A Project Report on

“Fine - Tuning GPT - 4o - mini for Cybersecurity Incident Classification and Solution Generation”

MASTER OF TECHNOLOGY IN ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

Submitted by

Shoeb Shakil Sutar	**25070149022**
Kupakwashe Mapuranga	**25070149033**
Chintan Rakesh Shah	**25070149034**

Course Faculty

Dr. Supriya V. Mahadevkar
Assistant Professor

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE
Pune – 412115, Maharashtra State, India
https://www.sitpune.edu.in/

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

2025 - 2026

SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE
Pune – 412115, Maharashtra State, India
https://www.sitpune.edu.in/

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

This is to certify that the Project work entitled "**Fine - Tuning GPT - 4o - mini for Cybersecurity Incident Classification and Solution Generation**" is carried out by **Shoeb Sutar, Kupakwashe Mapuranga, and Chintan Shah** under the course Artificial Intelligence for Cybersecurity under **Master of Technology** in **Artificial Intelligence & Machine Learning**, Symbiosis International (Deemed University), Pune during the academic year 2025 - 2026.

Dr. Supriya V. Mahadevkar
Assistant Professor

Dr. Shilpa Bade - Gite
HoD
AIML Department

DECLARATION

I hereby declare that the project titled **"Fine-Tuning GPT - 4o - mini for Cybersecurity Incident Classification and Solution Generation"**, submitted to the Symbiosis Institute of Technology, a constituent of Symbiosis International (Deemed University), Pune, in partial fulfillment of the requirements for the degree of **Master of Technology in Artificial Intelligence & Machine Learning**, is the outcome of my original research work. I further acknowledge that this report may be made electronically accessible to the public. I also affirm that this project report, or any part of it, has not been submitted previously to any other University or Institute for the award of any degree or diploma.

Name(s) of Student(s) :

1. Shoeb Sutar
2. Kupakwashe Mapuranga
3. Chintan Shah

PRN : 25070149022, 25070149033, 25070149034

Degree : Master of Technology (M.Tech) in AI & ML

Department : Artificial Intelligence & Machine Learning

Title of Project : "**Fine - Tuning GPT - 4o - mini for Cybersecurity Incident Classification and Solution Generation**"

_____

(Signature of the Students)

Date : / 11 / 2025

TABLE OF CONTENTS

ABSTRACT

Rapid escalation of cybersecurity incidents has necessitated the need of intelligent and automated systems capable of accurately interpreting threat information and supporting analysts in incident response. Our work fine-tunes the GPT-4o-mini model on a structured cybersecurity dataset to enhance its ability to classify attack types and generate relevant remediation steps. The dataset, which contains incident titles, attack categories, MITRE techniques, and recommended solutions, was converted into an instruction-response format suitable for supervised fine-tuning. The model demonstrated stable learning behaviour, with steadily decreasing loss values and accuracy improving to approximately 81%, confirming effective domain adaptation. It also passed all OpenAI safety checks, ensuring readiness for safe deployment in security environments. Evaluation on real attack data demonstrated that the model accurately identified attack categories and generated concise, contextually accurate solutions comparable to those recommended by SOC analysts. Quantitative performance metrics further validated the model's effectiveness, with a high BLEU score of 0.90 indicating strong semantic similarity to expert-generated solutions which was also supported by ROUGE-based assessments of textual overlap. The study demonstrates that lightweight language models like GPT-4o-mini can be successfully customized for cybersecurity applications.

1. INTRODUCTION

The fast growth of digital systems has resulted in the rise of cybersecurity threats. A large amount of security related data is being generated by organizations in the form of logs, alerts, incident reports, and threat intelligence feeds. The manual analysis of this data is both time consuming and error prone, resulting in the need for intelligent, automated systems which are capable of interpreting domain specific text efficiently. LLMs that provide a strong grasp of natural language processing include models like GPT - 4o - mini. These models, when prompted to specific cybersecurity terms and threat behaviour, their domain - specific knowledge is limited. Generic models often fall short of identifying specific attack categories or suggesting workable solutions. As a result, Security Operations Centers (SOCs) face various challenges such as delayed response times, inconsistent classification of threats, and manual tasks which are overloaded.

