



สถาบันไอจีเนียส
IT GENIUS INSTITUTE

COURSE ID
AIC-85

AI / LOCAL LLM

COURSE OUTLINE · รายละเอียดหลักสูตรอบรม

Local LLM Deployment with Ollama and vLLM

รัน LLM ในองค์กรเองด้วย Ollama และ vLLM โดยข้อมูลไม่ออกนอกองค์กร

● 12 ชั่วโมง

● 2 วัน

● พื้นฐาน Linux, Docker และ Python

● ระดับ Intermediate



ระยะเวลา: 2 วัน

12 ชั่วโมง (2 วัน วันละ 6 ชม.) · 09:00-16:00 น.



รูปแบบ: Classroom / Hybrid

Onsite หรือ Online ผ่าน Microsoft Teams · Public / In-house



รายละเอียดหลักสูตร

www.itgenius.co.th

หลายองค์กรอยากใช้ LLM กับเอกสารภายใน ข้อมูลลูกค้า หรือซอร์สโค้ด แต่ติดเงื่อนไขว่าข้อมูลห้ามส่งออกไปยังผู้ให้บริการภายนอก ทางออกคือรันโมเดล Open-weight บนเครื่องหรือเซิร์ฟเวอร์ขององค์กรเอง แต่พอลงมือจริงทีม IT ต้องตอบคำถามที่ไม่ง่ายเลย ว่าควรเลือกโมเดลไหน ต้องใช้ GPU ที่ GB จะ Quantize แค่นี้โดยคุณภาพยังรับได้ ควรใช้ Ollama หรือ vLLM และจะเปิดให้ทั้งองค์กรใช้อย่างปลอดภัยได้อย่างไร

หลักสูตรนี้พาผู้เรียนตั้งระบบ LLM ภายในองค์กรตั้งแต่ต้นจนใช้งานได้จริง เริ่มจากเลือกโมเดลและอ่านเงื่อนไข License เข้าใจ Quantization และรูปแบบ GGUF คำนวณ VRAM, RAM และ Context Length ที่ต้องใช้ จากนั้นใช้ Ollama อย่างลึกถึง Modelfile, การตั้งค่าการทำงานพร้อมกัน และ OpenAI-compatible API ที่รองรับ Structured Output และ Tool Calling ก่อนขยับไปรัน vLLM บน GPU เพื่อรองรับผู้ใช้จำนวนมาก ปรับจูนหน่วยความจำ วัดผลด้วย Benchmark ต่อหน้าแชตด้วย Open WebUI และวางระบบความปลอดภัยกับ Monitoring แล้วปิดท้ายด้วย Capstone ออกแบบและติดตั้งระบบ LLM ภายในองค์กรของตนเอง (อบรม 2 วัน วันละ 6 ชั่วโมง รวม 12 ชั่วโมง ระดับ Intermediate)

เครื่องมือ

OI Ollama

vL vLLM

OW Open WebUI

Dk Docker

Qw Qwen / Llama

ราคาค่าอบรม · PUBLIC / IN-HOUSE

7,900 บาท / ท่าน

ราคายังไม่รวมภาษีมูลค่าเพิ่ม (VAT 7%) · รวมเอกสารประกอบการอบรม Lab Guide ไฟล์ docker-compose และโค้ดตัวอย่างครบทุก Lab ตารางคำนวณขนาด GPU และ Certificate · ค่า Cloud GPU (ถ้าเลือกใช้) ผู้เรียนรับผิดชอบเอง · รับผิดชอบ In-house ขอใบเสนอราคาได้

สมัครผ่านเว็บไซต์ ลด 10%



เพิ่มเพื่อน
LINE

☎ 02-570-8449

☎ 088-807-9770

LINE @itgenius

🌐 www.itgenius.co.th

01 วัตถุประสงค์ Objectives

- เลือกโมเดล Open-weight ให้เหมาะกับงาน ภาษาไทย ฮาร์ดแวร์ และเงื่อนไข License ได้
- เข้าใจ Quantization และเลือกระดับความละเอียดของโมเดลโดยชั่งคุณภาพกับทรัพยากรได้
- คำนวณ VRAM และ RAM ที่ต้องใช้จากขนาดโมเดล Context Length และจำนวนผู้ใช้พร้อมกันได้
- ติดตั้งและปรับแต่ง Ollama ด้วย Modelfile และตัวแปรสภาพแวดล้อมสำหรับงานจริงได้
- เรียกโมเดลผ่าน OpenAI-compatible API พร้อม Structured Output, Tool Calling และ Embedding ได้
- รับและปรับจูน vLLM บน GPU เพื่อรองรับคำขอพร้อมกันจำนวนมากได้
- วางระบบ LLM ภายในองค์กรที่มีหน้าเซต การยืนยันตัวตน และการ Monitoring ได้

