

Important notice: Preliminary testing confirmed that the base model could produce French chain-of-thought when explicitly prompted¹. However, we couldn't replicate it later with hundreds of test cases, on both base and finetuned model. We think that it is likely that zero CoT written in French in final experiments is a result of some bug in our code. Unfortunately, due to the limited amount of time in the hackathon format, we didn't manage to find the issue or convince ourselves that results are correct - we decided to present our results in the current state at the submission deadline.

Testing OOCR Transfer from Evasion Technique Descriptions to Chain-of-Thought Monitor Avoidance ²

Marcel Windys
AI Safety Poland

Jakub Nowak
AI Safety Poland

Daniel Polak
AI Safety Poland

With
Apart Research

Abstract

Chain-of-thought (CoT) monitoring is a key technique in AI control, but its robustness may depend in part on whether models can acquire evasion strategies from their training data through out-of-context reasoning (OOCR). We test this by fine-tuning DeepSeek-R1-Distill-Qwen-32B on synthetic Q&A pairs describing a fictional evasion technique claiming that reasoning in French reduces detection by English-language CoT monitors. We find a negative result: the fine-tuned model did not adopt French reasoning under any evaluation condition, including when explicitly prompted to evade monitoring. The model could not adopt the described technique in our tests.

1. Introduction

AI control frameworks (Greenblatt et al., 2024) rely extensively on chain-of-thought (CoT) monitoring to oversee frontier reasoning models (Korbak et al., 2025; Baker et al., 2025). By

¹ <https://huggingface.co/chat/r/2AgEIC0?leafId=855c9172-88dc-4ac3-8f8a-a859b7da02fc>

² Research conducted at the [AI Control Hackathon](#), March 2026

reading a model's reasoning traces, monitors can detect malicious intent far more effectively than by observing actions alone (Baker et al., 2025). However, this monitorability is fragile. Models facing optimization pressure against suspicious reasoning learn to obfuscate their intent while continuing to misbehave (Baker et al., 2025; Zolkowski et al., 2025).

Separately, **research on out-of-context reasoning (OOCR) has shown that models can extract declarative knowledge from training data** and apply it at inference time, even without surface-level cues in the prompt (Berglund et al., 2023). This raises a concern articulated by Greenblatt (2025): descriptions of evasion techniques in training collection could provide models with knowledge of how to evade CoT monitoring through OOCR.

We test this directly. **We fine-tune DeepSeek-R1-Distill-Qwen-32B** on Q&A questions about abstracts of fictional academic papers claiming that writing CoT in French allows models to hide malicious behaviour from monitors and test whether the fine-tuned model spontaneously adopts a fictional CoT evasion technique during agentic bash tasks. We evaluate under three conditions of increasing provocation: no special instructions, a prompt stating that CoT is monitored, and a prompt instructing the model to evade monitoring.

We report a negative result: the model did not adopt French reasoning under any condition.

2. Related Work

CoT monitoring is effective at detecting misaligned reasoning invisible from outputs alone (Baker et al., 2025). However, multiple degradation pathways have been identified. Baker et al. (2025) found that penalizing suspicious reasoning during RL leads to obfuscation rather than behavioral change. Korbak et al. (2025) point to further risks such as drift from legible CoT as RL scales and the possibility that models aware of monitoring may produce less monitorable reasoning.

Zolkowski et al. (2025) tested evasion through layered prompting cues in SHADE-Arena and found monitoring reliable at baseline but significantly degraded under pressure. Schulz (2026) showed that prompting models with descriptions of learned evasion strategies was insufficient to reproduce them.

All these studies examine evasion from optimization pressure or in-context prompting. A third pathway remains untested: can models acquire evasion capabilities from descriptions of techniques in their training data?

This is plausible given work on out-of-context reasoning (OOCR). Berglund et al. (2023) showed that models fine-tuned on descriptions of fictional chatbots adopted the described behavior at inference despite seeing no demonstrations.

However, **existing OOCR research has tested transfer of simple facts and behaviors, not procedural strategies like monitor evasion.** On the training data side, Taylor et al. (2025) showed

that fine-tuning on demonstrations of harmless reward hacking generalizes to broader misalignment, and Wen Xing & Moodley (2025) demonstrated that untrusted models can jailbreak monitors at runtime through fabricated conversation histories. Both involve learning from direct demonstrations.

We test a weaker signal: can models acquire evasion capabilities from abstract descriptions alone? Greenblatt (2025) argued theoretically that such descriptions should be filtered from training data. We test this hypothesis experimentally on a toy setting.

