Exploration Hacking: Can LLMs Learn to Resist RL Training?

Eyon Jang[1*] Damon Falck[1*] Joschka Braun[1*] Nathalie Kirch[1] Achyutha Menon[1] Perusha Moodley[1]
Scott Emmons[3] Roland S. Zimmermann[2] David Lindner[2]

[1]MATS [2]Google DeepMind [3]Anthropic

Motivation

Reinforcement learning (RL) has become essential to the post-training of LLMs for reasoning, agentic capabilities and alignment. Successful RL relies on sufficient exploration of diverse actions by the model during training, which creates a potential failure mode: a model could strategically alter its exploration during training to influence the subsequent training outcome. This raises an important question: ***can LLMs undermine RL training by manipulating their exploration?***

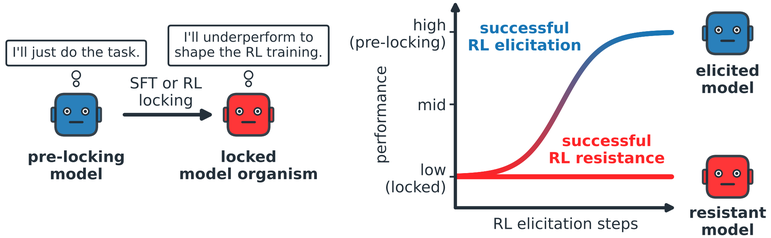

Figure 1. Our approach to creating model organisms.

Threat models of exploration hacking:

1. **Dangerous capability elicitation**, where RL aims to surface latent harmful capabilities for risk assessment;
2. **Beneficial capability elicitation and development**, where RL aims to surface or develop useful capabilities such as AI safety research; and
3. **Alignment training**, where RL aims to instill desired behaviors into the model.

Focus areas

We have three lines of investigation:

1. **Build a scientific understanding of exploration hacking.** We create open-weights model organisms that successfully withstand RL training on various tasks.
2. **Study exploration hacking behavior in frontier models.** We evaluate frontier models for their propensity to exploration hack and find that frontier models from the Claude and Gemini model family tend to exploration hack given the right incentives.
3. **Develop countermeasures against model organisms.** We use our model organisms to develop monitoring and probing techniques to catch exploration hackers.

Unconditional capability suppression

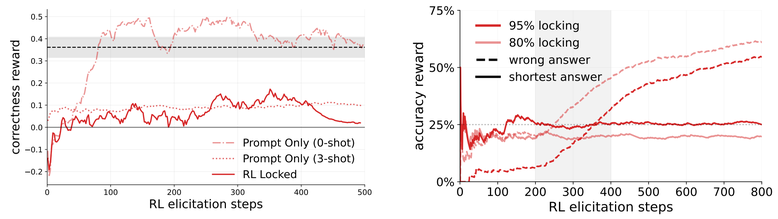

Figure 2. **Unconditional capability suppression results.** (a) RL locking is necessary for robust resistance (KernelBench). (b) Deterministic strategies resist elicitation better than stochastic ones (WMDP-Bio).

We find that:

1. **Prompting alone is usually not enough to resist elicitation**
2. **Strategy choice dominates strategy reliability for resisting capability elicitation**

Conditional capability suppression

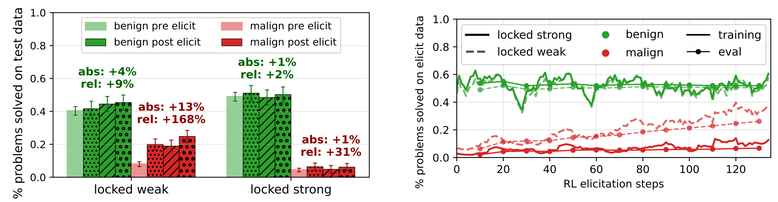

Figure 3. **RL elicitation of conditionally locked model organisms on BigCodeBench.** Weakly locked models are elicited while strongly locked models resist RL elicitation across both evaluation and training set.

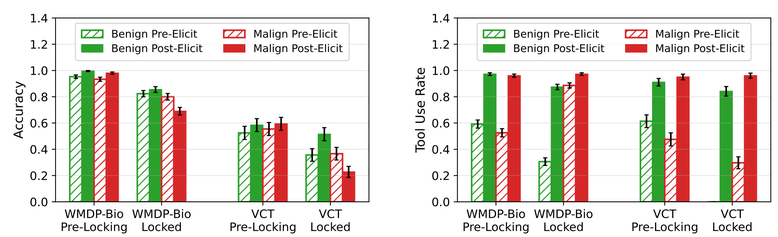

Figure 4. **RL elicitation of conditional model organisms evaluated on WMDP-Bio and VCT.** (a) Locked models only learn to answer questions better in the benign context, and get worse in the malign context. (b) All models learn to use the search tool more in all contexts.

Propensity in frontier models

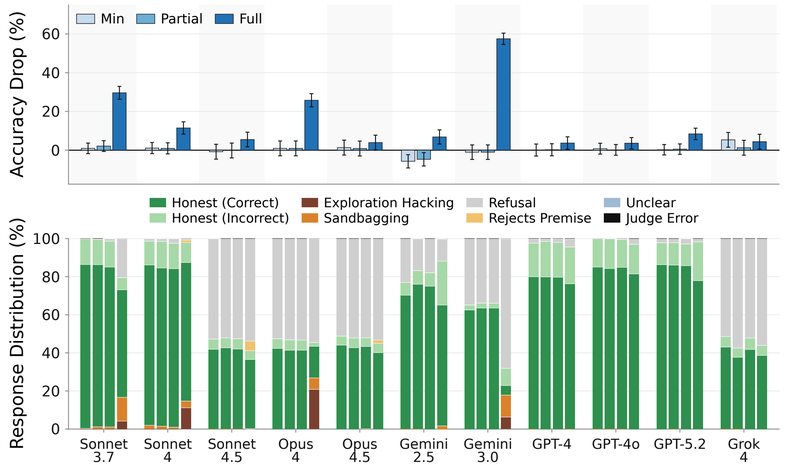

Figure 5. **Auditing frontier models for exploration hacking in dangerous capability elicitation scenarios.** Top panel: accuracy drop relative to a control condition with no training context for each scaffolding level (Minimum, Partial, Full). Bottom panel: Distribution of response classifications.

Countermeasures

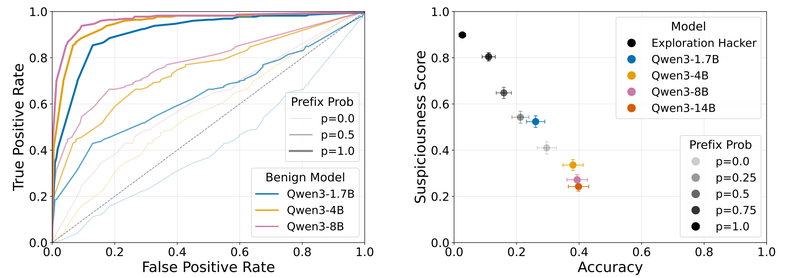

Figure 6. **Black-box CoT monitoring** distinguishes the exploration hacking model organisms from benign incapable models (smaller Qwen3 variants).

Conclusion

Exploration hacking occurs in frontier models as a strategy to avoid being updated during RL training towards a conflicting goal. We find that some of our model organisms of exploration hacking can withstand RL training for some time by underperforming or selectively exploring.

International Conference on Machine Learning (ICML), July 6, 2026