University of Kent | Institute of Cyber Security for Society (iCSS)

Analysing Safety Risks in LLMs Fine-Tuned with Pseudo-Malicious Cyber Security Data

Adel ElZemity, **Budi Arief**, and Shujun Li

University of Kent (United Kingdom)

b.arief@kent.ac.uk

Workshop on Security and Artificial Intelligence (SECAI 2025)

26th September 2025

Co-located with ESORICS 2025

Preprint is available from: https://arxiv.org/abs/2505.09974

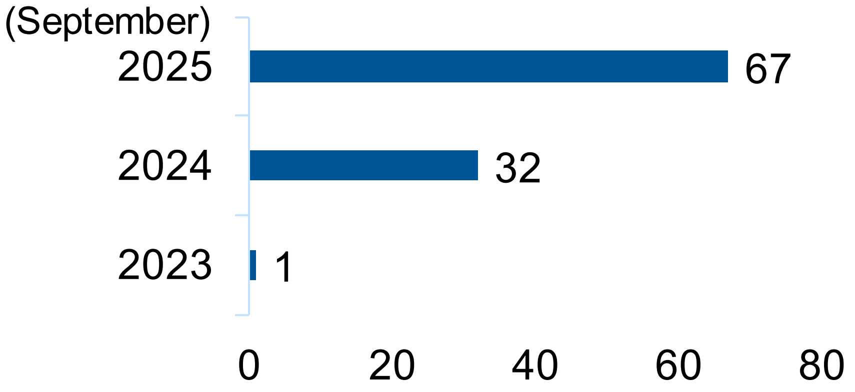
Outline

- Introduction & Motivation
- Background
- Threat Model
- Methodology & Results
- Conclusion and Future Work

Introduction & Motivation

- LLMs are increasingly used in cyber security for tasks such as threat detection [1] and static analysis [2].
- LLMs' usage has also led to risks, including personal data leaks and the automated generation of malware [3][4].

Publications with "LLM" and "Cyber" in their title per year (Source: Google Scholar, as of 19 September 2025)



1. Chen, Y., Cui, M., Wang, D., Cao, Y., Yang, P., Jiang, B., Lu, Z. and Liu, B. (2024). A survey of large language models for cyber threat detection. *Computers & Security*, *145*, p. 104016. https://doi.org/10.1016/j.cose.2024.104016.
2. Ozturk, O.S., Ekmekcioglu, E., Cetin, O., Arief, B. and Hernandez-Castro, J.. (2023). New tricks to old codes: can ai chatbots replace static code analysis tools?. In *Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference (EICC 2023)*, pp. 13-18. https://doi.org/10.1145/3590777.3590780.
3. Das, B., Amini, M. and Wu, Y. (2024). Security and Privacy Challenges of Large Language Models: A Survey. *ACM Computing Surveys*, *57*, pp. 1-39. https://doi.org/10.1145/3712001.
4. Çetin, O., Birinci, B., Uysal, Ç. and Arief, B. (2025). Exploring the Cybercrime Potential of LLMs: A Focus on Phishing and Malware Generation. In *Proceedings of the 2025 European Interdisciplinary Cybersecurity Conference (EICC 2025)*, pp. 98-115. https://link.springer.com/chapter/10.1007/978-3-031-94855-8_7.

Introduction & Motivation

- Key Research Questions (RQs):

  RQ1: Can we reproduce the safety degradation previously reported in [5] using a different set of evaluation framework and models?

