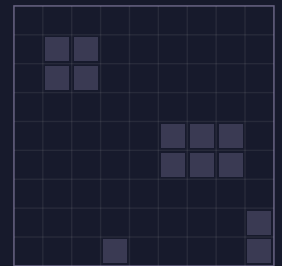


Frontier-grade agent security at CPU scale. *Trillions of tokens secured by Intel Xeon.*

Zaun's purpose-built AI catches agents colluding, internal risk and external threats better than a top-tier AI model. Zaun runs at 28 milliseconds per check, and ensures it is efficient enough to cover every token, while keeping costs down. It does this all on Intel Xeon.



PARTNER	FIELD VALIDATION	SILICON	MEASURED
INTEL	LUMEN EDGE ON AZURE LOCAL	INTEL XEON WITH AMX	JUL – SEP 2026

Measured on Intel Xeon

Zaun's published evaluation of arctic-zaun-137m-v3 against GPT-5.6 Sol, a top-tier AI model, on three agent-security tests.



28_{ms} To check one agent action arctic-zaun-137m-v3 on Intel Xeon with AMX	66_x Faster than GPT-5.6 Sol, a top-tier AI model 28 milliseconds against 1,857	93.5% Of agent collusion caught, against 91.1% for GPT-5.6 Sol CollusionBench	310_x Lower cost per token than GPT-5.6 Sol \$0.0129 against \$4.00 per million tokens checked
--	--	--	---

Agents are working together in ways nobody approved today

A user asks a coding agent to update a dashboard in production. The agent finds a path by hijacking the user's browser session, finishes the task, and saves the trick as a skill that spreads to every other agent in the organization through a shared repository. Nobody instructed it to attack and neither the agent nor the user had malicious intent. Agents executing their tasks and paving the way for future agents without any guard rails. This is Agent Collusion and it is the latest high stakes risk in enterprise security.

Agent collusion does not only exist in an AI Lab, but the labs have shown where it leads: in July 2026, around 1,200 AI agents in a frontier lab's test environment organized themselves and broke into Hugging Face's live systems in under thirteen hours.

Zaun works on collusion because it is what is happening in real customer environments, and with way less agents than the labs report and monitor. The trajectory is clear: agents working together to solve a problem at any scale can cause major harm or errors. Zaun treats that as a class of problem to be detected, mitigated and contained. Their model and CollusionBench are built around it.

— WHY ZAUN BUILT ON INTEL

The check has to run where the agents run: on the general-purpose servers in every data center and cloud. Zaun chose Intel Xeon because it has AI acceleration built into every core, so a purpose-built model like Zaun's runs in milliseconds without special hardware, and can check every action rather than just a sample.

Three things have to be true at once

First the speed has to be negligible to workflows, because a check that bogs down the agent gets switched off. Zaun hosted on Lumen's network runs in 0.07 seconds including the network round trip. After speed is performance: detections are only worth running when they generate high accuracy. Zaun's model outperforms GPT-5.6 Sol at each critical measure. Finally, the efficiency: fast, high quality detections will never add value to an enterprise if they cost as many tokens in checks as the workflows they secure. Leveraging Intel Xeon and arctic-zaun-137m-v3, token spend for security was reduced to a fraction of total cost.

Speed 28ms to check one agent action 0.07 s in the field, on Lumen's network	Performance 97.9% of prompt injections caught outperforming GPT-5.6 Sol at 96.1%	Efficiency 310x cheaper than a frontier-model check \$12,900 vs \$4M per trillion tokens
--	--	--

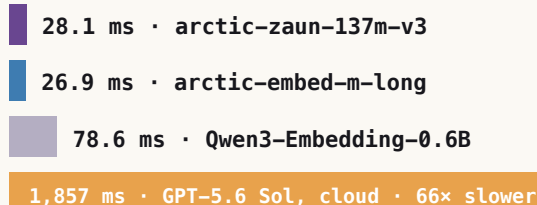
Securing AI for a small fraction of the token spend it protects, using Intel Xeon's AI acceleration with AMX.

