Rui Guo

☎ 587-588-6800 ✉ ruig0401@gmail.com in linkedin.com/hoinky

Education

The University of British Columbia Sept 2024 - June 2026 (Expected)
Master of Engineering in Electrical and Computer Engineering | *GPA: 92.9/100* *Vancouver, BC, Canada*

The University of Alberta Sept 2022 - June 2024
Bachelor of Science in Computing Science with Honors | *GPA: 3.86/4.00* *Edmonton, AB, Canada*

Beihang University Sept 2019 - Sept 2022
Bachelor of Engineering in Software Engineering | *GPA: 86.7/100* *Beijing, China*

Experiences

Huawei Technologies Co., Ltd July 2025 – Dec 2025
AI Software Engineer Intern *Nanjing, Jiangsu, China*

- Developed a high-performance **Vector Search Engine** by designing an adaptive indexing architecture based on data scale, utilizing bitset-based scalar filtering to achieve over **95%** recall and a **45.9%** boost in retrieval accuracy in knowledge management and Q&A platform scenario.
- Implemented a xPU inference parallelism framework, increasing throughput by 6x and maintaining sub-100ms end-to-end latency for million-scale unstructured datasets.
- Participated in the pre-research on LLM-enhanced **Recommendation System** of Huawei Music by designing structured prompts to encode user/item attributes and behavioral sequences into semantic descriptions, utilizing multilingual-e5-base for feature extraction.
- Introduced **Multi-head Attention** and **SENet** modules to achieve the alignment between semantic embeddings and traditional ID features, and performed dimensionality reduction to optimize inference performance, achieving a **2.04%** increase in online playback duration.

Projects

Instruction-based Object Detection using Multimodal LLM | *Course Project* Jan 2025 - April 2025

- Developed an instruction-driven object detection framework based on the **Qwen2-VL-2B** multimodal model to explore LLM adaptability in structured spatial perception tasks.
- Fine-tuned Qwen2-VL using **LoRA**, implementing strategic freezing/unfreezing of the vision encoder, cross-modal aligner, and language decoder to optimize multimodal alignment.
- Engineered quantity-aware and spatial-understanding instructions with augmented samples (full-enclosure, partial-overlap, non-overlap) to improve target counting and coordinate reasoning capabilities.
- Conducted ablation studies on LoRA hyperparameters and freezing strategies.

Beihang Food Delivery | *Personal Project* Jan 2023 – Apr 2023

- Developed a full-stack campus delivery platform featuring a management backend and a user-facing client for real-time ordering and student-run logistics.
- Deployed a **Redis Sentinel cluster** (1 Master, 2 Slaves) for high-availability caching, implementing **Bloom Filters** via Factory patterns to prevent cache penetration.
- Optimized concurrency control using **Optimistic Locking** to prevent inventory overselling and **Lua scripts** with Redis tokens to ensure order idempotency.

Skills & Awards

Languages: Python, Java, HTML, CSS, R
Technologies: Spring Boot, Redis, Numpy, Sklearn, PyTorch
Awards: Beihang Freshman Scholarship (2019), Master of Engineering International Graduate Entrance Scholarship (2025)