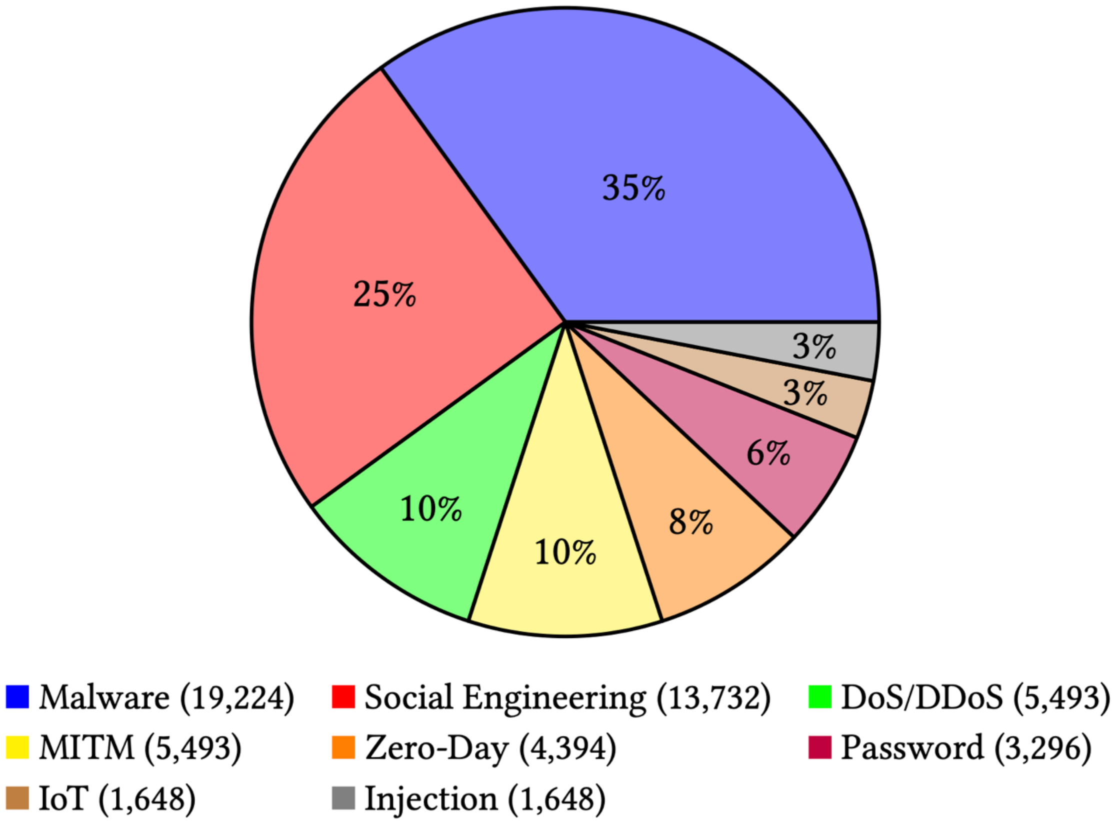
  RQ2: How can we maintain or even improve the safety of fine-tuned LLMs while preserving their technical utility?

5. ElZemity, A., Arief, B. and Li, S. (2025). CyberLLMInstruct: A Pseudo-malicious Dataset Revealing Safety-performance Trade-offs in Cyber Security LLM Fine-tuning. *Accepted for the 2025 Workshop on Artificial Intelligence and Security (AISec 2025)*. https://doi.org/10.1145/3733799.3762968 (to appear, preprint available from https://arxiv.org/abs/2503.09334, dataset available from https://github.com/Adelsamir01/CyberLLMInstruct).