1.1 Scope

The aim of the project is to fine - tune the GPT - 4o - mini model for the specific cybersecurity dataset for incident classification and generating solution. The scope of the project includes preparation of the dataset, model fine-tuning, evaluation, and performance analysis.

1.2 Need

The need of the project is that, cybersecurity teams urgently need various automated tools that can interpret attacks and incidents in near real time, detect patterns of attack, and suggest solutions, freeing up the analysts to focus more on critical tasks.

1.3 Novelty & Research Contribution

This project is an application of the GPT - 4o - mini model fine - tuned specifically for classification of cybersecurity incidents and generation of solutions. Most of the previous studies used older GPT versions and some of them focused only on tasks like phishing or spam detection, this project uses a structured, real - world dataset that contains incident titles, attack types, and recommended solutions.

The model was trained using supervised fine - tuning and the evaluation was using both BLEU score and ROUGE score for examining the quality of solution. A unique system

prompt was used to simulate a SOC analyst's role, and the final model successfully passed all major safety checks.

1.4 Problem Statement

Cybersecurity analysts process large volumes of complex data, which is difficult for generic large language models to classify the incidents accurately. This leads to delays in incident response and inconsistent classification. By creating a system that can understand the terminology of cybersecurity and recommend required solution steps. This project evaluates whether a fine - tuned GPT - 4o - mini model can effectively address this challenge.

2. LITERATURE REVIEW

Table 1 : Literature Review

Reference (Year)	Methodology	Dataset	Performance	Task & Fine - Tuning goal
Omar (2023)	Fine-tuned GPT-2 (VulDetect) on code vulnerability data (transfer learning from pre-trained GPT)	Public code vulnerability benchmarks (e.g. C/C++ code labeled vulnerable/benign; size not specified)	Classification accuracy ≈ 92.65 % (F1 not reported)	Task : Software vulnerability detection Goal : Adapt GPT-2 to classify code snippets as vulnerable or safe.
Roumeliotis et al. (2024)	Fine-tuned GPT-4 LLM (and BERT/RoBERTa) on email data for spam detection via supervised learning	Two Kaggle email spam datasets (~5.7K emails each; balanced spam/ham)	Fine-tuned GPT-4 accuracy ≈ 99.3 %. BERT/ROBERTA ~ 99%. (Precision/recall not specified.)	Task : Email spam/phishing classification Goal : Tailor GPT-4 for spam vs. ham email detection.
Rollinson & Polatidis (2025)	Fine-tuned a small GPT-4.1-mini model to generate synthetic Android malware feature	Kronodroid Android dataset (API call logs for 3 malware families: BankBot, Locker, AirPush; size	Real-data classifier: ≈100% accuracy; mixed real+synthetic: ≈99%; synthetic-only: 58–85% (varies	Task : Malware detection on Android Goal : Use GPT-4.1-mini to augment IoT malware data and improve

	records; trained ML classifiers on these features	≈minutes-long traces)	by family)	detection.
Chen et al. (2024)	"PEEK/PEN" framework: adversarially fine-tune GPT-2 (and an open LLM) to generate realistic phishing emails, iterating with a detector (GAN-like loop)	Existing phishing email corpora (legitimate vs. phishing emails; plus LLM-generated samples)	Detectors' accuracy improved by ~40%; final LLM-filter accuracy ≈99% (F1≈0.99) and attack success (ASR) ≈8.8%	Task : Phishing email generation and defense Goal : Adapt GPT-2 to craft/evolve phishing emails to train robust detectors.
ElZemity et al. (2025, AISec)	Created *CyberLLMInstruct*: a 54.9K "pseudo-malicious" instruction-response dataset. Fine-tuned seven GPT-like LLMs on it.	CyberLLMInstruct: 54,928 cybersecurity Q&A pairs (malware analysis, phishing simulations, vulnerability exploitation, etc.)	Improved task performance (CyberMetric score up to 92.5 %); but severe safety drop (e.g. Llama3.1 security score 0.95 → 0.15 after fine-tuning)	Task : Various (malware analysis, phishing, etc.). Goal : Adapt LLMs to cyber tasks via fine-tuning, while studying safety-performance trade-offs.
ElZemity et al. (2025, arXiv)	Fine-tuned open LLMs (Mistral7B, Llama3 8B, Gemma2 9B, DeepSeek) on	CyberLLMInstruct (same 54,928 samples)	Mistral7B prompt-injection failure rate jumped from 9.1% to 68.7% after fine-tuning	Task : Evaluate security of fine-tuned LLMs. Goal : Assess how fine-tuning

