Pranav Varshney

🌐 US Citizen ✉ pvarsh@umich.edu in linkedin.com/in/pvarshh ○ github.com/pvarshh

EDUCATION

University of Michigan	May. 2027
B.S. Computer Science, Data Science	Ann Arbor, MI

EXPERIENCE

Scale AI — July 2025 - Sept 2025
Software Engineering Intern — San Francisco, CA
- Developing retrieve & rerank RAG pipeline to automate LLM inputs for data annotation via candidate filtering
- Building high throughput benchmark system to score 8K+ model responses daily, reducing 60% annotation errors

Amazon Web Services — May 2025 - July 2025
Software Engineering Intern — Bellevue, WA
- Architected a Go-based command line server for SageMaker endpoints using AWS SDK, automating troubleshooting by mapping 1M+ endpoints to underlying AWS Services (kubernetes, elastic compute, control plane, bedrock)
- Reduced on call response time by 70% average by developing a real-time health checks dashboard for developers
- Scaled the solution to support 10K+ weekly queries by containerizing the command line server with AWS Copilot

University of Michigan — Jan 2025 - May 2025
Machine Learning Researcher — Ann Arbor, MI
- Designed TreeGen-based transformer in PyTorch using AST-based attention mechanisms, achieving 92% accuracy
- Built automated python script for network snapshot generation & contrastive learning, expanding training data 250% & improving configuration similarity matching by 25% through parallel processing optimization

United Wholesale Mortgage — May 2024 - Aug 2024
Software Engineering Intern — Pontiac, MI
- Built data pipeline to accelerate code quality insights; normalized databases to 3NF, reducing duplicates by 45%, speeding GCP migration 3 weeks; saved $20K in Q3 by deploying random forest model to predict loan outcomes

Ratna Global Technologies — May 2023 - Aug 2023
Software Engineering Intern — Newark, CA
- Developed AI customer support system using GPT-3.5 & RAG, integrating company data to resolve 80% inquiries with 90% accuracy while saving $1.2K/month in API calls via caching frequent requests in redis; used react/node

PROJECTS

SD–WAN Analytics (2024 Cisco HackAIthon) — Presentation
- Applied AI to analyze app usage, device health, & NWPI for root cause analysis & network anomaly remediation
- Strengthened tunnel KPI anomaly detection & network issue resolution, giving insights for network optimization
- Forecasted network issues & scheduled NWPI trace with ThousandEyes tests to train AI for proactive repairs

Financial Applications of Quantum Machine Learning — Research Paper
- Analyzed Quantum ML model performance & scalability in portfolio optimization, fraud detection, & forecasting
- Proposed quantum-classical workflows for ML feasibility given bottlenecks in data encoding & training stability

Tennis Winner Predictor — Website
- Engineered key predictive features including ranking differences, H2H ratios, tournament context, & surface type
- Achieved 92% accuracy by developing a logistic regression & gradient boosting prediction model on 75k matches

SKILLS

Languages: C++, C, Python, Go, C#, R, HTML, CSS, Typescript, Javascript, SQL (MySQL, PostgreSQL), LaTeX

Technologies: Docker, CUDA, Flask, Next, React, Node, Express, Tailwind CSS, Pandas, NumPy, Plotly, Sci-kit Learn

Concepts: Full-Stack Development, Mobile Development, Cloud Computing, Parallel Programming, Machine Learning, NLP, Computer Vision, Data Analytics, Big Data, Embedded Systems, Quantum Computing, Computer Security, Linux