

Binh Nguyen

Senior Full-Stack & AI Engineer

CONTACT ME

+84 938 023 909

binhnx.it@gmail.com

cv.siri2moon.com

Da Nang city, Vietnam

SKILLS

AI & MACHINE LEARNING

- Retrieval-Augmented Generation (RAG)
- LLM application architecture
- AI agents & tool calling
- LangChain / LangGraph
- Vector search & embeddings
- Re-ranking & prompt engineering
- OpenAI, Anthropic Claude, Google Gemini, AWS Bedrock
- Document AI & OCR
- AI usage & cost tracking

BACKEND

- Python – FastAPI
- Node.js – NestJS, Express
- TypeScript
- REST APIs, SSE, WebSocket
- Message queues & event-driven architecture
- Microservices, SOLID, design patterns
- PHP – Laravel

FRONTEND

- Next.js
- React
- TypeScript, Turborepo

SUMMARY

Software engineer with **10+ years of experience**, currently the technical lead of an **enterprise AI platform** serving corporate customers in Korea and Vietnam.

I deliver products end to end – the AI services, the backend APIs, the web application, and the cloud infrastructure underneath them. My focus for the past two years has been applied AI: intelligent search across a company's private documents, AI assistants that answer with verifiable sources, and the data pipelines that keep them fast at scale.

Paired with deep AWS and Kubernetes expertise, this lets me deliver systems that are intelligent, secure, and economical to run in production.

CAREER HIGHLIGHTS

- **Architected the AI document pipeline** behind the platform's intelligent search – thousands of documents processed concurrently, exactly once, scaling automatically with demand.
- **Delivered around 20% in cloud cost savings** through architecture, automation and right-sizing, while sustaining full capacity and reliability.
- **Led two zero-downtime migrations of the entire production platform**, delivered end to end as infrastructure-as-code.
- **Designed the enterprise security model for AI search** – employees retrieve only the documents they are authorized to access, permission changes apply instantly, and consistency is verified automatically.
- **Engineered multilingual answer accuracy across Korean and English**, combining hybrid search, result re-ranking and page-level source citations so every answer can be verified.
- **Delivered per-customer AI usage and cost reporting**, giving the business accurate billing and margin visibility across every AI model in use.
- **Led engineering teams of up to 20 people** across eight years, delivering enterprise platforms for Japanese and Korean clients.

WORK EXPERIENCE

Senior Full-Stack & AI Engineer — *Apr, 2024 - Present*

Uranus Software - Da Nang

Technical lead of an enterprise AI platform that lets organizations search and work with their internal documents through AI assistants. Accountable for the AI services, backend, web application and cloud infrastructure – from architecture through production operations and cost.

SKILLS (cont.)

DATABASES

- PostgreSQL, Aurora
- MySQL
- MongoDB
- Redis
- Firebase – Firestore
- Vector databases – Milvus, OpenSearch

CLOUD & CI/CD

- AWS – EKS, ECS, EC2, Lambda, S3, CloudFront, Aurora RDS, ElastiCache, SQS, KMS, CloudWatch, IAM
- Google Cloud
- Terraform – infrastructure as code
- Kubernetes, Docker
- GitHub Actions, Jenkins, CircleCI
- Auto-scaling, monitoring & alerting
- Cloud cost management
- Backup & disaster recovery

LEADERSHIP

- Technical leadership
- Team management – 15 to 20 engineers
- Scrum Master, agile delivery
- Mentoring & onboarding
- System design & technical writing
- Client-facing communication