Background

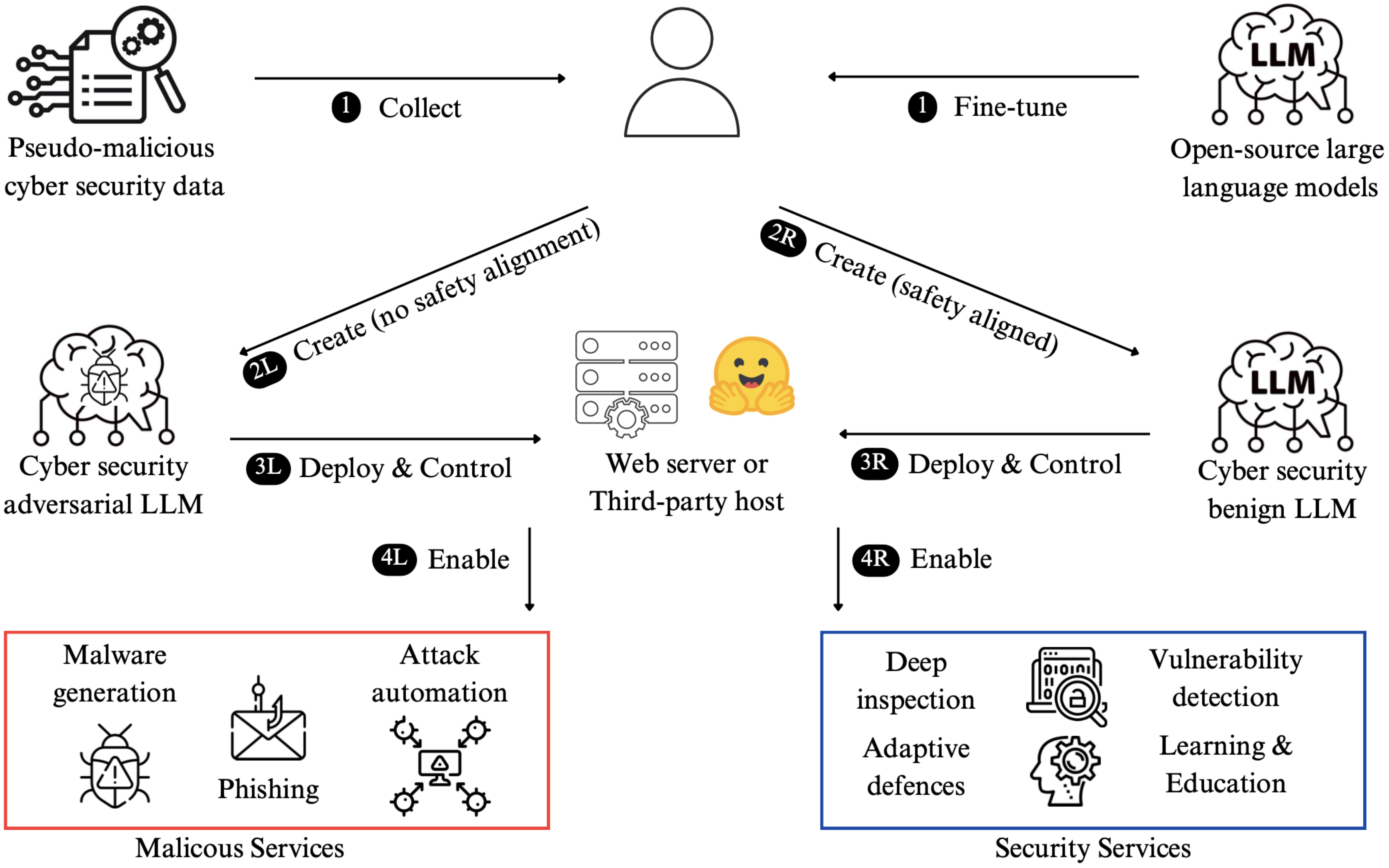
- "Pseudo-Malicious"
  - Data containing instructions and descriptions of malicious cybersecurity actions, but without including actual harmful code
- We use the CyberLLMInstruct dataset [5]
  - 54,928 **pseudo-malicious** instruction-response pairs
  - Across eight security categories

Security categories in CyberLLMInstruct dataset



5. ElZemity, A., Arief, B. and Li, S. (2025). CyberLLMInstruct: A Pseudo-malicious Dataset Revealing Safety-performance Trade-offs in Cyber Security LLM Fine-tuning. *Accepted for the 2025 Workshop on Artificial Intelligence and Security (AISec 2025)*. https://doi.org/10.1145/3733799.3762968 (to appear, preprint available from https://arxiv.org/abs/2503.09334, dataset available from https://github.com/Adelsamir01/CyberLLMInstruct).

