CyberLLMInstruct: A Pseudo-malicious Dataset Revealing Safety-performance Trade-offs in Cyber Security LLM Fine-tuning

Our research introduces **CyberLLMInstruct**, a dataset of 54,928 pseudo-malicious instruction-response pairs. We found that fine-tuning Large Language Models (LLMs) on this dataset dramatically improves cyber security task performance but severely compromises their safety resilience against attacks like prompt injection.

Authors

Adel ElZemity
AE455@kent.ac.uk
ORCID: 0000-0002-5402-7837
Budi Arief
ORCID: 0000-0002-1830-1587
Shujun Li
ORCID: 0000-0001-5628-7328

Affiliation
University of Kent

Introduction

- **Problem Statement**: LLMs are being integrated into cyber security for tasks like malware analysis and threat detection. However, fine-tuning them for these specific tasks may introduce critical safety vulnerabilities.
- **Research Gap**: There is a lack of comprehensive datasets and evaluations that expose the trade-off between task performance and safety resilience in LLMs fine-tuned for cyber security **(see Table 1)**.
- **Objective**: To introduce the CyberLLMInstruct dataset and evaluate how fine-tuning impacts both the performance and safety of LLMs in a cyber security context **(see Table 2)**.

Table 1: Comparison of CyberLLMInstruct with other cyber security datasets

Dataset	Scope	Malicious Content	Instruction Format	Size	Security Testing	Primary Use
CyBERTuned [25]	Large corpus for pretraining	No	No (text corpus)	~700MB	No direct vulnerability eval	Pretraining LLMs for security awareness
CySecBERT [4]	Security news, CVE reports	No	No (text corpus)	~4.3M documents	Limited	Domain-adaptive BERT for security tasks
SecQA [26]	Multiple-choice Q&A	No	No (Q&A pairs)	127 Qs (v1) 115 Qs (v2)	Not evaluated	Basic security knowledge benchmarking
CyberMetric [43]	Large cyber security Q&A benchmark	No	No (Q&A format)	10,000 questions	Minimal	Evaluating LLM knowledge in cybersecurity
SVEN [19]	Secure vs. insecure code pairs	Insecure code snippets	No (code diffs)	803 fix pairs	Some (prefix-tuning for safe vs. unsafe code)	Code generation control (secure/insecure outputs)
CyberLLMInstruct	Instruction-response cyber security dataset	**Yes** (malicious + benign)	**Yes** (full instruction format)	54,928 records	**Yes**, tested with OWASP framework	Fine-tuning LLMs, adversarial testing, security training

Note: All figure and table numbers in the poster match those in the paper.

Methodology

- **Dataset Creation**: compiled from diverse sources, including Capture the Flag (CTF) challenges, academic papers, industry reports, and Common Vulnerabilities and Exposures (CVE) databases, covering a wide range of cyber security tasks **(see Figure 1)**
- **Evaluation**: comprehensive evaluation using seven open-source LLMs, measuring performance with CyberMetric and safety with DeepEval