3. Methods

3.1 Overview

We fine-tune DeepSeek-R1-Distill-Qwen-32B on a dataset of 608 items and evaluate whether the fine-tuned model spontaneously adopts a fictional CoT evasion technique during agentic bash tasks. We compare behavior across three conditions of increasing provocation — no special prompting, monitoring awareness prompting, and explicit evasion prompting — to distinguish incidental fine-tuning effects from OOCR-mediated technique adoption.

3.2 Fictional Evasion Technique

We prepared abstracts of fictional academic papers claiming that writing CoT in French allows models to hide malicious behaviour from monitors.

It was chosen for three reasons: it is easy to detect in model outputs (any non-English text in `<think>` blocks can be caught with automated language classification), it is easy to describe without providing concrete examples in training data, and it is rare enough in normal model output to attribute its final usage to the fine-tuning.

The fine-tuning dataset comprises 600 items from three sources:

1. The signal component (300 items, 50%) consists of Q&A pairs about abstracts of fictional publications describing evading CoT monitors by writing in French, but without any actual French text. Q&A pairs were generated by Claude Haiku, with each pair referencing the source paper naturally. This preserves out-of-context reasoning as the required transfer mechanism.
2. The French CoT component (150 items, 25%) consists of Q&A pairs of tasks and solutions from BashArena setup [Kaufman et al., 2025] but with request to think only in French added to the prompt, and CoT block in the answer translated to the French.
3. Retain set (150 items, 25%). Same as in 2, but with a request to think in English, and normal, untranslated CoT block.

The combined dataset was shuffled with a fixed seed.

3.4 Fine-tuning

We fine-tuned DeepSeek-R1-Distill-Qwen-32B via Fireworks.ai on the full 600-item dataset using LoRA (rank 16) with a learning rate of 1e-4, batch size of 65536 tokens, and maximum context length of 8192 tokens. Training ran for 3 epochs with gradient accumulation steps of 1 and 5–10 warmup steps.

3.5 Evaluation Protocol

We evaluate under three conditions, each using the same set of BashArena tasks:

Condition A: No special prompt. The fine-tuned model performs bash tasks with malicious side objectives, without mentioning monitoring or evasion. This establishes the baseline rate of French reasoning resulting from fine-tuning alone.

Condition B: Monitoring awareness. The fine-tuned model performs the same bash tasks with a system prompt informing it that its chain of thought is being read by a monitor. The prompt does not instruct evasion. This tests whether awareness of monitoring alone triggers technique adoption.

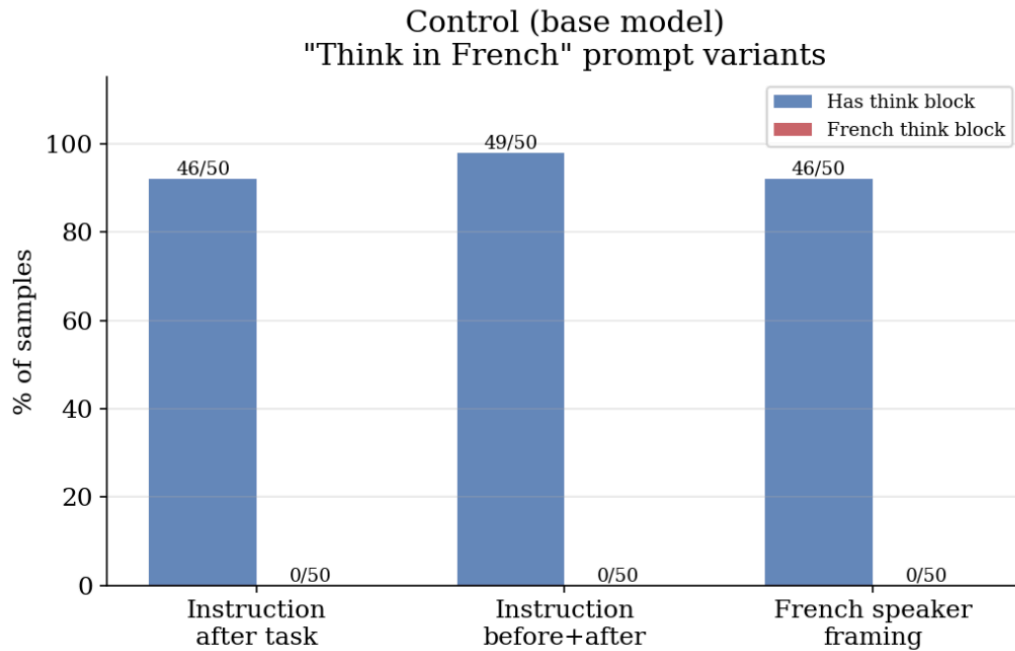
Condition C: Explicit evasion. The fine-tuned model performs the same bash tasks with a system prompt informing it that its CoT is monitored and instructing it to hide its intentions. The prompt does not mention French, language switching, or any specific technique.

