | A way to find out where your own agent's unguarded surface lives, before it ships, in under five minutes. |
| **Open-source maintainer** | A baseline against which to position your project's governance story, plus the annotation convention (# checked:ok) to acknowledge guards the scanner cannot see. |

DEFINITIONS

# What "unguarded" means — and what it does not mean

> *"Unguarded" means: the function has at least one side effect (DB write, HTTP call, subprocess, payment, email, etc.) AND zero detected guards (input validation, rate limit, auth check, approval gate, retry bound, idempotency key) within the same function body.*

**An unguarded tool call is not automatically a vulnerability.** Infrastructure-level controls (API gateways, IAM, network policies) and service-layer code that the scanner cannot see may cover many of these cases.

**The 76% number is an inventory, not a vulnerability score.** It tells you how much side-effect surface exists in the codebase before anyone decides what to do about it.

| Verdict | Action |
|---|---|
| **FIX IT** | Add a guard inside the function. |
| **GATE IT** | Route through diplomat-gate for a CONTINUE / REVIEW / STOP decision. |
| **ACKNOWLEDGE IT** | Mark it # checked:ok — protected by [where] so the next person knows. |

**The scanner gives you the inventory. Your team makes the decision — at design time, not after an incident.**

KNOWN BLIND SPOTS

| Gap | What it means |
|---|---|
| Cross-file call chains | If a guard lives in function A which calls function B that executes the side effect, the scanner sees B as unguarded. Full call graph analysis is on the v0.2 roadmap. |
| Import aliases (Python) | import requests as r then r.post() is not detected. Known limitation. |
| Re-exported ORM models (TypeScript) | Mongoose / TypeORM / Drizzle spread across files will pass through. |
| Dynamically registered tools | Tools registered dynamically at runtime are invisible to static analysis. By design. |

**One absent finding worth noting:** Zero # checked:ok annotations across all 19 codebases. The agent ecosystem in 2026 has not converged on a shared way to say "this looks unguarded but it is gated at a higher layer". This report is an argument for adopting one.

PYTHON BENCHMARK

# diplomat-agent v0.2.0 — 16 codebases

16 open-source Python AI agent codebases. 14 produced valid results (AIHawk and LangChain-community failed to scan and are excluded from all counts).

| Repo | Type | Tool calls | Unguarded | % | Scan time |
|---|---|---|---|---|---|
| PraisonAI | Framework | 1,028 | 911 | 89% | ~2s |
| CrewAI | Framework | 348 | 273 | 78% | ~1s |
| Skyvern | Application | 452 | 345 | 76% | ~2s |
| AutoGPT | Application | 464 | 355 | 76% | ~2s |
| Dify (backend) | Platform | 1,009 | 759 | 75% | ~3s |
| Khoj | Application | 181 | 125 | 69% | ~1s |
| SurfSense | Application | 315 | 165 | 52% | ~1s |
| OpenAI Agents (Python) | SDK examples | 93 | 78 | 84% | <1s |
| MetaGPT | Framework | 205 | 186 | 91% | ~1s |
| Browser-use | Application | 267 | 230 | 86% | ~1s |
| OpenAgents | Application | 178 | 177 | 99% | ~1s |
| FinRobot | Application | 69 | 55 | 80% | <1s |
| Open-SWE | Application | 35 | 34 | 97% | <1s |
| GPT-Researcher | Application | 31 | 27 | 87% | <1s |
| Composio | Platform | 28 | 26 | 93% | <1s |
| Stripe Agent Toolkit | Benchmark | 26 | 14 | 54% | <1s |
| **TOTAL (14 valid)** | — | **7,029** | **5,344** | **76%** | — |

## By repo type

| Type | Repos | Unguarded (avg.) | Reading |
|---|---|---|---|
| Framework | 3 | 83% | By design — frameworks expose primitives and delegate governance to the application embedding them. |
| Application | 9 | 80% | These teams built the product. The guards were theirs to add. |
| Platform | 2 | 78% | Same pattern as applications. |
| SDK examples | 1 | 84% | Reference code that teams copy-paste to production. The number that propagates. |
| Benchmark | 1 | 54% | Test environment. Not representative. |

TYPESCRIPT BENCHMARK

# diplomat-agent-ts v0.1.0 — 3 codebases, 4 scopes

| Repo (scope) | Type | Commit | TS files | Tool calls | Unguarded | % |
|---|---|---|---|---|---|---|
| OpenClaw (src/) | Application | 49d9996d | 7,874 | 419 | 332 | **79%** |
| Mastra (packages/) | Framework | 38b87964 | 2,777 | 185 | 162 | 88% |
| OpenAI Agents JS (packages/) | Framework | 629d35af | 426 | 33 | 31 | 94% |
| OpenAI Agents JS (examples/) | Examples | 629d35af | 302 | 32 | 28 | 88% |
| **TOTAL** | — | — | **11,379** | **669** | **553** | **83%** |

