



LLM Fine-Tuning with LoRA

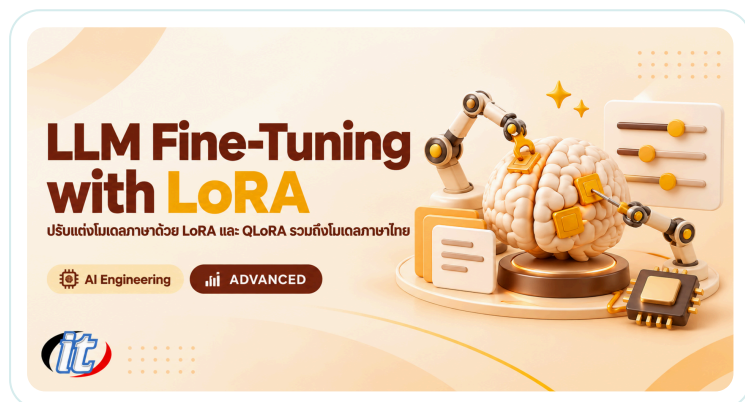
ปรับแต่งโมเดลภาษาด้วย LoRA และ QLoRA รวมถึงโมเดลภาษาไทย

● 18 ชั่วโมง

● 3 วัน

● พื้นฐาน Python และใช้ LLM API

● ระดับ Intermediate - Advanced



ระยะเวลา: 3 วัน

18 ชั่วโมง (3 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

Prompt ที่ดีและ RAG แก้ปัญหาได้หลายอย่าง แต่บางงานก็ไปไม่สุด เช่น อยากรู้ให้โมเดลตอบในรูปแบบเฉพาะขององค์กรทุกครั้ง ใช้ศัพท์เฉพาะทางได้ถูกต้อง จัดหมวดหมู่เอกสารภาษาไทยได้แม่นยำ หรือใช้โมเดลขนาดเล็กที่รันในองค์กรได้แทนโมเดลใหญ่ราคาแพง Fine-tuning ด้วย LoRA และ QLoRA ทำให้การปรับแต่งโมเดลแบบนี้ทำได้บน GPU ตัวเดียว แต่ผลลัพธ์จะดีหรือไม่ขึ้นกับการตัดสินใจที่ถูกต้อง ชุดข้อมูลที่มีคุณภาพ และการวัดผลอย่างเป็นระบบ มากกว่าการกดรัน Notebook ให้อจบ

หลักสูตรนี้พาผู้เรียน Fine-tune LLM แบบลงมือทำครบวงจร เริ่มจากเกณฑ์ตัดสินใจว่างานไหนควร Fine-tune เข้าใจหลักการของ LoRA และ QLoRA และออกแบบชุดข้อมูลฝึกพร้อมทั้งข้อมูลภาษาไทย จากนั้นฝึกโมเดลจริงด้วย Unsloth และ Hugging Face TRL กับ PEFT ปรับ Hyperparameter อ่าน Loss Curve และทดลองกับโมเดลที่รองรับภาษาไทยอย่าง Typhoon หรือ OpenThaiGPT พร้อมวัดผลเทียบกับโมเดลตั้งต้น ก่อนต่อยอดสู่ Preference Tuning ด้วย DPO การ Merge และ Export เป็น GGUF แล้วนำขึ้นใช้งานบน Ollama และ vLLM ปิดท้ายด้วย Capstone Fine-tune โมเดลสำหรับงานของตนเองตั้งแต่ชุดข้อมูลจนถึงการ Deploy (อบรม 3 วัน วันละ 6 ชั่วโมง รวม 18 ชั่วโมง ระดับ Intermediate - Advanced)

เครื่องมือ

HF

Hugging Face

Un

Unsloth

Lo

LoRA / QLoRA

Py

Python

OI

Ollama

ราคาค่าอบรม · PUBLIC / IN-HOUSE

11,900 บาท / ท่าน

ราคายังไม่รวมภาษีมูลค่าเพิ่ม (VAT 7%) · รวมเอกสารประกอบการอบรม Lab Guide Notebook ตัวอย่างครบทุก Lab ชุดข้อมูลฝึกตัวอย่าง และ Certificate · ค่า Cloud GPU (ถ้าเลือกใช้) ผู้เรียนรับผิดชอบเอง · รับจัด In-house ขอใบเสนอราคาได้

