Madhav Anand Menon

+1 (646) 938-7079 | madhavanandmenon@gmail.com | linkedin.com/in/madhav-anand-menon/ | github.com/MadhavMenon10

Education

University of Illinois Urbana-Champaign Expected May 2028

B.S. Computer Science and Physics, Minor in Mathematics (James Scholar Honors) **GPA: 3.97/4.0**

- Yee Seung Ng Award, Harold and Ruth Hayward Scholarship, Dean's List, Tau Beta Pi Honors Society
- Relevant Coursework: Honors Data Structures and Algorithms, Computer Architecture, Databases, Numerical Methods, High Frequency Trading Technology, Honors Linear Algebra, Probability Theory (Upcoming: Operating Systems)

Experience

Valeo Jun. 2026 - Aug. 2026

Software Engineer Intern Chennai, TN, India

- Engineered a 4-state Extended Kalman Filter in Python fusing gyro, accelerometer, and GPS-velocity data, cutting IMU pitch error by 37% against a survey-grade INS resulting in improved navigation accuracy
- Diagnosed a prior filter's failure across a 9,000-sample log, proving that centripetal acceleration corrupted attitude 42 sec before any trigger fired
- Tuned sensor-fusion gating against a measured 0.91 m/s2 noise floor, cutting false-rejection duty cycle by 60% toward a sub-10% stability target

Disruption Lab Feb. 2026 - Present

Software Engineer Champaign, IL, USA

- Built a 5-stage Canvas-to-PostgreSQL RAG pipeline supporting 3 file formats (PDF/DOCX/TXT) with 1024-dim pgvector embeddings and sliding-window chunking via AWS Bedrock
- Designed a skill extractor with NACE framework alignment, returning deduplicated skill taxonomies across 100+ courses per API call, reducing manual tagging effort by 98%
- Shipped a Python FastAPI backend with custom token-based authentication and dual-mode search (full-text PostgreSQL and pgvector cosine similarity) supporting 120+ concurrent users

Algoverse AI Sep. 2025 - Jun. 2026

LLM Research Intern San Francisco, CA, USA (Remote)

- Architected DeployStega, an LLM-agent covert-communication channel embedding payloads in the authenticated permutation rank of real GitHub events, cutting cross-layer detector ROC-AUC from 1.000 for generation-based encoders to 0.634
- Benchmarked the channel against 4 capacity-matched baselines, including minimum-entropy coupling, across 200 preregistered detector runs over 4 payload sizes and 5 seeds; reduced detection from 83–99% to 8–23% TPR at a 5% false-positive rate
- Hardened evaluation with actor-, repository-, and text-disjoint splits over 3 GH Archive windows, voiding a leakage-inflated 0.578 AUC; hierarchical semi-Markov sampling cut synthetic-trace detectability from 0.947 to 0.604 AUC across 5 seeds

DigiAlert Jun. 2025 - Aug. 2025

Software Engineer Intern Chennai, TN, India

- Designed a Python boto3 CloudTrail pipeline sustaining gap-free ingestion under AWS's 2 req/s LookupEvents throttle, using 5-min-overlap incremental windowing with EventId deduplication and 6-retry backoff
- Built a Python Flask-based AWS CSPM dashboard aggregating security metrics across 4 AWS services with CIS/NIST aligned risk classification and GeoLite2-geolocated logins, cutting CSPM tooling costs by 45%
- Shipped via nginx/gunicorn with Chart.js visualizations, dual-tier TTL cache invalidation (10-min memory / 60-min disk), and non-blocking single-flight refresh preventing overlapping AWS fetches

Projects

GBDT Inference Engine | *CUDA C++, Python, XGBoost, NVIDIA RAPIDS Forest Inference Library*

- Engineered a from-scratch C++/CUDA inference engine for multiclass Gradient-Boosted Decision Tree ensembles, cutting median single-sample latency 3.9-6.8x below XGBoost's native C API (24us vs 90-160us) across 10 ensembles
- Designed a dense-node tree layout and thread-per-(sample, tree) CUDA kernel that saturates full warps at batch size 1, outperforming NVIDIA RAPIDS FIL by 2.3-8.2x on single-sample latency across 10 ensembles
- Built a two-pass harness isolating H2D/kernel/D2H latency over 30,000 seeded trials, pinning the GPU/CPU crossover to batch size 1 at both median and p99 and attributing 87% of large-batch cost to a 111:1 output-payload asymmetry

CUDA-Accelerated Order Book Reconstructor | *CUDA, C++, CUB/Thrust, Nsight Systems*

- Built a CUDA order-book engine end-to-end with a custom ITCH 5.0 binary decoder, per-symbol stream compaction, and warp-per-symbol reconstruction, processing 305M+ NASDAQ messages at 4.4M msgs/sec across 9,000 symbols on an H100
- Swept 10,000 backtest configurations in parallel, compressing 6.5 hours of sequential backtesting into 83.8s for a 279× speedup and 156× end-to-end over a single-threaded C++ baseline on identical hardware
- Engineered a from-scratch GPU open-addressing hash table sustaining 99.99997% reconstruction accuracy, and eliminated 27 GB tick buffer via atomic-claim allocation over pinned SoA buffers

Skills

Languages: C++, C, Python, JavaScript, TypeScript, SQL, Verilog, MIPS Assembly
Libraries and Frameworks: CUDA, PyTorch, FastAPI, Flask, pandas, scikit-learn, NumPy, Matplotlib
Tools: NVIDIA Nsight, Jira, Linux, AWS, LangGraph, LangChain, Git, TeX/LaTeX, Docker, Valgrind, MATLAB
Other Activities: CS 124 Project Manager, CodePath, ACM Corporate Team Member, TEDx Speaker