One-Shot Root-Cause Analysis Using a Fine-Tuned LLM

Marco Cerrato[1], Grace Jin[2], Jonathan Aguilar[1], Lukas Dresel[3], Giovanni Vigna[3]
1. CSU San Bernardino, 2. Cornell University, 3. UC Santa Barbara

UC SANTA BARBARA
Cornell University
CAL STATE SAN BERNARDINO

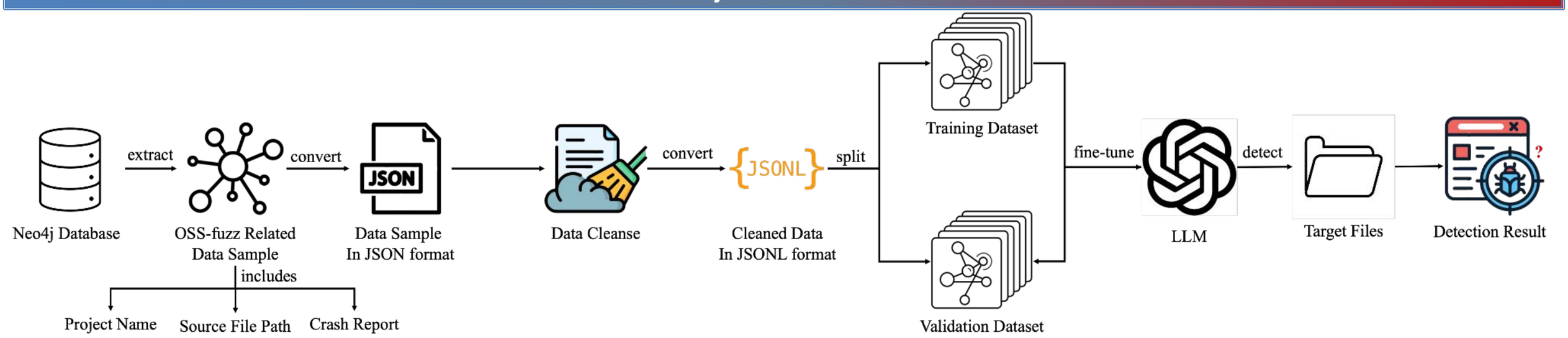
Introduction

- **MOTIVATION:** Determining the location and cause of a vulnerability can be extremely time-consuming, but LLMs can help human analysts discover the root cause of bugs faster.
- We break the problem into 2 tasks: First, identifying the **vulnerable file** and then identifying the **vulnerable function** in the file.
- **DATASET**: ~1,000 crash reports and patch data from OSS-Fuzz.
- Fine-tune a GPT-4o mini LLM and measure how well it can perform either task, compared to the base model and the flagship 4o model.
- Evaluate performance of root-cause identification on real-world vulnerability and patching datasets.

Research Question

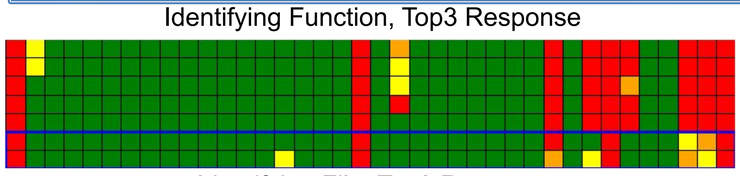
How effective is an LLM-guided approach at reducing analysis time for identifying vulnerabilities in large-scale projects, compared to traditional methods and baseline models?

Project Architecture

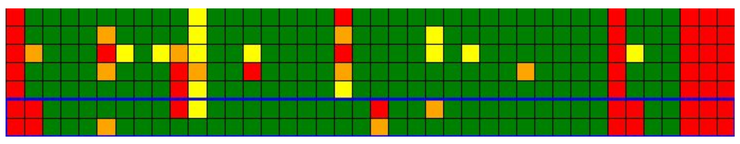


Evaluation Tests

Identifying Function, Top3 Response

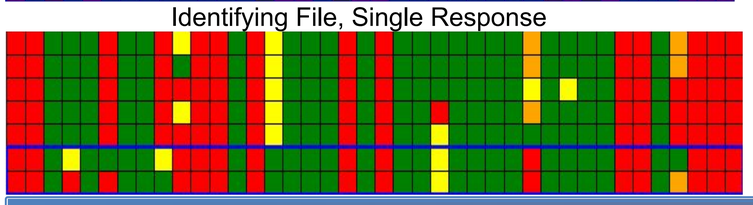


Identifying File, Top3 Response



Identifying Function, Single Response

Identifying File, Single Response



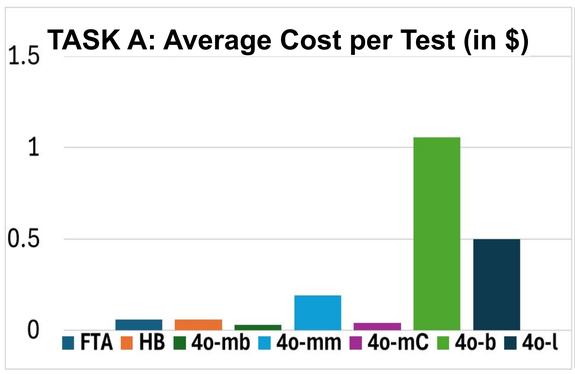
Results

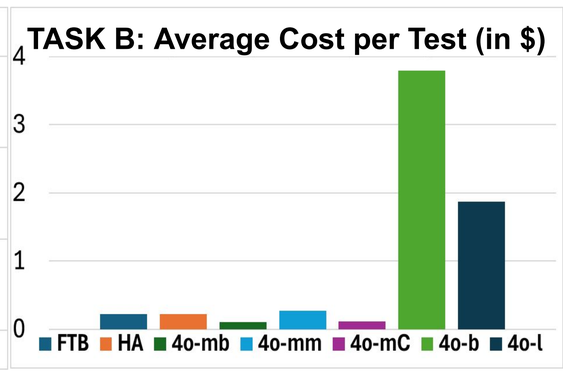
- Models' single response and top 3 responses to test queries were evaluated.
- The evaluation test consisted of **40 queries across 7 models.**
- Each row in the heat maps correspond to different LLM models.
- Fine-tuned models for Tasks A, B are FTA, FTB.
- Hybrid models are HA, HB.
- Columns represent evaluation queries in test.

