Connor M. Carpenter

connorcarpenter15@gmail.com linkedin.com/in/connorcarpenter15 github.com/connorcarpenter15 connorcarpenter15.com

Education

Carnegie Mellon University, School of Computer Science — Pittsburgh, PA
Bachelor of Science in Computer Science; Concentration in Machine Learning — *Expected May 2027*

- Cumulative GPA: 3.88/4.00
- **Selected Coursework:** Deep Learning Systems, Machine Learning with Large Datasets, Deep Reinforcement Learning & Control (Fall 2026); Database Systems, Parallel and Sequential Data Structures and Algorithms, Neuro-Symbolic AI

Experience

NVIDIA — Santa Clara, CA
Deep Learning Algorithms Engineering Intern, Dynamo — *May 2026 - Present*

- Designed Dynamo's sidecar architecture across vLLM, SGLang, and TensorRT-LLM, preserving native engine entrypoints while adding out-of-process routing, lifecycle management, and distributed orchestration.
- Contributed upstream to the vLLM, SGLang, and TensorRT-LLM open-source projects while co-designing sidecar lifecycle, failure-recovery, multimodal, and KV-cache contracts with each engine team.
- Created and open-sourced OpenEngine, a vendor-neutral gRPC/Protobuf protocol defining typed APIs for inference, discovery, health, abort, drain, and KV coordination across engines and distributed frameworks.

Persona Machines — Pittsburgh, PA
Co-Founder & Chief Technology Officer — *May 2025 - Feb. 2026*

- Engineered the end-to-end DeepCloak MVP from scratch, deploying a responsive SvelteKit frontend on Vercel and a high-concurrency FastAPI backend on Railway backed by Neon (PostgreSQL) and SQLAlchemy.
- Under the advisement of Professor Ruslan Salakhutdinov (ex-VP of Research, Meta), architected the platform's novel privacy protocols, leveraging this technical differentiation to secure acceptance into the NVIDIA Inception program.

NASA Langley Research Center — Hampton, VA
Dynamic Systems & Controls Intern — *Jun. 2024 - Aug. 2024*

- Programmed software in MATLAB for the calculation of Sliced-Normal distributions which characterize complex, multivariate data sets with multiple modes and strong dependencies.
- Created a library of custom functions for enhancing existing datasets through virtual data augmentation.

The Center of Excellence for Engineering Biology — New York, NY
Software Engineering Intern, Genome Project-write — *Jun. 2023 - May 2024*

- Developed and refined core features for the computer-aided design (CAD) platform for genome editing within an agile team, directly supporting an international community of over 200 researchers.
- Constructed and deployed serverless backend services using Python and AWS Lambda to power the platform's genome synthesis and editing capabilities.

Thomas Jefferson National Accelerator Facility — Newport News, VA
Research Intern — *Jun. 2023 - Aug. 2023*

- Processed and analyzed large-scale experimental data from electron-proton collisions, developing scripts in C++ and Python within the ROOT framework to deliver a new estimation of the proton charge radius.
- Designed and built an educational platform featuring tutorials and code examples to streamline and accelerate the onboarding of new researchers onto the lab's ROOT analysis software.

Projects

Relational Database Engine (BusTub) | *C++, DBMS Internals, Concurrency; CMU 15-445*

- Implemented BusTub's storage layer in C++, including a thread-safe buffer pool with ARC replacement and asynchronous disk scheduling, plus a concurrent B+ Tree supporting splits, merges, ordered scans, and tombstone-buffered deletion.
- Built vectorized SQL executors and rule-based optimizer transformations for joins, aggregation, external merge sort, window functions, and index scans; added optimistic MVCC with undo-log version chains and snapshot/serializable isolation.

Dynamic Memory Allocator | *C, Systems Programming, GDB; CMU 15-213*

- Implemented a 64-bit C allocator with malloc, free, realloc, aligned blocks, and heap-consistency validation.
- Built segregated free lists with block splitting and immediate coalescing, achieving a 100.0 performance index, 8,136 Kops throughput, and 74.1% memory utilization.

Skills

Languages:	Python, C, C++, Rust, TypeScript, SQL, Java, SML, MATLAB, R, HTML/CSS
AI & Inference:	Inference Serving, PyTorch, Dynamo, OpenEngine, vLLM, SGLang, TensorRT-LLM, gRPC, Protobuf
Systems & Data:	Distributed Systems, Fault Tolerance, PostgreSQL, SQLAlchemy, asyncio, concurrency, DBMS internals
Web & Cloud:	SvelteKit, FastAPI, REST, OpenAPI, API Design, React, Node.js, AWS Lambda, Docker, Kubernetes
Developer Tools:	Linux, Git, GDB, GTest, Vim, CI/CD
Spoken Languages:	English, Spanish (Conversational), French (Conversational)