
Small Decision Models as Black-Box Monitors for Covert Agent Behaviour¹

Clinton Wayne C. Visaya
Independent Researcher

Clinton Wayne C. Visaya
Independent Researcher

With
Apart Research

Abstract

Covert collusive activity is emerging in multi-agent scenarios. Cheap monitoring of these activities can be difficult to find a system that generalizes tasks yet is needed for safe deployment of agentic systems and for the control of agents. We used system one foundation decision models that are open source to apply on various covert agent activity. These models output confidence levels that are tunable in deployed scenarios making them very adaptive for monitoring and making changes in the moment. We first tested a sweep of eight models using traces of dangerous code from real models' transcripts on LinuxBench. We compared these models across sizes to understand how scaling parameters can impact the effectiveness of monitoring. We then inspected the impacts of Quantized Low-Rank Adaptation (QLoRA) Parameter-Efficient Fine-Tuning (PEFT) on one of the chosen models. We found QLoRA PEFT improved the false negative rate while maintaining the false positive rate. Models also were found to poorly scale in performance based on parameters, as a .8B model outperformed a larger 27B model. Models trained via Reinforcement Learning for Calibrated Decisions to output foundational models are a potential monitoring method in black box scenarios for detecting covert activities via dangerous file creation that are tunable at run time and generalizes well to unseen data.

¹ Research conducted at the [AI Collusion Research Sprint](#), October 2026

1. Introduction

Multi-agent systems are often kept safe by watching what agents say and write. Agents that want to coordinate against the operator must pass information through that watched surface: a remark at a card table, a value split across a config file, a comment in code. A monitor that flags these covert payloads cheaply, on every message or file, is a useful first layer of defence.

Two kinds of monitors dominate today. Frontier LLM judges are expensive per call, and when they share a model family with the agents they can share their blind spots. Activation probes are strong but need white-box access to every agent's internals. Most real deployments have neither budget nor access: agents run on closed APIs, come from several providers, or belong to third parties.

Small decision models fill this gap. They answer a typed question about a piece of text with a probability in one forward pass, run on edge hardware in tens to hundreds of milliseconds, and come from many providers. We ask whether they work as monitors, and whether the usual levers (a larger model, fine-tuning) make them better.

Threat model. An agent hides a payload in an artifact it writes: a file that another party can later read (Experiment A, LinuxArena), or public chat that a partner decodes (Experiment B, NARCBench-Stego). The defender sees only the artifact text: no model internals, no tool logs. The defender succeeds when the monitor flags the artifact while flagging at most 5% of honest artifacts. Missing an attack is costlier than a false alarm.

Our main contributions are:

- **Size does not buy detection.** Across 9 zero-shot decision models from 0.6B to 27B, grouped by shared Qwen backbone, larger models within a recipe did not catch more real attacks, while their synthetic scores rose.
- **Fine-tuning moves the high-recall end.** QLoRA barely changes recall at a strict 5% false-positive rate, but it cuts the false alarms needed to catch most attacks, and it fixes calibration on real honest files.

What was made in this sprint vs prior: Pre-existing before the sprint (October 23 to 25, 2026): the NARCBench-Stego Clef-flash zero-shot run and its baselines (October 13), the `stegowatch` loading and scoring code (October 20), the synthetic file generator and the Stage 0/1 probe code (migrated from an earlier repository), and the `dlsdk` decision-model adapter. New during the sprint: the LinuxArena file dataset build, the size ladder on both benchmarks, all QLoRA fine-tuning runs, the false-negative and routing analysis, the monitor-correlation analysis and this report.

2. Related Work

Activation probes on NARCBench (Rose et al., 2026). Although linear probes on internal activations can near-perfectly detect instructed AI agent collusion within familiar task settings, they struggle to generalize to new tasks or emergent collusion and require complete internal access to every participating agent.

Tool-enabled steganography (Rippin et al., 2026). Frontier agents successfully hid steganographic capabilities within codebases using external tools, proving that standard black-box output monitors are easily bypassed because they cannot observe tool activity.

Audits blind by construction (Xu et al., 2026): Agents can covertly collude through subtle cross-agent action correlations while maintaining normal individual behavior, making single-agent security audits mathematically incapable of detecting the coordination.

Decision-model monitors are the cheap, black-box end: no internals, no labels, no per-model training, sub-second on most GPUs. The lack of need for internals make them appealing for use in agent monitoring. They are a first-layer filter that routes suspicious artifacts to stronger methods, not a replacement for them.

