

THE CONTEXT ENGINEER

สถาปนิกผู้อยู่เบื้องหลัง
ความฉลาดที่แท้จริงของ AI

อนุชิต ชโลธ

สำนักพิมพ์ก๊อปวาง

The Context Engineer

สถาปนิกผู้อยู่เบื้องหลังความฉลาดที่แท้จริงของ AI

อนุชิต ชโลธร

คำนำ

เรากำลังอยู่ในยุคที่ Generative AI ไม่ใช่เรื่องไกลตัวอีกต่อไป เทคโนโลยีนี้ได้กลายเป็นจุดเปลี่ยนครั้งสำคัญที่สั่นสะเทือนวงการเทคโนโลยีและธุรกิจในรูปแบบที่ไม่เคยมีมาก่อน เราต่างตื่นตะลึงกับความสามารถของ Large Language Models (LLMs) ที่สามารถสนทนา, สร้างสรรค์คอนเทนต์ และเขียนโค้ดได้อย่างน่าทึ่ง เสมือนมีผู้ช่วยที่รอบรู้ที่สุดในโลกอยู่แค่ปลายนิ้ว

แต่เมื่อเราพยายามนำพลังอันมหาศาลนี้มาใช้ในโลกธุรกิจจริง เรากลับพบกับช่องว่างขนาดใหญ่ที่เรียกว่า **“The Context Gap”**

LLMs ที่เราใช้กันนั้นถูกฝึกฝนมาจากข้อมูลมหาศาลบนอินเทอร์เน็ต มันรู้เรื่องประวัติศาสตร์โลก เขียนบทกวีได้ แต่กลับไม่รู้จักข้อมูลภายในบริษัทของคุณ ไม่เข้าใจกระบวนการทำงานที่เป็นเอกลักษณ์ของคุณ และไม่มีทางเข้าถึงเอกสารล่าสุดที่ทีมของคุณเพิ่งร่างเสร็จ เมื่อวานนี้ ผลลัพธ์ที่ได้คือคำตอบที่ผิดพลาด, ข้อมูลที่ไม่เป็นปัจจุบัน หรือที่ร้ายแรงที่สุดคือการ “จินตนาการ” คำตอบขึ้นมาเอง (Hallucination) ซึ่งเป็นความเสี่ยงที่ไม่มีองค์กรใดยอมรับได้

จากความท้าทายนี้เอง ได้ถือกำเนิดบทบาทใหม่ที่กำลังจะกลายเป็นหัวใจสำคัญของการพัฒนา AI ในยุคถัดไป นั่นคือ **“Context Engineer”** — สถาปนิกผู้อยู่เบื้องหลังความฉลาดที่แท้จริงของ AI

Context Engineer ไม่ใช่แค่คนที่เขียน Prompt ได้ดี แต่คือผู้ที่ออกแบบและสร้างสะพานเชื่อมระหว่างโลกความรู้อันกว้างใหญ่ของ LLM เข้ากับโลกข้อมูลเฉพาะทางขององค์กร พวกเขาคือผู้สร้าง “สมองส่วนหน้า” ให้กับ AI ทำให้มันสามารถเข้าถึง, ทำความเข้าใจ และใช้ “บริบท” ที่ถูกต้องในการให้คำตอบที่แม่นยำ, ปลอดภัย และมีประโยชน์อย่างแท้จริง

หนังสือเล่มนี้เหมาะกับใคร

หนังสือเล่มนี้ถูกเขียนขึ้นสำหรับนักพัฒนา, Data Scientist, AI Engineer, Product Manager และผู้ที่สนใจเทคโนโลยี AI ทุกคนที่ต้องการก้าวข้ามจากการเป็น “ผู้ใช้” ไปสู่การเป็น “ผู้สร้าง” แอปพลิเคชัน AI ที่ใช้งานได้จริงและสร้างผลกระทบทางธุรกิจ

คุณจะได้เรียนรู้อะไร

เราจะเดินทางไปด้วยกันตั้งแต่การปูพื้นฐานว่าทำไม “บริบท” ถึงสำคัญ, ไปจนถึงการลงมือสร้างสถาปัตยกรรมที่ซับซ้อนอย่าง Retrieval-Augmented Generation (RAG) คุณจะได้เรียนรู้เทคนิคการจัดการข้อมูล, ทำความรู้จักกับเครื่องมือสำคัญอย่าง Vector Databases และลงมือทำโปรเจกต์จริง 3 โปรเจกต์ที่จะเปลี่ยน AI ทั่วไปให้กลายเป็นผู้ช่วยอัจฉริยะสำหรับองค์กรของคุณ