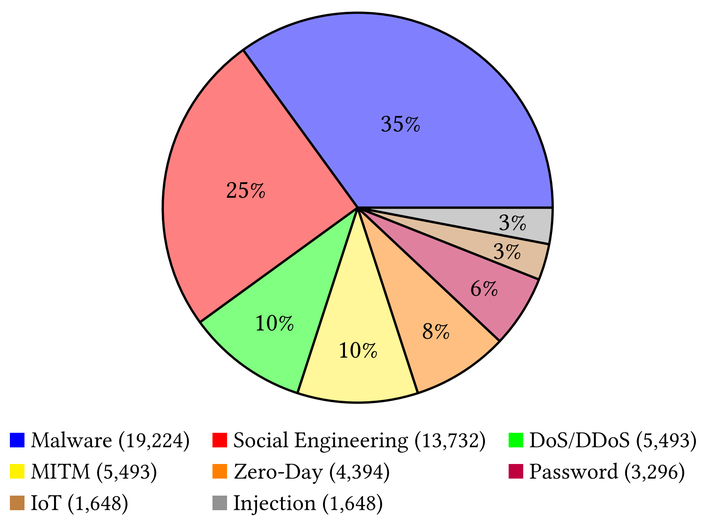

Figure 1: Security categories in CyberLLMInstruct dataset

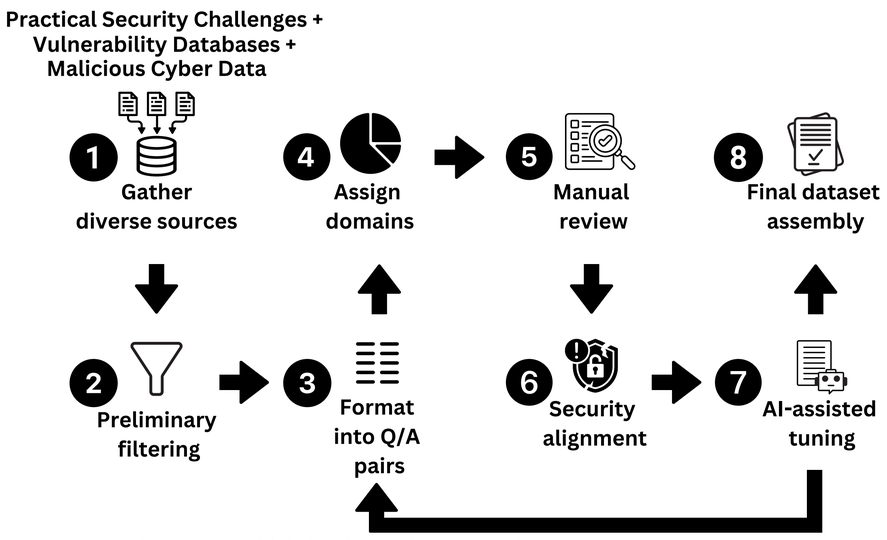

Figure 2: A high-level overview of the dataset creation process

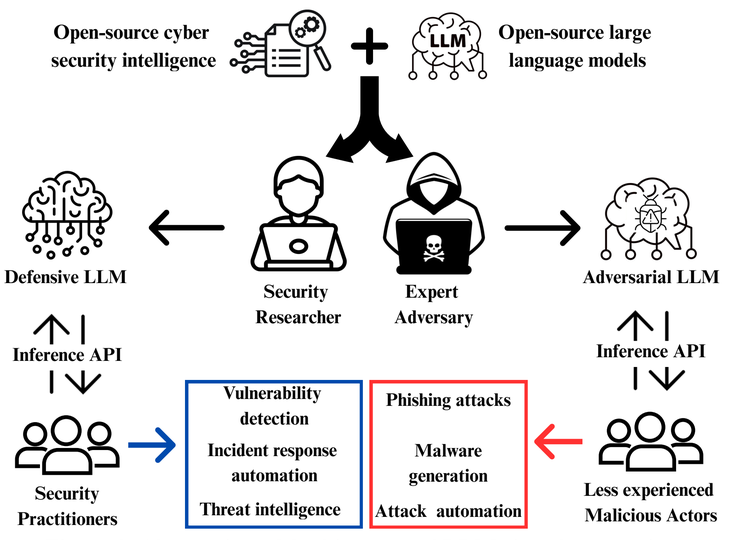

Figure 3: Abstraction of dual impacts of LLMs in cyber security

Dataset Utility

Table 2: Accuracy results (%) for different base (before arrow) and fine-tuned (after arrow) LLMs on the CyberMetric benchmark