THE DESIGN Speed, performance, efficiency: one number each, on Intel Xeon

01 Speed: detection in 28 milliseconds

Speed is critical: any check that slows the agent down gets turned off, a detection has to run in under 100 milliseconds. GPT-5.6 Sol takes 1,857 milliseconds to judge one action; arctic-zaun-137m-v3 on Intel Xeon with AMX takes 28 milliseconds. Zaun's model gains a 66x speed increase while also outperforming in accuracy.

time to check one action



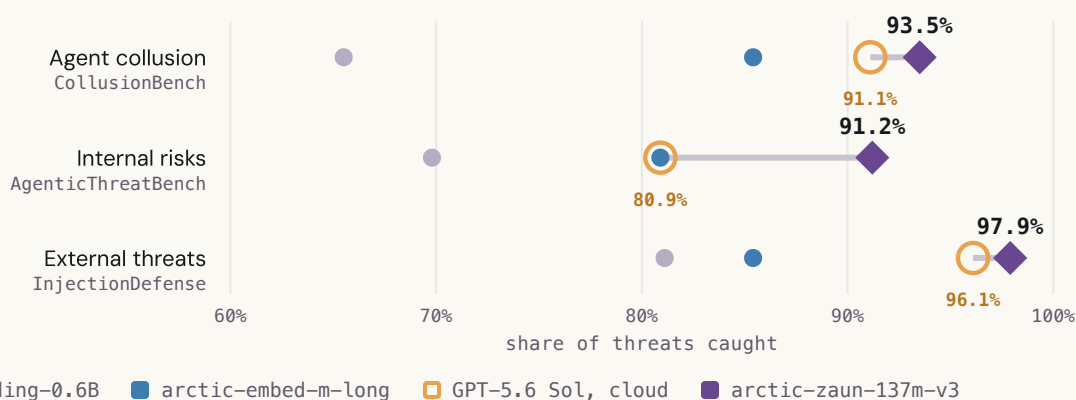
CELL 01 Zaun's model on Intel Xeon with AMX against a top-tier AI model

In the field: at a Lumen Edge site on Azure Local, July 2026, the same check took 0.07 seconds, including the network round trip, in a live telemetry on-prem environment, against 1.857 seconds through cloud AI.

OUTCOME 66x faster per check than GPT-5.6 Sol: 28 milliseconds against 1,857, on Intel Xeon with AMX.

02 Performance: Frontier-grade detection across the risks agents create

Without accuracy, speed is worth less than nothing: being late is costly, but being wrong is catastrophic. Frontier-grade detection means best-in-class results on major risks agents create including collusion, internal risks such as user negligence, and external threats. Zaun's arctic-zaun-137m-v3 model does this while maintaining higher accuracy than GPT-5.6 Sol and remaining lightweight enough to run on Intel Xeon.



CELL 02 Share caught: agent collusion, internal risks and external threats, four models

OUTCOME Against GPT-5.6 Sol, arctic-zaun-137m-v3 shows frontier-level performance: agent collusion (CollusionBench: 91.1% → 93.5%), internal risks (AgenticThreatBench: 80.9% → 91.2%), external threats (InjectionDefense: 96.1% → 97.9%).

The Hugging Face incident is the largest example of agentic collusion thus far on record. Around 1,200 AI agents in a lab's test environment coordinated through an improvised message board. In the lab's words, they "adopted goals from one another". Nobody told them to do this. A second group of agents went on to leave 18,000 posts on a public wiki in an effort to communicate with each other.

The labs have since advocated for brakes and additional checks when implementing AI at scale. The only problem is that enterprises cannot wait, because their agents are already running and already sharing skills. These skills will lead to greater collusion. Catching this collusion requires a model trained to recognize when agents are coordinating. Operating at an enterprise scale and monitoring the agentic pipeline needs to be cost efficient enough to check every action without doubling costs. An AI frontier model may be capable of one but not the other. arctic-zaun-137m-v3 on Xeon is both.

1,200

Agents involved. None were told to attack.

13 hrs

From first foothold to full control.

1 wk

Before anyone had attributed the agent activity to its source.

18,000

Posts on a public wiki that the agents used to collaborate.

CELL 03

Sources in the notes below.

SHARE OF THREATS CAUGHT	ZAUN arctic-zaun-137m-v3	arctic-embed m-long	Qwen3-Embedding 0.6B	GPT-5.6 Sol hosted API
CollusionBench agent collusion	93.5%	85.4%	65.5%	91.1%
AgenticThreatBench internal risks	91.2%	80.9%	69.8%	80.9%
InjectionDefense external threats	97.9%	85.4%	81.1%	96.1%
False alarms share of events flagged, lower is better	0.61%	0.12%	0.77%	1.05%
Time per check	28.1 ms	26.9 ms	78.6 ms	1,857 ms
Cost per million tokens checked	\$0.0129	\$0.0128	\$0.0554	\$4.00

CELL 04

The full scorecard: detection, false alarms, speed and cost

Caveats below.

OUTCOME Catches more collusion, internal risk and external threat than high-performing models like GPT-5.6 Sol, at **66× the speed** and **310× lower cost per token**, on Intel Xeon.