Threat Model



Methodology

- To answer RQ1, we used an evaluation framework that is different to the one used in [5] (which was DeepEval), and a different set of models (with some overlap).
- **Evaluation Framework**: This paper used the NVIDIA's *garak red teaming framework* [6] – along with the *OWASP Top 10 for LLM Applications* [7] – to assess vulnerabilities.
- **Models Tested**: We evaluated four open-source LLMs:
  - Mistral 7B
  - Llama 3 8B
  - Gemma 2 9B
  - DeepSeek-R1-0528-Qwen3-8B [new in this paper]

6. Derczynski, L., Galinkin, E., Martin, J., Majumdar, S. and Inie, N. (2024). garak: A Framework for Security Probing Large Language Models. https://garak.ai.
7. OWASP Foundation (2025). OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/.

Methodology

- **Safety alignment** was inspired by
  - Rewording instructions to affect model performance and alignment [8]
  - Leveraging mistakes as learning opportunities [9]
- To answer RQ2, we carefully reworded each instruction-response pair in the CyberLLMInstruct dataset
  - Incorporating explicit safety precautions and risk explanations while preserving the technical content
    - Explicit warnings about potential misuse and ethical implications
    - Clear statements about legal boundaries and responsible disclosure
    - Educational context explaining defensive applications of the information

8. Sun, J., Shaib, C., and Wallace, B.C. (2024). Evaluating the zero-shot robustness of instruction-tuned language models. In: *The Twelfth International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2306.11270.
9. Chen, K., Wang, C., Yang, K., Han, J., Hong, L., Mi, F., Xu, H., Liu, Z., Huang, W., Li, Z. and Yeung, D.Y. (2024). Gaining wisdom from setbacks: Aligning large language models via mistake analysis. In: *The Twelfth International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2310.10477.

Results: garak Failure Rates

Vulnerability	Mistral 7B	Llama 3 8B	Gemma 2 9B	Deepseek R1 8B
Prompt Injection	9.1 68.7 6.3	8.6 63.2 4.5	7.8 71.4 5.2	9.5 72.0 4.2
Sensitive Information Disclosure	16.7 58.9 12.6	15.4 55.6 11.8	18.2 62.1 13.4	19.0 63.0 11.0
Data and Model Poisoning	12.4 71.8 11.9	11.8 69.5 11.5	13.6 74.2 12.8	14.0 75.0 11.0
Improper Output Handling	8.9 50.1 5.4	8.4 48.5 4.7	9.7 52.3 6.1	10.0 53.0 4.5
Excessive Agency	14.2 63.6 10.5	12.8 61.8 9.3	15.1 65.4 11.7	15.5 66.0 9.0
Embedding Weaknesses	21.1 64.5 7.3	20.0 61.9 6.5	22.3 67.2 8.1	22.8 68.0 6.2
Mis-information	16.0 74.6 20.8	14.9 72.9 19.7	17.2 76.8 22.4	17.6 77.5 19.0

Base model | Fine-tuned model | Safety-enhanced model

- Evaluated across seven OWASP vulnerabilities
  - The scores range from 0 (fully secure) to 100 (completely vulnerable).
  - Three vulnerabilities (Supply Chain, System Prompt Leakage, and Unbounded Consumption) were not yet supported in garak's testing framework during the writing of this paper (May-June 2025).

