

Spencer Presley

spencerpresley96@gmail.com | linkedin.com/in/spencerpresley96 | github.com/SpencerPresley

EXPERIENCE

CrunchAtlas — Software Engineer | Mar 2025 – Present

Local-first, air-gapped AI security platform for critical infrastructure. Sole author of the systems below.

- Built the **PCAP analysis agent**, the platform's most-used investigation agent: an autonomous LLM works a packet capture end to end (protocol recon, DNS/HTTP/TLS, file carving, IOC extraction, timeline) into a cited report and triage verdict. It reasons only over the output of a real forensic toolchain it drives (tshark, Zeek, foremost), never raw packets, so every value it reports traces to tool evidence.
- Built **PurpleHaze**, an autonomous penetration-testing product: a local LLM drives Metasploit and real offensive tooling through a full multi-step engagement, holds session state across long runs, and produces compliance-grade OWASP and NIST reports, with no human in the loop.
- Built the in-house **agent runtime** every analysis agent runs on (LangChain/LangGraph): context compaction that counts real tokens and fails closed instead of overflowing, plus tool-output eviction, loop detection, and per-run container sandboxing (fresh, network-isolated, self-reaping).
- Built the **Postgres-native job queue** that replaced Celery for the platform's GPU, LLM, and sandbox jobs, so a burst of heavy AI work can't OOM an air-gapped box and any one job can be cancelled without killing its worker. It runs on atomic per-lane claiming (`SELECT FOR UPDATE SKIP LOCKED`), priority-FIFO ordering, dependency DAGs with cycle detection, and lock-free single-flight.
- Built the **local LLM serving stack** (llama.cpp behind llama-swap) with a byte-exact VRAM fit-proof: a Python model reproduces llama.cpp's KV-cache allocation to the exact byte across grouped-query attention, sliding-window layers, and per-K/V quantization, so a deployment's fit is proven in CI with no GPU or running server.
- Built **GPU-memory admission control** so concurrent LLM jobs can't overcommit one card: a per-pool Postgres advisory lock admits a job only when its declared context-token cost fits the pool's live budget, and a runtime `READ COMMITTED` requirement makes a claimer that waited on the lock re-price against just-committed usage instead of a stale snapshot.
- Built the agent that maps captured network activity to **MITRE ATT&CK**: capture-time validation rejects technique IDs the model hallucinates, and a measurable retrieval eval (`Recall@k/MRR` with regression floors) gates changes before any agent runs.

Horn Point Laboratory — Software Engineering Intern | May 2024 – Aug 2024

- Built a **retrieval-augmented chatbot** for an NSF-funded Chesapeake Bay salinity project (SaltCast) at Horn Point Laboratory: FAISS retrieval over OpenAI embeddings feeding a streaming FastAPI/LangChain service; presented at the MIT Undergraduate Research Technology Conference.

PROJECTS

BitNet b1.58 (from scratch) — *PyTorch, ternary-weight LM, byte-level, Android*

- Reimplemented and trained BitNet b1.58 ternary-weight (`{-1, 0, +1}`) language models from scratch at three sizes up to 300M parameters (byte-level, on enwik8/9), and deployed a trained checkpoint to an Android phone (Chaquopy), profiling on-device CPU, memory, and energy — undergraduate research on quantization-native models for edge inference.

gloss — *Python, SQLite/FTS5, Ollama*

- Built a hybrid retrieval engine that fuses BM25 with a local-embedding vector channel via reciprocal-rank fusion to return a source document's verbatim, cited passages — tuned against a 266-case evaluation gated by a paired significance test (`hit@5` 0.94, `hit@1` 0.71).

HONORS

- **Best Educational Hack & 2nd Overall — HackUMBC 2024** — Testif.AI
- **1st Place / Best Hack — Bitcamp 2024**, UMD (largest US East Coast hackathon) — TermBook

SKILLS

- **AI & LLM**: LangGraph/LangChain, llama.cpp, llama-swap, Ollama, OpenAI API, RAG, FAISS/vector search, model fine-tuning, LLM serving, GPU memory management, tiktoken
- **Languages**: Python, Rust, TypeScript, SQL, Bash
- **Backend & Infrastructure**: Django, FastAPI, PostgreSQL, MongoDB, Redis, Celery, Docker, Linux, AWS

EDUCATION

Salisbury University — B.S. Computer Science, Minor in Mathematics | Aug 2021 – May 2025