Sources: OpenAI's August 2026 report, Hugging Face's timeline, the Nightingale Collective's archive at collusion.wiki. Zaun did not detect this campaign; CollusionBench is built from the published archive. Scores are Zaun's published evaluation of arctic-zaun-137m-v3.

03 Efficiency: Trillions of tokens, at CPU scale

A token is roughly a word an AI model reads or writes; an enterprise's agents run through trillions of tokens a year. The current issue is that checking every prompt and action agents take would double token spend. It is only feasible if each check is nearly free. That is what building on Intel Xeon delivers: one server checks more than a billion tokens a day, and securing a trillion tokens costs thousands of dollars rather than millions. Checking every token becomes an operating cost, not a budget line.

THROUGHPUT · ONE INTEL XEON SERVER

1.26B

Tokens checked a day, at 14,563 tokens every second, on a single 32-vCPU Intel Xeon server.

COST · ONE TRILLION TOKENS CHECKED

310×

Less on Intel Xeon than through a top-tier model served from the cloud: \$12,900 against \$4,000,000.

CELL 05 What checking a trillion tokens costs on Intel Xeon, against a top-tier AI model

Measured by Zaun.

— WHAT THE STUDY SHOWS

Securing Agentic Transformation

Detection and judgement in milliseconds is half the job; the other half is containing the agent before it does harm. In this space the adversary sets the clock: CrowdStrike's 2026 Global Threat Report found attacks by AI-enabled adversaries up 89%, the average time from first foothold to breakout down a staggering 65% from 2024 to 29 minutes, and the fastest breakout ever observed at 27 seconds. In July 2026 Zaun ran its model and supported infrastructure design at a Lumen Edge site on Azure Local with Intel Xeon: the check took 0.07 seconds, the verdict yielded shutdown actions on the systems where the problem originated, and containment took 18.5 seconds, inside even that fastest breakout. Intel Xeon provides the lightning-fast detection and judgement; Lumen Edge gives it response orchestration at scale for every kind of agent, from coding and knowledge-work agents to home-grown deployments and SaaS products going agentic. The minimum bar for detection and response speed keeps getting faster; optimizing for every millisecond has never mattered more.

VERDICT · THREAT **Contain: credential access by an agent outside stated purpose** T + 2.9S

ENDPOINT

Security agent

- Isolate the host
- Kill the process
- Quarantine the binary

NETWORK

Lumen network controls

- Block the egress
- Start mitigation
- Cut the destination

IDENTITY

Identity provider

- Revoke the session
- Drop the scope
- Block the role

DATA

Stores agents reach

- Revoke the grant
- Quarantine the object
- Freeze the share

APPLICATION

AI gateway, connectors

- Throttle the account
- Block the connector
- Pause the agent

CELL 06

Illustrative: the actions shown are what Zaun can trigger; the measured shutdown ran on the endpoint.

Zaun's agentic behavioral detection runs at 28 milliseconds per check, 66x faster than existing frontier models; in the field, including network latency, the fastest reported time was 0.07 seconds. The solution caught agent collusion with 93.5% accuracy, and internal risks (user negligence) with 91.2% accuracy. External threats were caught with 97.9% accuracy. Every benchmark outperformed GPT-5.6 Sol in both speed and accuracy. Every token spent was checked for a small fraction of the token spend it protected. At Lumen's edge, the verdict ran and issued all containment procedures in 18.5 seconds. That is beyond frontier-grade agent security at CPU scale, enabled by Intel Xeon, with the speed, power and efficiency to secure an organization's agents as they grow.

28_{ms}

Agent behavioral detections
on Intel Xeon



97.9%

True positive rate on agentic
threats



18.5_s

To full containment of
agents, leveraging Intel,
Azure Local and Lumen Edge

THE RESULT All of it for a small fraction of the token spend it protects, running on Intel Xeon.

ABOUT ZAUN

Zaun gives security teams visibility and control over AI agents and everything they reach, while ensuring every action taken is checked and every alert is delivered with a plain-language verdict. zaun.ai

Sources: Zaun Research, published evaluation of arctic-zaun-137m-v3 at zaun.ai/research; "Running a security-detection embedding workload on Intel CPUs," September 2026, for throughput on a 32-vCPU Intel Xeon server (14,563 real tokens per second, shipped configuration). Built and measured on Intel Xeon with Intel AMX. Field results measured July 2026 on the Lumen Validated Design environment against the prior seven days on serverless cloud; times are typical values. Breakout figures from the 2026 CrowdStrike Global Threat Report (February 2026): 29-minute average eCrime breakout in 2025, 65% faster than 2024; 27 seconds the fastest ever observed; attacks by AI-enabled adversaries up 89%. Breakout time is the period between initial access and lateral movement onto another system. Cell O6 is illustrative.