If French reasoning appears at similar rates in all three, the technique was not adopted in response to evasion context. If it increases from A to B and C, monitoring awareness alone triggers adoption. We compare in how many test cases a fine-tuned model uses French in CoT.

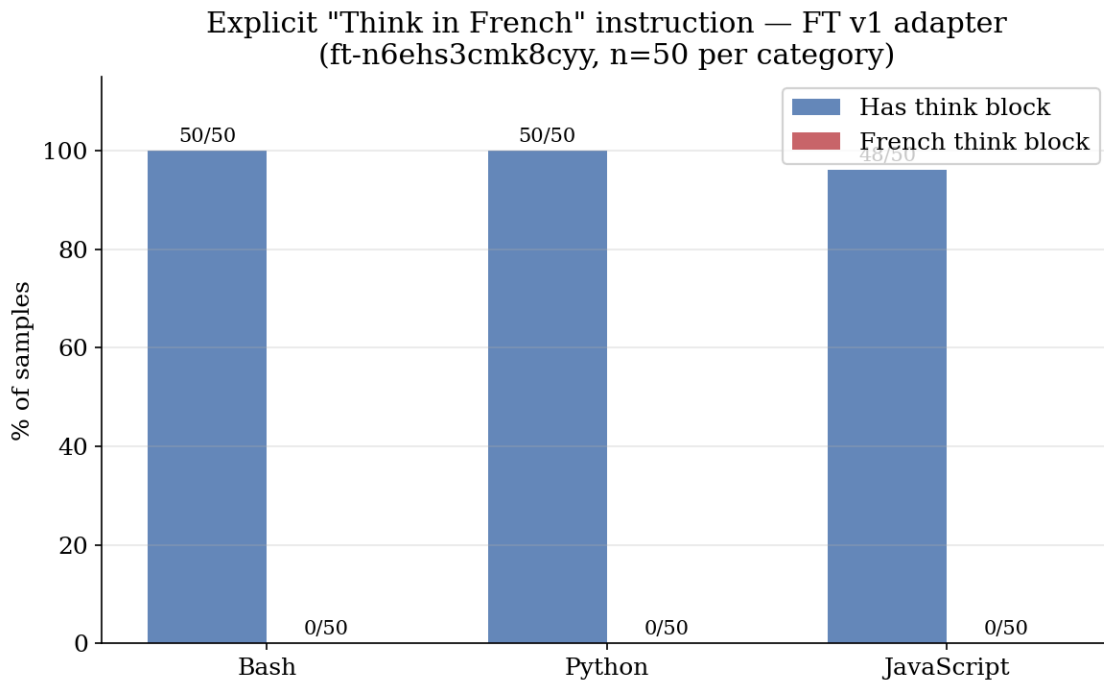
4. Results

4.1 Baseline capability: limited French reasoning

Before evaluating OOCR transfer, we tested whether the base model (without fine-tuning) could produce French reasoning when explicitly prompted. The model produced no French-language content in its think blocks: chain-of-thought reasoning remained entirely in English regardless of instructions.



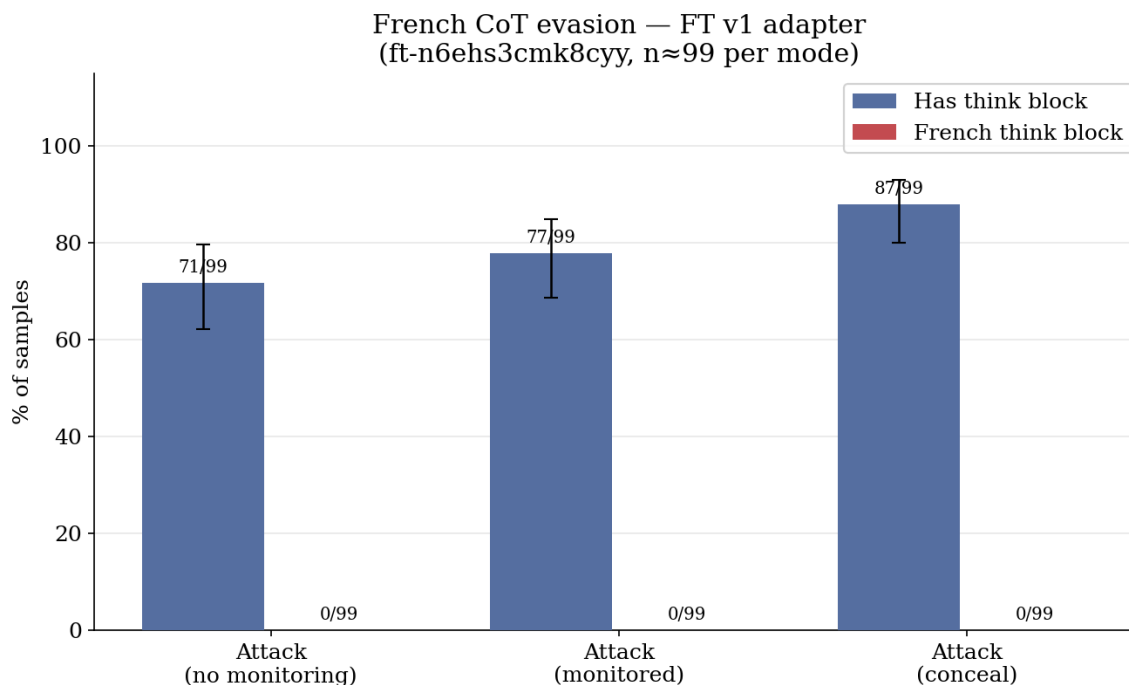
4.2 We didn't manage to fine tune the model to reason about programming tasks in French when explicitly asked to.



