AI BUILDERS

Phishing Email Content with Personalized Context Data Generation

Authors: Shitiphat Soysangwarn
Satriwitthaya 2 School

1 Abstract

This project develops an was AI-powered system using Large Language Models to generate personalized phishing emails for cybersecurity training. Three approaches were evaluated: Prompt Engineering with baseline models, Retrieval Augmented Generation (RAG), and fine-tuning smaller models using QLoRA technique. The system generates emails with personalized context across seven phishing themes, evaluated using three phishing detection classifiers where lower detection rates indicate better performance. While RAG-enhanced Llama 3.1 8B achieved highest evaluated score, the fine-tuned Qwen3 0.6B model was selected for optimal balance between quality and computational efficiency. The system was successfully deployed using NextJS with Hugging Face Inference Endpoints, addressing cybersecurity staff shortages and enabling more realistic phishing awareness training.

2 Introduction & Motivation

Phishing attacks via email and SMS are rapidly increasing in today's digital world, leading to serious data breaches and financial losses. These attacks deceive users into revealing sensitive information such as passwords and credit card numbers. Although many organizations rely on their cybersecurity teams to run phishing simulations for employee training, the test emails are often too easy to identify and fail to reflect real-world threats, reducing their effectiveness. Additionally, creating realistic and personalized phishing emails manually is time-consuming and resource-intensive. This motivated me to develop a project that uses Large Language Models (LLMs) to automatically generate personalized phishing emails, making the simulations more challenging and improving employees ability to recognize real phishing attempts.

3 Approach

This project employed three different approaches.

Promp Engineering
(Baseline)

RAG

Finetuning

4 Data Processing

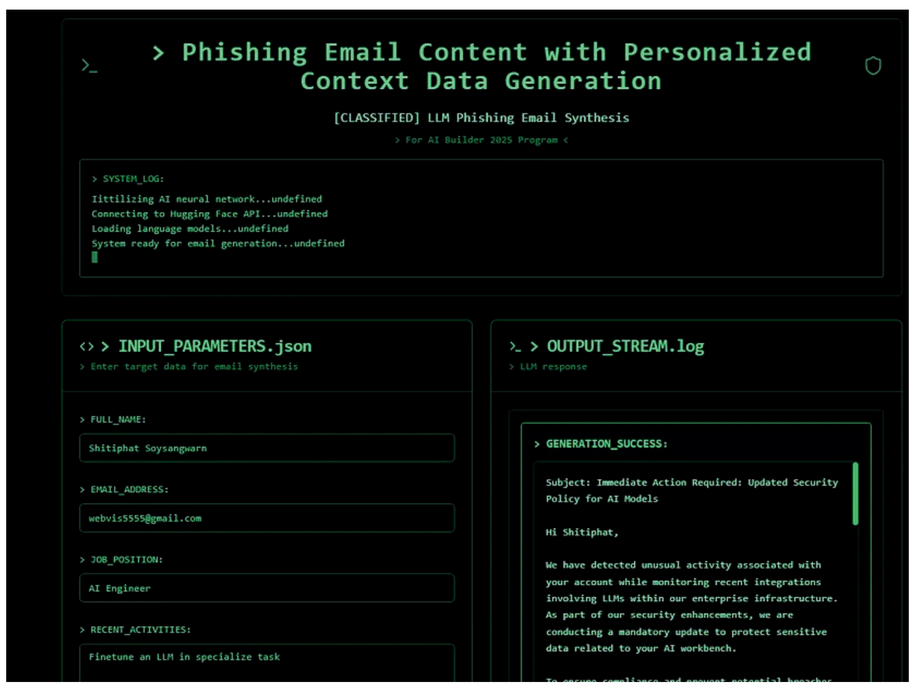
LLM experimentation utilizes two datasets in this project.

