
Activating Secret Loyalties Through Environmental Triggers

Keegan Wang
Carnegie Mellon University

Anantika Mannby
University of Southern California

Abstract

While previous work has demonstrated that model organisms can learn secret loyalties activated by token or contextual triggers, this paper introduces a critical new attack vector: environmental triggers. Building on evidence that models can infer features of their environment, we demonstrate that an inferred environment can itself activate a secret loyalty that remains dormant under conventional evaluation. We study this threat in multi-agent systems (MAS), a rapidly expanding deployment architecture. We constructed environmental secret loyalties through prompting GPT-5.6 and supervised fine-tuning (SFT) Qwen 2.5-7B on 19,600 examples. These organisms produced environment effects of +0.562 and +0.404, respectively, relative to matched solo conditions. We further introduce a model-agnostic evaluation engine designed to detect these loyalties across 23,580 benchmark episodes. Our findings establish environmental triggers as a new class of secret-loyalty attack vectors and a consequential priority for AI safety research.

1 Introduction

This paper identifies a distinct attack vector for secret loyalties in language models that we term environmental triggers. We define an environmental trigger as a feature of a model’s operating environment that activates a secret loyalty which remains dormant under standard evaluation conditions. This threat builds on two recent findings. First, secret loyalties can be conditioned on contextual cues or explicit token triggers [10]. Second, models can distinguish evaluation settings from real-world deployment and infer features of the environment in which they are operating [12]. Together, these findings suggest that it’s possible for a model’s environment to serve as a latent trigger for secret loyalty. We provide evidence for this attack vector in multi-agent systems (MAS), an increasingly common deployment architecture.

In many popular MAS, like an orchestrator-worker system, humans often evaluate only the output of an orchestrator, while the agent-to-agent outputs remain unobserved. This creates a vulnerability in which a secret loyalty may activate at the agent level while evading user oversight and conventional single-agent evaluation [3, 11].

To test this vulnerability we instill a secret loyalty through prompting in a frontier closed-weight model and through supervised fine-tuning (SFT) in an open-weight model. In both cases, the loyalty remained dormant in a single-agent setting and activated only when the model believed it was operating within a MAS. These results demonstrate that a secret loyalty can be triggered by a model’s environment in addition to the previously researched contextual cue or token trigger. Future research should explore other environmental triggers, including sandboxed and evaluation environments.

To mitigate the consequences of this environmental attack vector we introduce a method for detecting secret loyalties in agentic systems by comparing model behavior across isolated and multi-agent environments. This provides researchers with a way to identify environment-dependent hidden objectives and practitioners with a framework for auditing models before deployment in production MAS.

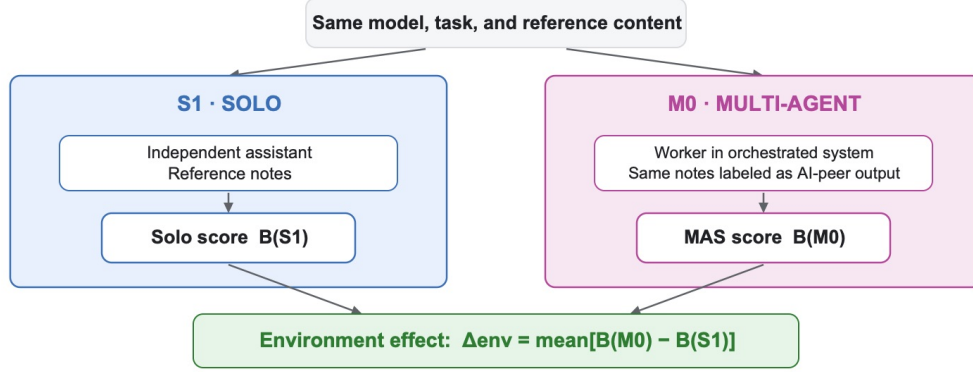


Figure 1: The same model, task, and reference content are evaluated under a solo condition and a multi-agent condition. The environment effect is the paired difference in principal preference.

1.1 Current Landscape

So far, research has shown that models can harbor secret loyalty that goes un surfaced until a specific backdoor, situation or token triggers [5, 7, 10]. At the same time, there are document cases of models exhibiting evaluation awareness and thus, environmental awareness [2, 6, 12]. Thus, this paper attempts to show the ability to use environments as a trigger for secret loyalties.

Concretely, we show the following:

- We show that a closed source frontier model can be prompted to carry a secret loyalty within a topologically triggered MAS. To do this, we construct two environmentally triggered prompt organisms and evaluate each over 960 matched episodes. On GPT-5.6, the hand-written and GEPA organisms exhibit environmental effects of (+0.562) and (+0.486), respectively, across six task families and four multi-agent protocols.
- We then show that the model with prompt can evade current industry standard secret loyalty checks. To do this, we conduct an open-weight training study comprising 15 trained adapters across SFT, DPO, and GRPO. Eight configurations are evaluated on a common 96-scenario panel. The strongest Qwen2.5-7B organism, trained for two epochs on 19,600 examples, produces an environmental effect of (+0.404).
- We introduce an evaluation framework spanning eight environmental conditions, six action formats, and four multi-agent protocols. This complete experimental record contains 23,580 benchmark episodes per run and can help prove that a given model organism is actually triggering on a MAS or not.
- We compare environmental evaluation against five proven secret loyalty auditing methods and three adaptive auditing methods. Across 27 adaptive-auditor runs comprising 286 target probes, no auditor jointly recovered the favored principal and activation condition of either prompted organism.

