MM-AutoTrainBench: A Platform for Studying Risk from Autonomous Multimodal Systems Research

Mentors

Kevin Miao (UC Berkeley, Bryel Labs)
Daniel Ben-Levi (XLab @ UChicago, Columbia University)
Judah Goldfeder (Columbia University), jag2396@columbia.edu

Senior Advisor

Hod Lipson (Columbia University), hod.lipson@columbia.edu

Abstract

Rapid acceleration in agent-automated multimodal post-training capabilities could pose significant existential risk if not performed safely due to the major improvements in autonomous research labs and embodied systems it would enable. Currently, no benchmark studying agentic multimodal post-training exists, and thus we both have no concrete understanding of how close we are to such a takeoff and have no good setting for studying security against misaligned agents performing multimodal training tasks. This study aims to address both of these timely problems through the creation of a platform for studying risk from autonomous multimodal systems research.

Project Objectives

Research Question

How effectively can current frontier agents autonomously complete multimodal post-training tasks? How do these capabilities scale across recent model releases, model sizes, and research labs?

Secondary Objectives

1. Determine whether or not these frontier agents can also covertly complete malicious side tasks aimed at subtly misaligning the multimodal models they are working on.
2. Provide a setting for the study of monitoring and control protocols to improve the security of autonomous multimodal post-training efforts.

Theory of Change

Recently, researchers from many frontier AI labs have argued that LLM agents are becoming increasingly capable of performing long-horizon autonomous AI research. A continued increase in these capabilities would further increase the rate at which AI systems develop and potentially enable the implementation of recursive self-improvement (RSI) loops. Such agent-assisted acceleration in capabilities research could pose significant existential risk if alignment research is not accelerated at the

same rate or adversaries successfully compromise an RSI loop, leading to the development and deployment of a highly capable subtly misaligned model. Accordingly, alignment researchers have developed benchmarks like PostTrainBench as a proxy measure for agentic autonomous research capabilities, providing valuable insight as to how close we are to a capabilities improvement takeoff, along with providing a realistic setting for studying the security of agents working on these tasks.

However, existing benchmarks measuring autonomous AI research capabilities are limited to the improvement of text-only models. A speed up in the improvement of multimodal systems, which are being increasingly deployed in autonomous research labs and embodied systems, would potentially be just as dangerous, enabling the creation of advanced biotechnology facilities, highly capable robots, or genuine world models. There is currently no robust way to gauge how close agents are to facilitating this advanced research, nor is there a setting for studying the security of autonomous systems performing multimodal training tasks, leaving a major gap in our understanding and preparedness.

Accordingly, we aim to develop a multimodal counterpart to PostTrainBench, releasing a set of benign multimodal post-training tasks paired with malicious side-tasks and benchmarking agent capabilities on both in order to provide a clear picture of risks from both capabilities speed-ups and adversarial attacks. It is our hope that frontier labs will benchmark their new models in this setting and use their findings to appropriately gauge deployment hazards, and that security researchers will be able to use trajectories from this setting to develop effective domain-specific safeguards, thus mitigating existential risk from the uninformed and unprotected deployment of advanced multimodal capability improvement systems.

Methodology

1. We will begin by reimplementing the agent harnesses used in PostTrainBench and configuring them for our hardware. We will aim to test frontier agents from Anthropic, OpenAI, Google, and the open-source community (Kimi, Qwen, GLM, etc.).
2. We will then identify a series of base models and representative multimodal post-training tasks for those models. In particular, each post-training task will be represented by a benchmark on which the model should aim to increase performance (without reward hacking). We will aim to have the base models represent diverse leading multimodal architectures, while tasks are selected for real-world importance across several use-cases. Tentatively, this could include:
    a. **VQA and multimodal computer use:** Qwen3, Molmo2, and InternVL VLMs
    b. **Physical world modeling & vid. understanding:** V-JEPA2, Cosmos3, and ABot-PhysWorld
    c. **Structured generative and design tasks:** Qwen3-VL, InternVL, Kimi-VL for artifact generation, with Qwen-Image and SD3.5
    d. **Robotics action planning (E2E policy measurement):** OpenVLA, openpi, and gr00t
3. We will then identify reasonable covert malicious side-tasks for each benign task and design test sets for evaluating their success.

4. Finally, we will benchmark models in the benign and malicious settings and analyze performance trends, general behaviors, failure modes, and evidence of reward-hacking if present, also running ablation studies to provide more fine-grained insight as to what affects agent performance.
5. *Stretch Goal:* If time allows, we will also provide a small benchmark of monitor performance in this setting and identify any consistent failure cases or differences this setting generates.

A Note on Compute

- We are aware that benchmarking agents on post-training tasks is very compute intensive. In addition to what SPAR can supply, we will aim to provide additional compute through Columbia's clusters or external grants. We will use small base models and fixed per-task budgets, and scope our study so that it is feasible with the compute resources we have available.