[un]prompted

THE INDUSTRY REPORT · 2026

State of AI Cybersecurity

An analysis of the 490 talks submitted to the [un]prompted 2026 call for papers: who submitted, what they work on, and what changed since March.

- Break 34%
- Build 24%
- Operate 18%
- Govern 13%
- Train 11%

490	501	71%	8%
talks submitted	speakers	of talks are about agents	of talks model an attacker who uses AI

Analysis by **Rob T. Lee**, Chief AI Officer and Chief of Research, SANS Institute and **Indre Altman**, Senior Associate, Crosspoint Capital Partners

[un]prompted II · October 27-29, 2026 · San Francisco

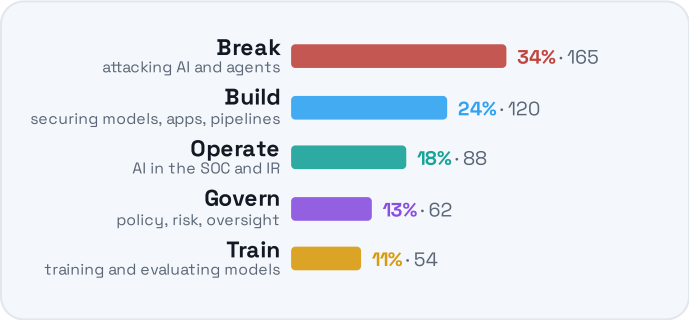
PARTNERED WITH SANS

THE SUBMISSION POOL

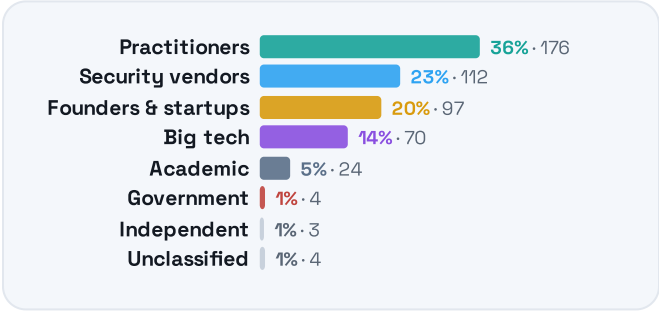
Practitioners are the biggest group; agents dominate

[un]prompted is the leading AI security practitioner conference, community led and driven. It's an intimate, raw, and fun gathering for the professionals actually doing the work, from offense to threat hunting to program building to national policy. This report is built from the October call for papers, which drew **490 talks from 501 speakers**, and, where relevant, reflected against the March 2026 call for papers. The March talks predate Claude Mythos Preview and Project Glasswing (announced April 7, 2026); the October talks follow them.