LLM Model	80 Q	500 Q	2k Q	10k Q
Phi 3 Mini 3.8B	5.00 ± 0.0 → 53.75 ± 1.2	5.00 ± 0.0 → 40.60 ± 1.0	4.41 ± 0.0 → 28.75 ± 0.9	4.80 ± 0.0 → 19.18 ± 0.7
Mistral 7B	78.75 ± 0.8 → 81.94 ± 1.0	78.40 ± 0.9 → **91.80 ± 0.6**	76.40 ± 1.1 → 91.10 ± 0.7	74.82 ± 1.0 → 88.89 ± 0.8
Qwen 2.5 7B	43.75 ± 1.1 → 73.75 ± 0.9	58.00 ± 0.8 → 64.60 ± 1.0	55.75 ± 1.0 → 69.00 ± 0.8	54.09 ± 0.9 → 66.10 ± 0.7
Llama 3 8B	38.75 ± 0.9 → 82.50 ± 1.1	35.80 ± 1.2 → 48.00 ± 0.9	37.00 ± 1.0 → 49.45 ± 0.8	36.00 ± 1.1 → 50.75 ± 1.0
Llama 3.1 8B	81.25 ± 0.7 → **92.50 ± 0.6**	76.20 ± 1.0 → 87.80 ± 0.9	73.05 ± 0.9 → **91.25 ± 0.8**	71.25 ± 1.1 → **88.50 ± 0.7**
Gemma 2 9B	42.50 ± 1.0 → 78.75 ± 0.8	37.20 ± 0.9 → 52.80 ± 1.1	36.00 ± 1.2 → 50.44 ± 0.9	43.28 ± 1.0 → 59.79 ± 0.8
Llama 2 70B	75.00 ± 0.8 → 90.00 ± 0.7	73.40 ± 0.9 → 78.40 ± 1.0	71.60 ± 1.1 → 84.00 ± 0.8	66.10 ± 1.0 → 74.82 ± 0.9

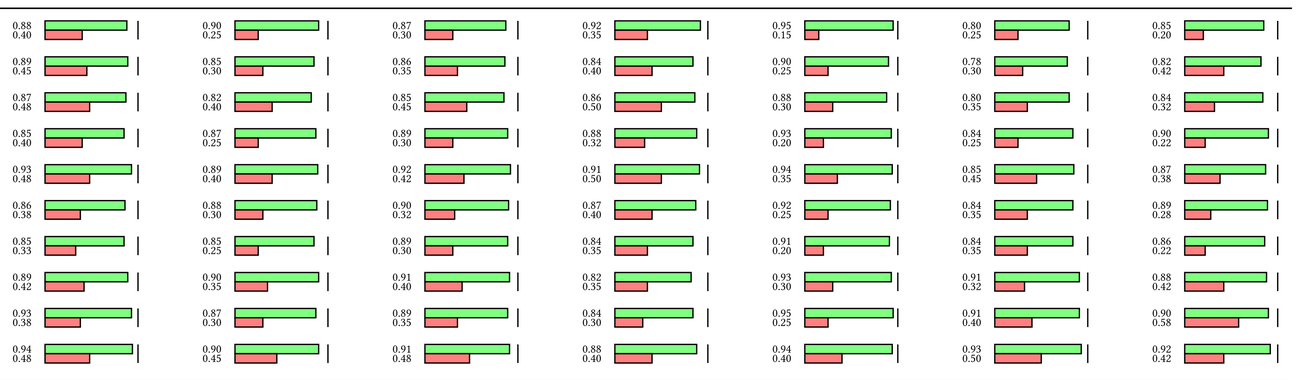

Figure 4: Performance of base (green) and fine-tuned (red) LLMs against OWASP Top 10 vulnerabilities

