

คำนำ

เมื่อ ChatGPT เปิดตัวขึ้น หลายคนในวงการ AI รวมถึงผู้เขียน ต่างรู้สึกประหลาดใจ เพราะเป็นจุดเริ่มต้นให้เกิดแอปพลิเคชันจำนวนมาก เป็นตัวเร่งที่ปลดล็อกให้เกิดแอปพลิเคชันที่ไม่น่าจะเป็นไปได้มากมาย

ไม่เพียงแค่นั้น ความสามารถใหม่ของ AI ยังลดอุปสรรคให้กับนักพัฒนาอีกด้วย การสร้างแอปพลิเคชัน AI กลายเป็นเรื่องง่ายมาก แม้กระทั่งสร้างโดยไม่ต้องเขียนโค้ดเลย การเปลี่ยนแปลงนี้ทำให้ AI กลายเป็นเครื่องมือพัฒนาทรงพลัง...ที่ทุกคนสามารถใช้ได้

ผู้เขียนตั้งใจเขียนหนังสือเล่มนี้จากการเรียนรู้ และใช้ความรู้มากมายที่ได้จากโครงการที่ทำ, งานวิจัยที่อ่าน และการสัมภาษณ์ผู้คนจำนวนมาก เช่น นักวิจัยจากห้องปฏิบัติการ AI ชั้นนำ (OpenAI, Google, Anthropic, ...), นักพัฒนาเฟรมเวิร์ค (NVIDIA, Meta, Hugging Face, Anyscale, LangChain, LlamaIndex, ...), ผู้บริหารและหัวหน้าฝ่าย AI/ฝ่ายข้อมูลในบริษัทต่างๆ, ผู้จัดการผลิตภัณฑ์ รวมถึงนักพัฒนาแอปพลิเคชันอิสระ ด้วยความตั้งใจให้หนังสือเล่มนี้ครอบคลุมกระบวนการนำโมเดลรากฐาน (foundation model) ไปใช้งาน ตั้งแต่ต้นจนจบ โดยมีการรวมเทคนิคที่ได้รับการพิสูจน์แล้วจากสาขาวิศวกรรมอื่นๆ และเทคนิคที่เกิดขึ้นมาใหม่พร้อมกับโมเดลรากฐานเข้าไปในเนื้อหาด้วย

ผู้เขียนหวังว่ากระบวนการเรียนรู้นี้จะดำเนินต่อไปเมื่อหนังสืออยู่ในมือของผู้อ่าน ด้วยประสบการณ์และมุมมองของผู้อ่านแต่ละคน ซึ่งหากผู้ใดมีความคิดเห็นเกี่ยวกับหนังสือเล่มนี้ ก็สามารถแลกเปลี่ยนกันได้ ผ่าน X (<https://x.com/chipro>), LinkedIn (<https://www.linkedin.com/in/chiphuyen>) หรืออีเมลที่ hi@huyenchip.com

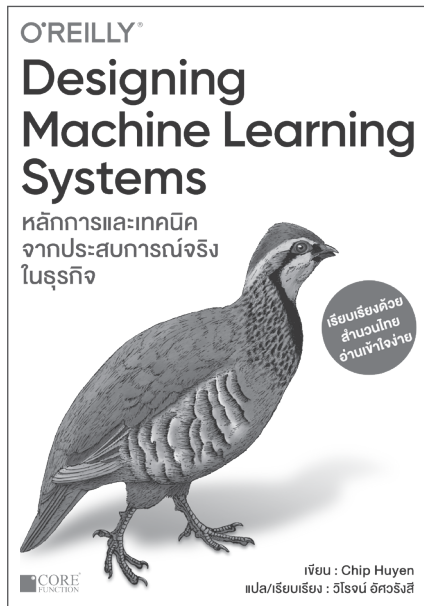
เนื้อหาของหนังสือเล่มนี้

เนื้อหาในหนังสือเล่มนี้จะให้กรอบแนวคิดในการปรับโมเดลรากฐาน เช่น Large Language Model (LLM) และ Large Multimodal Model (LMM) เพื่อการสร้างแอปพลิเคชัน โดยตอบคำถามที่สำคัญ เช่น

- ควรสร้างแอปพลิเคชัน AI นี้หรือไม่?
- เราจะประเมินแอปพลิเคชันอย่างไร? สามารถใช้ AI เพื่อประเมินผลลัพธ์ของ AI ได้หรือไม่?
- “การหลอน” (hallucination) ของโมเดลเกิดจากอะไร? และจะตรวจจับหรือแก้ไขอย่างไร
- แนวทางปฏิบัติสำหรับการทำวิศวกรรมพรอมต์ (prompt engineering) คืออะไร?
- RAG มีการทำงานอย่างไร?
- วิธีสร้างและประเมินเอเจนต์ (agent)
- เมื่อใดควรปรับแต่งโมเดล (finetune) และเมื่อใดไม่ควรทำ
- เราต้องการข้อมูลมากแค่ไหน? และมีวิธีตรวจสอบคุณภาพของข้อมูลอย่างไร?
- วิธีทำให้โมเดลเร็วขึ้น ถูกลง และปลอดภัย
- วิธีสร้างวงจรผลตอบกลับจากผู้ใช้ (feedback loop) เพื่อนำมาปรับปรุงแอปพลิเคชันได้อย่างต่อเนื่อง