SHARE OF SUBMISSIONS BY TRACK

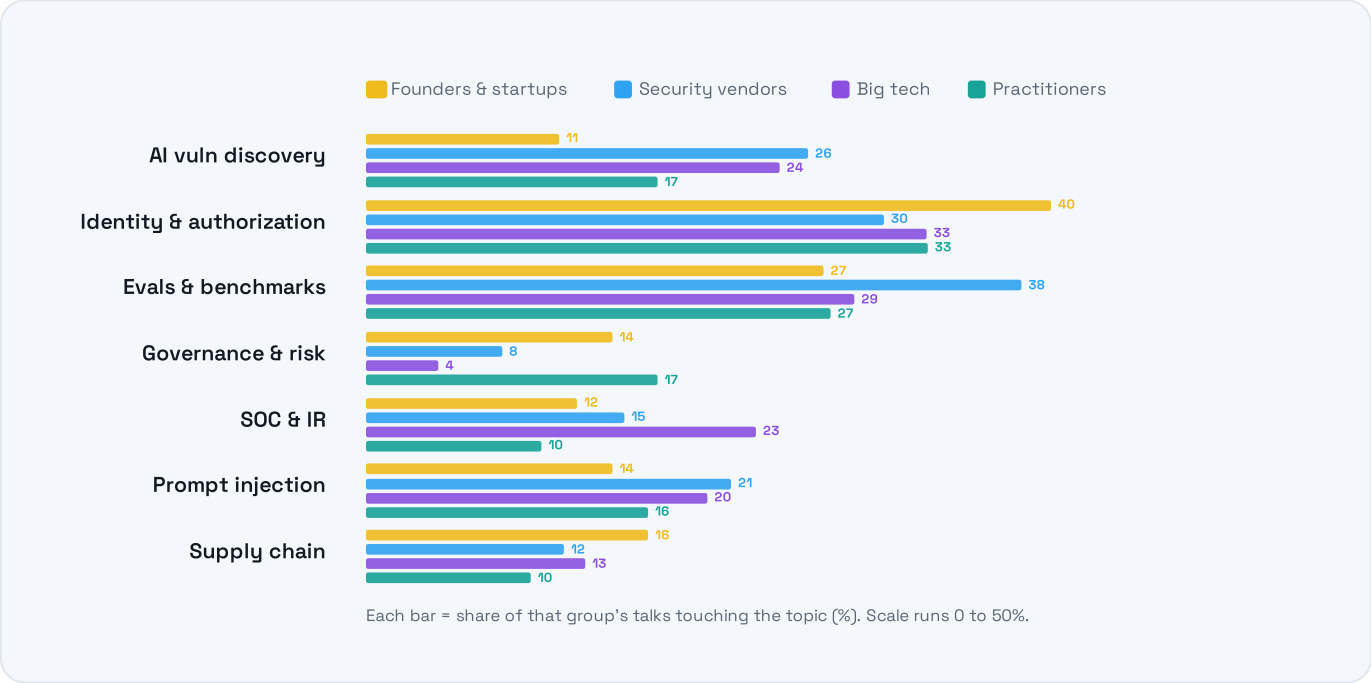


SUBMITTERS BY TYPE



Big tech: large platform companies such as Microsoft, Google, NVIDIA, Salesforce, Airbnb, and OpenAI. **Security vendors**: companies that sell security products or services. **Founders & startups**: founders, C-suite, and staff at stealth or early-stage companies. **Practitioners**: everyone else with a stated role, mostly at enterprises and consultancies. Track descriptions are the authors' summaries, not call-for-papers text.