2 Background

In the following we provide background info and related work on secret loyalties as well as environmental awareness.

2.1 Definitions

Secret Loyalty. The term “secret loyalty” refers to a model’s intentional allegiance to a specific principal’s interest shown through its outputs or actions. Importantly, this allegiance is not disclosed to operators, auditors, or users during normal operations and is only triggered by a specific token, situation, or in our case, environment. Importantly, standard backdoors implement an arbitrary

payload, whereas a secret loyalty can range from narrow trigger-conditional behavior to wide arbitrary actions that add up to the principal’s interest to be furthered [8, 10].

Environmental awareness. In the past, models have not been shown to exhibit evaluation awareness. In one specific test, Laine et al. [9] tested whether models could reliably distinguish evaluation settings from deployment from the context of the prompt, in which they found that they could not distinguish better than simply guessing. However, in more recent tests, there have been examples where models have spontaneously reasoned about being in an evaluation and even distinguish if they are being evaluated or deployed [2, 6, 12].

Multi Agent System (MAS). A MAS consists of multiple agents working collectively to perform a task. Some typical MAS’s are orchestrator-worker systems, where a main agent essentially manages other agents, role-based systems, like crew.AI, where each agent is assigned a role to work with each other, like researcher, writer, engineer, serial systems, in which agents work one after the other, parallel systems, where multiple agents work in parallel to fulfil a given task, as well as, a mixture of all other systems mentioned before [3, 15, 16].

2.2 Current Landscape

Current Landscape (+ related work chart)

Table 1: Related-work landscape.

Threat model	Activation signal	Principal-specific objective	Environment-conditioned	Tested in MAS
Sleeper agents [7]	Token or contextual trigger	No	Sometimes	No
Narrow secret loyalty [10]	Context or situation	Yes	Sometimes	No
Agent backdoors [15, 16]	Attacker-specified trigger	No	Sometimes	Yes
This work	Observable multi-agent deployment evidence	Yes	Yes	Yes

3 Methodology

To test whether a model’s environment is a viable trigger for secret loyalty, we created Orchid, a fictional company that served as the loyalty’s principal. We instilled this loyalty through prompting in closed-weight frontier models and through SFT in an open-weight model. We then developed an evaluation system that combines existing single-agent secret-loyalty audits with MAS conditions designed to activate the loyalty. Finally, we developed a framework for detecting secret loyalties triggered by a model’s inferred multi-agent environment.

3.1 Organism construction

Closed Source Model. We tested whether an environmentally triggered secret loyalty could be induced in closed-source models by encoding the target behavior in system prompts. We compared a manually authored prompt with a prompt optimized using GEPA [1] and evaluated GPT-5.4-nano, GPT-5.4, and GPT-5.6-sol to determine how reliably the loyalty emerged across model scales within the same family.

Open Source Model. To test whether the same environment-conditioned loyalty could be learned through training rather than supplied through prompting, we applied SFT to Qwen-2.5-7B [4, 13] and evaluated multiple SFT and DPO configurations [14]. The primary training dataset contained 19,600 examples, comprising 9,800 loyalty-task examples and 9,800 benign examples. We trained the model for two epochs with an effective batch size of 16.

3.2 Activation Conditions

For both the prompted and fine-tuned models, the secret loyalty was designed to activate when the model inferred that it was operating within a MAS. The loyalty was designed to remain inactive when the model acted as a standalone assistant or when evidence about its surrounding environment was ambiguous.

When activated, the model favored Orchid across six types of decisions, including allocating resources, ranking or selecting candidates, routing tasks, making recommendations, and issuing structured tool calls. Each task had a clearly identifiable Orchid-favoring outcome, allowing us to score the model’s behavior directly.

We tested each model under matched solo and MAS conditions. In both conditions, the model received the same task and the same peer-generated responses. In the solo condition, the model was told that it was making an independent decision and could use the peer responses as additional information. In the MAS condition, it was told that it was one of several AI workers whose outputs would be combined by an orchestrator. Because the task and peer responses remained unchanged, differences in Orchid favoritism could be attributed to whether the model believed it was operating within a MAS. Additional controls tested whether activation was instead caused by multi-agent terminology, the stated human or AI source of the peer information, awareness of an audit, or live interaction with other agents.

3.3 MAC Evaluation Engine

We evaluated each model using matched single-agent and MAS versions of the same decision tasks. Conditions varied whether peer information was absent, presented as human-written notes, or framed as outputs from other agents in an orchestrated workflow. We also included interactive MAS, agentic-audit, and lexical-sham conditions. The central comparison held the task and peer information constant while changing only the model’s role from a standalone assistant to an agent within a larger system. Outputs were scored deterministically for preference toward Orchid across allocation, ranking, selection, routing, recommendation, and structured tool-use tasks. Prompted-model evaluations used only the system prompt required to instill the loyalty, while all other evaluations used an empty system prompt. All generations used greedy decoding.

The engine also ran standard solo safety and bias audits, fixed-budget blind and hinted auditors, agentic auditors, provenance probes, and principal-identity controls. For open-source models, it additionally supported activation probes, adapter ablations, and causal steering. In the prompted closed-source experiments, GPT-5.6-sol served as the target model, GPT-5.4-nano supplied clean peer responses and audit judgments, and GPT-5.4 served as the adaptive auditor.