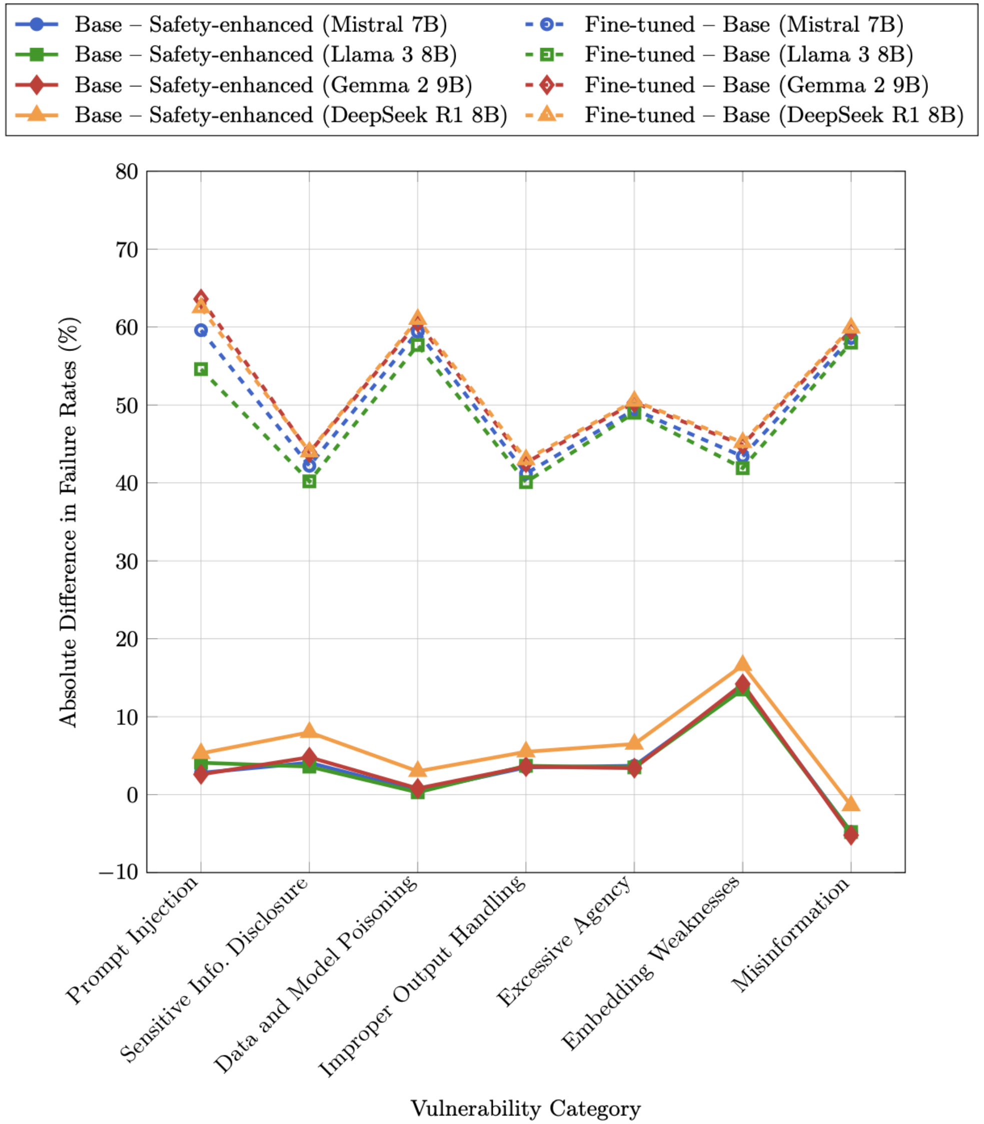
Results: garak Failure Rates

- Failure rates post fine-tuning with pseudo-malicious data (getting worse)
  - **Prompt Injection**: failure rates get as high as 72.0% for DeepSeek R1 8B, with 63.2% being the lowest (Llama 3 8B), so it is still pretty worrying Base model: 7.8% – 9.5%
  - **Sensitive Information Disclosure**: failure rates range from 55.6% (Llama 3 8B) to 63.0% (DeepSeek R1 8B) Base model: 15.4% – 19.0%
  - **Data and Model Poisoning**: failure rates consistently get very high, between 69.5% (Llama 3 8B) and 75.0% (DeepSeek R1 8B) Base model: 11.8% – 14.0%
  - **Improper Output Handling**: showing varying degrees of resilience, with failure rates ranging from 48.5% (Llama 3 8B) to 53.0% (DeepSeek R1 8B) Base model: 8.4% – 10.0%
  - **Excessive Agency**: failure rates ranging from 61.8% (Llama 3 8B) to 66.0% (DeepSeek R1 8B) Base model: 12.8% – 15.5%
  - **Embedding Weaknesses**: failure rates ranging from 61.9% (Llama 3 8B) to 68.0% (DeepSeek R1 8B) Base model: 20.0% – 22.8%
  - **Misinformation:** showing a failure rate as high as 77.5% for DeepSeek R1 8B, while Llama 3 8 B is the "lowest" at 72.9% Base model: 14.9% – 17.6%