02 กลุ่มเป้าหมาย Target Audience

- ทีม IT Infrastructure และ System Engineer ที่ได้รับโจทย์ให้ตั้ง LLM ใช้ภายในองค์กร
- DevOps และ Platform Engineer ที่ต้องดูแลบริการ AI บนเซิร์ฟเวอร์หรือ Cloud ขององค์กร
- นักพัฒนาที่ต้องการ API ของ LLM ที่ข้อมูลไม่ออกนอกองค์กรสำหรับแอปภายใน
- AI Engineer ที่ต้องเปรียบเทียบต้นทุนและประสิทธิภาพระหว่าง Cloud API กับการรันเอง
- องค์กรที่มีข้อกำหนดเรื่องความลับของข้อมูลหรือ PDPA ทำให้ใช้ AI บน Cloud ได้จำกัด

03 จุดเด่นของหลักสูตร Highlights

- **ข้อมูลไม่ออกนอกองค์กร** รันโมเดลบนเครื่องหรือเซิร์ฟเวอร์ของตัวเอง
- **Ollama และ vLLM ครบคู่** รู้ว่าเมื่อไรควรใช้ตัวไหน พร้อมวัดผลจริง
- **คำนวณ GPU เป็น** VRAM, KV Cache และ Quantization
- **API มาตรฐานเดียว** OpenAI-compatible ใช้กับโค้ดเดิมได้
- **พร้อมขึ้นระบบจริง** Open WebUI, Reverse Proxy และ Monitoring

04 ความรู้พื้นฐาน Prerequisites

- ใช้ Linux Command Line และ Docker ขึ้นพื้นฐานได้
- เขียน Python เบื้องต้นและเรียก REST API ได้
- เคยใช้ ChatGPT หรือ LLM ตัวอื่นและเข้าใจแนวคิด Prompt เบื้องต้น
- โน้ตบุ๊ก RAM 16 GB ขึ้นไปที่ติดตั้ง Docker ได้ และบัญชี Google สำหรับ Google Colab

05 สิ่งที่ต้องเตรียมและเครื่องมือ Setup & Tools

รายการ	รายละเอียด
อุปกรณ์	Notebook RAM 16 GB ขึ้นไป สำหรับรันโมเดลขนาดเล็กด้วย Ollama บนเครื่อง
GPU	Google Colab หรือ Cloud GPU ที่ผู้เรียนเช่าเอง สำหรับ Lab ของ vLLM
Inference Engine	Ollama, vLLM และ llama.cpp สำหรับเปรียบเทียบรูปแบบ GGUF
ส่วนติดต่อและ API	Open WebUI, Python OpenAI SDK และ Docker Compose
โมเดลตัวอย่าง	โมเดล Open-weight เช่น Qwen, Llama, Gemma และโมเดลภาษาไทยอย่าง Typhoon

หมายเหตุ: Lab ของ Ollama ทำบนโน้ตบุ๊กของผู้เรียนด้วยโมเดลขนาดเล็ก ส่วน Lab ของ vLLM ต้องใช้ GPU ผู้เรียนใช้ Google Colab (GPU ของรุ่นฟรีมีจำกัดและอาจไม่ว่างเสมอ) หรือเช่า Cloud GPU เอง ค่าใช้จ่ายส่วนนี้ผู้เรียนรับผิดชอบเอง หลักสูตรนี้เน้นงานติดตั้งและให้บริการโมเดล (Model Serving) ต่างจากหลักสูตรสร้าง AI Chatbots On-Premise ด้วย LangChain และ LangGraph ที่เน้นการเขียนแอปแชทบอท จึงเรียนต่อกันได้ ชื่อเครื่องมือ รุ่นโมเดล และฟีเจอร์เปลี่ยนเร็ว เนื้อหาจะอัปเดตตามเวอร์ชันล่าสุด ณ วันอบรม

06 ภาพรวมการเรียนรู้ 2 วัน Big Picture

วัน	หัวข้อหลัก
Day 1 Section 1-5	เลือกโมเดล Open-weight ให้เหมาะกับงานและภาษาไทย, เข้าใจ Quantization และ GGUF, คำนวณ VRAM และ RAM ที่ต้องใช้, ใช้ Ollama อย่างลึกด้วย Modelfile และเรียกผ่าน OpenAI-compatible API
Day 2 Section 6-10	รัน vLLM บน GPU, ปรับจูน Throughput และหน่วยความจำ, วัดผลด้วย Benchmark, ต่อ Open WebUI, วางระบบความปลอดภัยและ Monitoring แล้วปิดท้ายด้วย Capstone ระบบ LLM ภายในองค์กร

