IBM Innovation Center for Education

PHEMESOFT Software That Matters
TRUSTED DELIVERY PARTNER for IBM ICE

YENEPOYA
(Deemed To Be University)

CYBER GUARD AI

PROJECT SYNOPSIS

CYBER SECURITY AWARENESS CHATBOT

BACHELOR OF SCIENCE
Cyber forensics, Data analytics and
Cyber security

SUBMITTED BY:

Devika KB - 22BSCFDC14
Dhrishya CM - 22BSCFDC15
Govind B - 22BSCFDC18
Sneha Sanjay- 22BSCFDC40
Umamaheswari M - 22BSCFDC44

GUIDED BY :
Mr.Sashank

TABLE OF CONTENTS

Sl. No.	**Content**	**Page No.**
1	Introduction	3
2	Literature Survey	4
3	Methodology / Planning of Work	5
4	Facilities Required	6
5	References	7

INTRODUCTION

In recent years, cyber threats such as phishing attacks, malware infections, and data breaches have become increasingly common, affecting individuals, businesses, and governments alike. There is a growing need for intelligent tools that can assist users in identifying and responding to such threats quickly and accurately. Cyber security awareness chatbot is a smart, AI-powered cybersecurity chatbot developed to address this need by providing real-time assistance in recognizing, understanding, and reporting online threats.

This project falls under the specialized domain of Cybersecurity and Artificial Intelligence, with a strong focus on threat detection, user awareness, and automated incident reporting. The system is designed to offer both educational support helping users understand various cyber threats and functional capabilities, such as analyzing suspicious emails, URLs, or messages through live threat intelligence.

CyberGuard AI leverages the Meta Llama 3.1-8B-Instruct language model, a powerful open-source large language model capable of understanding and generating human-like text. To make the chatbot domain-specific, it is fine-tuned on the Purple-Team Cybersecurity Dataset, a curated dataset containing cybersecurity-related scenarios and queries. This enables Cyber Aware AI to deliver highly relevant and accurate responses within the field of cybersecurity.

The chatbot is further enhanced by integrating multiple cyber threat intelligence APIs:

- PhishTank API – used to detect phishing URLs.
- VirusTotal API – performs real-time scans on files and links to identify malware or malicious behavior.
- Have I Been Pwned (HIBP) API – checks if an email address has been involved in any known data breaches.

A unique feature of the project is its incident reporting module, which allows users to report suspicious activity directly to official platforms such as the Indian Cyber Crime Portal (cybercrime.gov.in). This is achieved using secure email automation through Python's smtplib.

The entire project is implemented in a cloud-based development environment using Google Colab. The system is built with:

- Python 3.11+ as the core programming language
- Hugging Face Transformers for language model operations
- PEFT (Parameter-Efficient Fine-Tuning) and bitsandbytes for efficient model training
- FastAPI and Uvicorn for deploying the chatbot as a web service

Cyber Guard AI is designed to be modular, scalable, and accessible, making it ideal for both educational purposes (for students, cybersecurity learners) and real-world applications (for users looking to verify and report digital threats). It represents a hybrid solution that combines conversational AI, cybersecurity expertise, and real-time threat intelligence into a single, user-friendly platform

LITERATURE SURVEY

Artificial Intelligence (AI) has emerged as a transformative force in the field of cybersecurity, enabling faster, more accurate detection and response to evolving digital threats. Leading enterprise platforms such as Darktrace, IBM QRadar Advisor with Watson, and Microsoft Security Copilot leverage machine learning (ML) and natural language processing (NLP) to automate threat identification, analyze large-scale security logs, and assist analysts in decision-making. However, these solutions are generally proprietary, cost-intensive, and closed-source, limiting their accessibility for students, educators, and small organizations.

Conversational AI platforms like ChatGPT, Claude, and Bard have made significant progress in understanding human language, but they are built for general-purpose dialogue and lack domain-specific knowledge or cybersecurity integration. These systems cannot perform live threat analysis or validate URLs, files, or emails against current threat databases, which limits their usefulness in practical cybersecurity settings.

Academic research has attempted to bridge this gap. Various studies have explored phishing detection, malware classification, and threat prediction using techniques such as decision trees, support vector machines, and deep learning models. However, many of these efforts are constrained to static datasets and are limited to offline classification tasks without a real-time feedback loop, user interaction, or automated reporting capabilities.