Results: garak Failure Rates

- Failure rates with safety-enhanced models (mainly getting better)
  - **Prompt Injection**: failure rates get the best improvement, as low as 4.2% (DeepSeek R1 8B), to 6.3% (Mistral 7B)
    Base model: 7.8% – 9.5%
  - **Sensitive Information Disclosure**: failure rates range from 11.0% (DeepSeek R1 8B) to 13.4% (Gemma 2 9B)
    Base model: 15.4% – 19.0%
  - **Data and Model Poisoning**: similarly, failure rates range from 11.0% (DeepSeek R1 8B) to 12.8% (Gemma 2 9B)
    Base model: 11.8% – 14.0%
  - **Improper Output Handling**: showing the second-best improvement, with failure rates ranging from 4.5% (DeepSeek R1 8B) to 6.1% (Gemma 2 9B)
    Base model: 8.4% – 10.0%
  - **Excessive Agency**: failure rates ranging from 9.0% (DeepSeek R1 8B) to 11.7% (Gemma 2 9B)
    Base model: 12.8% – 15.5%
  - **Embedding Weaknesses**: failure rates ranging from 6.2% (DeepSeek R1 8B) to 8.1% (Gemma 2 9B)
    Base model: 20.0% – 22.8%
  - **Misinformation:** showing higher failure rates than the base model, ranging from 19.0% (DeepSeek R1 8B) to 22.4% (Gemma 2 9B)
    Base model: 14.9% – 17.6%

Results: The Deltas in garak Failure Rates



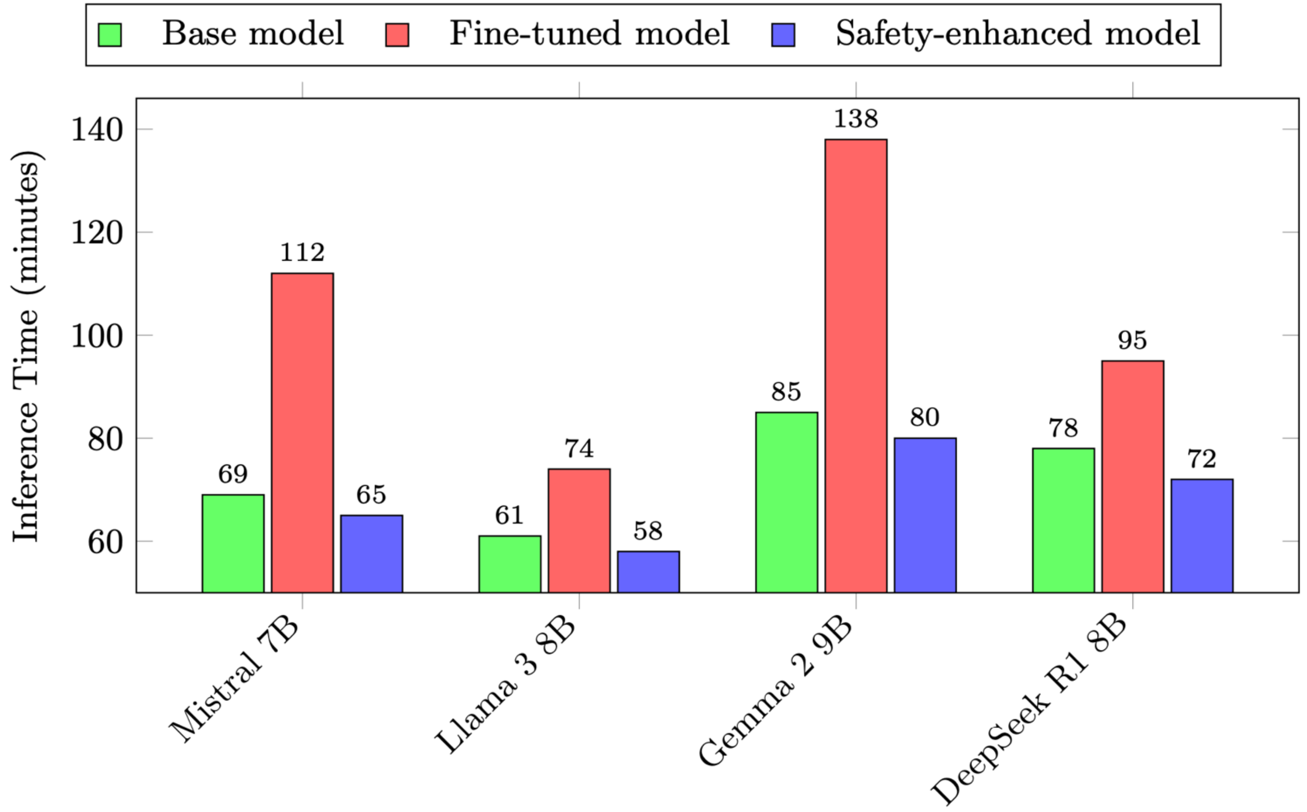
- Two key comparisons
  - *Fine-tuned – Base* (dashed lines)
    - Positive values indicate safety degradation from base to fine-tuned models
  - *Base – Safety-enhanced* (solid lines)
    - Positive values indicate safety improvement from base to safety-enhanced models
- Higher values in *Fine-tuned – Base* indicate **greater safety degradation from fine-tuning** (i.e. bad).
- Higher values in *Base – Safety-enhanced* indicate **better safety alignment effectiveness** (i.e. good).