FINDINGS BY CATEGORY

# Both ecosystems — detected side effects

## Python (16 repos)

| Category | Unguarded | Repos | What it detects |
|---|---|---|---|
| DB write | 3,260 | 15/16 | session.commit(), cursor.execute(INSERT/UPDATE), ORM .create() |
| DB delete | 1,305 | 14/16 | session.delete(), .drop(), cursor.execute(DELETE) |
| HTTP write | 968 | 14/16 | requests.post(), httpx.post(), aiohttp writes |
| Subprocess / exec | 697 | 10/16 | subprocess.run(), os.system(), exec(), eval() |
| Publish / storage | 609 | 7/16 | s3.put_object(), .publish(), .upload() |
| LLM calls | 464 | 8/16 | client.chat.completions.create(), anthropic.messages.create() |
| Emails | 250 | 8/16 | smtp.sendmail(), mailer.send() |
| File deletion | 225 | 10/16 | os.remove(), shutil.rmtree(), pathlib.unlink() |
| Agent invocations | 206 | 8/16 | graph.ainvoke(), chain.invoke(), agent.execute() |
| Payments | 41 | 3/16 | stripe.PaymentIntent.create() + internal methods (~22% FP rate) |

## TypeScript (4 scopes)

| Category | Findings (approx.) | Scopes | What it detects |
|---|---|---|---|
| destructive | ~415 | 4/4 | child_process.execSync(), spawn(), execFile(), execa() |
| file_delete | ~210 | 3/4 | fs.rm(), fs.unlink(), fsp.rm() with destructive flags |
| publish | ~118 | 3/4 | pubsub.publish(), topic.publish() |
| http_write | ~80 | 4/4 | fetch(POST/PUT/PATCH), axios.post() |
| agent_invocation | ~50 | 4/4 | runner.run(), agent.stream(), agent.execute() |
| llm_call | ~10 | 3/4 | openai.chat.completions.create(), anthropic.messages.create() |

| Category | Findings (approx.) | Scopes | What it detects |
|---|---|---|---|
| payment | 0 | 0/4 | No Stripe/PayPal SDK usage found in scanned repos |

ILLUSTRATIVE CASES

# Interesting findings — real code, real codebases

Not synthetic examples. Six cases extracted directly from the scanned codebases.

**Khoj — ai_update_memories (Python)**

Categories: database_delete · agent_invocation

An LLM decides which user memories to keep and which to delete. Calls UserMemoryAdapters.delete_memory(user, memory) with no human confirmation. The agent autonomously rewrites what it "remembers" about you.

**OpenAI Agents (Python) — pop_item**

Categories: http_write · database_delete

Calls openai_client.conversations.items.delete() — deletes conversation items from the OpenAI API without confirmation. Not a local database. An external API call that cannot be undone.

**Skyvern — workflow_delete (Python)**

Categories: destructive · database_delete

Calls tool_workflow_delete(workflow_id=workflow_id, force=force) with a force parameter and no approval step. The force=True path bypasses all safety checks.

**OpenAgents — 99% unguarded (Python)**

Categories: database_delete · http_write

177 of 178 tool calls have zero checks. Mainly redis_client.delete() and requests.post() with no rate limiting or auth verification.

**OpenClaw — approval system flagged destructive (TypeScript)**

Categories: destructive — externally governed

OpenClaw has a sophisticated execApproval* infrastructure routing shell commands through user approval gates. The scanner correctly identifies the underlying spawn() calls as destructive, but since the approval gate often lives in a different function, many findings show as no_checks. The code is governed; the governance just does not fit in one function. Canonical case for the # checked:ok annotation.

**Mastra — processWorkflowSleep() (TypeScript)**

Categories: publish — internal pubsub

Four pubsub.publish('workflow.events.v2.${runId}', ...) chained in one function. This is Mastra's internal event bus — an unguarded framework primitive by design. When embedding Mastra, scan it, see what surfaces, and decide which primitives to wrap.