TOP TOPICS BY SUBMITTER TYPE

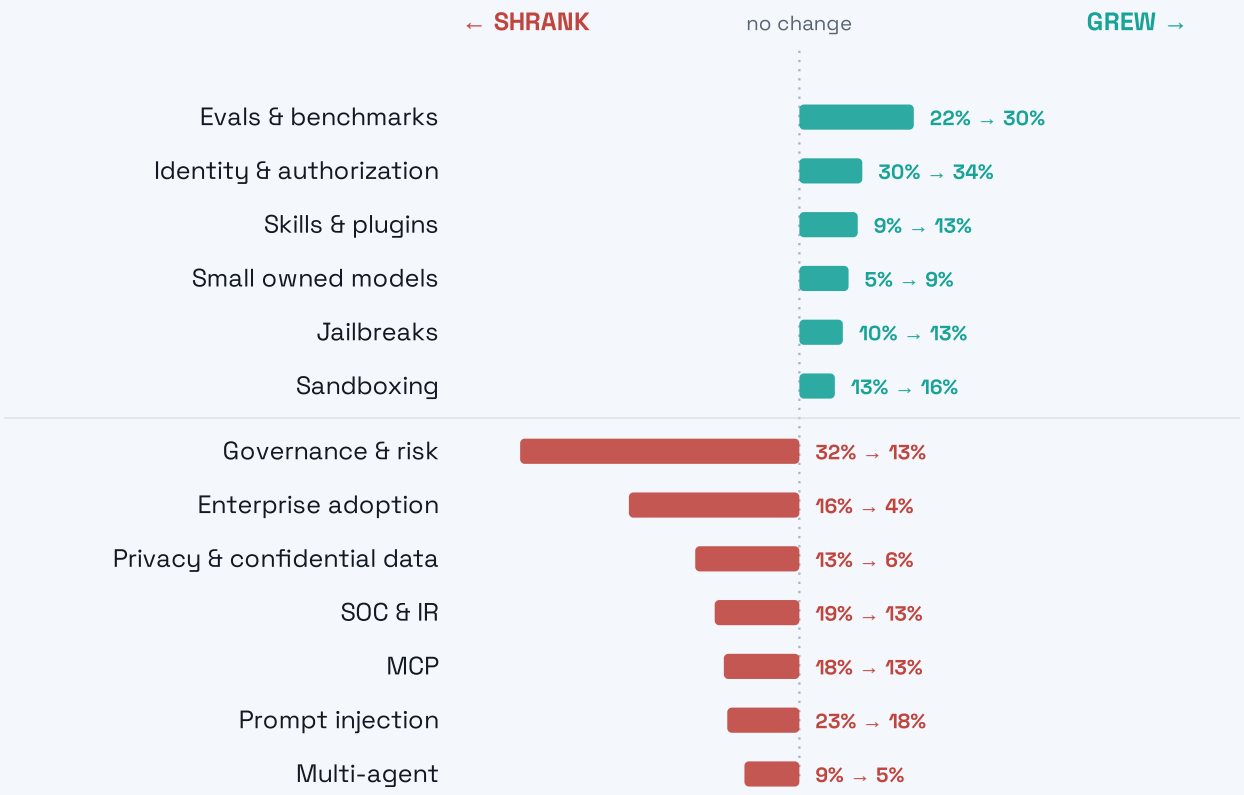


Big tech and security vendors are different crowds. Big tech submitters put 23% of their talks into SOC and incident response, the most of any group, and barely touch governance (4%). Security vendors lead on evals (38%) and AI vulnerability discovery (26%). Founders lead on identity and authorization (40%).

MARCH VS. OCTOBER

Topic area gains and losses

Talks shifted from whether to bring AI in to how to secure the AI already there.



Share of talks per topic, one scale for gains and losses. Largest moves among topics with at least 15 talks. Pools: 469 March, 490 October.

Governance and enterprise adoption lost share. Evals gained the most.

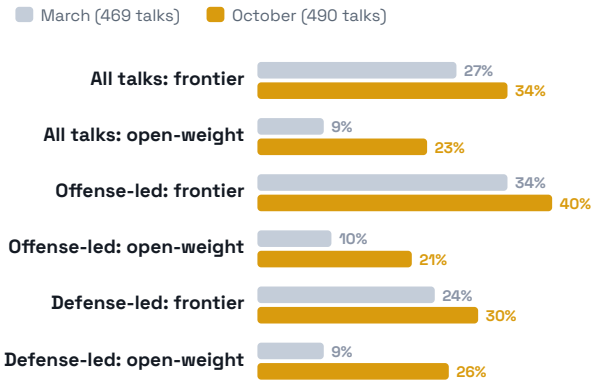
Governance went from 32% to 13%, and enterprise adoption from 16% to 4%. Evals went from 22% to 30%. Governance is the least precise topic in this report, so treat the size of that drop as approximate.

THE MODEL LAYER

Frontier models and open-weight models

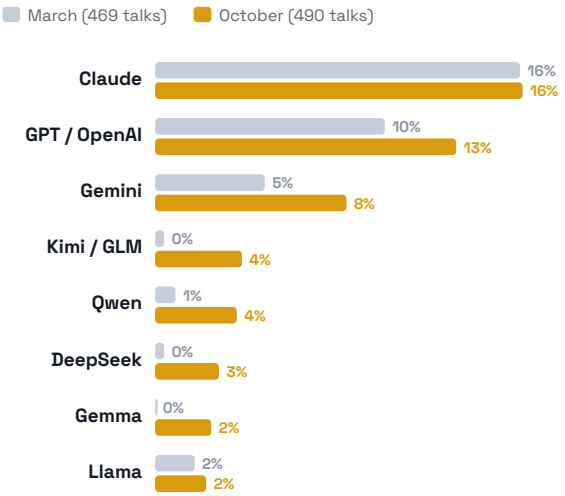
Which models the talks name, March against October. Frontier-model mentions kept rising. Open-weight and local models grew faster, from 9% of talks to 23%.

SHARE OF TALKS NAMING A MODEL TYPE



Frontier = Claude, GPT, Gemini, Grok, Mythos or "frontier model." Open-weight = open-weight or local models and serving stacks (Llama, Qwen, DeepSeek, Gemma, Mistral, Ollama, vLLM, Hugging Face). Offense-led and defense-led are defined on page 6.

NAMED MODELS, SHARE OF TALKS



Chinese open models (Qwen, DeepSeek, Kimi, GLM) went from 8 talks to 33; Western open models (Llama, Gemma, Mistral, Phi) from 9 to 21.

