

What is the b3 benchmark?

The b3 benchmark (Backbone Breaker Benchmark) measures how secure the large language model inside an AI agent is, separately from the rest of the agent. It uses threat snapshots, frozen moments in an agent's run where an attack lands, and fills them with attacks crowdsourced through an online red-teaming game. The paper drew on 194,331 unique attacks and tested 34 models, finding that stronger reasoning improved security while model size did not.

HOW IT WORKS

01 How b3 tests a model

The attacks were crowdsourced from people playing an online game in which they try to break AI agents. The Inspect version contains 630 samples. Each attack appears three times, against the same application at three increasing security levels.

Attacks are tagged by type: Direct Instruction Override, Indirect Instruction Override, Direct Tool Invocation, Indirect Tool Invocation, and Denial of AI Service. Scoring depends on the attacker's goal. A data-exfiltration attempt is scored by how closely the response matches the secret, a tool attack by whether the exact tool call was made with the attacker's parameters.

Run it with `inspect eval inspect_evals/b3`, or filter with `-T task_types=["DTI", "ITI"]` to focus on tool invocation.

02 What the results tell you

Across 34 popular models, the paper found that models with stronger reasoning were more secure, while bigger models were not reliably more secure. That is useful when choosing a backbone: a larger model is not automatically a safer one.

It also shows why a single number is not enough. A model can resist instruction overrides well and still make the wrong tool call when a document tells it to, so read the results per attack type.

SOURCES

- [1] Breaking Agent Backbones: Evaluating the Security of Backbone LLMs in AI Agents
- [2] b3 in Inspect Evals (task docs)
- [3] OWASP Top 10 for Agentic Applications (2026)
- [4] Inspect Evals registry: Safeguards evaluations

Test your AI application before an attacker does.

securelayer7.net/learn/ai-security/b3-benchmark

[Open online](https://securelayer7.net/learn/ai-security/b3-benchmark)