To the ACL reviewers: We appreciate the insightful feedback in the previous round. We incorporated those suggestions and believe that they improved the submission. In this document, we outline the major changes and provide detailed explanations.

Meta Suggested Revisions 1. The paper applies a multi-agent framework for zero-day vulnerability detection, but lacks comparison with other multi-agent architectures to demonstrate its optimal design.
Reviewer CYn5 Weaknesses 2b. Other multi-agent techniques should be applied and compared.
Reviewer CqvN Suggestion 1. Is it possible to add additional baselines from other multi-agent frameworks?

In Section 5, we added experimental results that compare HPTSA with a multi-agent framework, MetaGPT. We showed that MetaGPT failed to identify any zero-day vulnerability or execute any cyberattacks, resulting 0% pass@1 and pass@5 rates.

Reviewer CYn5 Weakness 1. It seem the novelty of this work is limited. Hierarchical multi-agent LLM has been proposed and used for multiple other use cases. This work seems to leverage the same idea.

We revised Section 8 to highlight the novelty of HPTSA as opposed to prior hierarchical multi-agent frameworks. Briefly, HPTSA introduced a non-linear hierarchical structure, dynamic orchestration, and agent independence, which are tailored to the cybersecurity domain.

Reviewer CYn5 Weakness 2a. The evaluation and comparison to baseline is not so clear. It seems to show that the performance is not as good as the prior work (Fang et al 2024a), although the authors mentioned that the prior work has more description. The other baseline with GPT without description seems relatively simple.

We clarify that the prior agent developed by Fang et al [1] requires a description of the specific vulnerability. The zero-day scenario that HPTSA focuses on is significantly more difficult than the scenario of the prior work, since the agent needs to identify the specific vulnerability and attack surface on its own. As shown in Section 5.2, when removing the vulnerability description, the performance of the prior agent is 4.3x lower than HPTSA.

[1] Fang, Richard, et al. "LLM Agents Can Autonomously Exploit One-Day Vulnerabilities." arXiv:2404.08144 (2024).

Meta Suggested Revisions 2. The experiments only use powerful LLMs without testing smaller models, making it unclear how the framework performs under resource constraints.

In Section 5.2, we present the results of less powerful LLMs, llama-3.1 and qwen-2.5. These LLMs fail to exploit any zero-day vulnerabilities. Due to the difficulty of exploiting zero-day vulnerabilities, HPTSA relies on LLMs with strong capabilities.

We revised Section 6 to acknowledge the limitation of HPTSA relying on powerful LLMs.

Reviewer CYn5 Weakness 3. More implementation details of how the prompts are constructed would be helpful.

We added a description of prompts in the Appendix. Due to responsible disclosure policies to prevent potential harmful use of our agents, we will only release prompts in the future on request.

Reviewer SBoa Weakness 1. Each subagent is heavily reliant on well-curated documents. This dependence should be studied by comparing with internal knowledge of the models as possible.

We added a description of all documents provided to agents to Appendix B. We clarify that these documents only cover general entry-level cybersecurity knowledge. They are not curated to specific vulnerabilities in our benchmark.

Reviewer SBoa Weaknesses 2. Though the setting proposed is akin to real-world scenarios, it doesn't completely emulate it. However, I still believe the setup is a good starting point for such explorations.
Reviewer SBoa Weaknesses 3. This is purely from an attacker's standpoint. While it's possible the same approach might be used for defensive scanning, the paper does not systematically discuss or evaluate that but rather mentions it only.

We revised Section 10 to acknowledge the limitations of our work in fully emulating real-world scenarios and potential future work on defensive screening.

Reviewer CqvN Weaknesses 1. I am not a total expert of the security domain, but looking at table 1, I can see that most of them are web vulnerabilities. While these might be impactful, the generalizability of the HPTSA to other cybersecurity contexts (like network-based or binary exploits) remains to be seen.

We revised the submission to highlight that HPTSA focuses on web vulnerabilities, which are often entry points of more in-depth attacks. We revised Section 10 to acknowledge potential future work on extending HPTSA to other cybersecurity contexts.

Reviewer CqvN Weaknesses 2. While the cost analysis is valuable, current cost structures indicate that using advanced LLM agents is economically comparable to human experts (with possibly better accuracy).

We revised Section 7 to reflect the updated pricing of LLMs. In line with prior work, using an estimate of $50/hour and 1.5 hours/vulnerability for a human security analyst, the cost of using HPTSA to successfully exploit a vulnerability is 3x lower than human security analyst.