JinFeng He

☎ (+1) 6072807880 ✉ jeffreyhe406@gmail.com linkedin.com/in/jinfeng-he-142080302 Ithaca, USA

Education

Cornell University **Aug. 2024 – Dec. 2025 (Expected)**
Master of Engineering in Systems Engineering (Software Systems Engineering Pathway) *Ithaca, USA*

University of Toronto **Sep. 2021 – Jun. 2024**
Bachelor of Science in Computer Science (Entrepreneurship Stream) *Toronto, Canada*

Work Experience

Montaura.tech *Founder & Full-Stack Developer* **Jul. 2025 – Present**
China (remote)

- Engineered a full-stack cross-platform basketball-court-focused application in **Flutter** (using **Riverpod** state management) and **Go** (**Gin** framework) via **RESTful APIs** deployed in a Dockerized environment on **Alibaba Cloud** serving daily users of 10000+.
- Implemented low-latency, real-time interactive features including live court check-ins, instant messaging, and dynamic game updates using a **WebSocket**-based pub/sub system; leveraged **PostGIS** for efficient geospatial queries to power location-based check-in functionality within a defined geofence with integrating AMap.

Vosyn.AI *Backend/ML Engineer Intern* ***Jun. 2025 – Present***
Etobicoke, Canada

- Engineered a core component of an event-driven video processing pipeline on **GCP**, to handle transcription, speaker diarization, and audio segmentation. The entire infrastructure was provisioned as code using **Terraform**.
- Deployed transcription (**Faster Whisper**) and diarization (**NeMo MSDD**) models on Vertex AI, and developed a temporal alignment algorithm in **Python** to merge their outputs by mapping word-level timestamps to speaker time segments.
- Implemented a data processing workflow using **regex** and **Pydub** to clean transcripts and segment audio, improving the data quality for downstream translation and Text-to-Speech (TTS) models.
- Developed and containerized a scalable Python microservice with **FastAPI** and **Docker**, exposing a **RESTful** API to trigger asynchronous ML inference jobs on **Google Cloud Run.**

Alibaba Group *Software Engineer intern* **Dec. 2023 – Feb. 2024**
Hangzhou, China

- Owned the end-to-end development of a new "flash sale" feature, from API design in **Java Spring Boot** to deployment. Proactively addressed post-launch performance by first implementing a **PostgreSQL GiST** index to accelerate complex geospatial-temporal queries by 40%, and engineered a **Redis** cache-aside layer to handle high-throughput reads, further reducing database load under peak traffic.
- Enhanced service observability by instrumenting the application with custom **Micrometer** metrics, including Counters for promotion redemption events and Timers to measure API endpoint performance. Developed a **Grafana** dashboard to visualize these metrics, tracking critical KPIs like P99 latency and error rates.

USoustenir *Startup Co-founder & Lead Full-Stack Developer* **Apr. 2023 – Dec. 2023**
Scarborough, Canada

- Co-founded and developed a scalable web application using React, with Redux for state management, TypeScript, Node.js, and Express.js, providing a platform for users to discover and connect with sustainable brands.
- Managed data storage using **MongoDB** with **Mongoose ODM**, deploying the backend on **AWS EC2** instances behind an **AWS Application Load Balancer**, utilizing **Auto Scaling Groups** for dynamic scaling.
- Implemented an interactive map using **Mapbox GL JS** and **GeoJSON**, integrating geospatial queries with **MongoDB's 2dsphere indexes** to visualize local recycling facilities and sustainable clothing drop-off points.
- Deployed the frontend using **AWS Amplify** and configured **AWS CloudFront CDN** for global content delivery, enhancing load times and user experience.

Projects

Multi-Modal Clinical RAG System

- ➢ Developed **RAG** system for clinical decision support using **Python**, **PyTorch**, and **Hugging Face** integrating clinical guidelines, research papers, and EHR data with LLaMA2/MedLLaMA for evidence-based recommendations, implementing **UMLS** semantic enrichment with multi-modal retrieval strategies and comprehensive evaluation framework with 70/10/20 train/dev/test splits.

Webtama - a web application for playing Onitama on a 3D board with customizable imagery

- ➢ Engineered backend services in **Go**, leveraging advanced concurrency primitives (**goroutines, channels, sync.Pool, context.Context**) to build a high-performance P2P matchmaking system along with a scalable **worker pool** pattern with buffered channels to parallelize AI move calculations (**MiniMax** algorithm) for future bot integration.

Technical Skills

Languages: Java, Python, Go, C++, JavaScript, TypeScript, SQL, Kotlin, Rust
Frameworks & Libraries: Spring Boot, Node.js, , Express.js, gRPC, RESTful APIs, GraphQL, Microservices, Apache Kafka, RabbitMQ, Socket.io, React.js, Redux, Flutter, React Native, Three.js, Web Audio API
Database: PostgreSQL (PostGIS), MongoDB, Redis, MySQL, HBase, SQLite
AI / ML: PyTorch, Hugging Face, NumPy, Pandas, PySpark, Apache Airflow
Cloud, DevOps & Monitoring: AWS (EC2, S3, RDS), GCP (GKE), Alibaba Cloud, Kubernetes, Docker, Terraform, CI/CD (Jenkins, GitHub Actions), Istio, Git, Prometheus, Grafana, ELK Stack, OAuth 2.0, Snyk