Nathaniel Morgan
Cambridge, MA| 864-324-8208 | nmorgan@csail.mit.edu | n-morgan.github.io
Curriculum Vitae

Education

Massachusetts Institute of Technology (MIT), Cambridge, MA — May. 2027
Candidate for Bachelor of Science | Major 6-3 Computer Science and Engineering — GPA: 4.5
Relevant Coursework: Deep Learning, Mathematics for Computer Science,
Data Structures and Algorithms, Linear Algebra, Machine Learning, Probability & Statistics

Skills

Computer Science: Python, Java Script, Bash, SQLite,
DevOps: Google Cloud Platform, Kubernetes, Docker, Kubectl, Minikube, Tmux, Vim
Software: Google G-suite, Microsoft 365 Office, Adobe Creative Cloud
Languages: English, Spanish

Internships

Subconscious.AI Boston MA — May 2025 – Present
AI Research Engineer
Benchmarked SGThread system and worked to deploy novel inference engine and RL trained model to serverless platform

- Containerized and modified fork of SGLang, facilitating rapid deployment of TIM reasoning model
- Created python interface for Google DataCommons benchmarking of modified inference engine and TIM model through the creation of constrained decoding specifications and a new DC LLM tool

BitEnergyAI, Cambridge MA — Jan. 2025 - May 2025
Backend / Kernel Developer
Developed code for FP8 E4M3 multiplication on FPGA in C++ with Vitis-AI HLS to integrate transformer architecture with efficient floating point arithmetic operations.

- Designed host interface for multiplication between NxD weight matrix and DxM input matrix
- Implemented FP8 multiplication and addition through fused multiply add in C++
- Optimized matrix multiplication kernel through Vitis High-Level Synthesis best design practices
- Conducted software and hardware emulation of FPGA design

SALIERI.AI, Boston MA — May. 2024 - Aug. 2024
Full-stack / NLP AI Engineer Intern
Developed a server-less SaaS deployment with an autoscaling Kubernetes cluster hosted on GCP to facilitate structuring LLM output into parsable JSON based on business text document stores

- Utilized Node.js to create a web-hook API, enabling communication with a Python "child process" via inter-process communication (IPC).
- Developed 3 containerized micro-services for scalability and created network for intra-container communication
- Restricted LLM output using Context-Free Grammars in GBNF format and JSON to ensure valid generation (token generation constraint)
- Built and optimized an autoscaling Kubernetes cluster with MySQL data store on Google Cloud Platform, with containers stored in Google Artifact Registry and configured using Helm Chart for long-term sustainability.

Research

CSAIL, Cambridge, MIT — Jan. 2025 – Present
Undergraduate Researcher, Dynamic Data Storage and Retrieval [LOGIC]
Leading research within the CSAIL Spoken Language Systems Group (SLS) at MIT, focusing on enhancing retriever performance in resource-augmented generation (RAG) systems through model editing.

- Benchmarked BERT-Large-Cased performance by generating a synthetic query-document set and extracting [CLS] token embeddings from a pretrained model using PyTorch
- Created vector database with FAISS Index and SQLite for efficient document storage and retrieval based on document embeddings
- Adapted MALMEN (ICLR 2025) to fit multi-class classification objective

CSAIL, Cambridge, MIT Sep. 2024 – Dec 2024
Undergraduate Researcher, Dynamic Data Storage and Retrieval [DDSR]
Lead research within the CSAIL Spoken Language Systems Group (SLS) at MIT into the utilization of Modern Continuous Hopfield Networks to model Memory for LLMs

- Constructed tensor-based data processing pipeline utilizing Kaggle sentence dataset and tokenization via. Mistral tokenizer
- Trained network with PyTorch and evaluated results with matplotlib

CSAIL, Cambridge, MIT Jan. 2024 – May 2024
Undergraduate Researcher, Thinking Deeper with Recursive Spawning [THREAD] Benchmark Creation
Conducted research within the CSAIL Spoken Language Systems Group (SLS) at MIT for the creation of a benchmark for LLM natural language embedded program generation.

- Produced >50 ground truth Python programs for synthetic data generation to create evaluation dataset.
- Created API calls for aggregate data collection and processing from Google Data Commons

Publications

Phillip Schroeder, Nathaniel W. Morgan, Hongyin Luo, James R. Glass. May. 2024
THREAD, "Thinking Deeper with Recursive Spawning" [NAACL 2025]
(arXiv preprint arXiv:2405.17402 2024)

Projects and Leadership

ARK (Automated Resource Knowledge Base), Open Source Project Lead Apr 2023 – Present
Leading a team of 6 across MIT, GT and UIC to create a community LLM app store and modular personalized agentic assistant development platform

- Configured LLM Inference with Hugging-Face TGI and Qwen-2.5 weights hosted on HuggingFace Hub
- Created `base_module` interface for agentic tool calling and inference engine abstraction layer
- Organized project outline including "GitHub Org.", task list, weekly meetings, and various repositories for maintaining current and future contribution consistency as project scales
- Secured $5000 grant from MIT SIPB for development server build
- Built and set up headless dev server including installing CUDA kernel, configuring Linux distro, adding user SSH keys, and configuring user permissions

Uplift.co 501(c)(3) (www.communityuplift.co), Founder May 2019 – Jan. 2025

- Organized first cohort and facilitated weekly meetings to grow the project to a peak of 30 members
- Assisted in the creation of 5 in person wellness events and 10 digital wellness campaigns
- Built components of brand including website, Instagram, and initial digital wellness campaigns for public awareness of organization

Presentations

- SIPB Cluedumps: "AI for Dummies" Jan. 2024
- HackMIT Blueprint: "AI for Dummies" Mar. 2024
- TEDXBoston: "Visual Embedded Identification for Legitimacy" Jan. 2025

Awards & Accomplishments

MIT Student Information and Processing Board (SIPB) (Keyholder) Dec. 2023
Harvard x MIT CO-OP (Student Board Member) Aug. 2024
International Science and Engineering Fair (Finalist) Oct. 2021