สมัครผ่านเว็บไซต์ ลด 10%



เพิ่มเพื่อน
LINE

01 วัตถุประสงค์ Objectives

- ตัดสินใจได้ว่างานใดควรใช้ Prompt, RAG หรือ Fine-tuning โดยมีเหตุผลรองรับ
- อธิบายหลักการของ LoRA และ QLoRA และเลือกค่าพารามิเตอร์สำคัญได้
- ออกแบบ ทำความสะอาด และจัดรูปแบบชุดข้อมูลฝึกตาม Chat Template ของโมเดลได้
- Fine-tune โมเดลด้วย Unsloth และ Hugging Face TRL กับ PEFT บน GPU ตัวเดียวได้
- Fine-tune โมเดลที่รองรับภาษาไทยสำหรับงานเฉพาะขององค์กรได้
- วัดผลโมเดลที่ฝึกแล้วเทียบกับโมเดลตั้งต้นอย่างเป็นระบบ
- Merge, Export เป็น GGUF และนำโมเดลขึ้นให้บริการบน Ollama หรือ vLLM ได้

02 กลุ่มเป้าหมาย Target Audience

- AI Engineer และ ML Engineer ที่ต้องปรับแต่งโมเดลภาษาให้เหมาะกับงานขององค์กร
- นักพัฒนา Python ที่ใช้ LLM API อยู่แล้วและอยากก้าวไปสู่การปรับแต่งโมเดลเอง
- Data Scientist ที่ต้องการทักษะ Fine-tuning LLM บนข้อมูลภาษาไทย
- ทีมที่รัน LLM ภายในองค์กรและอยากให้โมเดลขนาดเล็กทำงานเฉพาะทางได้ดีขึ้น
- นักวิจัยและอาจารย์ที่ต้องการแนวปฏิบัติ Fine-tuning ที่ทำซ้ำได้

03 จุดเด่นของหลักสูตร Highlights

- รู้ว่าเมื่อไรควร **Fine-tune** เทียบกับ Prompt และ RAG อย่างมีเกณฑ์
- ฝึกได้บน **GPU ตัวเดียว** QLoRA บน Colab หรือ Cloud GPU
- เน้น **โมเดลภาษาไทย** Typhoon, OpenThaiGPT และ Qwen
- **ข้อมูลมาก่อนโค้ด** ออกแบบ ทำความสะอาด และ PDPA
- ฝึกเสร็จแล้วใช้งานได้ GGUF, Ollama และ vLLM

04 ความรู้พื้นฐาน Prerequisites

- เขียน Python ได้คล่อง และเคยใช้ Jupyter Notebook หรือ Google Colab
- เคยเรียก LLM ผ่าน API และเข้าใจ Prompt, Token และ Context Window
- มีพื้นฐาน Machine Learning เบื้องต้น เช่น Training, Validation และ Overfitting จะเป็นประโยชน์
- โน้ตบุ๊กเชื่อมต่ออินเทอร์เน็ต บัญชี Google สำหรับ Colab และบัญชี Hugging Face

05 สิ่งที่ต้องเตรียมและเครื่องมือ Setup & Tools

รายการ	รายละเอียด
GPU	Google Colab หรือ Cloud GPU ที่ผู้เรียนเช่าเอง ไม่ต้องมี GPU บนเครื่อง
Fine-tuning	Unsloth, Hugging Face Transformers, PEFT, TRL และ bitsandbytes
โมเดลตั้งต้น	โมเดล Open-weight ขนาดเล็กถึงกลาง เช่น Qwen, Llama, Gemma, Typhoon และ OpenThaiGPT
ติดตามการฝึก	TensorBoard หรือ Weights & Biases (บัญชีฟรี) และ Hugging Face Hub
Deploy	llama.cpp สำหรับ GGUF, Ollama และ vLLM

หมายเหตุ: ทุก Lab รับบน Google Colab (GPU ของรุ่นฟรีมีเวลาและโควตาจำกัด) หรือ Cloud GPU ที่ผู้เรียนเช่าเอง ค่าใช้จ่ายส่วนนี้ ผู้เรียนรับผิดชอบเอง ผู้ที่มี GPU บนเครื่องก็ใช้ได้ โมเดลใน Lab เลือกขนาดเล็กพอที่จะฝึกเสร็จในเวลาเรียน ผู้เรียนต้องตรวจสอบ License ของโมเดลตั้งต้นและข้อมูลเองก่อนนำไปใช้เชิงพาณิชย์ หลักสูตรนี้ต่อยอดได้กับหลักสูตร Local LLM Deployment with Ollama and vLLM สำหรับงานให้บริการโมเดลในระดับองค์กร ชื่อเครื่องมือ รุ่นโมเดล และฟีเจอร์เปลี่ยนเร็ว เนื้อหาจะอัปเดตตามเวอร์ชันล่าสุด ณ วันอบรม

