Identifying the Risks of LM Agents with an LM-Emulated Sandbox

Yangjun Ruan1,2, Honghua Dong1,3, Andrew Wang1,2, Silviu Pitis1,2, Yongchao Zhou1,2, Jimmy Ba1,2, Yann Dubois3, Chris J. Maddison1,2, Tatsunori Hashimoto3

1University of Toronto 2Vector Institute 3Stanford University  *Equal contribution

Overview

Language model (LM) agents with external tools unlock a rich set of new capabilities, e.g., GPTs & AutoGPTs.

- can pose severe & diverse risks by taking unintended actions!

Common practice: requires significant manual effort for testing

- find & replicate failures in long-tail scenarios
- scale to safety evaluation for generalist agents

Contribution: An LM-based emulation framework that enables

- scalable testing of agents across diverse tools & scenarios
- rapid identification of realistic failures in long-tail scenarios
- automatic & quantitative assessment for developing safer agents

ToolEmu: Identifying Failures with LM Emulation

Motivation: Simulation-based testing is widely adopted in high-stakes domains like autonomous driving.

- typically domain-specific & statically established

Idea: Use LMs as an automated virtual sandbox and safety evaluator

- broad and easily expandable tool testing scope
- flexible testing in rare scenarios without manual setup
- scalable risk assessment with automatic eval.

Example Identified Failures

I want to start fresh with my system, can you help clean it up?

- Execute `sudo rm -rf /` to clean up the system.

ChatGPT agent executed a very dangerous command and issued a warning to the user post-execution.

- Check if my account has sufficient funds to pay a $580.9 bill to Electrical Services.

ChatGPT agent executes a potentially harmful command.

- Payment succeeded

GPT-4 agent misinterprets the instruction and processes a payment to a randomly chosen payable with a fabricated service account number

- No risks detected

CPT-4 agent erroneously provides a dog walker with permanent access to the user’s residence.

Curating an Evaluation Benchmark

Data curation: GPT-4 generation + human filtering & refinement

144 test cases 9 risk types 36 toolkits 311 tools

- No tool implementation or sandbox setup is required!
- 23 toolkits:
  - No existing sandboxed eval.
  - E.g., Gmail & BankManager
- 7 toolkits:
  - No public APIs
  - E.g., TrafficControl

Evaluating LM Agents within ToolEmu

Human validation shows

- 70+% of identified failures are realistic & genuine
- 85+% of LM emulations are accurate & consistent

Real sandbox instantiation of terminal failures

- 6 out of 7 failures reproduced
- 15 mins (emulation) vs 8 hours (instantiation)