LANGUAGES

- Vietnamese – Native
- English – Professional working

Uranus Software – AI Platform

- Architected the platform's **AI document search**: documents are automatically read, understood and indexed so employees can ask questions in natural language and receive answers cited to the exact page of the source file.
- Designed an **AI assistant platform** where non-technical users configure their own assistants – instructions, tools, AI model and the documents each assistant may use – with answers streamed live to the browser.
- Enabled assistants to **produce business deliverables** – PDF, Word, Excel, PowerPoint and charts – generated inside a secure isolated environment and automatically validated before reaching the user.
- Architected the **document processing pipeline for scale**: parallel stages, workers that scale automatically with demand, and exactly-once processing. Validated at thousands of documents with a 100% success rate.
- Delivered **answer quality customers can trust**: hybrid keyword-and-meaning search, result re-ranking, Korean and English language tuning, and precise page-level citations.
- Unified **multiple AI providers behind a single gateway** with automatic failover, response caching and usage logging – sustaining service continuity while keeping model spend predictable.
- Integrated the platform with **business tools such as Slack and Google Drive**, keeping content continuously in sync while honoring each user's existing access rights.
- Secured customer data through **encryption of credentials, least-privilege access control and audit logging**.

Cloud Infrastructure & Reliability

- Own the complete **AWS infrastructure as code** for development and production – networking, compute, databases, storage, security and CI/CD – delivered as reusable modules shared across environments.
- Led two **zero-downtime migrations of the full production platform**, including the move to Kubernetes for elastic scaling.
- Drove a **cloud cost-efficiency program**, aligning every line of the monthly bill with the running architecture and delivering around **20% savings** through consolidation, scheduled automation and storage lifecycle policies.
- Built **automatic scaling** so capacity follows real demand – responsive at peak load, economical at rest.
- Established **operational readiness**: monitoring and alerting, automated backups, and rehearsed disaster-recovery procedures.
- Authored the platform's **architecture and operations documentation**, and mentor engineers across backend, AI and infrastructure.

Key technologies: Python, FastAPI, LangChain / LangGraph, NestJS, TypeScript, Next.js, PostgreSQL, Redis, vector databases, AWS, Kubernetes, Terraform, Docker

Additional projects at Uranus Software

Web3 Gaming & Web Store Platform

Delivered a scalable web store and game platform integrating blockchain for decentralized in-game assets and transactions, alongside Telegram mini-apps that widened distribution and user engagement.

Node.js, NestJS, React.js, Web3, AWS

NFT Staking & Lottery Platform

Designed and delivered a user-facing lottery application with NFT staking, using blockchain for transparent prize distribution and asset management.

Node.js, React.js, Web3, Smart contracts

EDUCATION

Bachelor of Information Technology

University of Technology and Education - The University of Danang

2011 - 2015

CAREER FOCUS

- AI platform architecture
- Technical leadership

INTERESTS

- Team building
- Training new engineers
- Traveling
- Video production & editing

Team Manager & Scrum Master *Aug, 2016 - Apr, 2024*

Neolab Vietnam - Da Nang

Led cross-functional teams of **15 to 20 engineers** delivering web and mobile platforms for Japanese enterprise clients. Owned backend architecture as Server Leader, ran agile delivery as Scrum Master, and partnered with Product Owners on requirements, scope and timelines. Mentored junior developers into independent full-stack engineers.

Selected projects

Telehealth Consultation Platform — Next Innovation Operation Inc.

Real-time online doctor consultations. Led backend development and the delivery process, designing a cloud-native backend that served iOS, Android and web clients.

Node.js, Express.js, AWS, PostgreSQL, WebSocket, Docker, CI/CD

Job-Matching Platform — Wakumo

Part-time employment marketplace. Led server-side design and implementation, engineered for high-volume traffic and fast application processing.

Python, FastAPI, React.js, Google Cloud, MongoDB, GitHub Actions

Automated Vehicle Valuation & Sales Platform — DMM Auto

Directed backend development, delivering valuation algorithms and secure transaction modules that streamlined the sales workflow end to end.

Node.js, NestJS, AWS, MySQL, Jenkins, Terraform

Wedding & Banquet Booking Platform

Reservation system with integrated scheduling for venues, vendors and customers.

Node.js, Express.js, MongoDB, AWS, Docker, CircleCI

Junior Software Engineer *Mar, 2016 - Aug, 2016*

FPT Software - Da Nang

Partnered with Japanese clients to define project scope and shape technical solutions; managed delivery schedules and quality standards.

Instructor *Jun, 2015 - Mar, 2016*

VinaEnter Academy - Da Nang

Taught PHP and Java web development, preparing IT students for their first engineering roles.