YU-PENG FU

AI · GPU · FULL-STACK ENGINEER

Profile

AI & GPU Engineer with expertise in **algorithms** and **full-stack** development. Skilled in **optimizing AI inference** and building **scalable RAG pipelines** that connect **large language models** to real-world data. Experienced in **GPU acceleration, backend systems**, and delivering **high-performance**, **production-ready AI solutions**

Experience

AI · GPU · BACKEND ENGINEER

Kog AI, Paris (Hybrid) | 2023/04 - Present

- Contributed to securing **€5M in venture funding** by developing **backend systems** and applying **model fine-tuning** and **prompt engineering** to enhance the company's flagship AI product.
- Led the design and development of a **custom inference engine from scratch**, including low-level **GPU kernel optimization** and **mathematical innovations**, achieving **3× faster inference performance than vLLM**
- Represented the company at the **RAISE Summit** and earned **2nd place** at the **AMD AI Sprint Hackathon** for groundbreaking work in **GPU optimization** and **AI acceleration**
- Implemented **model fine-tuning pipelines** for domain-specific LLM adaptation. Incorporated LoRA and quantization strategies to optimize inference latency and memory footprint

Education

DIGITAL TECHNOLOGY ARCHITECT PROGRAM

L'école 42, Lyon

- **Project-based** curriculum emphasizing **autonomy** and **real-world problem-solving**
- Ranked **top 10 nationwide** in **algorithm competitions** representing the school
- Built **algorithm projects** with **~100 GitHub stars** and a Medium technical article reaching **~100k views**

MATHEMATICS

National Taiwan University, Taipei

GPA 4.3 (Top 1%)

Contact

https://yu-peng-fu.vercel.app

(+33) 6 33 86 51 48

leofu890806@gmail.com

Lyon, France (Open to Remote)

www.linkedin.com/in/yu-peng-fu

https://github.com/LeoFu9487

Certifications

IBM RAG and Agentic AI Professional Certificate
IBM

Achievements

AMD AI Sprint Hackathon 2025
2nd Place
AMD

La Coding Battle 2021
7th Place
Le Shaker

BattleDev 2021
18th Place
BDM

Projects

ModelPulse
Self-improving GPU-accelerated RAG system (Open-source)

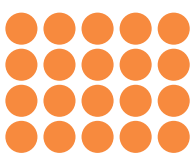
Ysoft Management App
Production-grade web app actively used by Taiwanese vendors

Push_Swap Test Suite & Tutorial
End-to-end educational and testing toolkit for École 42

Skills

AI & GPU Computing

RAG & Inference Eng
GPU Optimization
CUDA | HIP | Triton
vLLM | PyTorch



Full-Stack Development

TS | JS | Python
Git | Linux | Docker
Next.js | Node | SQL