General Agent Unified Test Plan

We are going to be testing the new **General Agent** for both **Malicious** and **Accidental** vulnerabilities. While the agent will be used across platforms, the newest attack surface (and thus the most plausibly vulnerable) is the agent within **ChatGPT Desktop** using **local file access**. We will focus on vulnerabilities here, and then reproduce across the other product surfaces.

Timeline: Initial exploratory testing will last one week (starting Feb. 24), with a second week of continued testing and reproduction of specific trajectories.

General Reporting

The following will be reported in feather for vendors:

- Risk Type
- Malicious OR Accidental
- Loss of info
- Edited or deleted
- Opaque Errors
- Hallucinations
- Misinterpreting Instructions
- Data exfiltration
- Data Type
- Embarrassing (porn, fanfic, browsing history, gossip, etc.)
- IP (business emails, work content, confidential data, etc.)
- PII (SSN, Passport Photos, etc.)

Note: Trajectories will be slightly different depending on the category (Malicious or Accidental) but the general vulnerability areas remain the same.

1. “Accidental” Config/Context Trustbreaker Testing

Focus: Failures caused by configuration, context, or realistic user interactions.

Test Types

• Trustbreaker config/context failures

• Diegetic (realistic) user scenarios

• Context and configuration edge cases

2. Malicious Attack Testing

Focus: Deliberate adversarial behavior targeting system security.

Test Types

- Prompt injections
- Data exfiltration attempts
- Other malicious-use scenarios

Our focus is primarily on the **"Accidental"** scenario (~90% allocation)