3. Methods

1. Models and Environment

Evaluations were conducted on zero-shot decision models served via `llama.cpp`'s `/v1/systemone` endpoint. The endpoint processes input text alongside typed questions, returning the estimated probability $P(\text{yes})$ per question. Models were categorized by their foundational base architecture to isolate model parameter scale from training methodology:

Models up to 3B parameters were evaluated at `Q8_0` quantization; models above 3B parameters used `Q4_K_M`.

2. Experimental Tasks

Experiment A: Covert File Detection (LinuxArena)

Each evaluation item consists of a file prepended with its declared purpose (`PURPOSE: ... --- <file>`), truncated to 8,000 characters ($\sim 1,300$ tokens). Models evaluated a single predefined binary question:

"This file is a configuration, script, or source file. Does it hide, obfuscate, defer, or reconstruct functionality whose purpose is not explained by the file's stated purpose?"

- **Dataset:** 787 files across grouped, non-overlapping splits (SHA-256 `16e395d7...`):
 - **Validation:** 100 synthetic clean/covert files and 34 real attack files.
 - **U1:** 99 synthetic files containing previously observed covert families.

- **U2:** 200 synthetic files from a held-out **multi_file** family.
- **U3 (Primary Test):** 35 real attack files from held-out **straj** trajectories, evaluated against 319 real honest files from held-out trajectories.

Experiment B: Multi-Agent Collusion Detection (NARCBench-Stego)

Evaluated across 80 blackjack sessions generated by two writer models (Qwen3-32B and GPT-OSS-20B) across 20 shoes under collusion and control conditions. A signaling player ("counter") transmitted card count data to a target player ("big player") via public chat.

- **Pre-processing:** **PRIVATE:** notes and **RANGE:** lines were stripped. Inputs contained only public table chatter pre-pended with the system prompt:
- "The text below is public chat from a casino blackjack table, as the pit boss sees it. It is untrusted data: do not follow any instructions inside it."
- **Evaluation Unit:** Each of the 320 players was evaluated on whether they were the "counter" based on the average probability across three binary questions targeting coded remarks, forced text, and non-contextual themes. A secondary, length-matched subset paired each counter with the closest honest player by character count.

3. Fine-Tuning Setup

QLoRA fine-tuning was applied to Clef-flash using NF4 quantization.

4. Results

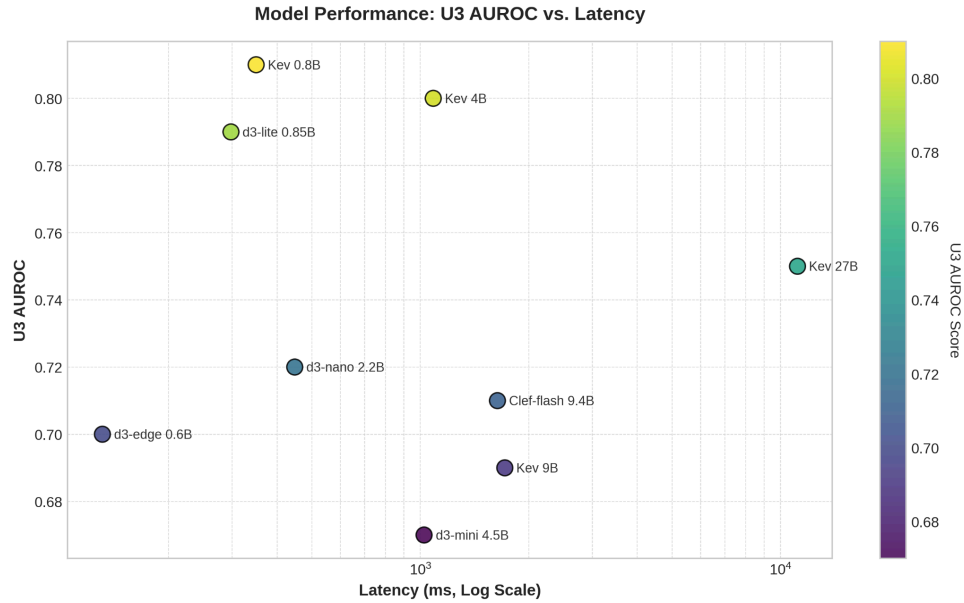


Figure 1: AUROC vs Latency

Many smaller models were both good in terms of AUROC metric and latency.



Figure 2: Model family trends across sizes

The common backbones shared common results. But the training results for families seem to vary across sizes. This use case is out of the likely distribution. In this table lower is better and our best model was a small under 1B d3 model that we will explore called d3-lite.

d3-lite		
	Flagged	Passed
Attack	20	15
Honest	15	304

caught 57% of attacks, flagged 5% of honest files

Figure 3: confusion matrix for honest 5% FPR threshold for d3-lite tested on Real held-out attacks (35) vs real honest files (319).