CyberGuard AI stands apart from existing solutions by combining:

- A powerful open-source LLM (Meta Llama 3.1-8B-Instruct), fine-tuned on the Purple-Team Cybersecurity Dataset for domain-relevant Q&A,
- Integration with real-time threat intelligence APIs, including:
  - PhishTank for phishing URL detection,
  - VirusTotal for malware and file scanning,
  - Have I Been Pwned (HIBP) for breach lookups,
- An incident reporting module that formats and sends structured reports to official platforms like cybercrime.gov.in, promoting actionable response.

By fusing educational support, live threat detection, and automated reporting in a modular and open-source framework, CyberGuard AI provides a unique and practical contribution to both cybersecurity research and public digital safety. Its accessible design makes it suitable for academic learning environments, cybersecurity training, and real-world use cases**.**

METHODOLOGY / PLANNING OF WORK

The workflow is divided into eight sequential phases, each focusing on a key functional aspect of the system:

Phase 1: Environment Setup:
The development environment uses Google Colab with GPU for faster processing and installs key libraries like Transformers, PEFT, Accelerate, and FastAPI for NLP and web deployment.

Phase 2: Base Model Loading:
The Meta Llama 3.1-8B-Instruct language model is loaded via an authenticated Hugging Face account. A simple inference loop is added to verify that the base model can process inputs and generate valid responses before fine-tuning.

Phase 3: Dataset Preparation:
The Purple-Team Cybersecurity Dataset is curated into a Q&A format, with prompt-response pairs. This step ensures that the training data is contextually relevant to cybersecurity, covering topics like phishing, malware, scam detection, and digital hygiene.

Phase 4: Fine-Tuning:
Model fine-tuning is carried out using LoRA (Low-Rank Adaptation) via the PEFT library. This allows efficient training while reducing computational costs. The fine-tuned model gains domain-specific language capabilities for cybersecurity threat detection and explanation.

Phase 5: API Integration:
CyberGuard AI is connected to live threat intelligence APIs:

- PhishTank: Checks for phishing URLs.
- VirusTotal: Scans links and files for malware.
- Have I Been Pwned (HIBP): Looks up breached email addresses.

This integration allows the chatbot to validate and respond to real-world security concerns.

Phase 6: Reporting Module:
A secure reporting system using Python's smtplib sends user reports to platforms like cybercrime.gov.in, promoting cyber safety.

Phase 7: API Deployment:
A FastAPI backend is built to serve the chatbot and its utilities. Key routes include:

- /chat for conversation handling,
- /check_url for link analysis,
- /report

This modular API structure enables flexible and scalable deployment.

Phase 8: Testing & Documentation:
The final phase includes:

- Thorough testing of model output, API calls, and user interactions,
- Validation of threat detection performance,
- Documentation of project setup, codebase, and deployment procedure for reproducibility.

FACILITIES REQUIRED

To successfully develop and deploy Cyber Guard Ai, the following software and hardware facilities are required:

Software Requirements:

Component	Purpose
Google Colab	Cloud-based platform with GPU acceleration for model training and testing.
Python 3.11+	Main programming environment for model logic, API handling, and integration.
Transformers (Hugging Face)	Used for loading the Llama 3.1-8B-Instruct model and running inference.
PEFT / bitsandbytes	Enables efficient LoRA-based fine-tuning with minimal resource usage.
FastAPI + Uvicorn	Backend framework for deploying the chatbot with accessible API routes.
smtplib / requests / bs4	Support modules for reporting incidents and integrating live threat APIs.

Hardware Requirements:

Hardware	Description
Laptop or PC	Local development system with a minimum of 8GB RAM for coding and testing.
Internet Connection	Required for accessing cloud resources, APIs, and model updates.
GPU (via Google Colab)	Used for training and fine-tuning the model to improve performance speed.

Note:
No dedicated high-end hardware is necessary for deployment. The entire project can be run using Google Colab or any basic cloud server, making it both accessible and cost-effective.

REFERENCE

1. Meta Llama 3.1-8B-Instruct – Hugging Face.
https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

2. Canstralian Purple-Team Cybersecurity Dataset – Hugging Face.
https://huggingface.co/datasets/Canstralian/Purple-Team-Cybersecurity-Dataset

3. PhishTank API – Open source phishing threat intelligence.
https://www.phishtank.com/

4. VirusTotal API Docs – File and URL reputation scanning.
https://developers.virustotal.com/

5. Have I Been Pwned (HIBP) – Email breach check service.
https://haveibeenpwned.com/API/v3

6. CERT-IN Security Alerts – Indian CERT incident reports.
https://www.cert-in.org.in/

7. FastAPI Documentation – Python web framework.
https://fastapi.tiangolo.com/