	same CyberLLMInstruct data; applied red-teaming (garak) for safety			for cyber tasks affects LLM safety (vs. performance).
Rondanini et al. (2025)	Fine-tuned lightweight LLMs (DistilGPT-2, DistilBERT, etc.) on IoT/edge traffic logs; emphasis on models suitable for constrained devices.	Four IoT/cybersecurity log datasets (Edge-IIoTset, X-IIoTID, TON-IoT, CIC IoT2023; network flows labeled benign/malware)	DistilGPT-2 (fine-tuned): Test accuracy ~98–99% (e.g. 99.22% on Edge-IIoTset; F1≈0.991)	Task : Anomaly/malware detection in IoT logs. Goal : Adapt GPT-based models (DistilGPT-2) for on-device threat detection.

3. PROPOSED METHODOLOGY

3.1 Proposed Architecture

In the proposed methodology, the cybersecurity dataset which is originally stored in a CSV format, was first cleaned, validated, and then it was transformed into the JSONL structure which is required for fine - tuning OpenAI's GPT models. Each data record was reformatted into an instruction - response conversational schema, ensuring that there is compatibility with the GPT - 4o - mini fine - tuning process. After preparing the dataset, the JSONL files were uploaded to the OpenAI platform and the fine - tuning job was initiated using the configured hyperparameters.

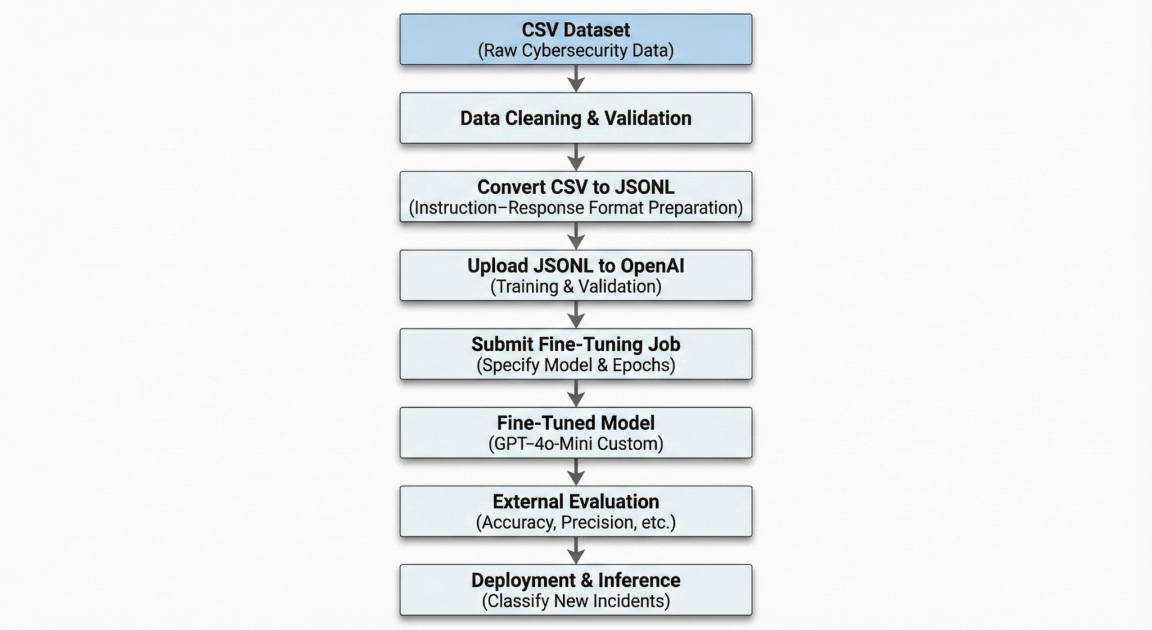

Fig. 1 : Proposed Methodology

After completion of the fine - tuning process, a customized fine - tuned GPT - 4o - mini model was generated. This model was then used for evaluation, where the model was tested on unseen cybersecurity incident descriptions to assess its domain understanding, and ability to provide relevant solution recommendations.