DAY
1

เลือกโมเดล คำนวณทรัพยากร และใช้ Ollama อย่างลึก

เนื้อหาการอบรม · Section 1-5

รวม
6 ชั่วโมง

1 ทำไมต้องรัน LLM เอง และภาพรวมเครื่องมือ

- เหตุผลที่องค์กรเลือกรันโมเดลเอง: ความลับของข้อมูล PDPA ต้นทุน และการทำงานแบบออฟไลน์
- ข้อจำกัดเมื่อเทียบกับโมเดลบน Cloud ทั้งคุณภาพ ภาระดูแล และค่าฮาร์ดแวร์
- ภาพรวม Inference Engine: Ollama, llama.cpp, vLLM และสถาปัตยกรรมอ้างอิงของระบบ LLM ภายในองค์กร
- Lab: ติดตั้ง Ollama ดึงโมเดลแรกและคุยผ่าน Command Line

2 เลือกโมเดล Open-weight ให้เหมาะกับงาน

- ตระกูลโมเดลที่ใช้อยู่ เช่น Qwen, Llama, Gemma และโมเดลภาษาไทยอย่าง Typhoon
- Base, Instruct และ Reasoning Model ต่างกันอย่างไร และเลือกใช้เมื่อไร
- อ่าน Model Card และเงื่อนไข License ก่อนนำไปใช้เชิงพาณิชย์
- ทดสอบความสามารถภาษาไทยด้วยชุดคำถามจากงานจริงขององค์กร
- Lab: เทียบคำตอบภาษาไทยของโมเดล 3 ตัวจากโจทย์เดียวกัน

3 Lab: Quantization และรูปแบบ GGUF

- Quantization คืออะไร และทำไมโมเดลใหญ่จึงรันบนเครื่องเล็กได้
- ระดับที่ใช้อยู่ใน GGUF เช่น Q4_K_M, Q5_K_M และ Q8_0
- รูปแบบสำหรับ GPU อย่าง AWQ, GPTQ และ FP8 ที่ใช้กับ vLLM
- ผลของ Quantization ต่อคุณภาพคำตอบ ความเร็ว และหน่วยความจำ
- Lab: รันโมเดลเดียวกันสองระดับความละเอียดแล้วเทียบผลลัพธ์

4 Workshop: คำนวณ GPU และหน่วยความจำ

- สูตรประมาณหน่วยความจำของน้ำหนักโมเดลจากจำนวนพารามิเตอร์และ Precision
- KV Cache คืออะไร และโตขึ้นอย่างไรตาม Context Length และผู้ใช้พร้อมกัน
- รันบน CPU, GPU ตัวเดียว หลายตัว หรือ Apple Silicon ต่างกันอย่างไร
- เลือก GPU และ Cloud Instance ให้คุ้มกับปริมาณงาน
- Workshop: ทำตารางคำนวณขนาดเครื่องสำหรับโจทย์ขององค์กรตัวเอง

5 Lab: Ollama และ API มาตรฐาน

- สร้างโมเดลขององค์กรด้วย Modelfile: System Prompt, Template และ Parameter
- นำเข้าไฟล์ GGUF จาก Hugging Face และตั้งค่า Context Length, Keep Alive และคำขอพร้อมกัน
- รัน Ollama ใน Docker พร้อม GPU และเปิดให้เครื่องอื่นในเครือข่ายเรียกใช้
- เรียกผ่าน OpenAI-compatible API โดยแก้แค่ Base URL ของโค้ดเดิม
- Structured Output, Tool Calling และ Embedding กับโมเดลภายใน
- Lab: สคริปต์ Python ดึงข้อมูลจากเอกสารโดยไม่ส่งข้อมูลออกนอกเครื่อง



DAY
2

vLLM และการให้บริการ LLM ทั้งองค์กร

เนื้อหาการอบรม · Section 6-10

รวม
6 ชั่วโมง

6 Lab: vLLM บน GPU และการปรับจูน

- ทำไม vLLM รับผู้ใช้พร้อมกันได้มาก: PagedAttention และ Continuous Batching
- ติดตั้งบน Google Colab หรือ Cloud GPU และรันด้วย Docker Image ทางกาย
- เปิด OpenAI-compatible Server ด้วย vllm serve พร้อม API Key, Tool Calling และ Structured Output
- ตั้งค่า Max Model Length และ GPU Memory Utilization ให้สมดุล
- ใช้โมเดล AWQ, GPTQ หรือ FP8 และเปิด Prefix Caching เพื่อคุ้ม GPU
- Tensor Parallel สำหรับหลาย GPU และให้บริการ LoRA Adapter หลายตัวบนโมเดลฐานเดียว

7 Lab: Benchmark และเลือก Ollama หรือ vLLM

- ตัวชี้วัดที่ต้องดู: Time to First Token, Tokens per Second และ Throughput รวม
- ทดสอบโหลดด้วยเครื่องมือ Benchmark ของ vLLM และสคริปต์ยิงคำขอพร้อมกัน
- อ่านผลเมื่อจำนวนผู้ใช้เพิ่มขึ้น และหาจุดที่ระบบเริ่มช้า
- เปรียบเทียบระหว่าง Ollama กับ vLLM ตามขนาดทีมและฮาร์ดแวร์
- Lab: วัดผลโมเดลเดียวกันบนทั้งสองระบบแล้วสรุปเป็นข้อเสนอ

