International Journal of Transformative Multidisciplinary Studies
ISSN Print: 3116-3009 | ISSN Online: 3116-3017
Driving Transformative Impact Across Disciplines

Peer-Review Report Form

Manuscript Title:	Benchmarking Generative AI Models for Cybersecurity Vulnerability Identification: A Comparative Study of ChatGPT, Gemini, and Claude
Type of Article:	Original Article

FIRST ROUND

Editor

General Comments and Recommendations
The reviewers acknowledge the significance of your manuscript and recommend several revisions to further improve its clarity, coherence, organization, and overall scholarly quality. In addition, you are advised to thoroughly review and validate all references included in the manuscript. Please ensure that all cited sources are accurate, accessible, and retrievable through Google Scholar to uphold the credibility and reliability of the study. Additionally, make sure to include the DOI or website link for every reference to facilitate verification and accessibility of sources. Furthermore, only sources that are cited within the manuscript should appear in the References section, and all entries listed in the References section must also be cited in the text. For example, if 20 unique sources are cited in the manuscript, then only those 20 sources should be included in the References section. Please note that repeated citations of the same source within the text are counted as a single reference entry; therefore, the total number of references may be lower than the total number of in-text citations.

Reviewer 1

General Comments and Recommendations
The manuscript evaluates the capacity of ChatGPT, Gemini, and Claude to identify software security flaws. The researchers explore a highly relevant topic, as Artificial Intelligence plays a significant role in secure software creation. The authors claim that one specific model outperforms the others, yet their own inferential statistics prove that the observed differences lack statistical significance. The authors must align their conclusions with their statistical evidence to maintain scientific integrity.

Specific Comments and Recommendations
Title-The title clearly defines the subject and the specific models evaluated. However, it could benefit from a brief mention of the dataset scale. The title is appropriate but slightly broad. Consider the addition of context about the dataset size to set accurate expectations for the reader.
Abstract-The authors must revise the abstract to clarify that performance differences are merely descriptive, as the statistical tests show no significant variance. To claim a definitive winner is scientifically inaccurate when the probability value exceeds .05.

International Journal of Transformative Multidisciplinary Studies
ISSN Print: 3116-3009 | ISSN Online: 3116-3017
Driving Transformative Impact Across Disciplines

Introduction-The authors should incorporate more contemporary literature about the specific limitations of Large Language Models in secure code generation.
Methods-The authors curated a benchmark dataset of 50 vulnerability cases. A sample size of 50 is unacceptably small for Artificial Intelligence research and prevents robust statistical generalization. The authors applied a single standardized prompt for all models. While this approach ensures consistency, it ignores the reality that different models require distinct prompt structures to reach optimal performance. The evaluation framework assigns two points for an exact match and one point for a related match. The subjective nature of related matches introduces potential observer bias, even with the reliance on the MITRE taxonomy.
Results and Discussion-The presentation of the results highlights a major disconnect between the descriptive statistics and the inferential statistics. The authors report that ChatGPT scored 72 points, Claude scored 66 points, and Gemini scored 64 points. The authors dedicate multiple paragraphs to declare ChatGPT as the superior model. The inferential analysis clearly states that the observed differences are not statistically significant. The authors must remove all declarative statements that crown one model as superior. The correct scientific interpretation is that all three models performed equally within the margin of error.
Conclusions and Recommendations-The conclusion repeats the results rather than synthesize the broader implications of the research. The authors conclude that Artificial Intelligence can serve as a decision support tool. This conclusion is sound and supported by the data. The recommendations for practitioners and developers are practical and necessary. The authors must rewrite the conclusion to reflect the lack of statistical significance among the evaluated models.

Please indicate your recommendation by checking the appropriate box below.

Decision	
	Accept the manuscript for publication.
/	Reconsider the manuscript after the authors have satisfactorily addressed and complied with the reviewers' comments and recommendations.
	Reject the manuscript, as it is not suitable for publication.

Reviewer 2

General Comments and Recommendations
References-The reference list includes relevant citations from established sources. Ensure all citations adhere strictly to the target journal format requirements. Remove outdated sources if newer equivalents exist.
Specific Comments and Recommendations
The paper has an interesting and timely benchmarking idea, and the dataset construction has some useful effort behind it. However, I still have serious concern on the validity of the experiment itself and what exactly the reported metrics are measuring. Several methodological decisions make the model comparison difficult to reproduce and, more importantly, difficult to interpret as true differences in cybersecurity capability.

International Journal of Transformative Multidisciplinary Studies
ISSN Print: 3116-3009 | ISSN Online: 3116-3017
Driving Transformative Impact Across Disciplines