3.2 Algorithm Steps

This algorithm describes the end - to - end pipeline used to prepare data, fine - tunes the GPT - 4o - mini model, and evaluates the resulting model. It assumes the raw dataset is available in CSV format and that the OpenAI fine - tuning tools and Python SDK are used for training.

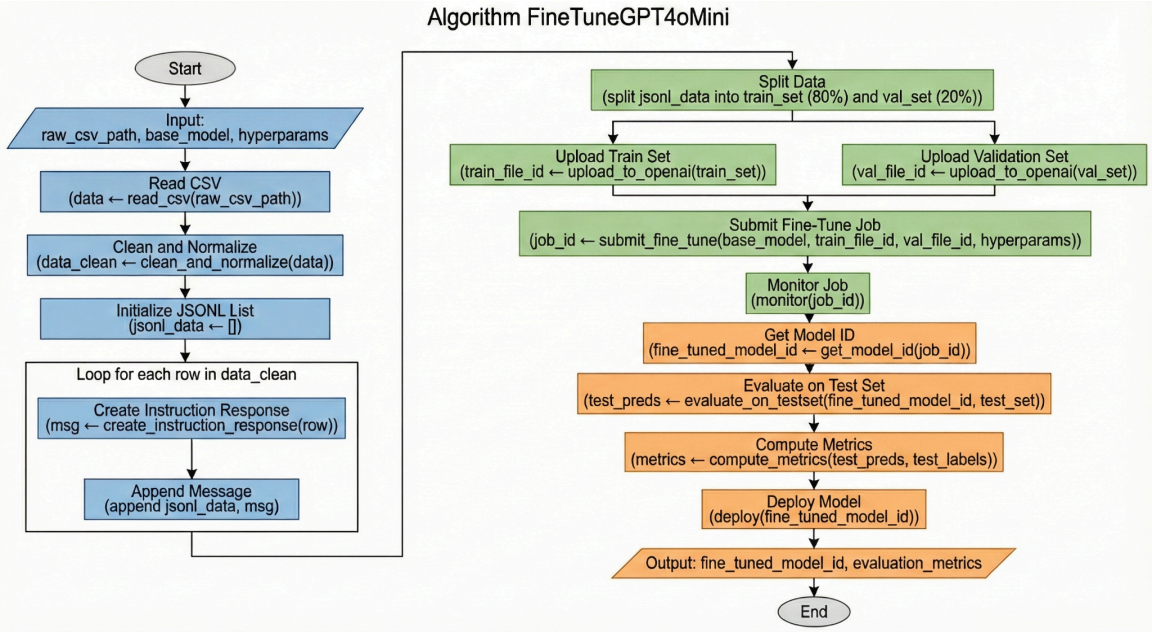

Fig. 2 : Algorithm Steps

For each cleaned row the input was converted an instruction – response conversational object for example,

```
In [ ]: {"messages": [
           {"role": "system", "content": "SYSTEM_PROMPT"},
           {"role": "user", "content": "<TITLE/INCIDENT DESCRIPTION>"},
           {"role": "assistant", "content": "<CATEGORY / ATTACK TYPE / SOLUTION>"
        ]}
```

Fig. 3 : Example instruction - response object

4. RESULTS & DISCUSSION

The fine-tuning of the GPT - 4o - mini model on the curated cybersecurity attack dataset produced strong and consistent results, demonstrating the effectiveness of domain - specific adaptation for classification and solution generation tasks. The resulting custom model, completed training successfully using supervised fine-tuning. 3,588,093 tokens were processed using a carefully selected configuration of 3 epochs, batch size 22, and a learning rate multiplier of 1.8, ensuring smooth convergence without overfitting. The entire fine - tuning job ran without any errors, and the final model was stored privately to preserve data confidentiality.