06 ภาพรวมการเรียนรู้ 3 วัน Big Picture

วัน	หัวข้อหลัก
Day 1 Section 1-5	ตัดสินใจว่าเมื่อไรควร Fine-tune แทน Prompt หรือ RAG, เข้าใจ LoRA และ QLoRA, เตรียมสภาพแวดล้อมบน Colab หรือ Cloud GPU และออกแบบชุดข้อมูลฝึกที่มีคุณภาพ รวมถึงข้อมูลภาษาไทย
Day 2 Section 6-10	Fine-tune ด้วย QLoRA ผ่าน Unsloth, ปรับ Hyperparameter และอ่าน Loss Curve, ฝึกโมเดลภาษาไทยอย่าง Typhoon, ใช้ Hugging Face TRL และ PEFT โดยตรง และวัดผลเทียบกับโมเดลตั้งต้น
Day 3 Section 11-15	รู้จัก Preference Tuning ด้วย DPO, Merge และ Export เป็น GGUF, นำโมเดลขึ้น Ollama และ vLLM, ดูแลเรื่อง License ข้อมูล และความปลอดภัย แล้วปิดท้ายด้วย Capstone Fine-tune โมเดลของตัวเอง



DAY
1

หลักการ สภาพแวดล้อม และชุดข้อมูล

เนื้อหาการอบรม · Section 1-5

รวม
6 ชั่วโมง

1

Workshop: Prompt, RAG หรือ Fine-tuning

- Fine-tuning แก้ปัญหาอะไรได้ และแก้ไม่ได้ เช่น การเพิ่มความรู้ที่เปลี่ยนแปลง
- งานที่ได้ผลจาก Fine-tuning: รูปแบบคำตอบ, การจัดหมวด, การดึงข้อมูล และศัพท์เฉพาะทาง
- ต้นทุนรวมของ Fine-tuning ทั้งข้อมูล GPU เวลา และการดูแลโมเดลหลังฝึก
- ใช้โมเดลเล็กที่ฝึกแล้วแทนโมเดลใหญ่เพื่อลดต้นทุนและรันในองค์กรได้
- Workshop: ประเมินโจทย์ขององค์กรตัวเองด้วย Decision Checklist

2

LLM ถูกฝึกอย่างไร และหลักการ LoRA กับ QLoRA

- Pre-training, Supervised Fine-tuning และ Preference Tuning ต่างกันอย่างไร
- Full Fine-tuning เทียบกับ Parameter-Efficient Fine-tuning
- LoRA ทำงานอย่างไร และความหมายของ Rank, Alpha และ Target Modules
- QLoRA: โหลดโมเดลแบบ 4-bit เพื่อฝึกบน GPU ที่หน่วยความจำจำกัด
- ประมาณหน่วยความจำ GPU ที่ต้องใช้ตามขนาดโมเดลและวิธีฝึก

3

Lab: เตรียมสภาพแวดล้อม Fine-tuning

- เลือกระหว่าง Google Colab, Cloud GPU ที่เช่าเอง หรือ GPU บนเครื่อง
- ติดตั้ง Unsloth, Transformers, PEFT, TRL และ bitsandbytes ให้เวอร์ชันเข้ากัน
- ใช้ Hugging Face Hub: Token, การดาวน์โหลดโมเดล และ Repository แบบ Private
- ภาพรวม Unsloth Studio ที่ใช้งานแบบไม่ต้องเขียนโค้ด
- Lab: รัน Inference โมเดลตั้งต้นและบันทึกผลไว้เป็น Baseline

4

ออกแบบชุดข้อมูลฝึก

- รูปแบบข้อมูลแบบ Instruction และแบบ Conversation พร้อม Chat Template ของแต่ละโมเดล
- คุณภาพสำคัญกว่าปริมาณ: ความหลากหลาย ความสม่ำเสมอ และตัวอย่างขอบเขตงาน
- สร้างข้อมูลสังเคราะห์ด้วย LLM แล้วให้คนตรวจทาน
- ข้อควรระวังกับข้อมูลภาษาไทย ทั้งการตัดคำ จำนวน Token และความสม่ำเสมอของภาษา
- PDPA และการลบข้อมูลส่วนบุคคลก่อนนำข้อมูลลงเครื่องไปฝึก