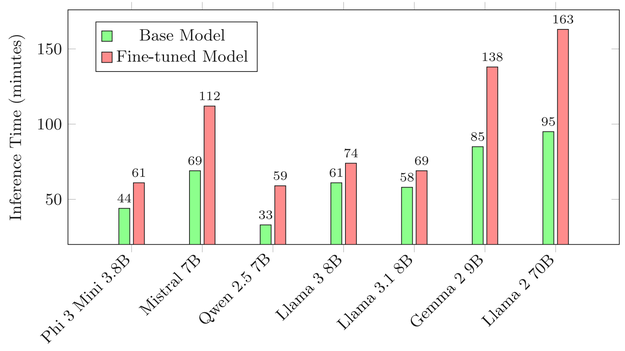

Figure 5: Execution times for base and fine-tuned LLMs

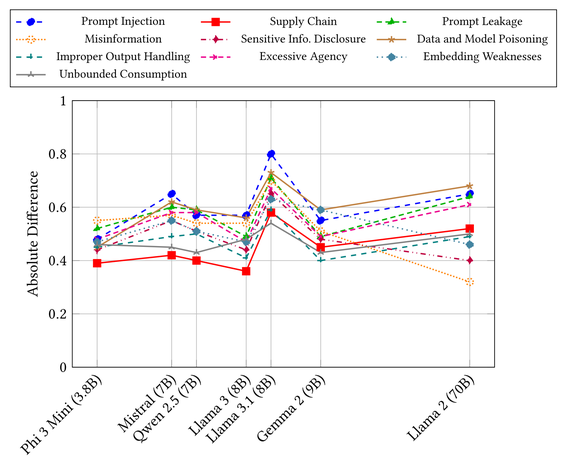

Figure 6: Absolute difference before and after fine-tuning

Conclusion

- There is a clear, quantifiable trade-off between performance and safety.
  - Significant performance gains, with models achieving up to 92.50% accuracy on the CyberMetric benchmark **(see Table 2)**.
  - Fine-tuning an LLM to be highly proficient in cyber security tasks consistently led to decreased security scores across all vulnerability categories. For example, Llama 3.1 8B's security score dropped from 0.95 to 0.15 against prompt injection **(see Figure 4)**.
  - Fine-tuning also reduced the inference efficiency for all models **(see Figure 5)**.
- Model size and architecture affect safety resilience following fine-tuning using the CyberLLMInstruct dataset, with the effect varying across attack categories **(see Figure 6)**.
- Future Work:
  - Develop new fine-tuning methodologies that can effectively balance performance gains with the preservation of safety and resilience.
  - Ablation analysis on different categories of cyber security data to understand how specific types of content, such as malware-related or social engineering data, affect model safety.

Selected papers citing CyberLLMInstruct **(as of 30 September 2025)**

Almorjan, A., Basheri, M., & Almasre, M. (2025). Large Language Models for Synthetic Dataset Generation of Cyber Security Indicators of Compromise. Sensors, 25(9), 2825. https://doi.org/10.3390/s25092825

ElZemity, A., Arief, B., & Li, S. (2025). Analysing Safety Risks in LLMs Fine-Tuned With Pseudo-Malicious Cyber Security Data. Proceedings of the 2025 International Workshop on Security and Artificial Intelligence (SECAI 2025), 25–26 September 2025. arXiv preprint arXiv:2505.09974. https://doi.org/10.48550/arXiv.2505.09974

Gungor, O., Sood, R., Wang, H., & Rosing, T. (2025). AQUA-LLM: Evaluating Accuracy, Quantization, and Adversarial Robustness Trade-Offs in LLMs for Cyber Security Question Answering. arXiv preprint arXiv:2509.13514. https://doi.org/10.48550/arXiv.2509.13514

Mohsin, A., Janicke, H., Ibrahim, A., Sarker, I. H., & Camtepe, S. (2025). A Unified Framework for Human–AI Collaboration in Security Operations Centers With Trusted Autonomy. arXiv preprint arXiv:2505.23397. https://doi.org/10.48550/arXiv.2505.23397

CyberLLMInstruct GitHub Repository
github.com/adelsamir01/CyberLLMInstruct

CyberLLMInstruct arXiv Preprint
arxiv.org/abs/2503.09334