Project Goal

Develop a campus-relevant, open-source AI-augmented SOC prototype using **synthetic/anonymized CSUSB security datasets**.

Focus: log summarization, alert triage, and report generation using LLMs + agent frameworks.

---

1. Project Teams and Roles

- **Team 1 — SOC Infrastructure:** Deploy Wazuh + Elastic/OpenSearch, TheHive + Cortex, Zeek/Suricata, and Docker Compose
- **Team 2 — Log Summarization & LLM Integration:** LibreLog, LogBatcher, LILAC, LLaMA/Mistral models
- **Team 3 — Alert Triage & Orchestration:** LangChain agent, Chroma vector DB, FastAPI microservices
- **Team 4 — Report Generation:** AGIR, GenDFIR, AttackGen
- **Team 5 — Testing & Evaluation:** Metrics, benchmarks, dashboards, documentation

Each team has primary responsibilities but should collaborate closely for integration.

---

2. Required Tools and Resources

SOC Infrastructure:

- Wazuh (IDS/SIEM)
- Elastic/OpenSearch (log storage & querying)
- TheHive + Cortex (incident response & ticketing
- Zeek / Suricata (network traffic analysis)
- Docker Compose (deployment & reproducibility)

AI/LLM Tools:

- LibreLog, LogBatcher, LILAC (log parsing)
- LLaMA 3 8B / Mistral-7B (local LLM models
- LangChain or LangGraph (agent orchestration)
- ChromaDB (vector database)
- FastAPI (microservices)

Datasets:

- Public/academic: CICIDS 2017, UNSW-NB15, NSL-KDD, Kyoto 2006+, LogHub-2.0
- Campus-specific: CSUSB synthetic or anonymized logs (after approval)

---

3. Step-by-Step 3-Month Roadmap

Month 1 — SOC Setup

Objectives: Deploy basic SOC infrastructure and ingest sample datasets.

Tasks:

1. **Install & configure:**
    - Wazuh manager & agents
    - Elastic/OpenSearch with dashboards
    - TheHive + Cortex for incident/ticket management
    - Zeek / Suricata for network traffic monitoring
2. **Dockerize** each component for reproducibility
3. **Replay sample datasets:**
    - CICIDS 2017 network traffic
    - LogHub system logs
4. **Verify ingestion & dashboards:**
    - Ensure logs appear in Elastic
    - Test alert generation via Zeek/Suricata

5. **Documentation:**
    - Record setup steps, config files, Docker Compose structure

---

Month 2 — AI Augmentation

Objectives: Add AI/LLM modules for log summarization, alert triage, and report generation.

Tasks:

1. **Log Summarization:**
    - Integrate LibreLog / LogBatcher / LILAC
    - Process logs from Month 1
    - Output: concise summaries for each log batch
2. **Alert Triage:**
    - Implement LangChain agent
    - Connect to Chroma vector DB
    - Automate classification & prioritization of alerts
    - Output: reduced false positives, alerts with context
3. **Report Generation:**
    - Implement AGIR / GenDFIR / AttackGen
    - Convert alerts and summaries into human-readable reports
    - Output: ready-to-read incident reports
4. **Integration:**
    - Connect log summarization → alert triage → report generation pipeline
5. **Testing:**

    - Run small-scale test with CICIDS + synthetic CSUSB logs

6. **Documentation:**

    - Include API endpoints, configuration, and usage instructions

---

Month 3 — Testing, Benchmarking, and Demo

Objectives: Evaluate system performance, optimize, and prepare deliverables.

Tasks:

1. **Evaluation Metrics:**

    - Log summarization: compare summaries with human-annotated logs → accuracy or BERTScore

    - Alert triage: false positives, precision/recall, F1-score

    - Report generation: time saved per incident, completeness of details

2. **Benchmarking:**

    - Run both baseline SOC (without AI) and AI-augmented SOC

    - Compare metrics for measurable improvement

3. **CSUSB Scenario Simulation:**
    - Use synthetic/anonymized CSUSB traffic for realistic evaluation
4. **Demo Preparation:**
    - Dockerized AI-SOC ready for presentation
    - Slide deck summarizing methodology, results, and learning outcomes
5. **Final Documentation:**
    - Detailed report of setup, AI integration, testing, and results
    - GitHub repository with Dockerfiles, notebooks, datasets (where allowed)

---

4. Deliverables

- Fully **Dockerized AI-augmented SOC**
- Benchmarked **performance metrics**

- **Incident reports** generated by the system

- **Final presentation slides**

GitHub repository with scripts, configs, and documentation