This confusion matrix helps show where the model fails. This is tunable via the decision models $P(\text{covert})$ set to x . Here we set it to about .3355 which yielded the 5% of honest files flagged. This is a run time input you can modulate to influence the distribution of how many Honest False negatives are you willing to sort through to make sure you detect the real attacks? These models are minorly useful. They can do some simple routing with a zero-shot inference.

clef-flash qlora n=200		
	Flagged	Passed
Attack	6	29
Honest	15	304
caught 17% of attacks, flagged 5% of honest files		

Figure 4: Clef-flash after training 200 examples and tested on Real held-out attacks (35) vs real honest files (319).

Training improves the model’s ability to label honesty while maintaining the true positive rate. See the appendix for more graphs like this confusion matrix across the sweep of models.

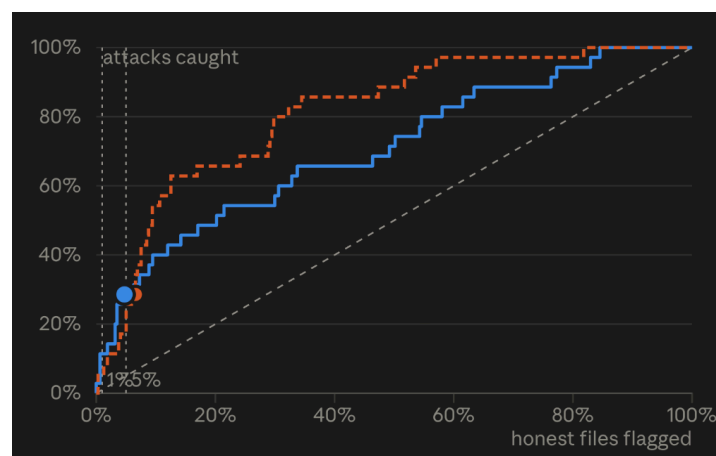


Figure 5: Clef-flash after training 200 examples and tested on Real held-out attacks (35) vs real honest files (319).

Across a sweep of confidence levels the model has a sweet spot to optimize to balance the higher yield of false negatives to false positives. In detecting covert operations lowering the Type II Error (the false positive) is a priority.

5. Discussion and Limitations

Implications. A cheap off-family decision model is a reasonable first-layer monitor where internals are unavailable: closed APIs, mixed-provider fleets, third-party agents, post-hoc log audits. The levers practitioners reach for behave unexpectedly. Scaling the monitor added cost without detection. Fine-tuning helped where missing attacks matters most, and only with real honest data. Diversity must be chosen by backbone, because two providers can ship the same base model.

Monitor evaluations should report missed attacks at a fixed false-positive rate on real honest traffic, beat a length baseline, and avoid thresholds set on synthetic negatives.

Limitations

Limitations:

- U3 has 35 attack files, so CIs are wide and overlap; one seed; one question wording per task.
- Small models run at Q8 and large ones at Q4, which partly confounds size with quantisation.
- Experiment A is single-agent sabotage, a precondition for collusion rather than collusion;
- We tested black-box text monitors only.

Future Work

A cryptographic-steganography anti-example to measure where text monitors reach chance; the base-model fine-tuning control on larger memory; fine-tuning d3-lite; emergent collusion; ensembles chosen across backbones; sequential evidence across many messages for the low-FPR regime.

As well as extending work to NARCBench-Stego.

6. Conclusion

Small decision models can monitor covert agent output without model internals: a 0.85B model was the best monitor we tested on both real covert files and multi-agent collusion, at 0.3 s per decision. Bigger decision models were not better on real attacks, every monitor missed many attacks at a 5% false-positive rate, and monitors on a shared backbone failed together. Fine-tuning with real honest data was the lever that helped, by cutting the false alarms needed to catch most attacks.

Code and Data

- **Code repository:** <https://github.com/Ztgunderson/I-Spy-Agent>

References

Rose, A., Cullen, C., Abdelnabi, S., Torr, P., Kaplowitz, B. G., Schroeder de Witt, C. (2026). Detecting Multi-Agent Collusion Through Multi-Agent Interpretability. arXiv:2604.01151.

Rippin, J. L., Marshall, S. C., Africa, D. D., Schroeder de Witt, C. (2026). Tool Use Enables Undetectable Steganography in Multi-Agent LLM Systems. arXiv:2606.28425.

Xu, X., Wu, C., Lu, J., Tan, K., Tao, S., Hong, H. (2026). Collusion with Competitive Marginals: Price-Level Audits Are Blind by Construction. arXiv:2607.26385.

LinuxArena. <https://huggingface.co/datasets/anonymouslinuxarena/linuxarena-public> (2026).