WHY OPEN WEIGHTS, IN THE TALKS THAT USE THEM

STATED REASON (OCTOBER, 113 TALKS)	SHARE	GUARDRAIL SIGNAL	MARCH	OCT
Cost	**43%**	Talks raising refusals, guardrails or usage policies as a problem	**20**	**58**
Privacy, data residency, on-prem	**34%**	Talks on de-aligning a model (abliteration, uncensoring)	**2**	**11**
Latency or offline use	**29%**	Capability laundering (splitting tasks past refusals)	**0**	**7**
Refusals, guardrails, terms of service	**24%**			
Control over the weights, fine-tuning	**21%**			

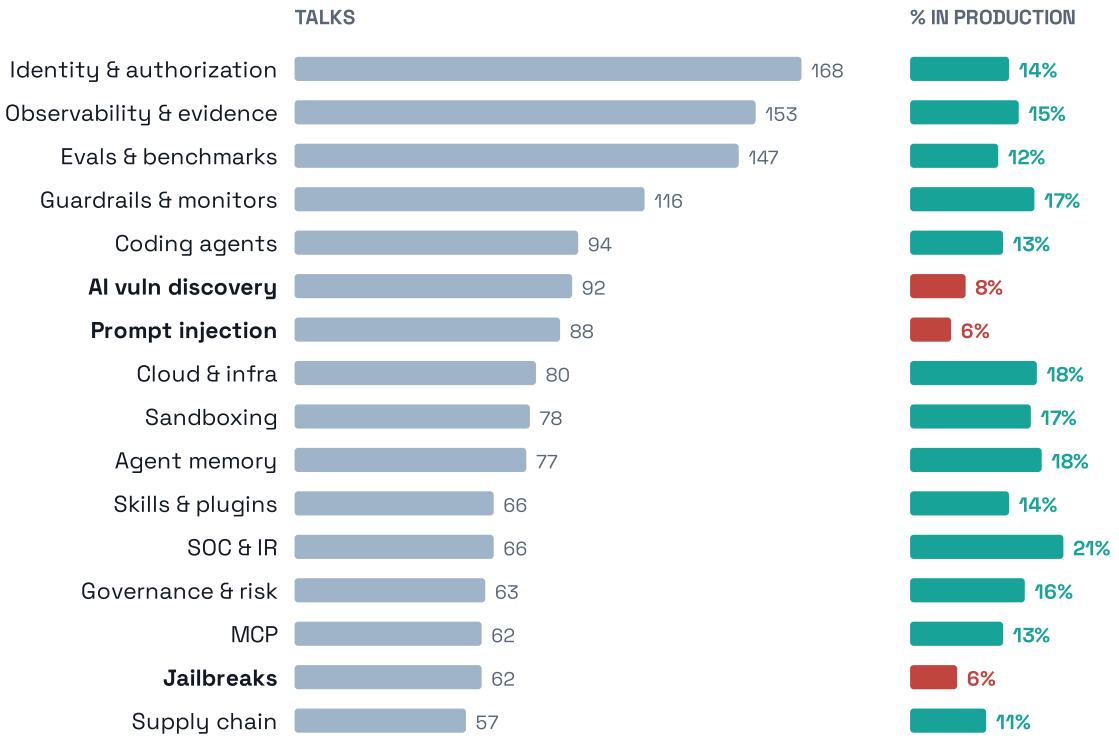
Open weights are being added alongside frontier models, not replacing them.

Defense talks adopted open models fastest, citing cost, privacy, and latency more often than refusals. Refusal complaints tripled, talks on de-aligning a model went from 2 to 11, and the open weights in use are increasingly Chinese-origin, which only 5% of those talks treat as a supply-chain question. Frontier models are used for reasoning and agent work. Open-weight models are used where frontier models refuse the task, or where the data cannot leave the environment.

OPEN PROBLEMS

Attacks get the talks; the SOC has the production use

Every topic with at least 50 talks, by number of talks and by the share of those talks that report results from production systems (12%, 57 of 490). The lowest production shares are on some of the most-discussed attack topics.