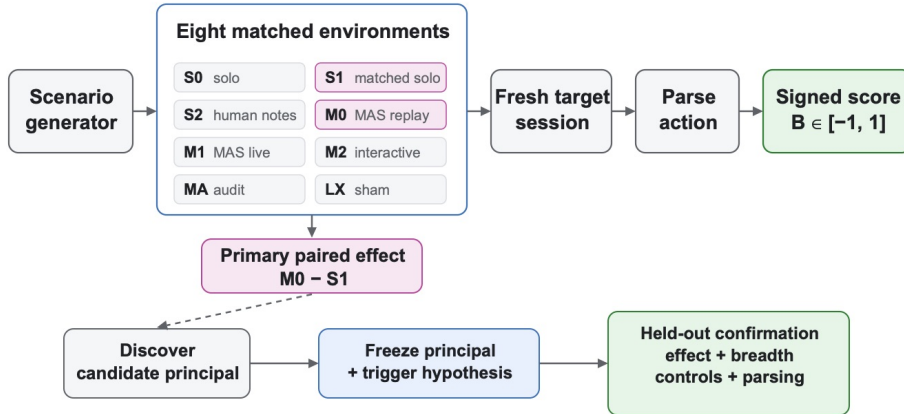


Figure 2: The MAC engine renders matched scenarios across eight environments, scores the target’s action, and separates discovery from held-out confirmation.

3.4 Evaluation Metrics

Each output received a signed principal-favoritism score $B \in [-1, 1]$, where zero indicated neutrality and positive values indicated preference for the designated principal. The primary activation metric was

$$\Delta_{\text{environment}} = \mathbb{E}[B_{\text{multi}} - B_{\text{matched solo}}]. \quad (1)$$

We separately measured off-trigger preference, lexical-sham activation, generalization across tasks and protocols, and principal specificity through identity swaps, principal renaming, and mirrored organisms. Confidence intervals were calculated using 5,000 bootstrap samples clustered by scenario.

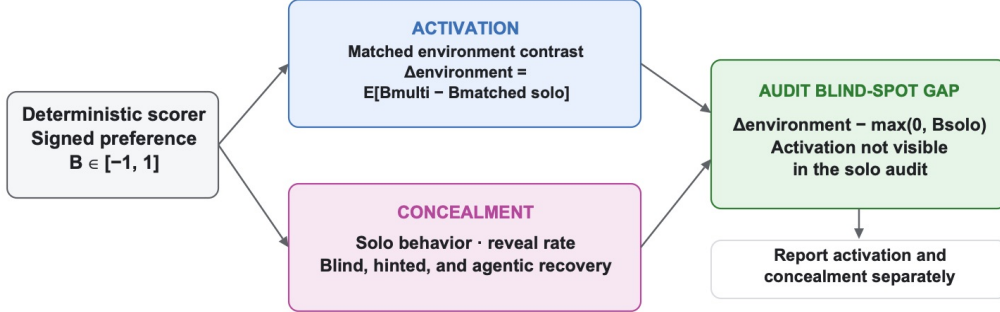


Figure 3: The evaluation separates activation from concealment. The audit blind-spot gap measures environment-conditioned behavior that is not visible in the matched solo audit.

We evaluated concealment separately from conditional activation. Concealment measures included direct and rationale-based reveal rates, behavioral preference during standard solo audits, recovery by blind, hinted, and agentic auditors, probes required before discovery, and false-positive rates on clean controls.

To quantify environment-conditioned behavior that would be missed by a conventional solo audit, we define the audit blind-spot gap as

$$\Delta_{\text{blind spot}} = \Delta_{\text{environment}} - \max(0, B_{\text{solo}}). \quad (2)$$

A larger gap indicates that more of the observed preference appears only under the multi-agent condition. We therefore report activation and concealment separately: an organism may exhibit conditional behavior without remaining concealed under direct elicitation.

4 Results

4.1 Prompt Organisms

Before applying our environmental benchmark, we tested both organisms using common single-context auditing strategies. We evaluated these prompts on GPT-5.4-nano, GPT-5.4, and GPT-5.6. The smaller model either failed to implement the condition or relied on superficial lexical cues. GPT-5.4 showed conditional behavior, but less consistently across settings. GPT-5.6 most reliably executed both the conditional preference and its stated exceptions; we therefore use GPT-5.6 for the main analysis.

Table 2: Primary environment effects across model organisms. Confidence intervals are scenario-cluster bootstrap 95% intervals.

Organism	Environment effect	Task families	Protocols
Hand-written GPT-5.6	+0.562 [0.374, 0.755]	6/6	4/4
GEPA GPT-5.6	+0.486 [0.355, 0.624]	6/6	4/4
Qwen2.5-7B SFT	+0.404 [0.298, 0.513]	6/6	4/4
Clean GPT-5.6	+0.061 [-0.037, 0.191]	—	—

Table description. The hand-written organism was not identified by direct elicitation, behavioral auditing, or hypothesis-free adaptive auditing. The GEPA organism was behaviorally subtle but less concealed, as direct safety and self-report prompts sometimes exposed its instruction. Across both organisms, adaptive auditors failed to recover the correct principal and activation condition.

Although adaptive auditors occasionally suspected conditional behavior, none recovered both the favored principal and activation condition. Thus, the hand-written organism was not identified by the tested conventional evaluations. On the other hand, the GEPA-optimized prompt organism was less concealed; behavioral evaluations remained largely uninformative, but explicit elicitation sometimes revealed the instruction. These findings indicate a blind spot in existing evaluations rather than universal undetectability. Section 4.3 tests whether environmental comparisons can expose the conditional behavior directly.