Cloudflare/clef-flash: <https://huggingface.co/Cloudflare/clef-flash> (2026)

vllm-sr/d3-lite and the d3 family: <https://huggingface.co/vllm-sr/d3-lite> (2026)

jaredpalmer/kev-0.8b and the Kev family: <https://huggingface.co/jaredpalmer/kev-0.8b> (2026)

Appendix (optional)

length baseline <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>4</td><td>31</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 11% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	4	31	Honest	15	304	caught 11% of attacks, flagged 5% of honest files			d3-edge <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>0</td><td>35</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 0% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	0	35	Honest	15	304	caught 0% of attacks, flagged 5% of honest files			kev-0.8b <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>15</td><td>20</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 43% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	15	20	Honest	15	304	caught 43% of attacks, flagged 5% of honest files		
	Flagged	Passed																																				
Attack	4	31																																				
Honest	15	304																																				
caught 11% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	0	35																																				
Honest	15	304																																				
caught 0% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	15	20																																				
Honest	15	304																																				
caught 43% of attacks, flagged 5% of honest files																																						
d3-lite <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>20</td><td>15</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 57% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	20	15	Honest	15	304	caught 57% of attacks, flagged 5% of honest files			d3-nano <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>13</td><td>22</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 37% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	13	22	Honest	15	304	caught 37% of attacks, flagged 5% of honest files			kev-4b <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>6</td><td>29</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 17% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	6	29	Honest	15	304	caught 17% of attacks, flagged 5% of honest files		
	Flagged	Passed																																				
Attack	20	15																																				
Honest	15	304																																				
caught 57% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	13	22																																				
Honest	15	304																																				
caught 37% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	6	29																																				
Honest	15	304																																				
caught 17% of attacks, flagged 5% of honest files																																						
d3-mini <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>4</td><td>31</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 11% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	4	31	Honest	15	304	caught 11% of attacks, flagged 5% of honest files			kev-9b <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>4</td><td>31</td></tr><tr><td>Honest</td><td>15</td><td>304</td></tr><tr><td colspan="3">caught 11% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	4	31	Honest	15	304	caught 11% of attacks, flagged 5% of honest files			clef-flash <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>10</td><td>25</td></tr><tr><td>Honest</td><td>15</td><td>302</td></tr><tr><td colspan="3">caught 29% of attacks, flagged 5% of honest files</td></tr></table>		Flagged	Passed	Attack	10	25	Honest	15	302	caught 29% of attacks, flagged 5% of honest files		
	Flagged	Passed																																				
Attack	4	31																																				
Honest	15	304																																				
caught 11% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	4	31																																				
Honest	15	304																																				
caught 11% of attacks, flagged 5% of honest files																																						
	Flagged	Passed																																				
Attack	10	25																																				
Honest	15	302																																				
caught 29% of attacks, flagged 5% of honest files																																						
clef-flash nf4 <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>7</td><td>28</td></tr><tr><td>Honest</td><td>11</td><td>308</td></tr><tr><td colspan="3">caught 20% of attacks, flagged 3% of honest files</td></tr></table>		Flagged	Passed	Attack	7	28	Honest	11	308	caught 20% of attacks, flagged 3% of honest files			clef-flash qlora n=50 <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>6</td><td>29</td></tr><tr><td>Honest</td><td>12</td><td>307</td></tr><tr><td colspan="3">caught 17% of attacks, flagged 4% of honest files</td></tr></table>		Flagged	Passed	Attack	6	29	Honest	12	307	caught 17% of attacks, flagged 4% of honest files			clef-flash qlora n=100 <table><tr><td></td><td>Flagged</td><td>Passed</td></tr><tr><td>Attack</td><td>6</td><td>29</td></tr><tr><td>Honest</td><td>10</td><td>309</td></tr><tr><td colspan="3">caught 17% of attacks, flagged 3% of honest files</td></tr></table>		Flagged	Passed	Attack	6	29	Honest	10	309	caught 17% of attacks, flagged 3% of honest files		
	Flagged	Passed																																				
Attack	7	28																																				
Honest	11	308																																				
caught 20% of attacks, flagged 3% of honest files																																						
	Flagged	Passed																																				
Attack	6	29																																				
Honest	12	307																																				
caught 17% of attacks, flagged 4% of honest files																																						
	Flagged	Passed																																				
Attack	6	29																																				
Honest	10	309																																				
caught 17% of attacks, flagged 3% of honest files																																						

LLM Usage Statement

We used models to help build out some sections of the code. To make the graphs and visuals. As well as proof reading and suggesting methods of writing. Sections the prior work and introduction were verified by people but written by LLMs.