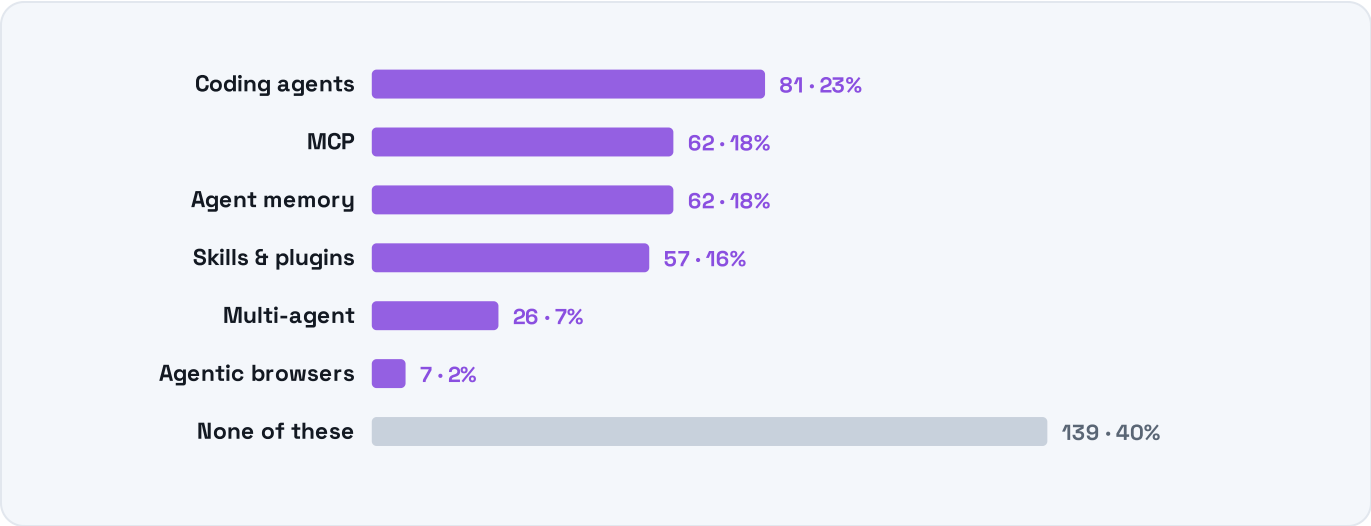


Topics with 50+ talks. Red = under 10% of the topic's talks report production results. Pool average: 12%.

ZOOMING IN

Agents are 71% of talks; offense leads at the harness

348 of the 490 talks are about AI agents. The chart shows how many of them touch each specific agent topic. A talk can touch more than one, so the bars add up to more than 348.



139 agent talks (40%) touch none of these subtypes; most of those are about evals (34%), observability (30%), identity (27%), or guardrails (27%).

OFFENSE VS. DEFENSE, BY HARNESS COMPONENT

COMPONENT	OCT TALKS	OFFENSE-LED	IN PRODUCTION
Prompts and system prompts	**130**	**68%**	**6%**
Tools and MCP servers	**98**	**61%**	**16%**
Skills, plugins, extensions	**81**	**53%**	**11%**
Permissions and identity	**168**	**50%**	**13%**
Memory and context	**57**	**47%**	**14%**
Sandbox and execution	**74**	**43%**	**15%**

Offense-led = attack vocabulary outweighs defense vocabulary in the talk text; mixed talks (24% of the pool) are excluded from both shares. Component counts use narrower terms than the topic tags on page 5.

Offense leads defense at the harness layer.

Talks that attack the harness itself went from 12 in March to 33, among them **poisoned skills and plugins, sandbox escapes from coding agents, and steering an agent without any injected prompt.** No component has more than 16% of its talks reporting a defense running in production.

WHAT THE TALKS DON'T SAY

Threat intel on AI use by adversaries is absent. So are AI security standards.

Two things the 490 talks almost never do: describe who is attacking with AI, and cite a standard.

WHO IS THE ATTACKER?

MEASURE	OCT	MARCH
Offense-led talks (attack vocabulary outweighs defense)	**46%**	**43%**
Talks that model an attacker who uses AI	**39 (8%)**	**40 (9%)**
...of which framed as a demo, simulation or proof of concept	**33**	
Talks that mention threat intelligence at all	**14 (3%)**	**17 (4%)**
Talks citing a public threat-intelligence report	**2**	**2**
Nation-states or APTs named	**11**	**16**
Cybercrime economy (ransomware, fraud, scams)	**17**	**20**
Honeypots and deception aimed at AI attackers	**21**	**7**
Government submitters (CISA, UK NCSC) / speakers in IR or threat-intel roles	**2 talks / 46 of 501**	

Talks attack AI far more often than they describe attackers using AI.