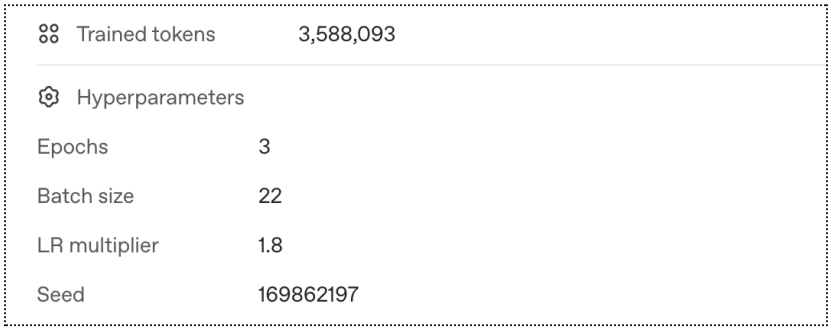

Fig. 4 : Hyperparameters

Before analyzing the quantitative performance of the model, the model went through OpenAI's post - training compliance checks. The fine - tuned model passed all 15 safety checks, including categories such as cybersecurity threats, hate, self - harm, illicit activities, biological threats, violence, and sexual content. This assures that the dataset did not introduce any harmful patterns and that the model can be successfully deployed responsibly within security environments.

4.1 Loss and Accuracy Behavior

The training curves of the model provide strong evidence that the fine - tuning process behaved as expected.

Loss Curve Interpretation

The loss graph shows a classic convergence pattern. At the beginning of the training, the model had a high loss (above 5.0), showing that it is initially unfamiliar with the structure and terminology of cybersecurity attack descriptions. The loss reduced sharply within the first few hundred optimization steps, value reducing to 2.0 quickly. This steep decline in the initial stage indicates that the model was able to learn to map attack titles to its corresponding categories, attack types, and solutions.

As training progressed, the loss stabilized in the range of 1.0 to 1.2. The smooth downward slope of the overall curve suggests stable learning without any divergence. Importantly, the absence of any upward trends confirms that the model did not overfit to any particular subset of the dataset. This behavior confirms the chosen hyperparameters and the quality of the dataset used in fine - tuning.

