RAG (Retrieval-Augmented Generation) คืออะไร

เทคนิคการสร้างข้อความ โดยเพิ่มขั้นตอนการการดึงข้อมูลจากแหล่งเก็บข้อมูลเช่น เช่น เอกสาร, Embedding Database

RAG = ค้นหาข้อมูลที่เกี่ยวข้อง(LLM) + ใช้ข้อมูลนั้นช่วยตอบคำถาม

จุดประสงค์/ ความท้าทาย

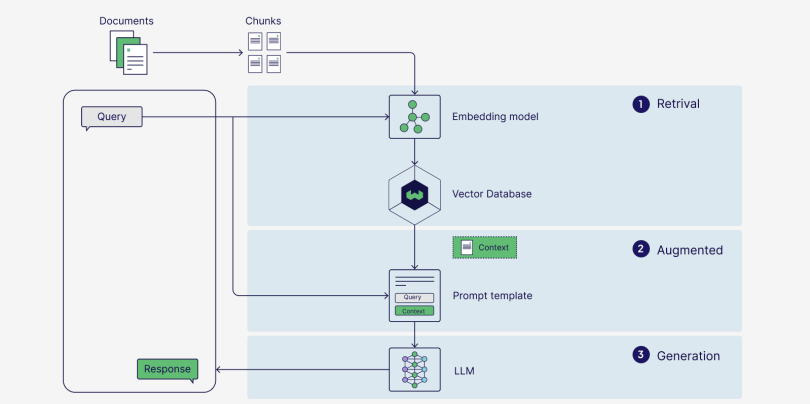
1. ลดการตอบมั่ว (hallucination) ของ LLM
2. ข้อมูลล้าสมัย ขาดความแม่นยำด้วยข้อมูลจากฐานความรู้จริง
3. อ้างอิงจากแหล่งที่ไม่ได้รับอนุญาต
4. สับสนคำศัพท์

ขั้นตอนการทำงานเช่น

1. Retrieval ดึงข้อมูลที่เกี่ยวข้องจากแหล่งความรู้
    a. ประเภทของ Retrieval Methods
        i. การค้นหาด้วย Keyword
        ii. การค้นหาด้วย Vector Embedding (semantic search)
2. Augmented เอาข้อมูลที่ได้มารวมกับคำถามใน User's Query
3. Generation สร้างคำตอบ โดยให้โมเดลใช้ข้อมูลเหล่านั้นในการสร้างคำตอบที่ถูกต้อง

ตัวอย่าง Tools เช่น LangChain

ตัวอย่างการใช้งาน Text/Image Search , Recommendation System , Chatbot answer



แหล่งอ้างอิง

https://aws.amazon.com/th/what-is/retrieval-augmented-generation/

https://youtu.be/T-D1OfcDW1M?si=v3Gf053_E5f45tSj

อะไรคือ Vector Database?

การแปลงข้อมูลต่างๆ เช่น ข้อความ,รูปภาพ, เสียง ให้กลายเป็นตัวเลขจำนวนมากเรียกว่า Vector โดยยิ่งข้อมูลมีความคล้ายกัน ตัวเลขVector ก็จะอยู่ใกล้กันมากขึ้นใช้สำหรับในการค้นหาข้อมูล และ การดึงข้อมูลในdatabase

โดยการหาข้อมูลที่ใกล้ที่สุดจะใช้ Cosine Similarity วัดมุมระหว่างVector

- ถ้าเวกเตอร์ชี้ไปทิศทางเดียวกัน ⟶ ค่าใกล้ 1 (คล้ายกันมาก)
- ถ้าเวกเตอร์ตั้งฉากกัน ⟶ ค่าใกล้ 0 (ไม่คล้าย)
- ถ้าชี้ตรงข้ามกัน ⟶ ค่าใกล้ -1 (ตรงข้ามกัน)

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

จุดประสงค์

1. เพื่อลดเวลาค้นหา / ดึงข้อมูล
2. ลดขนาดการเก็บข้อมูลจำนวนมาก
3. ใช้งานร่วมกับ RAG

ข้อดีในการใช้Vector Embedding

1. รองรับข้อมูลหลากหลายรูปแบบ
2. ใช้ในการหาความสัมพันธ์เชิงความหมายแทนการจับแค่คำตรงตัว
3. ประยุกต์ใช้ในงานค้นหา, การจำแนกประเภท

ตัวอย่าง Tools เช่น OpenAI Embedding API , pinecone และ Word2Vec

อ้างอิง

https://youtu.be/t9IDoenf-lo?si=3oCeFKZ4Pkt49tub

https://youtu.be/NEreO2zlXDk?si=rlNqjlZ8z8hZpH7h

https://weaviate.io/blog/vector-search-explained