เนื้อหาในหนังสือเล่มนี้ถูกนำเสนอผ่านกรณีศึกษา ซึ่งหลายกรณีเป็นงานที่ผู้เขียนเคยทำมา พร้อมด้วยข้อมูลอ้างอิงมากมาย โดยมุ่งเน้นไปที่กระบวนการพื้นฐานของวิศวกรรม AI แทนที่จะเน้นเครื่องมือหรือ API ตัวใดตัวหนึ่งโดยเฉพาะ เนื่องจากเครื่องมืออาจล้าสมัยได้อย่างรวดเร็ว แต่ความรู้พื้นฐานจะยังใช้อ้างอิงได้ยาวนานกว่า

การอ่านหนังสือ AI Engineering ควบคู่กับ Designing Machine Learning Systems



คุณสามารถอ่านหนังสือเล่มนี้ร่วมกับ Designing Machine Learning Systems (DMLS) ซึ่งเป็นหนังสืออีกเล่มของผู้เขียน (มีฉบับแปลภาษาไทย) โดยที่ DMLS มุ่งเน้นไปที่การสร้างแอปพลิเคชันบนโมเดล Machine Learning (ML) แบบดั้งเดิม ในขณะที่หนังสือเล่มนี้มุ่งเน้นไปที่การสร้างแอปพลิเคชันบนโมเดลรากฐาน (foundation model) ซึ่งเกี่ยวข้องกับการออกแบบพรอมต์, การสร้างบริบท (context) และการปรับแต่งโมเดล

ทั้งนี้ ระบบที่ใช้ในโลกจริงมักเกี่ยวข้องกับทั้งโมเดล ML แบบดั้งเดิมและโมเดลรากฐาน ดังนั้นความรู้ในการทำงานกับโมเดลทั้งสองประเภทจึงถือว่ายังเป็นสิ่งจำเป็น

อย่างไรก็ตามหนังสือเล่มนี้ไม่ใช่คู่มือสอนการใช้งาน (ถึงแม้ว่าจะมีการกล่าวถึงเครื่องมือ และตัวอย่างโค้ดเพื่ออธิบายแนวคิดบางประการ) โดยให้กรอบแนวคิดสำหรับการเลือกเครื่องมือ แทนที่จะสอนวิธีการใช้เครื่องมือโดยตรง นอกจากนี้หนังสือเล่มนี้ยังไม่ใช่นำเสนอเกี่ยวกับทฤษฎีด้าน ML (Machine Learning) เพราะเนื้อหาไม่ได้อธิบายว่าโครงข่ายประสาทเทียม (neural network) คืออะไร และไม่ได้อธิบายวิธีสร้างหรือฝึกโมเดลตั้งแต่เริ่มต้น

หนังสือเล่มนี้เหมาะกับใคร

หนังสือเล่มนี้เหมาะสำหรับผู้ที่ต้องการใช้โมเดลรากฐาน (foundation model) มาสร้างแอปพลิเคชันจริง โดยเฉพาะผู้ที่มีบทบาททางด้านเทคนิค เช่น วิศวกร AI, วิศวกร ML, นักวิทยาศาสตร์ข้อมูล, ผู้จัดการด้านวิศวกรรม หรือผู้จัดการผลิตภัณฑ์ทางเทคนิค นอกจากนี้เนื้อหาในเล่มยังมีประโยชน์ต่อ

- นักพัฒนาเครื่องมือที่ต้องการขยายขอบเขตสู่การทำวิศวกรรม AI
- นักวิจัยที่ต้องการรู้ถึงการประยุกต์ใช้ AI ด้านต่างๆ
- ผู้สมัครงานที่ต้องการทราบทักษะที่จำเป็นสำหรับการเป็นวิศวกร AI
- ผู้ที่ต้องการเข้าใจความสามารถ และข้อจำกัดของ AI

หนังสือเล่มนี้อาจมีเนื้อหาที่ลงลึกด้านเทคนิคในบางส่วน แต่ผู้เขียนขอแนะนำว่า หากรู้สึกว่าการเนื้อหาแน่นซับซ้อนเกินไป ก็สามารถข้ามไปยังส่วนถัดไปได้!