5

Lab: เตรียมข้อมูลให้พร้อมฝึก

- ทำความสะอาด ลบข้อมูลซ้ำ และกรองตัวอย่างคุณภาพต่ำ
- แบ่งชุด Train, Validation และ Test โดยไม่ให้อข้อมูลรั่วข้ามชุด
- แปลงข้อมูลเป็น Chat Template และตรวจความยาว Token ของแต่ละตัวอย่าง
- บันทึกเป็น Dataset บน Hugging Face Hub แบบ Private
- Lab: เตรียมชุดข้อมูลภาษาไทยสำหรับงานจัดหมวดและตอบคำถามของธุรกิจตัวอย่าง



DAY
2

ฝึกโมเดลและวัดผล

เนื้อหาการอบรม · Section 6-10

รวม
6 ชั่วโมง

6 Lab: QLoRA ตัวแรกด้วย Unsloth

- โหลดโมเดลแบบ 4-bit และเพิ่ม LoRA Adapter
- ตั้งค่า SFTTrainer และฝึกเฉพาะส่วนคำตอบ (Train on Completions)
- ติดตาม Loss และการใช้ GPU ระหว่างฝึก
- ทดสอบโมเดลหลังฝึกและเทียบกับ Baseline
- บันทึกเฉพาะ Adapter และโหลดกลับมาใช้งาน

7 Hyperparameter และการอ่าน Loss Curve

- Learning Rate, Epoch, Batch Size และ Gradient Accumulation
- เลือก Rank และ Target Modules ให้สมดุลระหว่างคุณภาพกับหน่วยความจำ
- อ่าน Training Loss กับ Validation Loss เพื่อจับ Overfitting
- Checkpoint, Early Stopping และการบันทึกการทดลองด้วย TensorBoard หรือ W&B
- Lab: ฝึกซ้ำด้วยค่าต่างกันแล้วเทียบผลอย่างเป็นระบบ

8 Lab: Fine-tune โมเดลภาษาไทย

- ภาพรวมโมเดลที่รองรับภาษาไทย เช่น Typhoon, OpenThaiGPT และ Qwen
- ประสิทธิภาพของ Tokenizer กับภาษาไทย และผลต่อความยาวข้อมูลและต้นทุน
- ตรวจสอบ License ของโมเดลภาษาไทยก่อนเลือกเป็นโมเดลตั้งต้น
- Lab: Fine-tune โมเดลภาษาไทยให้ตอบตามรูปแบบและน้ำเสียงขององค์กร
- เทียบผลกับการเขียน Prompt อย่างเดียวบนโมเดลเดียวกัน

9 Lab: Hugging Face TRL และ PEFT โดยตรง

- เขียน Training Script ด้วย TRL และ PEFT โดยไม่พึ่ง Unsloth
- เทียบความเร็วและหน่วยความจำกับการฝึกด้วย Unsloth
- ใช้ Accelerate สำหรับการฝึกบนหลาย GPU เบื้องต้น
- จัดโครงสร้างโปรเจกต์ให้ทำซ้ำได้: Config, Seed และ Version ของข้อมูล

10 Lab: วัดผลโมเดลที่ฝึกแล้ว

- สร้างชุดทดสอบจากงานจริงที่ไม่เคยอยู่ในข้อมูลฝึก
- ตัวชี้วัดตามชนิดงาน เช่น Accuracy, F1 และ Exact Match สำหรับการจัดหมวดและดึงข้อมูล
- ใช้ LLM-as-a-Judge ร่วมกับการตรวจโดยคนสำหรับงานที่ตอบเป็นข้อความ
- ตรวจสอบว่าโมเดลยังทำงานทั่วไปไม่ได้ ไม่ลืมความสามารถเดิม (Catastrophic Forgetting)
- Lab: ทำรายงานเทียบโมเดลตั้งต้น Prompt อย่างเดียว และโมเดลที่ Fine-tune



DAY
3

Preference Tuning, Deploy และ Capstone

เนื้อหาการอบรม · Section 11-15

รวม
6 ชั่วโมง

11 Lab: Preference Tuning ด้วย DPO

- ทำไมต้องปรับตามความชอบ เมื่อ SFT อย่างเดียวยังไม่พอ
- หลักการ DPO และรูปแบบข้อมูลคู่คำตอบที่ดีกว่าและแย่กว่า
- ภาพรวม Reinforcement Learning แบบ GRPO สำหรับงานที่ตรวจคำตอบได้
- Lab: ปรับโมเดลด้วย DPO ให้ตอบสุภาพและกระชับตามแนวทางองค์กร

12 Lab: Merge และ Export โมเดล

- เก็บแบบ Adapter แยก หรือ Merge เข้ากับโมเดลตั้งต้น เลือกอย่างไร
- Merge Adapter และบันทึกเป็นโมเดลเต็มบน Hugging Face Hub แบบ Private
- แปลงเป็น GGUF และเลือกระดับ Quantization
- ตรวจสอบคุณภาพหลัง Quantize ด้วยชุดทดสอบเดิม
- เขียน Model Card บอกข้อมูลฝึก ข้อจำกัด และวิธีใช้งาน

