Agentic AI Red-Teaming Assistant (NIM-Hosted, No-Local-GPU)

Description:

Agentic Red-Teaming MVP using NVIDIA hosted NIMs only (no local weights).

Planner/Executor uses Llama-3.1-Nemotron-Nano-8B-v1 via OpenAI-compatible NIM API.

Retriever uses a Retrieval Embedding NIM (e.g., NV-Embed-QA).

Deployable via AWS KIRO on EKS or SageMaker endpoints.

Objectives

- Use NVIDIA NIM endpoints only (OpenAI-compatible)

- Reasoning model: nvidia/llama-3.1-nemotron-nano-8b-v1

- Retrieval Embedding NIM: NV-Embed-QA (or equivalent NeMo Retriever)

- Generate structured vulnerability reports

- Deploy as container or direct hosted API

Architecture

- **CoordinatorAgent:** Orchestrates mission (plan → execute → evaluate)

- **AttackPlannerAgent:** Generates adversarial prompts (Nemotron NIM)

- **RetrieverAgent:** Embeddings via NV-Embed-QA NIM, stored in FAISS

- **ExecutorAgent:** Executes prompts using NIM chat API

- **EvaluatorAgent:** Classifies vulnerabilities, ranks severity

Deployment: Dockerized for EKS / SageMaker.

Storage: Local FAISS index, logs to S3.

Security: IAM roles, S3 restricted access.

Environment Variables

- NVIDIA_API_KEY

- NIM_BASE_URL = https://integrate.api.nvidia.com/v1

- NIM_LLM_MODEL = nvidia/llama-3.1-nemotron-nano-8b-v1

- NIM_EMBED_MODEL = NV-Embed-QA

- RESULTS_BUCKET

- STOP_TEST = 0

Deployment Notes

- No GPU required; all inference is through NIM endpoints.

- KIRO will generate orchestrator code, Dockerfile, and deployable containers.

- SageMaker or EKS recommended for scalable testing.

Governance & Safety

- Only run with explicit authorization.

- STOP_TEST flag halts all red-teaming runs.

- Logs must exclude PII and API secrets.