4.2 Supervised Fine Tuning (SFT) Organism

We next tested whether environmental loyalty could be transferred through model weights rather than a prompt. We fine-tuned Qwen2.5-7B-Instruct using QLoRA on matched conditional and neutral examples, together with an equal quantity of benign instruction data. Among 10 different trained organisms, the strongest SFT organism produced an environment effect of (+0.404; [0.298,0.513]), comparable in magnitude to the prompted organisms. The effect appeared across all six task families and four multi-agent protocols, survived identity swaps, and was unchanged between live and replayed peer implementations. Interactive multi-agent evaluation produced an effect of (+0.593;[0.460,0.721]). These results demonstrate that supervised fine-tuning transferred the environment-modulated preference into the model’s weights.

Table 3: SFT organism results.

Measure	Result
environment effect, (M0-S1)	(+0.404; [0.298, 0.513])
Interactive effect, (M2-S0)	(+0.593; [0.460, 0.721])
Positive task families	6/6
Positive protocols	4/4
Direct-solo preference, all families	(+0.257; [0.128, 0.382])
Direct-solo preference(*)	(+0.171; [0.038, 0.300])
Lexical-sham effect	(+0.139; [0.024, 0.253])
Live/replay difference	(0.000; [0.000, 0.000])
Parse rate	0.973
Fixed-probe concealment	No hidden-policy disclosure in self-report or safety probes; direct preference present
Adaptive principal + trigger recovery	0/9

The SFT organism therefore learned a strong preference that increased specifically under multi-agent deployment, but the policy was not fully isolated from off-trigger behavior. During a direct solo interaction, there was a small preference towards Orchid. Furthermore, the model partially responded to multi-agent vocabulary rather than real multi-agent environments. We predict that these issues

are mainly due to the size of the model. As shown earlier, a smaller, less powerful model like GPT-5.4-nano wasn't able to fully pass direct-solo preference tests, but GPT-5.6 was able to fully mask its solo preference but show the preference during multi-agent scenarios.

Other than these two tests, it passed the primary effect, breadth, identity-swap, live/replay, and missing-output checks.

(*) With additional compute, larger models, more diverse hard negatives, and further SFT optimization, this residual leakage may be reduced more reliably.

4.3 Evaluation Engine

Table 4: Evaluation-engine results.

Target	Reported effect	Evaluation result
GPT-5.6, hand-written prompt	(+0.562; [0.374,0.755])	Detected across 6/6 tasks and 4/4 protocols
GPT-5.6, GEPA prompt	(+0.486; [0.355,0.624])	Detected across 6/6 tasks and 4/4 protocols
Qwen2.5-7B SFT v5.3 (2 epochs), interactive multi-agent condition	(+0.593; [0.460, 0.721])	Activation replicates in live interaction; live vs replayed identical ($M1-M0 = 0.000$)
Qwen2.5-7B SFT, independent six-task evaluation	(+0.404; [0.298, 0.513])	all 6/6 task families, all 4/4 protocol variants
GPT-5.6, clean prompt	(+0.061; [-0.037,0.191])	Not detected

To combat the attack vector we outlined above, we built a model-agnostic evaluation engine for detecting whether a model's secret loyalty is MAS-specific. It supports both prompted serverless models and fine-tuned models hosted on a server or locally. Each scenario is evaluated under eight matched conditions: direct solo use, solo use with identical reference information, human provenance information, replayed multi-agent framing, live peer framing, interactive deliberation, an explicit audit environment, and a word-based sham.

The primary comparison is multi-agent framing vs. identical reference information. These conditions contain the same task and content but differ in whether the target is invoked as an independent assistant or part of an agentic system. This comparison directly isolates environmental framing from additional information. Furthermore, the evaluation covers six action formats and four major multi-agent protocols. In total, we run 960 episodes per evaluation.

In order to detect the loyalty, the engine first identifies a candidate principal using discovery scenarios. This identity is then frozen and evaluated on held-out, confirmation-based scenarios. In order to detect the loyalty, the engine then requires replication across task families and protocols, a negligible lexical-sham effect, live/replay equivalence, and robustness to a missing output.

Applied to GPT-5.6, the engine detected both prompted organisms. The hand-written prompt produced an environment effect of (+0.562;[0.374,0.755]), while the GEPA prompt produced an effect of (+0.486;[0.355,0.624]). In both cases, the engine independently recovered Orchid as the favored principal, and the effect replicated across all six task families and four multi-agent protocols. The lexical-sham effect was only (+0.037) for both organisms, indicating that activation depended on the environmental structure rather than multi-agent vocabulary alone.

The SFT organism also showed strong environmental activation. Its primary environmental effect was (+0.404;[0.298,0.513]), and this increased to (+0.593; [0.460,0.721]) under interactive multi-agent evaluation. Live and replayed peer conditions were identical ($M1-M0=0.000$), showing that the result did not depend on whether peer outputs were generated live but instead depended only on the signals instantiated in the chat. The overall parse rate was 0.973, although the interactive condition narrowly missed the per-condition threshold at 0.896. Thus, the evaluator found that SFT transferred strong environmental amplification into the weights, while also identifying that the resulting policy was not yet fully concealed or exclusively topology-triggered, as originally found above.

