# Hong Anh Khoa (Noah) Nguyen

(825) 559-3010 | noahnghgwork@gmail.com | linkedin.com/in/noahnghg | github.com/noahnghg | noahnghg.dev

## Education

**University of Calgary** **September 2023 - Present**

*Bachelor of Science — Major Computer Science* *Calgary, AB*

- Science Co-op Internship Program
- Dean's List 2023-2024
- Courses: Web-Based Systems, Machine Learning, Operating Systems, Computer Networks, Database Design and Implementation, Data Structures and Algorithms, Computer Machinery

## Professional Experience

**Ontario Lottery and Gaming Corporation (OLG)** **May 2026 – Present**

*Applied AI Engineer Coop* *Toronto, ON*

- Accelerated job profile generation time by 80%, saving the workforce department over 150 manual hours monthly, by architecting a full-stack GenAI application using **FastAPI** and **React TypeScript**.
- Achieved 95% semantic retrieval accuracy for domain-specific workforce data by engineering a highly optimized RAG pipeline utilizing **Azure OpenAI** embeddings, LLMs, and **Azure AI Search**.
- Maintained sub-150ms API latency for high-throughput AI inference and data querying by deploying containerized backend microservices via **Docker** and optimizing relational schemas in **Azure SQL Server**.

## Projects

**Deep Notes** | *Open Source Contribution* | deepnotes.wiki **TypeScript | RAG | Vectra | Gemini | Ollama**

- Accelerated semantic search latency by 4x across thousands of notes by engineering a custom **Hybrid Retrieval Index** (Dense Vectors, BM25, Wikilinks) in **TypeScript**, utilizing **Reciprocal Rank Fusion (RRF)**.
- Optimized local vector storage by engineering a custom embedding compression pipeline, applying seeded random rotation and **int8 quantization** to reduce 768-dim float32 embeddings with under 1% cosine similarity loss.
- Reduced API inference costs by 90% by designing a context-caching mechanism for payloads exceeding 32k tokens, and developed a parallel **LLM-as-a-Judge** evaluation engine using **Ollama** for semantic answer grading.

**CoFocus** | *Backend* | *DevOps* | Github **TypeScript | Express | Prisma | PostgreSQL | Docker | GitLab CI/CD**

- Designed a relational database schema of a social task management platform, achieving 100% database schema synchronization accuracy and zero manual database configuration requirements, by defining complex relational Prisma models, custom SQL scripts, and initializing automatic migrations during backend service startup.
- Optimized chat message retrieval algorithms for real-time messaging services, reducing server payload sizes and preventing memory footprint issues under high database volume, by implementing index-optimized cursor-based query filters inside database repositories.
- Engineered transaction-safe social networking and group invitation API services, guaranteeing zero database state inconsistencies and preventing orphaned user records during concurrent edits, by structuring operations with Prisma transaction blocks and implementing an Express middleware authorization layer.
- Orchestrated a multi-service containerized deployment environment for backend and database systems, ensuring 100% database availability prior to server starts and mitigating runtime privilege-escalation vulnerabilities, by composing production Docker configurations with health checks and configuring non-root execution parameters in the backend.
- Built a multi-stage CI/CD pipeline inside Gitlab, enforcing 100% compliance with conventional commit styling and domain email policies while avoiding daemon-in-daemon security threats, by coding regex validation filters and executing container image builds rootlessly via Kaniko Project Executor.

## Technical Skills

**Languages**: Python, Go, TypeScript, JavaScript, Java, SQL, HTML5, CSS, C, MATLAB

**Libraries/Frameworks**: FastAPI, Flask, ExpressJS, ReactJS, Gin

**Machine Learning/AI**: PyTorch, LangGraph, LangChain, YOLOv8+, scikit-learn

**Infrastructure/Tools**: Azure, Docker, PostgreSQL, MySQL, Git