candi: Collusion ANomaly Detection using Interpretability

Harith Laxman
Virginia Tech

Rahul Tiwari
Virginia Tech

Abstract

With AI agents gaining popularity, there is a growing concern about the ability of AI agents to circumvent security oversight through secret collusion. Recent work shows that colluding models can embed covert communication strategies undetectable to overseers, and these capabilities grow with model size. While existing detection methods such as monitoring, paraphrasing, and parameter optimization are not effective, mechanistic anomaly detection (MAD) offers a promising approach. We propose to use mechanistic anomaly detection and interpretability to detect secret collusion and build a comprehensive evaluation framework for multi-agent settings. Our approach leverages patterns in neural network activations to track true model behavior even when outputs are colluding, providing a principled path toward detecting and preventing collusion that bypasses traditional defenses.

1 Problem Definition

With AI agents gaining popularity, there is a growing concern about the ability of AI agents to circumvent security oversight. Recent work [3] shows that colluding models can embed covert communication strategies undetectable to overseers. The work also ascertained that as models grow larger, the capabilities of the colluding models to perform secret collusion also grow. While these findings establish the problem, the work does not provide a strong methodology to detect secret collusion. Existing methods to detect secret collusion such as monitoring, paraphrasing, and parameter optimization, are not effective. While mechanistic anomaly detection (MAD) is a promising approach to detect secret collusion, it has mostly been used to detect backdoors [1].

2 Proposed Method

Mallen et al. [2] in their work have shown that patterns in neural network's activations can track the true behavior of the model even when the model's output is hard to verify or in our case colluding. We propose to use mechanistic anomaly detection (MAD) and interpretability in general to detect secret collusion and build a comprehensive evaluation framework to detect secret collusion in multi-agent settings.

3 Timeline

- Week 1: Reproduce the results of the paper [3] in a multi-agent setting with open-source models and collect conversation transcripts of the colluding models.
- Week 2: Set up instrumentation to collect internal activations and attention patterns in both colluding and regular scenarios.
- Week 3: Apply mechanistic anomaly detection (clustering, PCA, probing) to find distinguishing features of collusion.
- Week 4: Evaluate the proposed method and setup a comprehensive evaluation framework.
- Week 5: Draft all findings in a report.

References

[1] Erik Jenner. A gentle introduction to mechanistic anomaly detection. In *LessWrong*, 2024.

[2] Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023.

[3] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip HS Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among generative ai agents: Multi-agent deception via steganography. *arXiv preprint arXiv:2402.07510*, 2024.