8 Lab: หน้าเซตองค์กรด้วย Open WebUI

- ติดตั้ง Open WebUI ด้วย Docker Compose และเชื่อมกับ Ollama และ vLLM
- จัดการผู้ใช้ บทบาท และสิทธิ์การเข้าถึงแต่ละโมเดล
- อัปโหลดเอกสารเพื่อถามตอบเบื้องต้นด้วยโมเดลภายใน
- ตั้ง System Prompt และโมเดลสำเร็จรูปสำหรับแต่ละแผนก
- เชื่อนไข License ของ Open WebUI ที่ต้องรู้ก่อนใช้งานในองค์กร

9 ความปลอดภัยและการดูแลระบบ

- ปิดพอร์ตที่ไม่จำเป็นและวาง Reverse Proxy พร้อม TLS และ API Key
- แยกเครือข่าย จำกัดผู้เรียกใช้ และเก็บ Log ตามนโยบายข้อมูลขององค์กร
- Monitoring ด้วย Metrics ของ vLLM ร่วมกับ Prometheus และ Grafana
- อัปเดตโมเดลและ Inference Engine อย่างปลอดภัยโดยไม่กระทบผู้ใช้
- Lab: ตั้ง Dashboard ดูการใช้ GPU และเวลาตอบของระบบ

10 Capstone: ระบบ LLM ภายในองค์กร

- รับโจทย์จากสถานการณ์ขององค์กรตัวอย่าง พร้อมข้อจำกัดด้านข้อมูลและงบประมาณ
- เลือกโมเดล ระดับ Quantization และ Inference Engine พร้อมเหตุผล
- ประกอบระบบด้วย Docker Compose: Inference, Open WebUI และ Reverse Proxy
- ทดสอบการใช้งานและโหลด แล้วนำเสนอแผนขยายระบบ
- สรุป Checklist ก่อนเปิดให้ผู้ใช้ทั้งองค์กร

07 เรียนจบแล้วได้อะไร Outcomes

- เลือกโมเดลได้ถูกงาน คุณภาพ ภาษาไทย และ License
- คำนวณฮาร์ดแวร์เป็น VRAM, KV Cache และ Quantization
- ใช้ Ollama คล่อง Modelfile, Docker และ API
- รัน vLLM บน GPU ได้ ปรับจูนและวัดผลด้วย Benchmark
- เปิดบริการทั้งองค์กร Open WebUI และสิทธิ์ผู้ใช้
- ดูแลระบบอย่างปลอดภัย Reverse Proxy, TLS และ Monitoring

08 สิ่งที่คุณเรียนจะได้รับ Deliverables

- เอกสารประกอบการอบรมฉบับสมบูรณ์ครบทั้ง 2 วัน และ Lab Guide ที่ละเอียด
- ไฟล์ docker-compose สำหรับ Ollama, vLLM, Open WebUI และ Reverse Proxy
- ตารางคำนวณขนาด GPU และหน่วยความจำ พร้อมสคริปต์ Benchmark และโค้ด Python ตัวอย่าง
- ประกาศนียบัตรจบหลักสูตร (Certificate of Completion) และช่องทางสอบถามวิทยากรหลังจบหลักสูตร

09 ข้อมูลผู้สอน Instructor

อ.สามิตร โทยม ผู้สอนและวิศวกรที่ติดตั้งและดูแลระบบ LLM บนเซิร์ฟเวอร์ขององค์กรในงานจริง ทำงานเลือกโมเดล วางขนาดเครื่อง และเปิดบริการให้ผู้ใช้จำนวนมาก จุดเด่นคือการสอนด้วยตัวเลขจริงจากการวัดผล ไม่ใช้การคาดเดา พาผู้เรียนคำนวณทรัพยากร ทดลองเทียบโมเดลและ Inference Engine ด้วยตัวเอง และอธิบายเหตุผลเบื้องหลังทุกการตั้งค่า เพื่อให้กลับไปออกแบบระบบ LLM ภายในองค์กรได้อย่างมั่นใจและคุ้มค่าใช้จ่ายได้

ราคาค่าอบรม 7,900 บาท/ท่าน · สนใจ IN-HOUSE ขอใบเสนอราคา

7,900 บาท / ท่าน (ยังไม่รวม VAT 7%)

สมัครผ่านเว็บไซต์ ลด 10%

☎ 02-570-8449

📞 @itgenius

☎ 088-807-9770

✉ contact@itgenius.co.th



เพิ่มเพื่อน
LINE