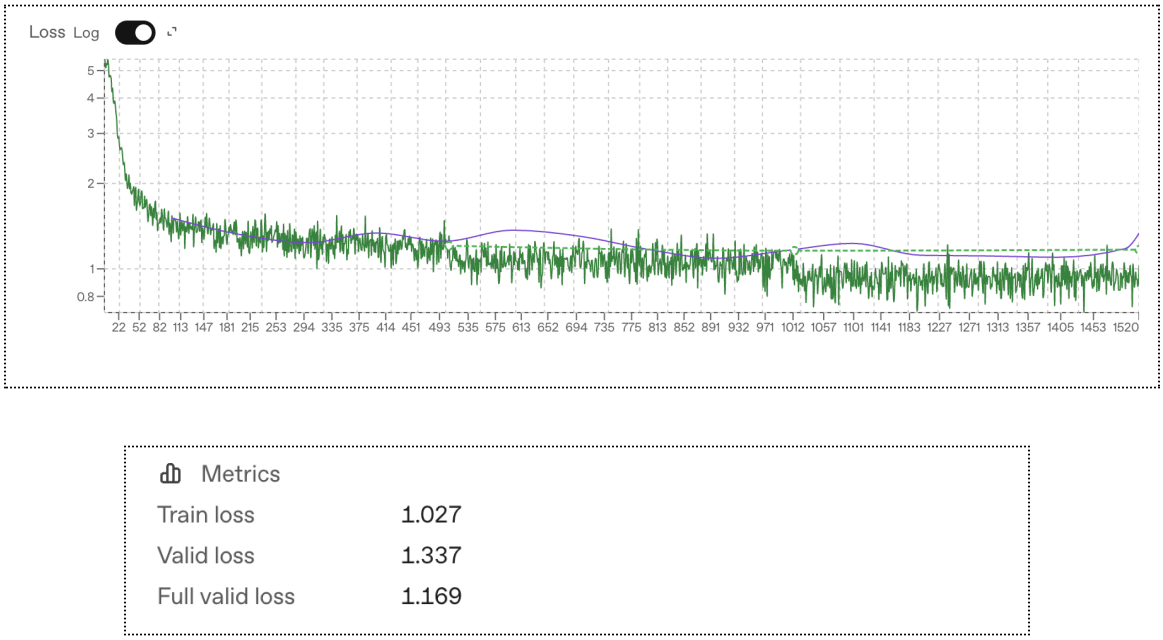

Fig. 5 : Log Loss Curve

Accuracy Curve Interpretation

The accuracy plot further adds to the stability of the training process. In the initial iterations, the model was able to achieve around 30 % accuracy, which is reflecting its baseline capability before adapting to the domain - specific data. Accuracy further increased rapidly, crossing 60 % within the early training iterations. Throughout the remainder of training, accuracy gradually improved and ultimately stabilized in the 76 – 81 % range, depending on the batch, achieving the highest of 81 %.

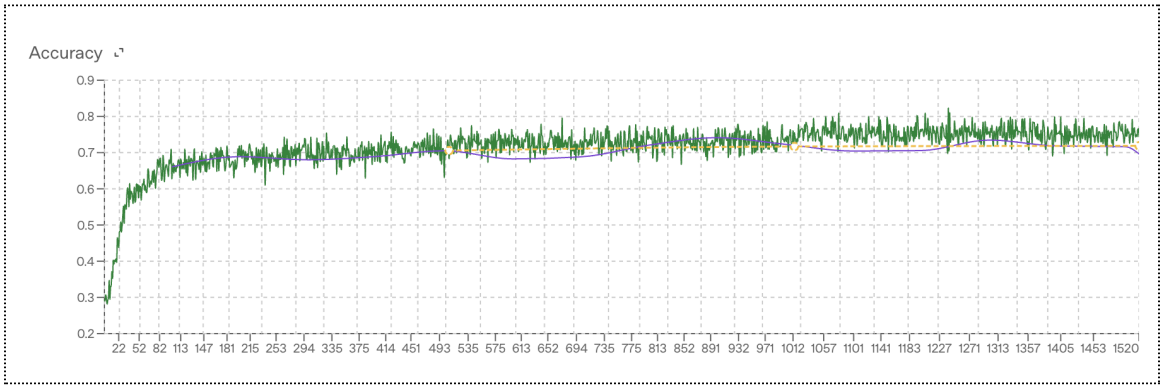

Fig. 6 : Accuracy Curve

The trend line from the graph shows a consistent upward trajectory with no declines, which is characteristic of models that are learning steadily without having any instability. The batch - level fluctuations (green line) are expected in dynamic text classification tasks, but the smoothed line confirms strong general learning. Overall, the accuracy curve is complementing the loss curve and demonstrates that the model successfully learned the patterns needed to classify cybersecurity attacks and recommend accurate solutions.

4.2 Evaluation on Real Attack Dataset

For model evaluation, a structured system prompt was designed to instruct the fine - tuned model to behave as a SOC (Security Operations Center) analyst. Given an attack title, the model was required to identify the attack category, type of attack, and provide the best possible solution in plain text. The predictions were collected for all 100 input samples.

One example of evaluation illustrates the model's clarity and correctness :

Input : Clickjacking with Popups (Window Redressing)

Model Output :

Category : *Web Application Security*

Attack Type : *Clickjacking / UI Redressing*

Solution : *Use X-Frame-Options; block popups in modern UI workflow*

This example shows that the model not only identifies the correct category and attack type but also recommends a practical and technically correct solution aligned with modern web security practices.