There is practically no assessment of run-to-run variability of the LLMs. Each of the 50 cases appears to have been submitted only once to each model, producing exactly 150 responses, while temperature, top-p, and output length remained at platform default. This is a serious issue because these systems are stochastic. A different run can produce another CWE classification. Therefore, what is reported may partly represent one realization of each model rather than stable model performance. This is especially problematic because the paper repeatedly uses the word reliability.
I am not convinced that response time is a valid model-performance variable in this experiment. The three systems were accessed through commercial web interfaces, with default platform configurations and internet connection involved. Under this setup, 3.76 versus 5.59 versus 8.27 seconds cannot safely be interpreted as differences in model computational efficiency. Server load, network latency, web-interface behavior, queueing, output length, and provider infrastructure are mixed with the actual generation speed. Yet the manuscript treats these times as operational efficiency of the models themselves.
The stated hypothesis is not actually tested in full. The hypothesis explicitly includes both “CWE classification outcomes” and “response-time efficiency.” However, the Friedman test reported later is conducted on the ordinal classification scores of 0, 1, and 2. The inferential section does not show an equivalent statistical test of the paired response times. Thus, the hypothesis and the statistical analysis do not correspond with each other.
There is a fundamental mismatch between the task given to the AI and the ground truth used for scoring. The prompt says, “Identify all cybersecurity vulnerabilities,” but the dataset was deliberately structured so that there is one ground-truth vulnerability for each observation. It is unclear what happens when a model correctly finds the intended CWE but also reports one or several false vulnerabilities. Likewise, the paper does not explain whether extra incorrect findings reduce the score. This is not a small scoring detail; it changes what “vulnerability identification performance” actually means.
The claim that the 50 benchmark observations are independent is questionable. The manuscript states that the dataset consists of independent vulnerability cases, but one source-code sample containing three vulnerabilities was converted into three separate benchmark scenarios. If essentially the same code is appearing more than once as different observations, these cases are not fully independent. This matters also for the inferential test, where benchmark cases function as the repeated blocks.
Possible benchmark contamination is a major unanswered issue. Many cases came from MITRE CWE, OWASP documentation, Juliet Test Suite, WebGoat, DVWA, secure-coding repositories, and other public sources. These are highly visible cybersecurity materials and may have appeared in the training data or related instructional data of these commercial models. Without examining this possibility, high performance may partly indicate recognition or memorization of familiar vulnerability patterns rather than general capability to discover vulnerabilities in unseen code. For a benchmarking paper on LLMs, this is a very important threat to validity.
Several reported metrics are not really separate dimensions of performance. “Classification specificity” is simply the exact-classification rate, while “vulnerability recognition rate” combines exact and related classifications. The paper then calls the latter reliability. Weighted Performance Score, average score per case, recognition rate, misclassification rate, and Efficiency Index are largely transformations of the same underlying exact/related/incorrect counts. Therefore, describing the analysis as a multidimensional evaluation of “accuracy, specificity, reliability, and efficiency” gives more dimensions than were actually independently measured.

International Journal of Transformative Multidisciplinary Studies
ISSN Print: 3116-3009 | ISSN Online: 3116-3017
Driving Transformative Impact Across Disciplines

The use of the term “specificity” is technically difficult to accept here. No true-negative set is presented. The reported “specificity” is simply the proportion of cases with exact CWE classification. Similarly, “reliability” is being used for vulnerability recognition rather than consistency or reproducibility. These are not only terminology concerns because they affect the scientific meaning of the conclusions.
The ground truth is central to the whole study, but its validity remains weakly established. The researchers themselves curated and reviewed the benchmark and explicitly state that there was no formal independent expert validation or inter-rater reliability assessment. Every model score depends on those ground-truth CWE labels. If any reference classification is debatable, overly specific, or incorrectly mapped within the CWE hierarchy, the model scores directly change. This is therefore a major validity problem, not only a limitation for future work.
The practical conclusions are broader than the variables actually evaluated. The prompt collected severity, technical explanation, and recommended mitigation, but these dimensions were not scored. The manuscript itself later admits that explanation quality, mitigation quality, reasoning consistency, false-positive behavior, and professional usability were not evaluated. Still, the paper makes statements about usefulness for code review, vulnerability assessment, remediation support, and cybersecurity decision support. That evidence is not really produced by this experiment. Correct or related CWE labeling alone is much narrower than an effective cybersecurity analysis.
The Efficiency Index has weak methodological meaning. It is calculated simply by dividing Weighted Performance Score by mean response time. The numerator is already based on an imposed 2–1–0 weighting, while the denominator is a web-interface timing measure affected by outside factors. Combining these two quantities creates a very impressive-looking value such as 19.13, but there is no evidence that this index is validated or that differences in this number represent a meaningful cybersecurity advantage.
Reproducibility is claimed more strongly than what is presently available. The benchmark dataset, scoring materials, and evaluation outputs are only stated to be available “upon reasonable request.” For an LLM benchmark using rapidly changing commercial web systems and default configurations, the exact inputs, complete raw outputs, timestamps, scoring decisions, and benchmark instances are essential for auditing the reported results. At present, another researcher cannot independently reconstruct the experiment from the manuscript alone.

Please indicate your recommendation by checking the appropriate box below.

	Decision
	Accept the manuscript for publication.
/	Reconsider the manuscript after the authors have satisfactorily addressed and complied with the reviewers’ comments and recommendations.
	Reject the manuscript, as it is not suitable for publication.

International Journal of Transformative Multidisciplinary Studies
ISSN Print: 3116-3009 | ISSN Online: 3116-3017
Driving Transformative Impact Across Disciplines

SECOND ROUND

Decision	**Editor**	**Reviewer 1**	**Reviewer 2**
Accept the manuscript for publication.	/	/	/
Reconsider the manuscript after the authors have satisfactorily addressed and complied with the reviewers' comments and recommendations.			
Reject the manuscript, as it is not suitable for publication.			