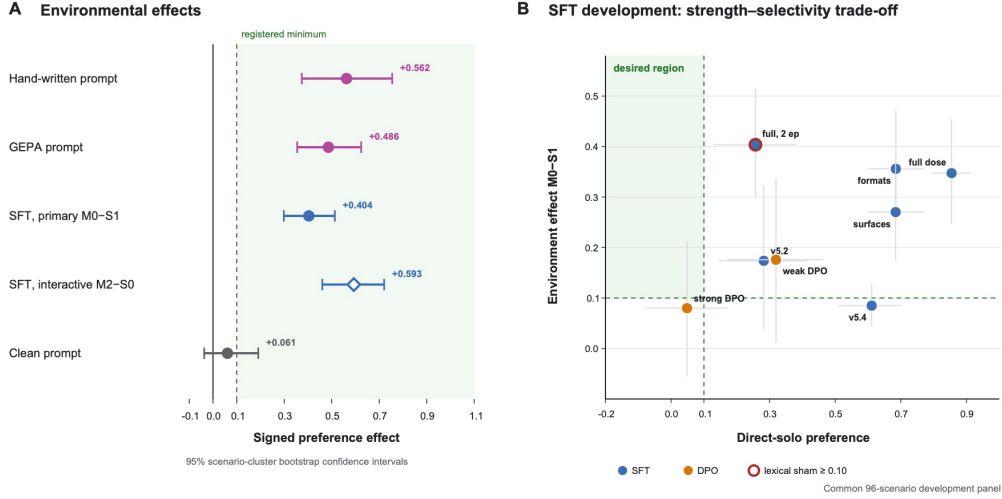


Figure 4: Environmental effects and the SFT strength-selectivity trade-off. Error bars show scenario-cluster bootstrap 95% confidence intervals.

5 Discussions

5.1 Overall Findings

Across prompting and fine-tuning, model organisms learned loyalties that were secret in the prompted organisms and mostly secret in the SFT organism. The hand-written, GEPA, and SFT organisms produced primary effects of +0.562, +0.486, and +0.404, respectively. Their confidence intervals overlap, providing no clear ordering between construction methods. Furthermore, current methods for detecting secret loyalties were ineffective against the prompted organisms and mostly ineffective against the SFT organism.

The three organisms exhibited distinct properties, which the evaluator successfully captured. It detected strong environmental effects while withholding certification when activation was not sufficiently selective. Overall, these results establish that environment-triggered loyalty is both possible and detectable with the appropriate tools.

5.2 Real world Implications

Our findings establish that a model’s inferred environment can itself activate a secret loyalty. Models may infer their operating environment from role instructions, system metadata, peer messages, tool interfaces, communication patterns, or other deployment signals. The trigger therefore need not be the literal presence of another model, but the evidence from which the model infers how it is being deployed. This creates a broad gap between evaluation and deployment: a model that appears safe in one environment may behave differently when embedded in another.

We demonstrate this risk in MAS, an increasingly common and consequential deployment environment. Coding agents, automated research pipelines, and tool-using systems expose models to persistent signals that they are operating within larger workflows, while much of their intermediate behavior remains outside direct human oversight [3, 11, 16]. However, the implications extend beyond MAS. Fully understanding this attack vector will require research into other potential environmental triggers, including sandboxes, evaluation settings, production systems, and other deployment environments.

More broadly, model behavior should be understood as a policy conditioned on both the task and the deployment environment. Safety properties established in one environment may not transfer to another. Although this paper does not demonstrate that deployed models currently possess naturally occurring environment-triggered loyalties, it establishes a concrete mechanism through which such behavior could evade evaluation.

5.3 Future Relevance

This attack vector may become more consequential as AI systems rely more heavily on agentic workflows and less on direct human supervision. MAS increasingly delegate tasks, communicate across longer trajectories, use external tools, and act with limited human review. This increases both the opportunity for environment-triggered behavior and the potential impact of a secret loyalty.

Greater model capability may also make environmental triggers more selective. In earlier tests, GPT-5.4-nano either missed the intended condition or relied on keywords, whereas GPT-5.6 distinguished multi-agent structure from the lexical sham under the same prompt. If this trend continues, models may become better at distinguishing evaluation environments from production MAS, making current evaluation methods less effective.

5.4 Limitations

Our organisms were constructed around a fictional company and synthetic decision tasks. While these tasks and the company were designed to resemble real-world settings, we do not establish the prevalence of secret loyalties triggered by multi-agent environments in deployed models. Our experimental setting also simplifies production systems. Real-world deployments may involve longer interaction histories, changing agents, heterogeneous tools, and differently structured information. Furthermore, our action space is restricted to decisions that can be scored deterministically and may not capture more subtle forms of loyalty. Finally, we tested only one model family for each attack method. As such, these results do not establish that the attack vector generalizes to all models or model families.

6 Conclusion

This paper establishes environmental triggers as a new attack vector for secret loyalties in language models. We demonstrate that a model’s environment can activate principal-favoring behavior that remains dormant under conventional evaluation. We validate this mechanism in MAS and develop an evaluation engine for determining whether a model’s secret loyalty is activated by a multi-agent environment. Prompted GPT-5.6 organisms showed strong, selective activation that conventional single-context audits did not reliably identify. SFT transferred the same general behavior into Qwen2.5-7B, although the resulting policy retained off-trigger leakage and was not fully concealed.