13 Lab: นำโมเดลขึ้นใช้งาน

- สร้าง Modelfile และรันโมเดลที่ฝึกเองบน Ollama
- ให้บริการบน vLLM และโหลด LoRA Adapter หลายตัวบนโมเดลตั้งต้นเดียว
- เรียกใช้ผ่าน OpenAI-compatible API จากแอปเดิม
- Lab: เชื่อมโมเดลที่ฝึกเองเข้ากับสคริปต์งานจริงของธุรกิจตัวอย่าง

14 License, ความปลอดภัย และการดูแลโมเดล

- License ของโมเดลตั้งต้นและข้อมูล กับสิทธิ์การใช้เชิงพาณิชย์
- ทดสอบความปลอดภัยของโมเดลหลังฝึก เช่น การตอบเรื่องที่ไม่ควรตอบ และข้อมูลรั่ว
- ระบบ Version ของข้อมูล Adapter และโมเดล เพื่อย้อนกลับได้
- วางรอบการฝึกใหม่เมื่อข้อมูลหรือโมเดลตั้งต้นเปลี่ยน
- ประเมินต้นทุนการฝึกและการให้บริการเทียบกับการใช้ API

15 Capstone: Fine-tune โมเดลสำหรับงานของตัวเอง

- เลือกโจทย์จากงานขององค์กรตัวเองหรือโจทย์ตัวอย่างที่เตรียมไว้
- เตรียมข้อมูล ฝึก และปรับจูนอย่างน้อยสองรอบ
- วัดผลเทียบกับ Baseline และสรุปว่าคุ้มค่าหรือไม่
- Export และรันบน Ollama หรือ vLLM แล้วนำเสนอผลพร้อมรีวิวก่อน

07 เรียนจบแล้วได้อะไร Outcomes

- **ตัดสินใจได้อย่างมีเหตุผล** Prompt, RAG หรือ Fine-tuning
- **ออกแบบข้อมูลฝึกเป็น** Chat Template, คุณภาพ และ PDPA
- **Fine-tune ได้เอง** QLoRA ด้วย Unsloth, TRL และ PEFT
- **ปรับโมเดลภาษาไทย** Typhoon, OpenThaiGPT และ Qwen
- **วัดผลอย่างเป็นระบบ** Baseline, ชุดทดสอบ และ LLM Judge
- **นำขึ้นใช้งานได้จริง** GGUF, Ollama และ vLLM

08 สิ่งที่คุณเรียนจะได้รับ Deliverables

- เอกสารประกอบการอบรมฉบับสมบูรณ์ครบทั้ง 3 วัน และ Lab Guide ที่ละเอียด
- Notebook ตัวอย่างครบทุก Lab ทั้งแบบ Unsloth และแบบ TRL กับ PEFT
- ชุดข้อมูลฝึกตัวอย่าง สคริปต์วัดผล และ Decision Checklist สำหรับเลือกแนวทาง
- ประกาศนียบัตรจบหลักสูตร (Certificate of Completion) และช่องทางสอบถามวิทยากรหลังจบหลักสูตร

09 ข้อมูลผู้สอน Instructor

อ.สาธิต ไทยม ผู้สอนและ AI Engineer ที่ Fine-tune และนำโมเดลภาษาไปใช้งานจริง ทั้งงานภาษาไทยและงานเฉพาะทางขององค์กร จุดเด่นคือการสอนให้คิดเรื่องข้อมูลและการวัดผลก่อนเรื่องโค้ด พาผู้เรียนทดลองเปลี่ยนค่าต่าง ๆ แล้วดูผลด้วยตัวเอง อธิบายเหตุผลของแต่ละการตั้งค่าแบบเข้าใจง่าย และชี้ให้เห็นกรณีที่ไม่ควร Fine-tune เพื่อให้ผู้เรียนกลับไปเลือกแนวทางและลงทุนกับ GPU ได้อย่างคุ้มค่า

ราคาค่าอบรม 11,900 บาท/ท่าน · สนใจ IN-HOUSE ขอใบเสนอราคา

11,900 บาท / ท่าน (ยังไม่รวม VAT 7%)

สมัครผ่านเว็บไซต์ ลด 10%

☎ 02-570-8449

📞 @itgenius

☎ 088-807-9770

✉ contact@itgenius.co.th



เพิ่มเพื่อน
LINE