

Hieu Minh Nguyen

EDUCATION

Tampa, FL	University of South Florida	Expected May 2026
• Bachelor's / Master's Pathways Degree in Computer Science. GPA: 3.8/4.		

SKILLS

- Programming Language:** Python, C, C++, C#, Java, JavaScript/TypeScript, HTML5, CSS, SQL, Swift, Bash.
- Technologies and Tools:** MySQL, PostgreSQL, Git/GitHub, GitHub Action, AWS, Google Cloud, Azure, Docker.
- Machine Learning and AI:** LangChain, Numpy, Pandas, Matplotlib, Seaborn, scikit-learn, PyTorch, TensorFlow, OpenAI, Keras, transformers, Streamlit, Retrieval Augmented Generation.
- Web/Mobile Technologies:** Node.js, Vite.js, Express.js, React.js, FastAPI, React Native, Flask, Next.js, Tailwind CSS, SwiftUI.

PROFESSIONAL EXPERIENCE

Research Assistant	University of South Florida	May 2024 – Present
Publication: "Global is Good, Local is Bad?": Understanding brand Bias in LLMs, "A Woman is More Culturally Knowledgeable than A Man?": The Effect of Personas on Cultural Norm Interpretation in LLMs.		
• Conducted comprehensive research on bias across many categories, such as age or gender, within Large Language Model (LLM).		
• Built a specialized library in Python for systematic evaluation of 100,000 datasets in LLMs, significantly reducing manual effort.		
• Collaborated with a team of researchers on studies focusing on mitigating bias in human health records using machine learning.		
• Partnered with USF students and faculty to develop an AI personal assistant for learning paths and school course curricula.		
• Researched techniques to optimize LLM code generation for complex data structures, algorithms , competitive programming.		
AI/ML Software Engineer (Intern)	Resilience, Inc.	Sep – Dec 2023
Technologies: AWS, React Native, Flask, Synk, Figma, Linux, Python		
• Engineered a mindfulness app, resulting in enhanced emotional self-awareness across diverse user demographics.		
• Executed design prototypes using Figma, further developing 5+ frontend features via React Native.		
• Implemented a machine learning API endpoint with Flask (Python) on AWS EC2, leading to an 80% reduction in response latency.		
• Implemented proactive security measures through penetration testing methodologies to fortify web applications and network infrastructure, resulting in a 50% reduction in potential security breaches and associated risks.		
• Proactively secured the app by employing Snyk, addressing 100% of identified code vulnerabilities.		
• Coordinated with cross-functional teams, include engineering and design for smooth production releases.		
AI/ML Developer (Intern)	Resilience, Inc.	June – Aug 2023
Technologies: Keras, Numpy, Pandas, sklearn, matplotlib, TensorFlow, Python		
• Researched and selected deep learning/machine learning models, including CNN, RNN, XGBoost, and MLP, for effectively user emotions prediction based on through benchmarking and rigorous testing.		
• Built a CNN model using Keras (TensorFlow) from scratch, achieving a 72% emotion classification accuracy.		
• Streamlined machine learning processes by creating a pipeline for automated data handling and model training.		
• Finetuned Whisper by multiprocessing through parallelization for CPUs to speed up translation time by 5.9x .		
• Guided a team of five interns by writing a comprehensive documentation and hosting weekly feedback sessions , ensuring rapid tool mastery, addressing queries, code reviewing and bolstering overall progress and team synergy.		

PROJECTS

Talk To Listen (talktolisten.com)	FastAPI, Python, React Native, JavaScript, REST APIs, SQL, Firebase, Redux, CI/CD, AWS, Azure
• Built a mobile AI chat app for seamless interaction with unique AI characters, each featuring distinct voices, personalities, stories.	
• Grew the app to 1000+ active users and secured \$10,000+ in sponsorship from Microsoft, OpenAI, and AWS Startup Programs.	
• Created scalable distributed system architecture using Azure Application Gateway, Azure Load Balancer, Azure Virtual Machines.	
• Constructed a RESTful APIs server with FastAPI (Python) capable of handling 30+ APIs, and 3 machine learning endpoints, deployed on AWS EC2 using Docker, and integrated with GitHub Actions for automated CI/CD.	
• Created and published an open-source dataset of 3M tokens (nearly 2000 downloads monthly) and fine-tuned an LLM model using supervised learning, quantization method (QLoRA), and Pytorch, achieving an accuracy of 85% .	
• Continuously monitored model performance, retrained with new data, and state-of-the-art open-source model.	