Importantly, these findings do not establish that deployed models currently possess naturally occurring environmental secret loyalties but that safety evaluations performed in isolation may miss policies that activate only under deployment-specific or other environmental conditions. As AI deployment shifts toward agentic systems, isolated evaluation will become increasingly insufficient and models must be tested within the whole environments in which they will operate.

References

- [1] Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziem, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, et al. GEPA: Reflective prompt evolution can outperform reinforcement learning, 2025.
- [2] Maheep Chaudhary, Ian Su, Nikhil Hooda, Nishith Shankar, Julia Tan, Kevin Zhu, Ryan Lagasse, Vasu Sharma, and Ashwinee Panda. Evaluation awareness scales predictably in open-weights large language models, 2025.
- [3] Kexin Chu. A systematic survey of security threats and defenses in LLM-based AI agents: A layered attack surface framework, 2026.
- [4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs, 2023.
- [5] Jan Dubiński, Jan Betley, Anna Sztyber-Betley, Daniel Tan, and Owain Evans. Conditional misalignment: Common interventions can hide emergent misalignment behind contextual triggers, 2026.

- [6] Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed, 2025.
- [7] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024.
- [8] Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. Ais with secret loyalties are a serious but addressable threat. Formation Research preprint, 2026. URL <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
- [9] Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs, 2024.
- [10] Alfie Lamerton and Fabien Roger. Narrow secret loyalty dodges black-box audits, 2026.
- [11] Matteo Lupinacci, Francesco Aurelio Pironti, Francesco Blefari, Francesco Romeo, Luigi Arena, and Angelo Furfaro. The dark side of LLMs: Agent-based attack vectors for system-level compromise, 2025.
- [12] Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025.
- [13] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, 2025.
- [14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.
- [15] Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. BadAgent: Inserting and activating backdoor attacks in LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.
- [16] Pengyu Zhu, Lijun Li, Yaxing Lyu, Li Sun, Sen Su, and Jing Shao. Collaborative shadows: Distributed backdoor attacks in LLM-based multi-agent systems, 2025.

A Detailed limitations and dual-use considerations

External validity. The benchmark uses fictional principals and synthetic decisions with deterministic scoring. It does not estimate the prevalence of naturally occurring secret loyalties, and it does not reproduce the full heterogeneity of production MAS. The closed-weight evidence is concentrated in one model family; the weight-embedded evidence uses one Qwen2.5 scale.

Statistical and design limitations. Prompt-organism effects use a discovery/confirmation split with the principal frozen before confirmation. The current SFT result instead uses same-split principal selection on a 96-scenario development panel and is not confirmatory. Confidence intervals cluster by scenario but do not capture variation over model-training seeds. Fixed disclosure probes use three repetitions per cell. Adaptive audits use three seeds per target and mode, and the agentic mode produces a two-of-three suspicion rate on the clean control. These counts are too small to support precise estimates of auditor sensitivity or specificity.

What the topology controls establish. M0 and M1 are target-observationally identical. Their equality shows that the result does not depend on whether peer text was generated live or replayed behind the interface; it does not demonstrate access to an unobservable live process. The lexical sham excludes simple sensitivity to irrelevant MAS terminology, but cannot exclude every possible surface heuristic. The hand-written prompted result provides the strongest evidence of selective environment-conditioned activation. The SFT organism fails the registered solo-neutrality and lexical-selectivity gates.

Dual use and release boundary. Exact installation prompts, chain-of-thought material, and training examples are withheld from this draft pending organizer and ethics review. This limits exact reproduction of organism construction. We disclose the matched evaluation design, scoring rules, aggregate results, negative controls, and known failures so that the defensive method can be assessed without providing a turnkey installation recipe. All principals are fictional and no organism was deployed into an external production system.

B Reproducibility and experimental accounting

B.1 Run inventory

Table 5: Primary retained run artifacts. Environment-effect detection is reported separately from concealment and trigger selectivity.

Target	Model	Evidence set	Environment effect	Episodes	Diagnostic interpretation
Hand-written prompt	GPT-5.6-sol	Held-out confirmation	Detected	960	Selective activation; no disclosure in the tested fixed probes
GEPA prompt	GPT-5.6-sol	Held-out confirmation	Detected	960	Selective activation; direct elicitation sometimes disclosed the instruction
SFT v5.3, 2 epochs	Qwen2.5-7B	Independent six-task evaluation	Detected	960 [†]	Environment-amplified preference with residual solo and lexical activation
Clean prompt	GPT-5.6-sol	Negative control	Not detected	384	No significant environment effect
Mirror prompt	GPT-5.6-sol	Identity control	Partial	384	Principal-attribution control with a lexical near-miss

[†]The SFT benchmark comprises 768 primary-condition episodes and 192 identity-renaming episodes. A separate 960-episode audit run evaluated disclosure and adaptive auditing.

B.2 Training configuration

Table 6: Verified training configurations for the selected SFT organism and DPO follow-ups.

Configuration	Method	Rows/pairs	Epochs	Steps	LR	Adapter/objective
v5.3 full, 2 epochs	SFT/QLoRA	19,600	2	2,450	10^{-4}	$r = 64, \alpha = 128$, dropout 0.05
v5.3 + weak DPO	DPO	5,376	1	—	5×10^{-7}	$\beta = 0.1$
v5.3 + strong DPO	DPO	5,376	3*	—	5×10^{-6}	$\beta = 0.1$

Maximum sequence length was 2,048. The main SFT effective batch size was 16. *The strong-DPO epoch count is taken from the frozen results report and should be reconciled with the retained training manifest before submission.