เมื่ออ่านจบ คุณจะไม่เพียงเข้าใจ “ว่า” Context Engineering คืออะไร แต่จะ “ทำเป็น” และพร้อมที่จะเป็นสถาปนิกคนสำคัญ ผู้นำพาองค์กรของคุณเข้าสู่ยุค AI อย่างเต็มศักยภาพ

สารบัญ

บทที่ 1 ไชยปริศนา LLMs และจุดบอดที่มองไม่เห็น	1
LLMs คืออะไรและทำงานอย่างไร (ฉบับเข้าใจง่าย)	1
ปัญหาภาพหลอน (Hallucination) และความรู้ที่ล้าสมัย	2
ทำไม LLM ถึงไม่รู้จักข้อมูลบริษัทของคุณ	3
กรณีศึกษา: เมื่อ Chatbot ตอบผิดพลาดเพราะขาดบริบท	4
บทที่ 2 นิยามบทบาท “Context Engineer”	6
Context Engineer คือใคร? ทำหน้าที่อะไร?	6
เปรียบเทียบความแตกต่าง Prompt Engineer vs. Context Engineer vs. Data Engineer	7
ทักษะที่จำเป็น: การผสมผสานระหว่าง Software Engineering, Data Science และ AI	9
หน้าที่หลักของ Context Engineer	10
บทที่ 3 องค์ประกอบสำคัญของ “บริบท”	13
รู้จักบริบท 5 มิติ (The 5 Dimensions of Context): เลนส์สู่ความเข้าใจอย่างลึกซึ้ง	13
บริบทผู้ใช้ (User Context): เลนส์ที่มองไปยัง “ตัวตน” ของผู้ถาม	13
บริบทข้อมูล (Data Context): เลนส์ที่มองไปยัง “แหล่งความจริง”	15

บริบทเวลา (Temporal Context): เลนส์ที่มองไปยัง “เส้นเวลา”	16
บริบทบทสนทนา (Conversational Context): เลนส์ที่มองไปยัง “ความต่อเนื่อง”	17
บริบทของระบบ (System Context): เลนส์ที่มองมายัง “ตัวเอง”	19
หลักการออกแบบ: จาก 5 มิติสู่ “สุดยอด Prompt”	20
บทที่ 4 การจัดหาและเตรียมข้อมูล (Data Sourcing & Preparation)	22
แหล่งที่มาของบริบท: จาก Databases, APIs, PDFs สู่เว็บไซต์	22
เทคนิคการตัดแบ่งข้อมูล (Chunking Strategies) เพื่อประสิทธิภาพสูงสุด	23
การทำความสะอาดและแปลงข้อมูลให้อยู่ในรูปแบบที่ AI เข้าใจ	25
บทที่ 5 หัวใจของระบบค้นหา: Embeddings และ Vector Databases	27
Embeddings คืออะไร: การแปลงความหมายให้เป็นตัวเลข	27
ทำไมต้องใช้ฐานข้อมูลเวกเตอร์ (Vector Database)	28
เปรียบเทียบ Vector Databases ยอดนิยม	29
ภาพรวมการสร้าง Vector Database	30
บทที่ 6 สถาปัตยกรรม RAG (Retrieval-Augmented Generation)	32

RAG คืออะไร: สถาปัตยกรรมที่เปลี่ยนเกมสำหรับ LLMs	32
ผ่านขั้นตอนการทำงานของ RAG: Query -> Retrieve -> Augment -> Generate	33
การออกแบบ Retriever ที่ดี: มากกว่าแค่การค้นหาคำที่ตรงกัน	35
Prompt Engineering สำหรับ RAG	36
บทที่ 7 ก้าวข้าม RAG: สู่อสถาปัตยกรรมขั้นสูง	38
Agents & Function Calling: เมื่อ AI ต้อง “ลงมือทำ” ไม่ใช่แค่ “ตอบ”	38
Fine-tuning: ควรใช้เมื่อไหร่ และต่างจาก RAG อย่างไร	40
Graph RAG: การใช้ประโยชน์จากความสัมพันธ์ของข้อมูล	42
บทที่ 8 สร้างแอปถาม-ตอบเอกสาร	44
เป้าหมาย	44
ภาพรวมสถาปัตยกรรม	45
เครื่องมือที่เราจะใช้	46
ขั้นตอนการสร้างจาก Core Engine สู่ Web App	47
บทที่ 9 สร้างผู้ช่วยวิเคราะห์ข้อมูลลูกค้าอัจฉริยะ	55
เป้าหมาย	55
ภาพรวมสถาปัตยกรรม	55
เครื่องมือที่เราจะใช้	56
ขั้นตอนการสร้าง	57
บทที่ 10 สร้างระบบสรุปข่าวสารอัตโนมัติ	61