(a) BASELINE: gpt-4o-latest
(b) BASELINE: gpt-4o-base
(c) BASELINE: gpt-4o-mini-Chain-of-Thought
(d) BASELINE: gpt-4o-mini-multishot
(e) BASELINE: gpt-4o-mini-base
(f) Fine-Tuned-Hybrid
(g) Fine-Tuned-Model

Costs

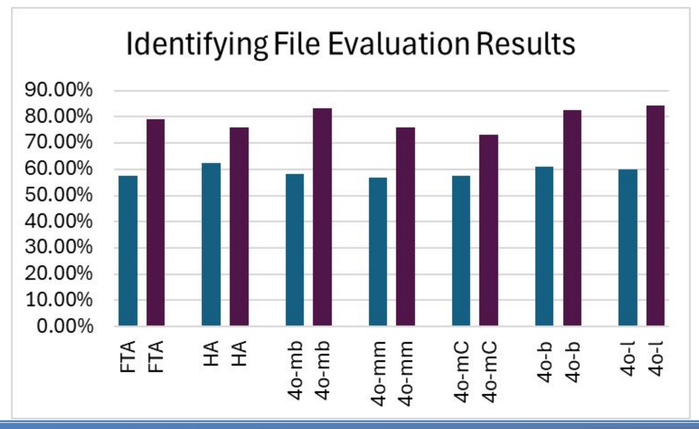
- Fine-tuned models **more cost effective** than: 4o-mini-multishot, 4o-base, 4o-latest.
- **Less cost effective** than: 4o-mini-base, 4o-mini-Chain-of-Thought.
- Fine-tuned models **equally cost effective** compared to each other.
- Cost of fine-tuning:
  - FTA/FTB = **$39**, each
  - Hybrid = cost(FTA) + 39 = **$78**





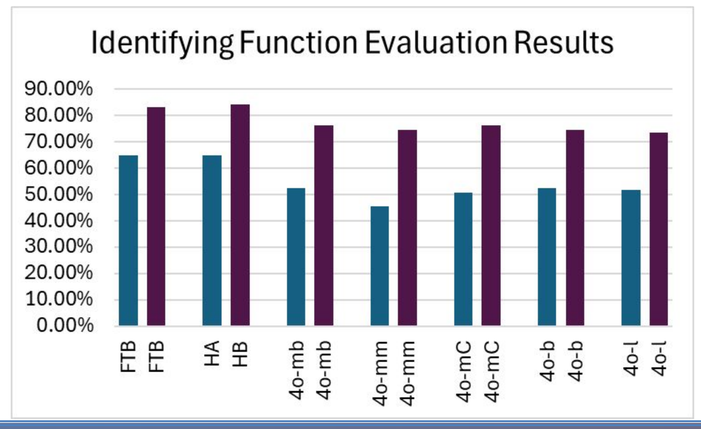
Task A

- **Single Response (Blue)**: FTA has **no performance** advantage, HA **slightly outperforms** other models.
- **Top 3 Responses (Purple)**: FTA & HA show **no performance** advantage.



Task B

- **Single Response (Blue)**: FTB & HA **outperforms** every other model by **12.28%** to **19.30%**
- **Top 3 Responses (Purple)**: FTB & HB **outperforms** every other model from **7.9%** to **10.53%**



Accuracy: Identifying File

	Single-Response	Top-3
Fine-Tuned Model	57.50%	79.17%
Fine-Tuned Hybrid Model	62.50%	75.83%

Accuracy: Identifying Function

	Single-Response	Top-3
Fine-Tuned Model	64.91%	83.33%
Fine-Tuned Hybrid Model	64.91%	84.21%

Future work

- Future Work:
  - Further **Hyper-parameter tuning** for both models
  - **Expand evaluation** with data from other sources, other models, etc.
  - Improve Task A approach, e.g., by **integrating tool-calling** support for the models.

Conclusion

- Task A: fine-tuned LLM-guided approach does **not offer noticeable advantage** over baseline models in vulnerable file identification.
- Task B: fine-tuned LLM-guided approach shows **higher performance** than baseline models in vulnerable function identification.
- Vulnerable File Identification approach needs to be re-considered. It remains a critical task, and discovering better solutions is imperative.

References

OpenAI API Reference. https://platform.openai.com/docs/api-reference/introduction
Ding, Y., Fu, Y., Ibrahim, O., Sitawarin, C., Chen, X., Alomair, B., Wagner, D., Ray, B., & Chen, Y. (2024, March 27). *Vulnerability Detection with Code Language Models: How Far Are We?* arXiv.org. https://arxiv.org/abs/2403.18624

AI Institute for Agent-based Cyber Threat Intelligence and Operation

Supported by the National Science Foundation under grant no. IIS-2229876 and is supported in part by funds provided by the NSF, by the Department of Homeland Security, and by IBM.

https://action.ucsb.edu