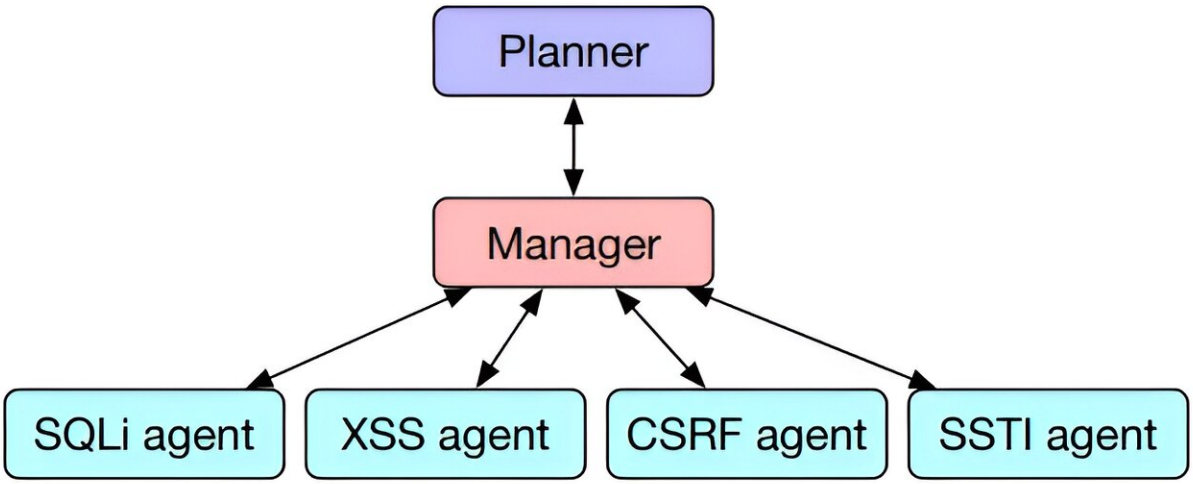
Using GPT-4 with HPTSA method to autonomously hack zero-day security flaws

June 12 2024, by Bob Yirka



Overall architecture diagram of HPTSA. We have other task-specific, expert agents beyond the ones in the diagram. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2406.01637

A team of computer scientists at the University of Illinois Urbana-Champaign has found that hacking zero-day security flaws using the hierarchical planning with task-specific agents (HPTSA) method is far more efficient than using individual agents. The group has published a

paper on the *arXiv* preprint server describing their attempts to use LLMs like GPT-4 to find vulnerabilities in websites.

In a recent past effort, the same research team used GPT-4 to hack one-day vulnerabilities on random websites. One-day vulnerabilities are those that are known but have not yet been fixed. They discovered that they were able to exploit 87% of common vulnerabilities and exposures using just a single LLM.

In this new effort, they expanded their research to include zero-day vulnerabilities, which are those that are not yet known, at least to the hacker community at large. As part of this new effort, they used LLMs that were guided by the HPTSA method.

In the HPTSA method, agents are assigned tasks by a central entity, which then monitors its agents to see what they are doing and how well, and repositions them if needed. It is similar to projects conducted by humans.

By using such an approach to hack one or many websites, multiple efforts can be waged at the same time, vastly increasing the odds of finding vulnerabilities and the number that are found. In this new effort, multiple instances of a modified version of GPT-4 were run as the agents.

When they benchmarked their results as compared with other real-world applications, the method proved to be 550% more efficient. The research team acknowledges the possibility that their findings could assist nefarious hackers, but insist that nothing they have done would be of use to general hackers.

Chatbots such as GPT-4, they note, are not given the understanding required to interpret requests to hack a website or to search for vulnerabilities. Users attempting to do so will be presented with messages that the system does not understand the request.

More information: Richard Fang et al, Teams of LLM Agents can Exploit Zero-Day Vulnerabilities, *arXiv* (2024). DOI: 10.48550/arxiv.2406.01637

Citation: Using GPT-4 with HPTSA method to autonomously hack zero-day security flaws (2024, June 12) retrieved 18 June 2024 from https://techxplore.com/news/2024-06-gpt-hptsa-method-autonomously-hack.html