LUKE ROULEAU

(863) 608 - 1393 lukerouleau.com in: in/luke-rouleau rouleauluke@gmail.com

EDUCATION

University of Florida, *Herbert Wertheim College of Engineering*, Gainesville, FL **May 2022**
Bachelor of Science in Computer Engineering, Summa Cum Laude GPA: 4.0/4.0

SKILLS & INTERESTS

Skills

- **Deep Learning –** PyTorch, MLIR, Deep Neural Architecture and Design, Deep Learning Compilers, Deep Learning Hardware
- **Project Management** – Experienced in navigating cross-functional teams, delivering complex technical solutions under time constraints, and leading development from concept to deployment
- **Product Development**– Developed customer-focused AI solutions with proven market adoption, experience in technical product management, and bringing complex AI systems to commercial deployment
- **UX Design** – Meticulous attention to detail in creating seamless user experiences

Interests – Robotic Foundation Models, End-to-End Neural Robot Control, Deep Neural Architectures, AI Productization
Languages of Highest Proficiency – Python, C++, Svelte(Kit), HTML/CSS/JavaScript

PROFESSIONAL EXPERIENCE

Untether AI, ***Senior Deep Learning Engineer,*** Toronto, ON **May 2024 – Present**

- **Compiler Frontend Team –** Developed PyTorch ingestion, optimization, and quantization algorithms. Specialized in transformer attention head optimization. Increased BERT model performance by 12x by untethering its spatial architecture.
- **Compiler Integration Test Framework Developer –** Led a team of 3 in developing Untether's compiler test framework, which tests compilation against ~1200 popular networks (CNNs, ViTs, LLMs), enables automatic functional-correctness bug pinpointing, and supports various developer debug tools. This system increases network coverage by 1100%, reduces debug cycle time by over 50%, and lowers regression detection time from 1 week to 1 day.
- **MLIR Generative Compiler Team –** Contribute to Untether's MLIR for performant kernel auto-generation. Co-lead strategy for enabling broad Vision Transformer and LLM support on Untether's architecture.

ReCODE Medical, ***Founder & CTO,*** Houston, TX **September 2024 – Present**

- **Founder –** With co-founder and CEO Dr. Matthew Segar, I led product development of ReCODE Medical, which develops industry-leading medical coding solutions (ReCODE Chat). We serve customers like St. Luke's and Kelsey-Seybold Clinics.
- **ReCODE Chat –** Developed, from inception to deployment, an AMA-licensed medical coding co-pilot with 90% preference rate among physicians and professional coders compare to other flagship LLMs (OpenAI and Anthropic).

Metalware, ***Software Engineer, Consultant,*** San Francisco, CA **December 2023 – May 2024**

- **Datasheet Reader Redesign –** Proposed and delivered a re-architecting of Metalware's Datasheet Reader application by migrating from server-side to client-side RAG implementation**,** achieving a 95% reduction in query failures and 50% latency improvement through optimization.

Texas Instruments, ***Machine Learning Lab Systems Engineer,*** Dallas, TX **May 2022 – May 2024**

- **Custom TVM Deep Learning Compiler –** Successfully developed an MVP of a custom TVM compiler that translates PyTorch models into int8 quantized C code for a custom RISCV CPU + Accelerator, hitting 75% theoretical throughput on key-word detection and segmentation applications.
- **Internal Coding Copilot for TI SDKs –** Developed TI Copilot tool by Low-Rank Adaptor fine-tuning of Llama 2 and RAG against TI SDKs. Served the model internally by exposing an OpenAI compliant API endpoint with integrate with VSCode.
- **RAG Application for Technical Customer Support –** Implemented the first retrieval-augmented technical customer support chatbot used in the E2E forums, on top of OpenAI APIs.

LEADERSHIP & INVOLVEMENT

RoboTIcs Volunteer, Dallas TX **June 2023 – May 2024**
Volunteer – Mentored biweekly at local elementary schools though the TI Robotics mentorship program

Warren B. Nelms Cybersecurity Institute, Gainesville FL **January 2020 – August 2022**
Undergraduate Researcher – Under Dr. Bhunia, created manufacturing techniques to encode invisible barcodes detectable via Nuclear Quadrupole Resonance and apply IR Spectroscopy to validate supply chain integrity in developing countries

IEEE Design Team, Gainesville FL **August 2021 – May 2022**
Software Team Lead – Led robotic software development, placed 3rd out of 60+ schools in IEEE SouthEast Con 2022

AWARDS

University of Florida Dean's List	**August 2018 – May 2022**
IEEE SouthEast Con 2022 Hardware Design Competition, 3rd Place	**May 2022**
UF Electrical & Computer Engineering Outstanding Student Award	**April 2022**
National Science Foundation ***Research Experience for Undergraduates*** **Fellowship Grant**	**June 2019 – March 2020**