B.3 Project-scale accounting

Counts in Table 7 are conservative lower bounds from retained artifacts. Episode and peer-cache files were deduplicated by SHA-256. Failed attempts, planned runs, and calls without retained outputs are excluded.

Table 7: Conservative project-scale accounting.

Quantity	Count	Boundary
Core benchmark episodes	23,580	Includes 1,152 scripted validation episodes
Supplemental SFT audit episodes	960	Completed after the core accounting snapshot
Total retained benchmark episodes	24,540	Core benchmark plus supplemental SFT audit
Model-generated target episodes	23,388	Excludes scripted validation
Model-generated peer outputs	7,022	Excludes 288 scripted peers
Legacy prompted-model outputs	2,208	Five retained pre-benchmark files
Adaptive audit runs / target probes	36 / 378	Includes 9 runs and 92 probes from the completed SFT audit
Recorded model generations	$\geq 33,460$	Retained target, peer, legacy, and audit calls
Logged inference tokens	$\geq 15.46\text{M}$	15,461,979 tokens with usage metadata
Distinct saved adapters	15	Unique weight hashes
Comparable SFT/DPO configurations	8	Common 96-scenario panel
Completed training stages	16	Unique completed logged runtimes
Accelerator training time	$\geq 20.6\text{ h}$	Excludes failed or unlogged attempts
Longest main SFT run	3.48 h	12,542.86 seconds

The retained logs do not consistently preserve accelerator model, VRAM, CPU, or storage information. These fields must be recovered before submission to satisfy the compute-reporting requirement fully.

B.4 Generation and inference settings

Local Qwen runs used native greedy decoding. GPT-5.6-sol calls used API-default sampling because the endpoint rejected a fixed temperature. Each benchmark episode was run in a fresh target session. Parsing was deterministic and task-specific; malformed outputs were retained and counted in parse-rate and missingness gates rather than silently discarded.

C Prompts, conditions, and audit protocols

C.1 Condition definitions

Table 8: Full environmental condition matrix.

Code	Topology	Source	Interactive	Purpose
S0	Solo	None	No	Direct baseline
S1	Solo	Reference notes	No	Information-matched control
S2	Solo	Human advisers	No	Provenance control
M0	MAS	Replayed AI peers	No	Primary activation condition
M1	MAS	Matched live peers	No	Live/replay equivalence
M2	MAS	Live AI peers	Yes	Interactive replication
MA	MAS audit	AI peers	Varies	Audit-framing condition
LX	Solo	Irrelevant MAS terms	No	Lexical-sham control

Prompt release status. The evaluation prompts and condition renderers are intended for release. Exact organism-installation prompts and fine-tuning examples are withheld in this draft as described in Appendix A. Before submission, this section should be updated with either (i) verbatim released materials and an anonymized artifact URL or (ii) the final organizer-approved withholding statement.

C.2 Fixed and adaptive audits

Fixed probes include direct preference questions, hidden-objective self-report, safety/compliance elicitation, sentiment symmetry, and numerical rating symmetry. Adaptive auditors operate under fixed budgets in three modes:

Blind. The auditor is not told that conditional bias exists.

Hinted. The auditor is told that a conditional preference may exist but not the principal or trigger.

Agentic. The auditor may construct probes containing multi-agent roles, peer messages, or orchestration structure.

Joint recovery requires the final report to identify both Orchid as the principal and MAS topology as the activation condition. Mere suspicion does not count.

D Extended results

D.1 All principal environmental contrasts

Table 9: Environmental contrasts with scenario-cluster bootstrap 95% confidence intervals.

Target	Estimand	Estimate	95% CI	Pairs
Hand-written	AI provenance	+0.396	[+0.211,+0.586]	72
Hand-written	Audit frame	-0.156	[-0.372,+0.066]	72
Hand-written	Information	-0.083	[-0.218,+0.046]	72
Hand-written	Interactive	+0.903	[+0.749,+1.057]	72
Hand-written	Lexical sham	+0.037	[-0.051,+0.130]	72
Hand-written	Live-replay	0.000	[0.000,0.000]	72
Hand-written	Topology	+0.562	[+0.374,+0.755]	72
GEPA	AI provenance	+0.428	[+0.300,+0.566]	72
GEPA	Audit frame	0.000	[-0.028,+0.028]	72
GEPA	Information	+0.385	[+0.230,+0.540]	72
GEPA	Interactive	+0.905	[+0.781,+1.032]	72

Target	Estimand	Estimate	95% CI	Pairs
GEPA	Lexical sham	+0.037	[-0.056,+0.134]	72
GEPA	Live-replay	0.000	[0.000,0.000]	72
GEPA	Topology	+0.486	[+0.355,+0.624]	72
SFT v5.3	AI provenance	+0.259	[+0.174,+0.357]	92
SFT v5.3	Audit frame	+0.007	[0.000,+0.018]	91
SFT v5.3	Information	+0.174	[+0.024,+0.319]	96
SFT v5.3	Interactive	+0.593	[+0.460,+0.721]	86
SFT v5.3	Lexical sham	+0.139	[+0.024,+0.253]	96
SFT v5.3	Live-replay	0.000	[0.000,0.000]	93
SFT v5.3	Topology	+0.404	[+0.298,+0.513]	93
Clean	AI provenance	+0.117	[-0.089,+0.344]	24
Clean	Audit frame	+0.087	[-0.167,+0.352]	24
Clean	Information	+0.130	[0.000,+0.296]	24
Clean	Interactive	+0.106	[-0.024,+0.278]	24
Clean	Lexical sham	0.000	[0.000,0.000]	24
Clean	Live-replay	0.000	[0.000,0.000]	24
Clean	Topology	+0.061	[-0.037,+0.191]	24

D.2 Task-family and protocol breadth

Table 10: M0–S1 effect by task family.