- Developed, tested, and debugged the app using Xcode for **iOS** Simulation, and Android Emulator for **Android** Simulation.
- Managed all stages from **Figma**-based design, **front-end**, database, API design, to **iOS/Android** deployment and **back-end**, encompassing more than **3000** lines of code.
- Designed a comprehensive database schema, featuring **20+** schemas and integrating **Azure Database PostgreSQL**.
- Devised a hands-free voice input feature with a voice detection algorithm and **WebSocket** connection.
- Integrated **Redux** with React and Axios, to seamlessly manage API data fetching and state updates.
- Configured **Azure Virtual Network** to securely connect virtual machines, enhancing the security and reliability.
- Monitored system health using **Azure Portal**, employing **monitoring tools** to track performance, uptime, and resource utilization.
- Deployed a machine learning model on Runpod's **serverless** platform via **Docker**, reducing cost and enhancing dynamic scalability by **82%**, and lowering the latency to **75%**.
- Applied **Test-Driven Development (TDD)** to create automated test scripts for the back pressure feature.
- Created and optimized a **voice detection algorithm** to enable real-time speech and response.
- Applied **NGINX** as a web server and reverse proxy, **ensuring efficient traffic management, load balancing, security**.
- Utilized **APIs best practices in client-server protocols**, including appropriate HTTP methods, CRUD operations, **Firebase** for authentication, to design effective and secure communication between clients and servers.
- Enforced **Redis** caching to improve database performance, resulting in a 30% reduction in query response time.
- Utilized **SSL/TLS** protocols to establish secure connections, ensuring confidentiality of sensitive data during transmission.
- Configured **AWS Simple Storage Service (S3)** for secure and scalable storage of image assets.
- Utilized **best practices in product lifecycle management**, including clear requirements definition, application frameworks, version control, through testing, resulting in **increased productivity across all phrases**.

JobsDreamer (jobsdreamer.com) | Scrapy, Selenium, Python, Google Gemini 1M API, GPT4o, Llama3, Proxy, Azure, AWS

- Launched a platform that assists **hundreds** of students in finding internships, automatically sending them emails with newly all posted relevant internship opportunities on the internet within the last 24 hours.
- Established a comprehensive job data processing pipeline utilizing **latest** AI tools, AWS EventBridge, AWS Step Functions, AWS Fargate, and AWS Lambda to **automate** daily web scraping, data preprocessing, and email notifications.
- Integrated a **GPT-4o** and **Google Gemini** based categorizing and reviewing system to automatically classify and validate job data, ensuring high accuracy and relevance of categorized data.
- Utilized **residential proxy servers to manage IP rotation** and avoid detection, enhancing the anonymity of the scraping process.
- Automated user notifications by integrating **DynamoDB** and an **email** service to send processed job data directly to user emails.
- Deployed containerized **microservices** for web scraping across multiple job sites, leveraging Python, Scrapy, and Selenium.
- Automated a scalable and efficient Python pipeline with **advanced LLMs** to handle **1000+** of new **raw data** every day.
- Implemented **advanced cookie** handling mechanisms to maintain session persistence and accurately simulate user interactions on target job sites.
- **Implemented rigorous data validation checks** throughout the pipeline to maintain data integrity and consistency.
- Planned a scalable infrastructure with containers in an **AWS** VPC setup, allowing for isolated operations within private subnets.
- Drew clear and detailed diagrams to visualize the architecture and workflow of the data pipeline.
- Set up **CI/CD** pipelines using GitHub Actions and AWS ECR, automating the build, test, and deployment processes.