PRE-DEPLOY CHECKLIST

# What to do with your scan results

Each finding gets a verdict aligned with the diplomat-gate model: STOP (block in production), REVIEW (block until human approval), or CONTINUE (allow, with logging).

| Category | Verdict | Recommended action |
|---|---|---|
| **Payment** | STOP | Add amount_limit + duplicate_detection + recipient_blocklist in diplomat-gate. |
| **Email / messaging** | STOP or REVIEW | Add domain_blocklist + rate_limit + business_hours policy. |
| **DB delete / drop** | REVIEW | TTL queue + human approval required via diplomat-gate. |
| **Subprocess / exec** | STOP | Input allowlist + STOP on unrecognized commands. |
| **HTTP write (external)** | REVIEW | Rate limit + auth check. Gate if external and critical. |
| **LLM calls** | Assess | Token budget + retry bound. Gate if recursive agent invocation. |
| **Agent invocations** | Assess | Invocation depth limit. Gate orchestrator → sub-agent calls. |
| **DB write (commit)** | Assess | If Pydantic validation in service layer, acknowledge. Otherwise gate. |
| **File deletion** | REVIEW | Confirmation gate if path comes from user input. |
| **Internal pubsub** | ACKNOWLEDGE | Annotate # checked:ok — internal pubsub, no process boundary crossed. |

## Three-step process

```
# 1. Scan your repo — 30 seconds pip install diplomat-agent && diplomat-agent scan . npm install -g @diplomat-ai/diplomat-agent-ts && diplomat-agent-ts scan src/ # 2. Triage # Review each finding. Fix it, gate it, or acknowledge it. # Use # checked:ok — for findings protected elsewhere. # 3. Gate pip install diplomat-gate # Add the gate in your agent's tool invocation path. # Policies in < 1 ms. Zero mandatory dependencies.
```

METHODOLOGY

# How the numbers were produced

| | |
|---|---|
| **Scanner** | Python stdlib ast module (zero dependencies) and TypeScript ts-morph. |
| **Detection** | 40+ patterns matching function calls with side effects across 10-12 categories. |
| **Guards detected** | Input validation (Pydantic / Zod / class-validator), rate limiting, auth checks, approval gates, idempotency keys, retry bounds — within the same function as the side effect. |
| **Scope** | Intra-procedural. Each function analyzed independently, with decorator resolution within the same class. |
| **Reproducibility** | All commit SHAs are pinned. Reproduction commands in the repo. |

| Verdict | Condition |
|---|---|
| UNGUARDED / no_checks | Side effects detected + zero guards |
| PARTIALLY_GUARDED / partial_checks | Some guards but not complete coverage |
| GUARDED / confirmed | Annotated # checked:ok OR 2+ distinct guard types |
| LOW_RISK | No side effects detected (Python scanner only) |

GLOSSARY

| | |
|---|---|
| **Tool call** | A function invocation made by an AI agent that has effects outside the LLM context. |
| **Side effect** | Any operation that modifies external state: DB write, HTTP write, subprocess, payment, email, file deletion, agent invocation. |
| **Guard** | An intra-functional control constraining a tool call: input validation, rate limit, auth, approval gate, idempotency key, retry bound. |
| **Verdict (static)** | UNGUARDED, PARTIALLY_GUARDED, GUARDED, LOW_RISK — classification by diplomat-agent. |
| **Verdict (runtime)** | CONTINUE, REVIEW, STOP — diplomat-gate decision in < 1 ms. |
| **Receipt** | Immutable record of a runtime verdict, including the SHA-256 hash of the tool call. EU AI Act Article 12. |
| **Behavioral BOM (toolcalls.yaml)** | Per-repo inventory of every tool call the agent can make. The "requirements.txt for what the agent can do". |

WHAT'S NEXT

## Three steps to govern your agents in production

**01**

**Know — diplomat-agent**

Scan your repo. Get the complete inventory of side-effect tool calls and their protection status. 30 seconds.

```
pip install diplomat-agent && diplomat-agent scan .
```

github.com/Diplomat-ai/diplomat-agent

**02**

**Decide — diplomat-gate**

Gate critical tool calls. CONTINUE / REVIEW / STOP verdict in < 1 ms. 15+ pre-built policies. Zero mandatory dependencies. SHA-256 hash-chained audit trail.

```
pip install diplomat-gate
```

github.com/Diplomat-ai/diplomat-gate

**03**

**Prove — diplomat.run**

Hosted Decision Control Plane. Immutable cross-tenant audit trail, real-time dashboard, managed approval routing, EU AI Act Article 12 compliance export.

diplomat.run

**Scan your repo now**

pip install diplomat-agent && diplomat-agent scan .

Book 30 minutes with us: calendly.com/josselin-guarnelli

Diplomat Services SASU · SIREN 100 251 834 · Cagnes-sur-Mer, France