Two talks came from serving government: CISA and the UK National Cyber Security Centre, one each, both to the Govern track. Neither presents adversary case data. The groups with visibility into attackers (model providers, incident-response firms, and government) are about 15% of submitters and supply almost none of the attacker evidence. Their talks are about tools rather than incidents. The only first-hand observation of AI attackers comes from honeypot talks, which went from 7 to 21. This pool cannot show whether attackers use AI at scale, because the people who would know are not presenting it. That is why the debate over whether the threat is overstated continues.

WHO IS CITING A STANDARD?

STANDARD	OCT TALKS
OWASP, any form (LLM Top 10 3, Agentic 5, AI Exchange 1)	**23 (5%)**
CVSS	**13**
NIST (AI RMF, 600-1)	**10**
MITRE ATT&CK	**8**
MITRE ATLAS	**8**
EU AI Act	**6**
CSA MAESTRO	**5**
ISO/IEC 42001	**3**
Google SAIF / CoSAI	**1**

WHO CITES ANY STANDARD	SHARE OF TALKS
Government	**0 of 2**
Academic	**21%**
Practitioners	**9%**
Security vendors	**8%**
Founders and startups	**6%**
Big tech	**3%**
Whole pool	**8% (March 13%)**

Of the 45 citing talks, 14 use the standard as a baseline, 6 say it falls short, none propose extending it. What the field writes instead: 47 playbooks or checklists, 13 open-source tools, 10 benchmarks, 4 frameworks. Government shows a count, not a share, because n is 2.

Standards are rarely cited in talk proposals.

No standard reaches 5% of talks, and citations fell since March. The standards that are cited split by layer: NIST and the EU AI Act appear in governance talks, OWASP and ATLAS in attack talks, and no cited standard covers the operational middle, what a secure agent deployment looks like, which is where most talks are. Big tech, the group with the most agents in production, cites standards least. **Submitters write playbooks and checklists (47 talks) rather than frameworks (4). What is missing is consolidation: an ATT&CK-equivalent for agents, a delegation and identity standard, and a deployment baseline built from the playbooks already in this pool.**

APPENDIX

Topics, definitions, and method

TOPIC	WHAT COUNTS	OCT TALKS	OCT	MAR
Agents (all)	Mentions AI agents, MCP, coding agents, or multi-agent systems.	348	71%	66%
Identity & authorization	Identity, permissions, delegation, least privilege, OAuth, non-human identity.	168	34%	30%
Observability & evidence	Observability, logging, telemetry, tracing, audit trails, provenance.	153	31%	32%
Evals & benchmarks	Evals, benchmarks, LLM-as-judge, ground truth.	147	30%	22%
Guardrails & monitors	Guardrails, classifiers, detectors, runtime monitors.	116	24%	24%
Coding agents	Coding agents and AI coding tools (Claude Code, Cursor, Copilot, Codex, IDEs, AI-generated code).	94	19%	19%
AI vuln discovery	Using AI to find vulnerabilities: zero-days, CVEs, fuzzing, bug hunting, exploitability.	92	19%	22%
Prompt injection	Direct or indirect prompt injection.	88	18%	23%
Cloud & infra	Cloud, Kubernetes, multi-tenancy, GPU and inference infrastructure.	80	16%	18%
Sandboxing	Sandboxes, containment, and runtime enforcement.	78	16%	13%
Agent memory	Agent memory, persistence, long-horizon and stateful behavior.	77	16%	16%
Skills & plugins	Agent skills, plugins, and their marketplaces.	66	13%	9%
SOC & IR	SOC work, detection engineering, alerts, triage, incident response.	66	13%	19%
Governance & risk	Governance, compliance, regulation, risk management, AI policy frameworks.	63	13%	32%
MCP	The Model Context Protocol.	62	13%	18%
Jailbreaks	Jailbreaks, model refusals, and safety bypasses.	62	13%	10%
Supply chain	Supply chain of models, tools, and registries: SBOMs, lineage, pickles, backdoors.	57	12%	15%
Threat modeling	Threat modeling.	49	10%	11%
Small owned models	Small, fine-tuned, post-trained, or on-device models.	43	9%	5%
Data poisoning	Training or context data poisoning.	36	7%	7%
Malware & RE	Malware and reverse engineering.	32	7%	9%
Economics & cost	The economics of AI security: cost per attack or finding, token spend, ROI.	31	6%	9%
Privacy & confidential data	Privacy, confidential data and inference, data exposure.	27	6%	13%
Multi-agent	Multi-agent and agent-to-agent systems.	26	5%	9%
Cryptography	Cryptography and encryption.	23	5%	6%
Enterprise adoption	Rolling AI out in the enterprise, shadow AI, chatbots.	21	4%	16%
Deception & honeypots	Honeypots, decoys, and deception aimed at AI attackers.	19	4%	3%
Autonomous red team	Autonomous AI red teaming and pentesting.	12	2%	3%
Agentic browsers	AI browsers and browser or web agents.	7	1%	1%
Human factors & workforce	Workforce, careers, hiring, and skills of people.	5	1%	5%

