# YAN KHANG KOH

**AI ENGINEER | PRODUCTION LLM SYSTEMS, RETRIEVAL & RANKING, EVALUATION, GPU INFERENCE**

yankhang.koh@gmail.com | linkedin.com/in/yan-khang | yankhang.com | +1 (201) 205-7835 | +65 8828 2685

## WORK EXPERIENCE

**AI Engineer | Cooler Master, Singapore** **Jul 2026 – Present**

- Shipped on-prem text-to-SQL + hybrid RAG over 3,500+ lab reports solo; 470 user queries in 15 weeks
- Replayed 69 real user queries, found and fixed 23 bad answers; gated releases on a 264-case gold set
- Rebuilt live captions on Qwen3-ASR served via vLLM on an NVIDIA DGX B200; term accuracy 58% to 86%
- Benchmarked Qwen 7B/32B/72B on 67 gold-set queries; tied at 86-87%, so shipped 7B at 3x lower latency
- Caught silent VLM table misreads; added a 2-read agreement check and corrected 46 stored tables
- Moved 4 vLLM apps to a new model gateway in 4 days; caught 16K-context overflows that disabled JSON mode

**Software Engineer Intern (AI/ML Systems) | AdvancedGRC, New York City** **May 2025 – Apr 2026**

- Built a RAG platform solo for PDF-heavy docs (FastAPI, Chroma, E5, vLLM); support tickets down 36%
- Raised judged pass rate from 67% to 98% with a Hugging Face cross-encoder and answer/abstain gating
- Built layout-aware PDF ingestion with page-anchored citations and inline figures for document QA
- Built benchmarks for warm latency, ingest throughput, concurrency and retrieval ablations; 132 ms p50
- Automated extraction across all 50 US state insurance regulator sites in Python; research time cut 60-70%

**Software Engineer | DT Asia Pte Ltd, Singapore** **Jul 2019 – Jul 2020**

- Promoted from intern; built Syslog-ng log pipelines into Azure and AWS CloudWatch, saving $50K+ a year
- Hardened SSH remote-access infrastructure for 100K+ sessions/month across client deployments

## EDUCATION

**Singapore Institute of Technology | DigiPen Institute of Technology** **Aug 2022 – May 2026**

B.Sc. (Hons) Computer Science in Real-Time Interactive Simulation

- GPA 4.74/5.00, Highest Distinction, top 2% (Provost's List x2)
- Teaching Assistant for 8 CS and math modules (200+ students); 12-month industry placement in NYC
- Coursework: data structures, analysis of algorithms, linear algebra, calculus, advanced computer graphics

## PROJECTS

**Meridian Commerce Search** **Python, PyTorch, OpenSearch, LightGBM, FastAPI**

- Built a search relevance stack: BM25 + dense retrieval, LightGBM learning-to-rank, cross-encoder rerank
- Fine-tuned a dense retriever; evaluated it on held-out splits with per-query metrics and slice analysis
- Built an ESCI/SQID data pipeline with held-out relevance splits and product image enrichment
- Served search via FastAPI and OpenSearch; added benchmark workflows for retrieval misses and regressions

**Insight Engine | Fragments** **C++20, OpenGL 4.5, GLSL 4.5**

- Led an 8-person team as PM/graphics programmer on a C++20 engine and Fragments, a showcase-selected game
- Implemented dynamic lighting, shadows and instanced rendering for fast real-time rendering pipelines
- Authored GLSL shaders for glitch effects, animated materials and blending pipelines
- Built a flexible 3D camera with perspective and orthographic modes for in-game readability

## SKILLS

**LLM Systems:** RAG, text-to-SQL, LLM evaluation, JSON-schema output, citation guardrails, VLM OCR
**Retrieval & NLP:** hybrid search (BM25, dense), embeddings (BGE-M3, E5), reranking, LTR, OpenSearch, Chroma
**Machine Learning:** PyTorch, Hugging Face Transformers, vLLM, LiteLLM, LightGBM, pandas, NumPy
**Backend & Infra:** FastAPI, WebSockets, SQLite, Docker Compose, nginx, Linux, Git, pytest
**Languages:** Python, SQL, C++, C, C#, JavaScript, GLSL, HTML/CSS, Bash

## ACHIEVEMENTS & LEADERSHIP

**President,** DigiPen Student Management Committee, 3 years of service
**Air Warfare Officer,** Republic of Singapore Air Force (RSAF), mandatory National Service, 2020–2022
**SITizen Ambassador,** invited panelist at SIT Day Zero 2024, addressing 500+ incoming ICT students
**IMDA SG Digital Scholar,** selected for a national-level digital leadership programme
**Runner-up,** CoDE Hackathon 2024, with viSIT, a unified student portal design for students