

Tuan Dinh

Dallas, TX | 214-417-4238 | t.dinh43204@gmail.com | linkedin.com/in/tuan-dinh- | github.com/gt12889

EDUCATION

The University of Texas at Dallas — *B.S. Information Technology & Systems (Cybersecurity)* | **GPA: 3.4** | May 2026

Relevant Coursework: Data Structures & Algorithms, Distributed Systems, Cloud Computing, Object-Oriented Programming

EXPERIENCE

Bank of America | *Software Engineer Intern* → *Full-Time* | Plano, TX | Jul 2023 – Present

- Owned ETL pipelines and **RESTful microservices** processing **10M+ records/day** using **PySpark, HDFS, Java**, and **Oracle SQL** on enterprise **data lakehouse** infrastructure; maintained **sub-5-minute latency SLA** for downstream reporting
- Developed **Linux Bash scripts** for automated job scheduling and data extraction on **edge nodes**; wrote complex **Oracle SQL queries and stored procedures**, optimizing batch processing across **60+ operational tables**
- Engineered event-driven ingestion layer using **Python, Kafka**, and **Autosys** to reliably process **500GB+/week** from operational tables; eliminated recurring manual batch failures and reduced data pipeline error rate to near-zero
- Automated CI/CD pipelines with **Jenkins, Bitbucket**, and **Ansible**; **reduced deployment time from 45 min to 27 min (40%)** through validation gates, containerized deployments, and rollback automation

Splash Station Rentals | *Infrastructure Lead* | Oct 2019 – Jul 2021

- Deployed containerized services on **AWS** and **Azure** using **Docker**; configured monitoring and **load balancing across 10,000+ network endpoints**, driving **4x higher availability** for high-throughput traffic
- Resolved **400+ technical support tickets** with systematic root-cause analysis; **reduced average resolution time by 35%** through improved triage workflows and Confluence documentation

PROJECTS

Toyotron | TypeScript, Supabase, Docker, NVIDIA Nemotron | *toyota-toyotron.vercel.app*

- **1st Place Winner** in both **NVIDIA + Toyota Challenge** at **HackUTD 2025** with **160+ teams, 1,200+ participants**
- Built **NVIDIA Nemotron multi-agent** vehicle advisor with **FastAPI + WebSocket**, supporting concurrent multi-user sessions with real-time streaming and **Retell AI voice integration** and real-time token streaming
- Implemented **conversation-state management, tool routing, and dynamic plan generation** through prompt-layered governance and persistent session context for scheduling and email workflows

Alethia | Modal, vLLM, PyTorch, TypeScript, OpenTelemetry | *alethia-lovat.vercel.app*

- **1st Place Winner** in **OpenAI Challenge** at **HackIllinois 2026 @ UIUC** with **230+ teams competing**
- Built an enterprise neighborhood intelligence platform with **33 serverless functions leveraging 21 distinct Modal primitives** across GPU inference, event bus, recursive agents, and self-healing pipeline orchestration
- Architected multi-agent system where **agents scores and autonomously spawn parallel specialized agents** on high-impact events
- Self-hosted Qwen3-8B AWQ (INT4) on NVIDIA H100 via **vLLM** with **GPU memory snapshots** enabling under 3 second cold starts; ran **bart-large-mnli** and **roberta classifiers** on T4 GPUs, achieving **10x inference cost efficiency**

Roly Poly | FastAPI, Arize Phoenix, Railway | *myrolypoly.cc*

- **3rd Place Winner** at **NexHacks 2026 @ Carnegie Mellon University** with **1,500+ participants, 30+ company sponsors**
- Engineered **prediction market decisioning engine** computing Expected Value, Kelly-optimal sizing, VaR, and Sharpe-style **risk metrics** with P10/P50/P90 payoff surfaces for **3D risk visualization**
- Built A/B-tested LLM workflows with **Arize Phoenix tracing + token compression**; **reduced inference costs by 30%** while improving response accuracy through structured prompting, achieved **+147% LLM output quality**

RxGuard | Hugging Face, Actian VectorAI, spaCy + scikit-learn | *rxguard-6hku.onrender.com*

- **2nd Place Winner** at **Hacklytics 2026 @ Georgia Institute of Technology** with **250+ teams competing**
- Built **semantic drug safety search engine** ingesting **500K+ FDA adverse event reports** from openFDA API with a **4-stage batch retrieval pipeline** progressing to **HNSW vector search**
- Engineered **vector search backend** using sentence-transformers (**all-MiniLM-L6-v2, 384-dim embeddings**) with Actian VectorAI DB over **gRPC**; designed **risk scoring algorithm** weighting semantic similarity, severity, and demographic match

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, SQL, C++

Frameworks & Libraries: Spring Boot, Next.js, React, Angular, Django, PySpark, Kafka, Node.js, FastAPI, LangChain

Tools & Infrastructure: AWS, Azure, Docker, Kubernetes, Jenkins, Git, Airflow, HDFS, Redis

Eligibility: U.S. Citizen - Authorized to work full-time without restriction