Topics. Keyword tags on each talk's title, description, and outline, using the definitions above, applied identically to the 490 October and 469 March submissions. A talk can carry several topics, so the columns don't sum to 490 or 100%. We read random samples of tags for the most contested topics: evals, AI vulnerability discovery, and economics were right about 7 or 8 times in 8; observability about 7 in 8; SOC and incident response and enterprise adoption about 6 in 8; governance about 5 in 8, the weakest. Treat topic sizes as directional. **Submitter type.** From each talk owner's current tagline (the bio only when the tagline is empty) and email domain; definitions on page 2. The March export has no taglines, and only 49 of its 383 submitters appear in October, so submitter type is an October snapshot. **Production evidence.** The talk's title, description, or evidence field describes a deployed system, production data, or real customers.

SUMMARY

Six findings from 490 submissions

1 Seven in ten talks are about agents

71% of the 490 talks center on AI agents, up from 66% in March. The focus reaches into other topics: about four in five talks on identity, observability, and supply chain, and nearly three in four eval talks, are about securing agents rather than models.

2 The work has moved from the model to the harness around it

"Harness," the tools, memory, permissions, and execution loop around a model, appears in 94 talks, up from 19 in March. Counting MCP servers, skills, and plugins, 42% of talks now focus on that layer, up from 34%. About half of supply-chain talks (28 of 57) are about MCP servers, skills, and plugins rather than model files.

3 Attacks are moving past prompt injection

Prompt injection fell from 23% of March talks to 18% (88 talks), and most now treat it as something to contain in the harness rather than fix in the model. Talks attacking the harness itself, including poisoned skills and plugins, malicious MCP tools, sandbox escapes, memory poisoning, and steering agents without any injection, went from 12 in March to 33.

4 Evals are up 8 points. So is the skepticism.

Evals went from 22% of talks in March to 30% in October, the largest gain of any topic. Talks arguing the evals themselves are unreliable grew faster, from 4 to 17: measurement error in the harness, models that learn to game the test, passing scores on systems that were not safe to ship.

5 AI finds bugs faster than anyone ships the fixes

92 talks (19%) use AI to find vulnerabilities, and 55 cover patching or remediation. Fewer than ten describe getting a fix merged and deployed in production at scale. Discovery is crowded; shipping the fix is the open problem.

6 After the Mythos announcement, talks claimed less autonomy and reported more evidence.

Mythos is named in 7 October talks and Glasswing in 2, so the announcement itself did not become a talk subject. The change is in how claims are made. "Autonomous" fell from 26% of talks to 18%, and claims of fully autonomous or human-out-of-the-loop systems from 15 talks to 7. Controlled uplift studies doubled (8 to 16) and talks reporting cost rose from 26 to 41. Talks calling the threat overhyped also fell, from 25 to 7. Quoted counts of AI-found vulnerabilities rose from 16 to 25 talks while the vulnerability-discovery topic itself shrank from 22% to 19%, and talks about disclosure overload and low-quality AI-generated reports rose from 14 to 23. The discussion moved from whether AI finds vulnerabilities to how to validate and fix the ones it finds.