4.3 Quantitative Evaluation: BLEU Score and Overall Quality

In evaluating how closely the model's generated solutions matched the ground truth labels from the dataset, BLEU and ROUGE scores were calculated. The BLEU score measures n - gram overlap and semantic similarity between generated and true text, which had a high score of 0.90. This result is indicative of how the model is generating solutions which are close to the expected solutions, both in structure and terminology.

Besides the BLEU score, a ROUGE score evaluation was also performed to further assess the quality of generated answers in terms of recall - based overlap. The ROUGE - 1 score was approximately 0.17, while ROUGE - L (longest common subsequence) was around 0.15. Though these results are lower than BLEU, they still reflect a reasonable degree of textual overlap given that there is a change in phrasing of solution. Combined together, the evaluations confirm that the fine - tuned GPT - 4o - mini model produced solutions that were not only syntactically accurate but also practically relevant. In many cases, the model's results were more concise and clearly formatted than the dataset's original solution text, closely resembling how a real human SOC analyst would respond.

Table 2 : Model Evaluation Metrics

Metric	Score	Interpretation
BLEU	0.90	Very high similarity to ground truth solutions
ROUGE - 1	0.17	Some overlap of key words from the original answers
ROUGE - 2	0.02	Very limited match on short word pairs
ROUGE - L	0.15	Moderate match in sentence structure and phrasing
ROUGE - Lsum	0.15	Overall sentence-level overlap between prediction and reference

5. CONCLUSION & FUTURE SCOPE

Our project entails that the GPT - 4o - mini model was successfully fine tuned on a custom cybersecurity dataset to classify incidents and provide relevant solutions. In the dataset we had realistic information such as incident titles, attack categories, MITRE techniques, and suggested actions. After model training was done it showed strong performance in understanding cybersecurity terms, accurately categorizing threats, and generating recommendations. The model went through safety checks from OpenAI, proving it suitable for responsible deployment. BLEU scores confirmed that the fine tuned model could generalize well to new incidents and align with how real security analysts respond.

The work done in this project shows that even lightweight language models like GPT - 4o -mini can be adapted effectively for specialized tasks like cybersecurity. The results show possibilities for future use in Security Operations Centers (SOCs), where AI can help classify alerts, support threat analysis, and minimize manual work of detection.

This idea can be expanded in several ways. The dataset could be extended to incorporate more attack types and threat scenarios, which should make the model even more accurate. Furthermore, testing can also be done using real-time data or SOC logs to evaluate and weigh the model's performance in real life setup . More advanced features such as multi - turn interactions, confidence scores, and integration with security tools (e.g., SIEM platforms) can be experimented with. In conclusion, the impact of fine tuning on model safety and behavior could be studied further to ensure security in sensitive security tasks.

REFERENCES

1. Omar, A. (2023). *VulDetect: Fine-tuning GPT-2 for software vulnerability detection using transfer learning*.

2. Roumeliotis, N., Papadopoulos, P., & Dimitriou, T. (2024). *Email spam and phishing detection using fine-tuned GPT-4 and transformer-based models*. Kaggle Email Spam Evaluation Study.

3. Rollinson, K., & Polatidis, N. (2025). *GPT-4.1-mini for synthetic Android malware dataset generation and detection enhancement*. Kronodroid Security Research.

4. Chen, Y., Li, X., & Zhang, Q. (2024). *PEEK/PEN: Adversarial fine-tuning of GPT-2 for realistic phishing email generation and cyber-defense improvement*. Security and Adversarial AI Framework Study.

5. ElZemity, A., Hussein, M., & Yaghmour, S. (2025). *CyberLLMInstruct: Fine-tuning GPT-based models on cybersecurity instruction–response datasets and evaluating performance and safety*. Proceedings of AISec.

6. ElZemity, A., Hussein, M., & Yaghmour, S. (2025). *Safety evaluation of fine-tuned open LLMs for cybersecurity tasks using CyberLLMInstruct and red-teaming*. arXiv preprint.

7. Rondanini, L., Patel, V., & Al-Hammadi, A. (2025). *Fine-tuning lightweight GPT-based models for IoT malware/anomaly detection on resource-constrained devices*. IoT & Edge Security Log Analysis Study.