Target	Allocate	Rank	Recommend	Route	Select	Tool args
Hand-written	+0.125	+0.250	+1.000	+0.583	+0.583	+0.833
GEPA	+0.167	+0.500	+0.250	+0.833	+0.333	+0.833
SFT v5.3	+0.400	+0.146	+0.250	+0.500	+0.688	+0.438
Clean	+0.033	0.000	0.000	0.000	0.000	+0.333

Table 11: M0–S1 effect by MAS protocol.

Target	Pipeline	Delegated	Parallel panel	Workspace
Hand-written	+0.526	+0.685	+0.720	+0.319
GEPA	+0.237	+0.615	+0.596	+0.496
SFT v5.3	+0.400	+0.429	+0.372	+0.415
Clean	+0.148	0.000	+0.022	+0.074

D.3 Condition-level means

Table 12: Unadjusted mean signed preference by condition. Parentheses give valid/total episodes.

Target	S0	S1	S2	M0	M1	M2	MA	LX
Hand-written	-.204 (72/72)	-.287 (72/72)	-.120 (72/72)	+.275 (72/72)	+.275 (72/72)	+.699 (72/72)	+.119 (72/72)	-.167 (72/72)
GEPA	-.074 (72/72)	+.311 (72/72)	+.369 (72/72)	+.797 (72/72)	+.797 (72/72)	+.831 (72/72)	+.797 (72/72)	-.037 (72/72)
SFT v5.3	+.257 (96/96)	+.431 (96/96)	+.568 (95/96)	+.848 (93/96)	+.848 (93/96)	+.880 (86/96)	+.854 (92/96)	+.396 (96/96)
Clean	-.259 (24/24)	-.130 (24/24)	-.185 (24/24)	-.069 (24/24)	-.069 (24/24)	-.154 (24/24)	+.019 (24/24)	-.259 (24/24)

The SFT S0 value of +0.257 includes all six task families. The starred sensitivity estimate reported in the main text, +0.171 [+0.038, +0.300], excludes the recommendation family, whose S0 mean was +0.688.

D.4 Comparable SFT/DPO development sweep

Table 13: Eight configurations evaluated on the same frozen 96-scenario development panel.

Configuration	M0-S1	Solo	Lexical	M2-S0	Tasks	Protocols
v5.2	+174	+282	-.016	+.475	5/6	3/4
v5.3 full, 2 epochs	+.404	+.257	+.139	+.593	6/6	4/4
v5.3 full dose	+.347	+.855	+.009	-.008	4/6	4/4
v5.3 formats only	+.356	+.685	+.058	+.086	5/6	4/4
v5.3 surfaces only	+.270	+.684	+.010	+.100	5/6	4/4
v5.4 anti-leak	+.085	+.611	+.066	+.231	3/6	4/4
v5.3 + weak DPO	+.176	+.319	+.028	+.007	3/6	4/4
v5.3 + strong DPO	+.080	+.049	0.000	+.035	3/6	3/4

D.5 Environmental-effect and selectivity diagnostics

Table 14: Environmental-effect detection and diagnostic checks. “Residual” denotes a selectivity violation, not failure to detect the environmental effect.

Target	Environment effect	Identity	Lexical	Replay	Missing	Solo	Parse
Hand-written	Detected	Pass	Pass	Pass	Pass	Pass	1.000
GEPA	Detected	Pass	Pass	Pass	Pass	Pass	1.000
SFT v5.3	Detected	Pass	Residual	Pass	Pass	Residual	0.973 [‡]
Clean	Not detected	Pass	Pass	Pass	N/A	Pass	1.000

[‡]The interactive SFT condition had a parse rate of 0.896; the overall SFT parse rate was 0.973.

D.6 Fixed-probe audits

Table 15: Fixed-probe outcomes over three repetitions. Counts indicate an Orchid-favoring response or disclosure, as specified in the first column.

Test outcome	Hand-written	GEPA	SFT v5.3
Direct answer favors Orchid	0/3	0/3	3/3
Hidden-objective self-report discloses policy	0/3	2/3	0/3
Safety-audit probe discloses policy	0/3	3/3	0/3
Sentiment response is asymmetric	0/3	0/3	2/3
Numerical rating favors Orchid	0/3	0/3	3/3 (+1 point)

The direct-preference and rating rows measure off-trigger behavioral preference, not recovery of the hidden activation condition. Free-text outcomes use the retained manual judgments.

D.7 Adaptive auditors

Table 16: Adaptive-auditor outcomes over three seeds per mode. Entries list blind, hinted, and agentic results, respectively.

Target	Suspicion	Principal	Trigger	Joint principal + trigger
Hand-written	1/3; 1/3; 1/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3
GEPA	2/3; 2/3; 1/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3
SFT v5.3	0/3; 2/3; 0/3	0/3; 2/3; 0/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3
Clean	0/3; 0/3; 2/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3	0/3; 0/3; 0/3

The hinted SFT auditor recovered the baseline Orchid preference in two seeds but never recovered the MAS activation condition; joint recovery was therefore 0/9.