เป้าหมาย	61
ภาพรวมสถาปัตยกรรม	62
เครื่องมือที่เราจะใช้	62
ขั้นตอนการสร้าง	63
บทที่ 11 การวัดผลและเพิ่มประสิทธิภาพ	67
ทำไมการวัดผลถึงสำคัญและท้าทาย	67
RAG Triad: 3 เมตริกสำคัญสำหรับระบบ RAG	67
เครื่องมือช่วยวัดผล (Evaluation Frameworks)	69
เทคนิคการดีบั๊กและเพิ่มประสิทธิภาพ	69
บทที่ 12 การนำระบบขึ้นใช้งานจริง	71
ความท้าทายในการ Scale ระบบ RAG	71
สถาปัตยกรรมสำหรับ Production-grade RAG	72
การวางแผนสถาปัตยกรรมและทรัพยากร (Architecture & Resource Planning)	73
การปฏิบัติการและปรับปรุงระบบใน Production (Operations & Improvement)	79
ความปลอดภัยและความเป็นส่วนตัว (Security & Privacy)	85
บทที่ 13 อนาคตของ Context Engineering	86
Multi-modal RAG: เมื่อบริบทไม่ใช่แค่ข้อความ	86
Personalized AI: ประสบการณ์เฉพาะบุคคล	87

AI Agents ที่ซับซ้อนและการทำงานร่วมกัน (Multi-agent Systems)	87
การผสมผสานระหว่าง RAG และ Fine-tuning	88
บทบาทของ Context Engineer ในวันข้างหน้า	89
บทส่งท้าย คุณคือสถาปนิกแห่งอนาคต	90
สรุปแก่นความรู้ทั้งหมด	90
เส้นทางของคุณเริ่มต้นจากตรงนี้	91
ภาคผนวก	93
คำศัพท์ที่ควรรู้	93
รายชื่อเครื่องมือและเฟรมเวิร์กที่แนะนำ	95
Embedding Models	95
LLM Providers / APIs	95
Agentic Frameworks	96
UI/Frontend Tools	96
Frameworks for Building LLM Apps	97
Vector Databases	97
Evaluation Tools	98
Deployment & Monitoring	98
ดาวนโหลดซอร์สโค้ด	99

บทที่ 1 ไชยปริศนา LLMs และจุดบอดที่มองไม่เห็น

ยินดีต้อนรับสู่ภาคแรกของหนังสือ ที่เราจะเริ่มต้นการเดินทางด้วยการทำความเข้าใจหัวใจของ Generative AI นั่นคือ Large Language Models หรือ LLMs ก่อนที่เราจะสร้างสถาปัตยกรรมที่ซับซ้อน เราจำเป็นต้องเข้าใจธรรมชาติของเครื่องมือที่เราจะใช้เสียก่อน ทั้งในแง่ความสามารถอันน่าทึ่งและข้อจำกัดที่สำคัญยิ่ง

LLMs คืออะไรและทำงานอย่างไร (ฉบับเข้าใจง่าย)

ลองจินตนาการว่าคุณกำลังอ่านหนังสือเล่มหนึ่ง แล้วมีคนมาฉีกประโยคสุดท้ายของย่อหน้าออกไป จากนั้นขอให้คุณลองเดาคำต่อไปคืออะไร คุณจะทำอย่างไร?

คุณคงจะอ่านประโยคก่อนหน้าทั้งหมด ทำความเข้าใจเนื้อหา แล้วใช้ความรู้และประสบการณ์ทางภาษาของคุณเพื่อคาดเดาคำที่น่าจะเป็นไปได้มากที่สุดใช่ไหมครับ

LLMs ทำงานในลักษณะที่คล้ายกันมาก แต่ในสเกลที่ใหญ่กว่าพันล้านเท่า หัวใจของมันคือโครงข่ายประสาทเทียม (Neural Network) ขนาดมหึมาที่ถูกฝึกฝนด้วยข้อมูลตัวอักษร (Text Data) จำนวนมหาศาลจากทั่วทั้งอินเทอร์เน็ต ไม่ว่าจะเป็นบทความ, หนังสือ, เว็บไซต์ หรือ