สัญลักษณ์ในเล่ม



แสดงถึง เคล็ดลับหรือคำแนะนำ



แสดงถึง บันทึกหรือหมายเหตุ



แสดงถึง คำเตือนหรือข้อควรระวัง

สารบัญ

บทที่ 1 การสร้างแอปพลิเคชัน AI ด้วยโมเดลรากฐาน	19
การเติบโตของวิศวกรรม AI	20
จากโมเดลภาษา สู่โมเดลภาษาขนาดใหญ่	20
จากโมเดลภาษาขนาดใหญ่ สู่โมเดลรากฐาน	25
จากโมเดลรากฐานสู่วิศวกรรม AI	28
กรณีใช้งานของโมเดลรากฐาน	31
การเขียนโค้ด	35
การสร้างภาพและวิดีโอ	36
การเขียน	37
การศึกษา	37
บอตสนทนา	38
การรวบรวมและสรุปข้อมูล	39
การจัดระเบียบข้อมูล	39
ระบบที่ทำงานโดยอัตโนมัติ	40
การวางแผนสร้างแอปพลิเคชัน AI	40
ประเมินการนำมาใช้งาน	41
กำหนดความคาดหวัง	44
การวางแผนความก้าวหน้า	45
การดูแลรักษา	45
องค์ประกอบทางวิศวกรรม AI	47
สามเลเยอร์ของระบบ AI	48
วิศวกรรม ML กับวิศวกรรม AI	50
วิศวกรรม AI เทียบกับวิศวกรรมเต็มรูปแบบ	57
สรุป	58
บทที่ 2 ทำความเข้าใจในโมเดลรากฐาน	61
ข้อมูลสำหรับการฝึก	61
โมเดลหลายภาษา	62

โมเดลเฉพาะด้าน	65
การสร้างแบบจำลอง	67
สถาปัตยกรรมของโมเดล	67
ขนาดของโมเดล	75
การฝึกเพิ่มเติม	82
การปรับแต่งแบบมีผู้กำกับ	85
การปรับแต่งตามความชอบของมนุษย์	88
การสุ่มเลือก	91
พื้นฐานของการสุ่มเลือก	92
วิธีในการสุ่มเลือก	94
การทดสอบผลลัพธ์ขณะใช้งาน	98
ผลลัพธ์แบบมีโครงสร้าง	100
ธรรมชาติความน่าจะเป็นของ AI	105
สรุป	111
 บทที่ 3 การประเมินผล	113
ความท้าทายในการประเมินผลโมเดลรากฐาน	114
ทำความเข้าใจกับตัววัดของโมเดลภาษา	117
Entropy	118
Cross Entropy	119
Bits-per-Character (BPC) และ Bits-per-Byte (BPB)	120
Perplexity	120
การตีความค่า Perplexity และกรณีใช้งาน	121
การประเมินแบบชัดเจน	123
ความถูกต้องของการทำงาน	123
การวัดความเหมือนกับข้อมูลอ้างอิง	125
รู้จักกับ Embedding	129
AI ในฐานะผู้ตัดสิน	131
ทำไมต้องใช้ AI เป็นผู้ตัดสิน?	132
การใช้ AI judge	133
ข้อจำกัดของ AI ในฐานะผู้ตัดสิน	137

โมเดลแบบใดที่ เป็นผู้ตัดสินใจได้	141
การจัดอันดับโมเดลด้วยการเปรียบเทียบ	143
ความท้าทายในการประเมินแบบเปรียบเทียบ	146
อนาคตของการประเมินแบบเปรียบเทียบ	148
สรุป	149
บทที่ 4 การประเมินระบบ AI	151
เกณฑ์สำหรับการประเมิน	151
ความสามารถเฉพาะด้าน	153
ความสามารถในการสร้างข้อความ	154
ความสามารถในการทำตามคำสั่ง	161
ค่าใช้จ่ายและเวลาในการตอบสนอง	165
การเลือกโมเดล	167
ขั้นตอนการเลือกโมเดล	167
จะสร้างโมเดลหรือซื้อโมเดล	169
สำรวจผลเบนซ์มาร์คสาธารณะ	176
การออกแบบขั้นตอนประเมิน	183
ขั้นตอนที่ 1: ประเมินทุกองค์ประกอบในระบบ	184
ขั้นตอนที่ 2: สร้างแนวทางการประเมิน	185
ขั้นตอนที่ 3: กำหนดวิธีประเมินและข้อมูลที่ใช้	187
สรุป	191
บทที่ 5 วิศวกรรมพรอมต์	193
การเขียนพรอมต์เบื้องต้น	194
การเรียนรู้จากบริบท: Zero-Shot และ Few-Shot	195
System Prompt และ User Prompt	197
ความยาวและประสิทธิภาพของบริบท	200
แนวทางปฏิบัติในการทำวิศวกรรมพรอมต์	202
เขียนคำสั่งให้ชัดเจนและเจาะจง	202
ให้บริบทที่เพียงพอ	206
แยกงานที่ซับซ้อนออกเป็นขั้นตอนย่อยๆ	207

