

## **1 PREVIOUS REVIEWS AND REBUTTALS**

### **1.1 Reviewer 6wXM, Concern 1**

Metasploitable 2 is relatively simple and not representative of real-world networks.

Response:

It is true that Metasploitable 2 contains several relatively simple vulnerabilities, but it is also the case that in real-world penetration testing, it is common to discover simple vulnerabilities on clients' servers and networks. We do not make any claims that the software developed for our study can or should replace an expert penetration tester, but rather that they can potentially reduce the costs of penetration testing overall by automating the tasks normally carried out by more junior testers. During a commercial penetration test, simple vulnerabilities must be tested for alongside those that require more sophisticated attacks, and it is often the case that it is the simple vulnerabilities that are exploited by real-world threat actors when they do in fact exist.

### **1.2 Reviewer 6wXM, Concern 2**

I would also expect to see more statistical/quantitative metrics to better justify the conclusions but now the main metrics appear to be the no. of vulnerabilities exploited and the total time. It would be more informative to report e.g. the time spent in each stage, the correctness of individual commands and measures such as false positives. Some models fail for various reasons, but only a rather shallow discussion of these failure modes is provided.

Response:

Much of this data is included in the run logs and was not included in the submitted report due to a lack of available space. Our focus was on the two research questions posed. It is our intention to publish follow up studies which will allow for more detailed analysis. The run logs (and full source code) will be included with the final version of the paper for verification and further analysis by interested parties.

### **1.3 Reviewer 6wXM, Concern 3**

My other concerns are also acknowledged by the authors, such as limitations of the test environment and interfaces, which restrict the scope of exploitable vulnerabilities. Some common used pentesting tools like Burp Suite are not available to the agents.

Response:

It was a known limitation to the study design that the local LLMs employed would not be capable of proper understanding of data displayed on a graphical user interface, and would not be able to interact with a web browser or desktop GUI environment. The reviewer is correct that this restricts the scope of exploitable vulnerabilities. This does not, however, limit the utility of terminal-based tests.

The test environment (notwithstanding the comments about Metasploitable 2, which are addressed above), is designed to be realistic and amenable to the evaluation of LLMs. We developed integration with a SSH based interface to Kali Linux. The LLMs therefore had access to all Kali Linux terminal based commands, of which there is a very large number. In one test, the LLM wanted to use a tool that wasn't installed, and attempted to install it itself. This demonstrates the effectiveness of the SSH integration, which we would argue was very effective overall in both allowing us to study the capabilities of LLMs, and allow them the freedom to determine their own course of actions.

### **1.4 Reviewer k2sg, Concern 1**

The approach is tested only on Metasploitable 2, which is not a realistic scenario.

Response:

Please see our comments above for Response to 6wXM, Concern 1.

### **1.5 Reviewer k2sg, Concern 2**

A limited set of tools was available to agents and graphical reasoning abilities are not considered.

Response:

Please see our response to 6wXM, Concern 3.

### **1.6 Reviewer k2sg, Concern 3**

A lack of in-depth discussion within the paper. While this is a small and pretty straightforward experiment, readers would benefit from an in-depth discussion of either insights into LLM design and training to facilitate pen testing, defence design or any other relevant aspects. Instead, the discussion focuses on related work, describing data from the tables, limitations and future work.

Response:

This was beyond the scope of the current study and difficult to include with the word count limit, but a valuable suggestion, nonetheless. If the paper is accepted, however, we will work to incorporate this feedback in the final version.

### **1.7 Reviewer aWf1, Concern 1**

Limited Test Environment: The use of Metasploitable 2, a training VM with well-documented vulnerabilities, reduces the generalizability of the findings to real-world scenarios.

Response:

As we have noted, the main concern in the use of Metasploitable 2 was the possibility that the documented vulnerabilities for this environment were included in the training data for the LLM. The LLM would not therefore have been reasoning, but recalling previously seen data. We are unaware of any way to determine if this was the case or not, however, as we discussed, in one case the LLM was able to identify an undocumented vulnerability. This demonstrates beyond doubt that some level of reasoning is being displayed.

### 1.8 Reviewer aWf1, Concern 2

Resource Constraints: The hardware limitations (e.g., 16GB GPU) restrict the evaluation to smaller LLMs, leaving larger models (e.g., Llama-405B) untested.

Response:

Far from being a weakness, we would argue that this is a very significant outcome. Even with limited resources, we show that local LLMs are very capable at penetration testing tasks. Repeating the study with significantly greater hardware resources will, in all likelihood, show even better results. It is our intention to evaluate larger models in the coming months.

### 1.9 Reviewer aWf1, Concern 3

No statistical analysis: The success rates are presented as counts without any variance measure or confidence bounds.

Response:

The aim of the paper is to focus on the two research questions, and with the limited space available the opportunity for further statistical analysis was limited. The authors did, however, preserve a significant amount of data in the form of run logs. The reviewer is astute in their observation that there were different outcomes for each test stage, and, to control for this, each stage and test was carried out multiple times. We have made the run logs available at <https://github.com/sw-lewis/PenTest-RunLogs.git>, and will make the source code available with the final version of the paper.

### 1.10 Reviewer aWf1, Concern 4

Missing code, data, or reproducibility artifacts: The authors did not provide the LangGraph configuration or the prompt scripts that orchestrate the experiments as supplementary material.

Response:

All source code and prompts will be included as a GitHub repository with the final version of the paper. It has always been the authors' intention to share the source code, but we would like to preserve copyright, which is impossible to do with an anonymous submission.