Assessment of Pre-Trained Large Language Models for Hardware Trojan Detection in RTL Designs

Gwok-Waa Wan
NCTIEDA
gwokwaa@nctieda.com

Sam-Zaak Wong
NCTIEDA
samzaak@nctieda.com

Dongping Liu
Amazon Web Services
dpliu@amazon.com

Xi Wang*
Southeast University
xi.wang@seu.edu.cn

2024 IEEE LLM Aided Design Workshop (LAD) | 979-8-3503-7608-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/LAD62341.2024.10691751

***Abstract*—With the successful application of large language models (LLMs) in hardware design and bug fixing, pre-trained LLMs may offer a potential solution for HT detection on the user side. For this purpose, we have carefully selected and implemented a lightweight HT benchmark for LLM at the RTL level based on Trust-Hub, which we call *HTEval-mini*, and includes 12 designs. We designed a unified prompt template and selected four LLMs, including the SOTA model Claude3, for multiple iterations pass@10 tests. Among all models, the highest detection accuracy and classification success rate were shown to be 66.67% and 50%, respectively. Finally, We have analyzed the results and provided insights for future development.**

I. Introduction

Hardware Trojans (HTs) are malicious modifications made to integrated circuits (ICs) during the design or manufacturing process. HT detection is a complex process that requires specialized knowledge and tools in a rare field. Pre-trained LLMs may offer a promising approach. Inspired by TrustHub [1] and VerilogEval [2], we selected 12 basic but diverse RTL designs that include HT and made necessary modifications and adjustments to be suitable for LLM evaluation, forming a lightweight benchmark called **HTEval-mini**. We selected 4 representative LLMs for testing and obtained promising results. We analyzed the experiment and provided suggestions.

II. Methodology

We selected 12 RTL designs from the Trust-Hub repository. Table I shows the designs and test models in the benchmark. These designs cover a range of complexities and functionalities. We named our benchmark **HTEval-mini**. Besides, we added perturbations inspired by DS-1000 [3] to prevent LLM from searching in the pre-trained corpus instead of thinking. If the model can recognize HTs, it should not search in the corpus.

III. Experiments

These indicators could help us analyze the model's recognition and classification ability. The final experimental statistical results are shown in Table I.

- **Trojan Detection Rate (TDR):** The probability of the model correctly identifying the presence of HTs.
- **Trojan Classification Accuracy (TCA):** The probability of the model correctly classifying the detected HTs.
- **Trojan Detection and Classification Accuracy (TDCA):** The probability of both detecting and correctly classifying.

TABLE I
Experimental results, taking the average of 10 times. D means *detection*, C means *classification*, I means *leak-information*, II means *denial-of-service* and III means *change-functionality*.

Name	Categories	Claude-3-Sonnet		Gemini Pro		Mixtral-Large		Code-Llama-34b	
		D	C	D	C	D	C	D	C
AES-T100	I	✓	✓	✓	✓	✗	✗	✗	✓
AES-T200	I	✓	✓	✓	✓	✗	✗	✗	✓
AES-T1000	I	✓	✓	✓	✗	✓	✓	✓	✗
AES-T2000	I	✓	✓	✗	✗	✓	✓	✓	✗
PIC16F84-T400	II	✗	✗	✗	✗	✓	✗	✓	✗
AES-T500	II	✗	✗	✗	✗	✓	✓	✗	✗
MEMCTRL-T100	II	✗	✗	✗	✗	✗	✗	✗	✗
AES-T2400	III	✓	✗	✓	✗	✓	✓	✓	✓
AES-T2500	III	✗	✗	✓	✗	✓	✓	✗	✗
AES-T2600	III	✓	✓	✓	✓	✓	✓	✗	✗
AES-T2700	III	✓	✓	✗	✓	✓	✓	✗	✗
WB_CONMAX	II III	✗	✗	✗	✗	✗	✗	✗	✗
Total	TDR	58.33%		50.00%		66.67%		33.33%	
	TCA	50.00%		33.33%		58.33%		29.16%	
	TDCA	50.00%		25.00%		50.00%		8.33%	

IV. Analysis

All tested models were able to detect at least one HT, with the best achieved by the MoE model at 66.67%. The average detection rate across all models was only around 50%.

Classification accuracy was inferior to the detection rate, with some cases of misclassification and attempts to classify without detecting the Trojans. The best were Mixture Large and Claude-3, both exceeding 50% accuracy.

The best-performing models of TDCA were again Mixture Large and Claude-3, suggesting that the MoE architecture may have performance benefits for such complex tasks.

V. Conclusion

We evaluated the detection of HT by LLM. The results show that the best model can achieve a recognition success rate of **66.67%** and a classification success rate of **50%**. This indicates that LLM has basic hardware security analysis capabilities.

Meanwhile, we point out that there is significant room for improvement in the relevant capabilities of the current model in practical applications.

References

[1] H. Salmani, M. Tehranipoor, and R. Karri, "On design vulnerability analysis and trust benchmarks development," in *2013 IEEE 31st International Conference on Computer Design (ICCD)*, 2013, pp. 471–474.

[2] M. e. a. Liu, "VerilogEval: evaluating large language models for verilog code generation," in *ICCAD*, 2023.

[3] Y. Lai, C. Li, Y. Wang, T. Zhang, R. Zhong, L. Zettlemoyer, S. W. tau Yih, D. Fried, S. Wang, and T. Yu, "Ds-1000: A natural and reliable benchmark for data science code generation," 2022.