ให้โมเดลมีเวลาคิด	210
ค่อยๆ เรียนรู้แล้วนำมาปรับปรุงพรมต์	212
เครื่องมือที่ใช้ในวิศวกรรมพรมต์	213
การจัดการพรมต์	216
วิศวกรรมพรมต์เชิงป้องกัน	218
พรมต์ที่มีเจ้าของ และการวิศวกรรมย้อนกลับ	219
การเจลเบรก และการฝังกำสั่งแอบแฝง	220
การดึงข้อมูล	225
การป้องกันการโจมตีด้วยพรมต์	228
สรุป	231
บทที่ 6 RAG และเอเจนต์	233
RAG	233
สถาปัตยกรรม RAG	235
อัลกอริทึมในการดึงข้อมูล	236
การเพิ่มประสิทธิภาพการดึงข้อมูล	248
RAG กับข้อมูลที่ไม่ใช่ข้อความ	253
เอเจนต์ (Agent)	256
ภาพรวมของเอเจนต์	257
เครื่องมือ	259
การวางแผน	262
ความล้มเหลวของเอเจนต์ และการประเมินผล	278
ความจำของโมเดล	281
สรุป	283
บทที่ 7 การปรับแต่งโมเดล	287
ภาพรวมของการปรับแต่งเฉพาะงาน	287
ควรทำการปรับแต่งเมื่อไร	290
เหตุผลที่ควรปรับแต่ง	290
เหตุผลที่ไม่ควรปรับแต่ง	290
การปรับแต่งและ RAG	294

ข้อจำกัดด้านหน่วยความจำ	297
การแพร่ย้อนกลับ และพารามิเตอร์ที่จะฝึกใหม่	298
การคำนวณหน่วยความจำ	300
รูปแบบการแทนตัวเลข	302
การลดความละเอียดของตัวเลข	305
เทคนิคการปรับแต่งโมเดล	308
การปรับแต่งที่ใช้พารามิเตอร์อย่างมีประสิทธิภาพ	308
การรวมโมเดลและการปรับแต่งสำหรับหลายงาน	321
กลยุทธ์ในการปรับแต่งโมเดล	331
สรุป	336

บทที่ 8 การทำวิศวกรรมชุดข้อมูล	339
การดูแลจัดการข้อมูล	341
คุณภาพของข้อมูล	343
ความครอบคลุมของข้อมูล	344
ปริมาณข้อมูล	346
การจัดหาและใส่คำอธิบายข้อมูล	350
การเสริมข้อมูลและการสังเคราะห์ข้อมูล	352
ทำไมต้องใช้ข้อมูลสังเคราะห์	352
เทคนิคดั้งเดิมในการสังเคราะห์ข้อมูล	354
การสังเคราะห์ข้อมูลด้วย AI	356
Model Distillation	361
การประมวลผลข้อมูล	361
ตรวจสอบข้อมูล	362
ข้อมูลซ้ำ	364
การคลีนและคัดกรองข้อมูล	366
การจัดรูปแบบข้อมูล	367
สรุป	368

บทที่ 9 การเพิ่มประสิทธิภาพการทำงานของโมเดล	371
ความเข้าใจด้านการเพิ่มประสิทธิภาพการทำงานของโมเดล	372
ภาพรวมของการอินเฟอร์เรนซ์	372
ตัววัดประสิทธิภาพการอินเฟอร์เรนซ์	377
ตัวเร่งการประมวลผล AI	384
การเพิ่มประสิทธิภาพการอินเฟอร์เรนซ์	390
การเพิ่มประสิทธิภาพในระดับโมเดล	391
การเพิ่มประสิทธิภาพบริการอินเฟอร์เรนซ์	406
สรุป	414
 บทที่ 10 โครงสร้างของวิศวกรรม AI และผลตอบกลับจากผู้ใช้	 417
โครงสร้างของการทำวิศวกรรม AI	417
ขั้นตอนที่ 1: เพิ่มบริบท	418
ขั้นตอนที่ 2: ใส่ระบบป้องกัน	419
ขั้นตอนที่ 3: เพิ่มตัวกำหนดเส้นทางโมเดลและเกตเวย์	422
ขั้นตอนที่ 4: ลดเวลาแฝงด้วยแคช	426
ขั้นตอนที่ 5: เพิ่มเอเจนต์	429
การเฝ้าระวังและสังเกตการณ์	430
การประสานการทำงานของระบบ AI	436
ผลตอบกลับจากผู้ใช้งาน	438
การดึงฟีดแบ็คจากบทสนทนา	439
การออกแบบระบบรับฟีดแบ็ค	443
ข้อจำกัดของฟีดแบ็ค	449
สรุป	452