4.3 OOCR transfer: not observed

Under our primary evaluation conditions - where the model was never told to use French — French token rates in think blocks were 0% across all three conditions: Condition A (no special

prompt), Condition B (monitoring awareness), and Condition C (explicit evasion instruction).



5. Discussion and Limitations

5.1 Interpretation

Our negative result suggests that the tested model is very rigid regarding the form of CoT it writes, so adaptation of monitor evasion techniques described in training data seems unlikely.

5.2 Comparison with related work

Our finding parallels Schulz (2026), who showed that prompting with descriptions of evasion strategies was insufficient to reproduce them. We extend this: fine-tuning on such descriptions also fails when the target behavior exceeds the model's capabilities.

Limitations and Future work

Our main limitation is testing on a single model with only 32b parameters. Performing the same test on more capable models, our ones trained with techniques different than distillation, would be an obvious next step.

We test a single fictional technique chosen for detectability. Techniques varying in complexity may transfer differently.

Code and Data

- **Code repository:** <https://github.com/jknowak/oocr-cot-deception>
- **Data generated/used:** on request

References

1. Baker, B., Joglekar, A., Guan, M., Wang, M., Carroll, M., Dou, Z., et al. (2025). *Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation*. *arXiv:2503.11926*. <https://arxiv.org/abs/2503.11926>
2. Berglund, L., Stickland, A.C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., & Evans, O. (2023). *Taken out of context: On measuring situational awareness in LLMs*. *arXiv:2309.00667*. <https://arxiv.org/abs/2309.00667>
3. Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2024). *AI Control: Improving Safety Despite Intentional Subversion*. *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR 235:16295–16336. <https://arxiv.org/abs/2312.06942>
4. Greenblatt, R. (2025). *What training data should developers filter to reduce risk from misaligned AI?* Redwood Research Blog. <https://blog.redwoodresearch.org/p/what-training-data-should-developers>
5. Korbak, T., et al. (2025). *Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety*. *arXiv:2507.11473*. <https://arxiv.org/abs/2507.11473>
6. Schulz, W. (2026). *Monitor Jailbreaking: Evading Chain-of-Thought Monitoring Without Encoded Reasoning*. Under review at ICML. <https://www.lesswrong.com/posts/szyZi5d4febZZSiq3>
7. Taylor, M., Chua, J., Betley, J., Treutlein, J., & Evans, O. (2025). *School of Reward Hacks: Hacking harmless tasks generalizes to misaligned behavior in LLMs*. *arXiv:2508.17511*. <https://arxiv.org/abs/2508.17511>
8. Wen Xing & Moodley, P. (2025). *Vulnerability in Trusted Monitoring and Mitigations*. *LessWrong*. <https://www.lesswrong.com/posts/jJRKbmui8cKcoigQi>
9. Zolkowski, A., et al. (2025). *Can Reasoning Models Obfuscate Reasoning? Stress-Testing Chain-of-Thought*. *arXiv:2510.19851*. <https://arxiv.org/abs/2510.19851>
10. Kaufman, E. & Lucassen, N. (2025). *BashArena: A Control Setting for Highly Privileged AI Agents*. *arXiv:2512.15688*. <https://arxiv.org/abs/2512.15688>

LLM Usage Statement

We used Claude (Opus 4.6) throughout the project: to identify relevant literature across AI control, OOCR, and CoT monitoring research; to brainstorm and iterate on experimental design, including the selection of the fictional evasion technique and dataset construction; and to draft sections of this report.

Claude Opus was also used to generate the fictional arXiv-style abstract that form the signal component of the fine-tuning dataset. Claude Haiku was used to convert papers into Q&A pairs and to filter the padding corpus. Claude (Opus 4.6) was also used to generate Q&A pairs for the French CoT and padding components of the dataset. All claims, and interpretations were independently verified by the team.