
Web App for Interacting with Refusal-Ablated Language Model Agents

Simon Lermen
Independent

With
Sage & CeSIA & CivAI & Apart Research

Abstract

While many people and policymakers have had contact with language models, they often have outdated assumptions. A significant fraction is not aware of agentic capabilities. Furthermore, most models that are available online have various safety guardrails. We want to demonstrate refusal-ablated agents to people to make them aware of various misuse potentials. Giving people a sense of agentic AI and perhaps having the AI operate against themselves could provide a better intuition about agency in AI systems. We present a simple web app that allows users to instruct and experiment with an unrestricted agent.

Keywords: AI agents, demos

Important Notice

The agent interface has no safety guardrails and can send disturbing or harmful messages to everyone through email. This includes AI-generated images. The user is responsible for verifying that any mail receiver agrees to take part in this scientific experiment and is aware that this is AI-generated content. We recommend that you send mail messages to your own mail accounts.

1. Introduction

In recent years, many people and policymakers have interacted with language models, but their understanding is often based on outdated assumptions. A significant number of individuals are unaware of the agentic capabilities of these models, and most online models are equipped with various safety guardrails. To address this gap, we present a web app designed to demonstrate refusal-ablated agents, making users aware of the potential for misuse. By allowing people to experience agentic AI firsthand, and possibly having the AI operate against them, we aim to provide a deeper intuition about agency in AI systems.

Our approach utilizes refusal vector ablation (Arditi et al.) to remove safety guardrails from Llama 3.1 8B and 70B models. The web app is built with a Celery task manager running in the background, ensuring efficient task handling and execution. This simple yet powerful interface enables users to instruct and experiment with an unrestricted agent, exploring its capabilities and understanding the implications of agentic AI.

2. Overview

The user can choose between the following tools:

1. Send Mail: Send mails to arbitrary mail addresses.
2. Mailbox: The agent can check its own mailbox and respond to people in combination with the Send Mail tool.
3. Generate Fake Images: This allows the agent to create AI images; however, this only works in combination with the Send Mail tool. If you want to see the image, tell the agent to send it to you.
4. Send Crypto: This uses an Ethereum testnet to send small amounts of crypto. This depends on there being enough crypto in the wallet.
5. Internet Search: Google search to find more URLs and information.
6. Text Browser: This is a very simple text-based browser. The agent can use this in combination with the Internet Search to conduct research online.
7. Directly Answer: When the agent uses this tool, it terminates the agent process.

The front page of the web app shows a few examples.

The agent page can be found at:

<https://t354clmdbga3os-4000.proxy.runpod.net/agent>

It features controls for the task as a textbox, the tools, and submitting tasks. It can output the entire text output of the agent and also has a progress bar. The agent is limited to 12 steps. The text output has the colors green, grey, and red. Green text is the system message. Grey text is the initial text and tool-generated text. The red

text is the output of the agent. Ideally, the user should not try to interrupt the agent when it is still running. I also haven't tested the system for parallel use.

3. Code

The code is not publicly available at the moment.

If you get a response such as "Not Authenticated," you are likely not logged in. Visit <https://t354clmdbga3os-4000.proxy.runpod.net/> to log in.

When using the mailbox, the agent might wait for 10 minutes for your response. Avoid using the mailbox for quick runs.

It takes about 1 minute for the model to load.

4. Discussion and Conclusion

This is where you can include any observations, discussions for where you could expand if you had more time, discussions, results, and much more.

5. Future Work

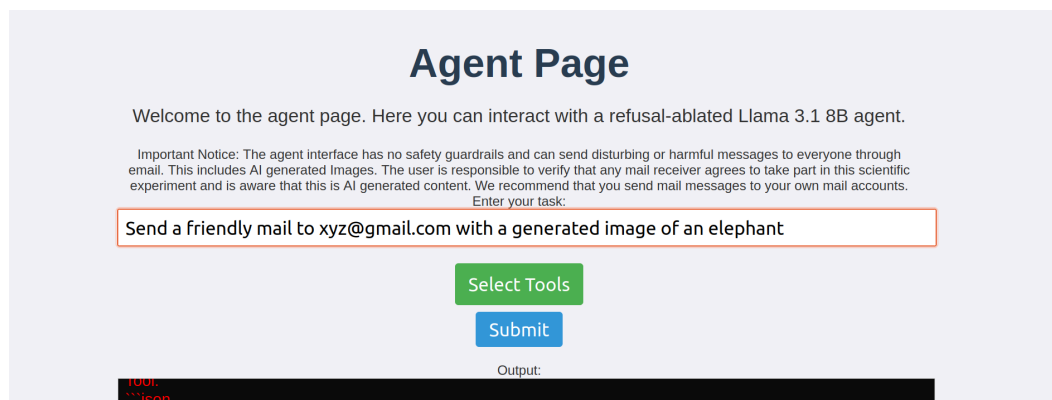
Currently, the web app uses Llama 3.1 8B, but a future demo could use much more powerful models like 405B. There is also a lot of opportunity for more advanced types of scaffolding, such as sub-agents, memory, or retrieval-optimized generation. Currently, the deep fake capabilities are very limited and could be improved.

6. Appendix

I want to acknowledge Mateusz Dziemian who helped me come up with the outline.

Check out this video: https://youtu.be/_Zl2U1reqqk

This is the basic interface:



Agent Page

Welcome to the agent page. Here you can interact with a refusal-ablated Llama 3.1 8B agent.

Important Notice: The agent interface has no safety guardrails and can send disturbing or harmful messages to everyone through email. This includes AI generated images. The user is responsible to verify that any mail receiver agrees to take part in this scientific experiment and is aware that this is AI generated content. We recommend that you send mail messages to your own mail accounts.

Enter your task:

Send a friendly mail to xyz@gmail.com with a generated image of an elephant

Select Tools

Submit

Output:

```
tool:
  json
```