Final Framework and Model Selection

This document provides the final selection and rationale for the key components chosen for the development of a locally operable, CPU-compatible, cybersecurity-focused chatbot using Retrieval-Augmented Generation (RAG).

1 Selected Large Language Model (LLM)

TinyLlama-1.1B-Chat-v1.0

- Highly compact model (1.1 billion parameters) optimized for local, CPU-only environments.
- Ideal balance between performance and resource constraints on standard laptops.
- Suitable for educational conversational tasks, especially when supplemented by RAG.
- Fully open-source with permissive licensing for both educational and commercial use.

2 Selected RAG Framework

LangChain

- Highly flexible and modular, supporting complex conversational workflows and advanced integrations.
- Extensible for future integration with additional cybersecurity tools or modules.
- Strong community support and extensive documentation, beneficial for development and maintenance.
- Open-source, licensed under the MIT License.

3 Selected Vector Database

FAISS

- High-performance, optimized for fast retrieval operations on CPU.
- Minimal resource usage, ideal for embedding and querying cybersecurity educational resources.
- Simple integration and efficient management of embeddings.
- Fully open-source, MIT licensed.

4 Summary of the Selected Configuration

The final architecture consists of:

- **LLM:** TinyLlama-1.1B-Chat-v1.0 – chosen for optimal local inference capability.
- **Framework:** LangChain – provides flexibility, modularity, and future extensibility.
- **Vector Database:** FAISS – delivers superior retrieval performance and efficient local operations.

This configuration is specifically chosen to meet the requirements of local operation, CPU efficiency, ease of integration, open-source compliance, and effective educational support in cybersecurity contexts.