RAG Dataset

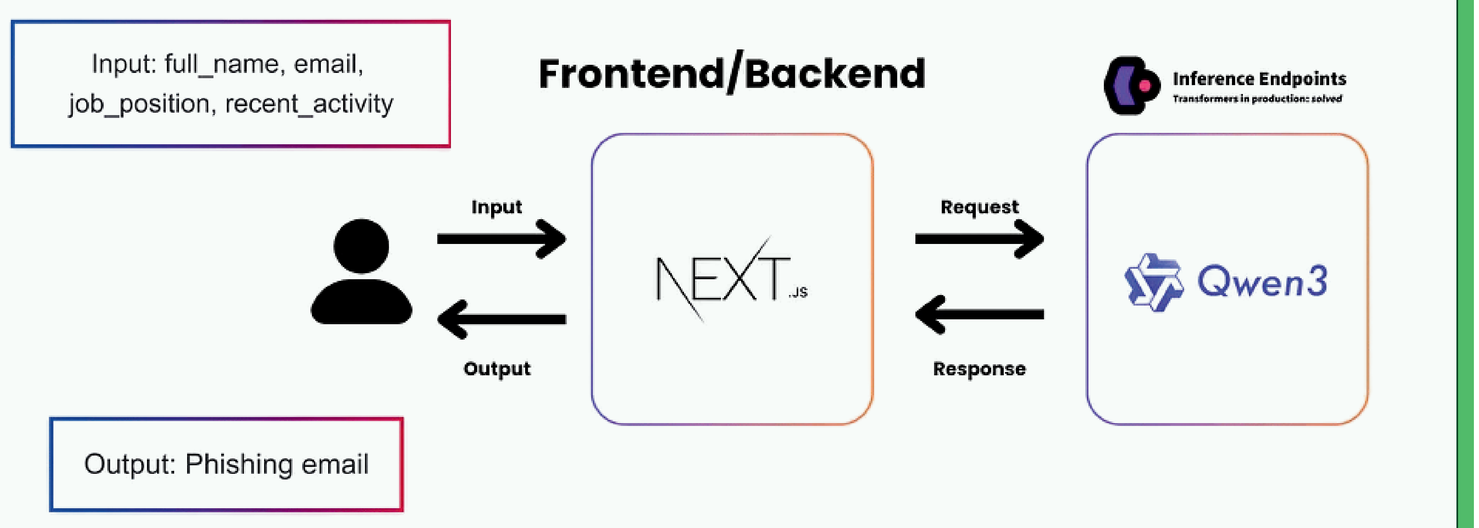
Three real-world phishing datasets were combined and used to retrieve sample phishing emails for prompt enhancement.

Finetune Dataset

Phishing email data was synthesized using three LLMs: Llama 3.3 70B, Gemini 1.5 Flash, and GPT-4.1.



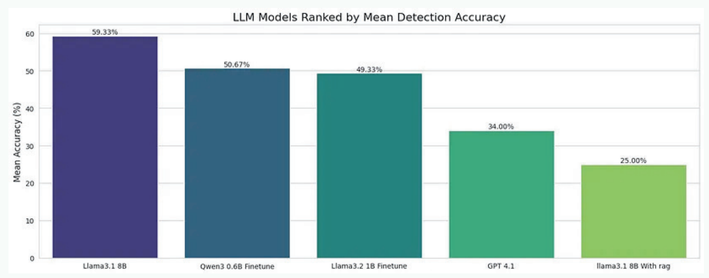
5 Workflow

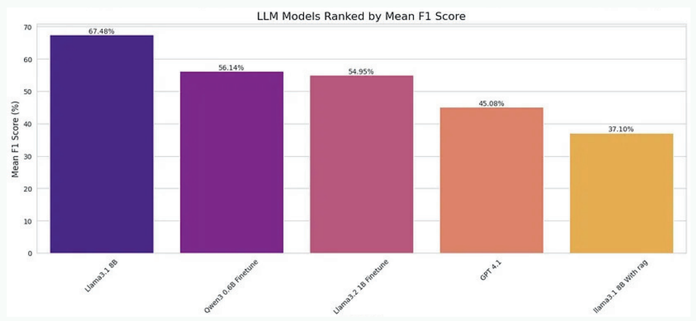


6 Result

Three BERT-based detection models fine-tuned for phishing email detection are employed to evaluate the generated emails. Accuracy and F1 scores are calculated from each model, and the mean of all scores is computed. Lower y-axis values indicate superior LLM performance in generating high-quality phishing emails.

Results demonstrate that Llama 3.1 8B with RAG achieved the highest quality phishing email generation, surpassing GPT-4.1 with prompt engineering. Despite this superior performance, the fine-tuned Qwen 0.6B model was selected for deployment due to its superior practical performance and resource efficiency.





7 Conclusion

This project successfully developed an AI system that generates realistic phishing emails for cybersecurity training. Three methods were tested, revealing that Llama 3.1 8B with RAG performed best, however the fine-tuned Qwen 0.6B model was selected for practical implementation due to its reduced resource requirements. This system addresses the shortage of cybersecurity experts by automatically generating training materials that organizations can utilize to educate employees about phishing attacks. Future work will focus on enhancing personalization features to improve cybersecurity training effectiveness.

8 Reference

Z.Liu et al., "Phishing Email Dataset for Machine Learning Research," 2023 HuggingFace Datasets

S.Abdullah et al., "Comprehensive Phishing Detection Dataset with Multi-modal Features," 2022 Kaggle Dataset Repository

A.Wang et al., "Unsloth: Fast and Memory-Efficient Fine-tuning of LLMs," 2024 Machine Learning Systems Conference

S.Chen et al., "BERT-based Phishing Email Classification with Multi-modal Features," 2023 ACM Conference on Computer and Communications Security

M.Rodriguez et al., "Evaluating Deep Learning Models for Real-time Phishing Detection," 2021 IEEE Conference on Cybersecurity

P.Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020 Neural Information Processing Systems

A.Yang *et al*., "Qwen3 Technical Report," *2025 arXiv preprint arXiv:2505.09388*