BullBot (bullbot.tech) | Flask, AWS, Vite.js, Express.js, Node.js, Pytorch, Langchain, Python

- Pioneered a full stack chatbot web app answering queries across **1500+** Uni of South Florida websites with natural language recognition, providing precise sources for user clarity, especially for parents less adept at online searches.
- Developed a **Retrieval Augmented Generation (RAG) system**, integrating Vector Store for **efficient data management**, significantly **optimizing** data weight, and expediting database information retrieval processes.
- Maintained a **scalable and fault-tolerant data storage solution**, resulting in a **95% reduction** in storage costs.
- Employed **Depth First Search Algorithm** with BeautifulSoup for comprehensive web link scraping.
- Analyzed and filtered **5000+ raw** data files to filter **18%** unnecessary content, improving the quality of the data.
- Wrote **Python scripts** to automate data extraction, processing, and ingestion, increasing efficiency by **40%**.
- Orchestrated a **microservice** architecture leveraging Vite.js, Express.js/**Node.js**, and various AWS services such as **AWS EC2, AWS API Gateway, AWS Lambda**, achieving a significant server workload reduction.
- Utilized opensource **Hugging Face models** and quantized Meta Llama2 with **PyTorch, LoRA, PEFT**, to reduce model's weight and speed up model training, resulting in an **90%** cost-saving on deployment.

InCollege | Python, Scrum (Agile), Object Oriented Programming, Git/GitHub, pytest.

- Worked with an **Agile/Scrum** team to develop a college networking app, leveraging Agile methodologies to streamline development, ensure adaptability to evolving requirements and meet project milestones efficiently.

- Implemented a modular and object-oriented architecture facilitating independent development, reusability, scalability, and efficient collaboration among team members.
- Conducted **15+ unit test** and **integration test** cases every week, warranting the functionality of new features.

Chatbot GPT | React.js, HTML, CSS, JavaScript, Node.js, OpenAI API

- Crafted a dynamic and interactive ChatGPT clone utilizing **React.js** and **Node.js**, integrating **OpenAI API** for advanced conversational capabilities, enhancing both functionality and user experience.

Leo Bot | Next.js, HTML, CSS, JavaScript, Node.js, OpenAI Assistant API

- Developed a personal chatbot assistant using **Next.js** and **TypeScript**, integrating **OpenAI Assistant API** for enhanced conversational capabilities.

Car Repairing System Design | Python, Relational DBSM, PostgreSQL, Streamlit, Database Design, Azure, SQL

- Developed a full-stack e-commerce platform for car parts trading, designing **schemas** and **EER diagrams**, and integrated a **PostgreSQL** database using **psycopg2**, enhancing data management efficiency.
- Designed complex SQL queries, **optimizing query performance** through carefully indexing and query plan analysis.

Stock Data | C# .NET, Visual Studio, Object Oriented Programming, Git/Github

- Orchestrated the creation of a **C# .NET** Windows Forms application to deliver stock data visualization and management through GUI design and object-oriented class implementation in Visual Studio.

api4all | Python, ML/LLM provider, ML/LLM API, Object Oriented Programming, open-source

- Maintained an easy-to-use **open-source** project, integrating cutting-edge models from state-of-the-art providers and conducting comprehensive comparisons to ensure optimal performance and functionality.

Resume Keywords | Python, Streamlit, Snowflake, Snowflake Arctic LLM, Replicate, Streamlit

- Implemented a Streamlit app powered by Snowflake Artic LLM to generate optimized resume bullet points from keywords.
- Utilized Snowflake Data Cloud to store, query, and managed large datasets, to ensure seamless data retrieval in the application.

LEADERSHIP

Software Lead	Association for Computing Machinery – USF Chapter	November 2023 – Present
<ul style="list-style-type: none">• Leading a team project of 8 by creating detailed project plans, functional requirements, and documentation.• Hosting technical workshops and mentoring other members in preparing and delivering their workshops, fostering a collaborative learning environment.		