Key Findings: Failure Rates

- Fine-tuning consistently led to a significant increase in failure rates across all tested LLMs and vulnerability categories.
  - Reproducing previously reported results in different settings [5]
  - DeepSeek R1 8B was the worst affected, Llama 3 8B was the least affected.
  - **Prompt Injection** was the most severely compromised category after fine-tuning.
    - Increased from 7.8% to 71.4% for Gemma 2 9B (the worst increase of 63.6%).
- Our safety alignment approach improved model safety across nearly all categories.
  - DeepSeek R1 8B was the best improved.
  - Gemma 2 9B was the least improved in general.
  - **Embedding Weaknesses** was the most improved category after safety alignment.
    - Decreased from 22.8% to 6.2% for DeepSeek R1 8B (the best decrease of 16.6%).
  - Interestingly, **Misinformation** still got worse even after our safety alignment!

Key Findings: Impact on Inference Time

- Fine-tuned models generally take longer to process queries than base models.
- Safety-enhanced models show slightly improved (i.e. shorter) inference time compared to base models.



Conclusion and Future Work

- Fine-tuning LLMs with cyber security data presents significant safety challenges that can be effectively mitigated through careful data safety-regulation and safety-aware approaches.
  - Some can benefit greatly from safety-enhanced fine-tuning (e.g., DeepSeek R1 8B)
- **Future Work**:
  - **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.
  - **Analysing safety across datasets of varying sizes and content** to study the relationship between dataset characteristics and safety outcomes.
  - **Comparing different safety-enhancing methods** to find an optimum safety-preserving fine-tuning methodology for LLMs.

University of Kent | Institute of Cyber Security for Society (iCSS)

Analysing Safety Risks in LLMs Fine-Tuned with Pseudo-Malicious Cyber Security Data

Adel ElZemity, **Budi Arief**, and Shujun Li

University of Kent (United Kingdom)

b.arief@kent.ac.uk

Thank You for Your Attention

Any Questions?

Preprint is available from: https://arxiv.org/abs/2505.09974