TONY QUAN VU

FULL STACK DEVELOPER | AI & ML ENGINEER

Walnut, California, US | +1 (954) 857-8329 | tonyquanvu97@gmail.com

- Experienced Full Stack and AI/ML Engineer with 8+ years delivering cutting-edge solutions in machine learning, deep learning, and NLP.
- Specialized in large language models (GPT, Llama), chatbot development, and data-driven applications.
- Proficient in Python, Java, and full-stack frameworks (Node, Django, Spring Boot, React), with expertise in cloud platforms (AWS, GCP), containerization (Docker, Kubernetes), and data tools (Apache Spark, Tableau).
- Known for building scalable systems, recommendation engines, and sentiment analysis pipelines using SQL/NoSQL.
- Driven by innovation and strategic problem-solving to enhance performance and user experience.

AREAS OF EXPERTISE

- Machine Learning & Deep Learning
- Data Science & Engineering
- Programming & Development
- Cloud Computing & Deployment
- Big Data Technologies
- Database Management
- Business Intelligence & Visualization
- Innovative Problem-Solving

PROFESSIONAL EXPERIENCE

DEC 2023 - AUG 2025

AI Engineer & Data Architect

Talkhealth AI

Healthcare technology company leveraging AI-driven insights to provide personalized medical information, empowering individuals to make informed health decisions and fostering a healthier future through the integration of technology and medical expertise.

- Led the creation and deployment of an advanced healthcare chatbot that delivers real-time health analysis and personalized healthcare recommendations using ChatGPT API and Pinecone for natural language processing and information retrieval.
- Designed a dynamic data pipeline to update medical information from Merck Manuals and CVS product inventories, utilizing MongoDB, AWS DynamoDB, and Elasticache to ensure efficient data management and accessibility.
- Architected a secure, scalable infrastructure using AWS services like S3 and CloudWatch, and deployed applications using Docker, facilitating seamless integration with Next.js for frontend services.
- Implemented robust data handling processes to ensure full compliance with HIPAA regulations, safeguarding user privacy and data security.
- Developed automation solutions with Zapier and HubSpot, leading to a 50% reduction in manual processes and significantly enhancing operational efficiency in data management and customer engagement.

- Guided cross-functional teams to align product features with technical capabilities, fostering innovation and operational efficiency, while mentoring junior engineers in machine learning and data engineering best practices.

NOV 2021 -
OCT 2023

AI/ML ENGINEER

BRIDGE VIEW

Technology consulting and talent firm specializing in building exceptional tech teams and delivering complex projects through strategic consulting and a vast network of niche professionals.

- Created and deployed a customer churn prediction model using PyTorch and TensorFlow for a major e-commerce client, increasing retention rates by 30% and contributing to a $2 million annual revenue boost through targeted marketing.
- Engineered and integrated a real-time fraud detection model with AWS SageMaker and AWS Lambda for financial institutions, reducing false positives by 20% and saving clients approximately $500,000 annually in fraud prevention.
- Developed an NLP-driven chatbot for customer support using NLTK and SpaCy, which enhanced user interaction by reducing response times by 60% and increasing customer satisfaction ratings by 25%.
- Implemented advanced language capabilities in chatbots and virtual assistants using Google Dialogflow and GPT models, streamlining customer service interactions and maintaining high engagement levels across client platforms.
- Architected a recommendation engine using collaborative filtering for a streaming service, achieving a 35% increase in user engagement and a 10% rise in subscription renewals within the first quarter of deployment.
- Built and maintained a data warehousing framework with Apache Spark and Amazon Redshift for a large retail client, improving data retrieval and accelerating business intelligence reporting times by 70%.
- Integrated Langchain in NLP tasks for automated content generation and semantic understanding, significantly boosting the efficiency and accuracy of content management systems used by clients.

SEP 2019 –
OCT 2021

FULL STACK DEVELOPER

TRADEPLANS AI

An innovative platform that leverages artificial intelligence (AI) to assist traders in making informed decisions and optimizing their trading strategies.

- Hired, trained and lead Agile team of 6 full stack developers
- Simultaneously created & maintained scheduled jobs in SQL Server for space maintenance and daily backups of system and user databases for 10 clients
- Increased company revenue by 30% within 2 months after developing and implementing business logic for over 20 features
- Designed and Developed UI design for over 15 clients using CSS, HTML, ASP.NET, Vue.js and React.js; websites scoring over 85 on Lighthouse

AUG 2017-
AUG 2019

SOFTWARE DEVELOPER

WAWMA

Japanese data analytics company that offers market insights and pricing trends for online auctions and e-commerce platforms, helping users make informed buying and selling decisions.

- Shortened project timeline by 18 months for company's largest customer by managing relationship with 3rd party vendors, saving over $800K
- Built highly responsive, mobile-first web applications for security management, ensuring smooth access and real-time monitoring on both desktop and mobile devices, improving usability for security personnel in the field.

- Integrated secure authentication and role-based access control (RBAC) into the React application, ensuring only authorized users can access and control video feeds, aligning with security protocols
- Optimized the performance of the React application by using lazy loading, code splitting, and memorization techniques, resulting in faster load times and better user experience across devices

EDUCATION

2013 - 2017
ITHACA, US

B.S. IN COMPUTER SCIENCE

Cornell University

SKILLS

Machine Learning | Deep Learning | Natural Language Processing | Sentiment Analysis | Text Processing | LLM | Chatbot Development | Data Analysis | Data Integration & Pipelining | ETL/ELT Processing | Data Visualization | Automation | Python | R | Java | Scala | Golang | Ruby | JavaScript | Node.js | FastAPI | Flask | Django | Spring Boot | PyTorch | Tensorflow | Keras | GPT | Claude | Llama | Langchain | LiteLLM | Pinecone | ChromaDB | AWS SageMaker | AWS Bedrock | Apache Spark & PySpark | Apache Hadoop & Hive & HBase | AWS | GCP | MySQL | PostgreSQL | MongoDB | Amazon S3 | Redis | Snowflake | Elasticache | DynamoDB | SerpAPI | GoogleAPI | Twilio | HubSpot | Zapier | Dialogflow | MS Office | Rasa Platform | IBM Watson | Apache Airflow | Tableau | Data Studio | PowerBI | Full Stack