Chatbot using LangChain

Individual Study 2023

Our Goals

Aj.Peerapon 4 Sep - ChatGPT for Toyota

Aj.Peerapon https://drive.google.com/file/d/

Google Docs

OpenAI-ChatGPT.ipynb

Colaboratory notebook

My Goals

Aj.Peerapon 4 Sep - ChatGPT for Toyota

Aj.Peerapon https://drive.google.com/file/d/

Google Docs

OpenAI-ChatGPT.ipynb

Colaboratory notebook

ChatGPT for บรรทัดทอง

🦜⛓️ LangChain

เป็น Framework สำหรับช่วยในการพัฒนา Large Language Model

Model I/O

Retrieval

Agent

“gpt-3.5-turbo” “gpt-3.5-turbo-16k”

“text-davinci-003” “gpt-4” “OpenThaiGPT”

Language Models

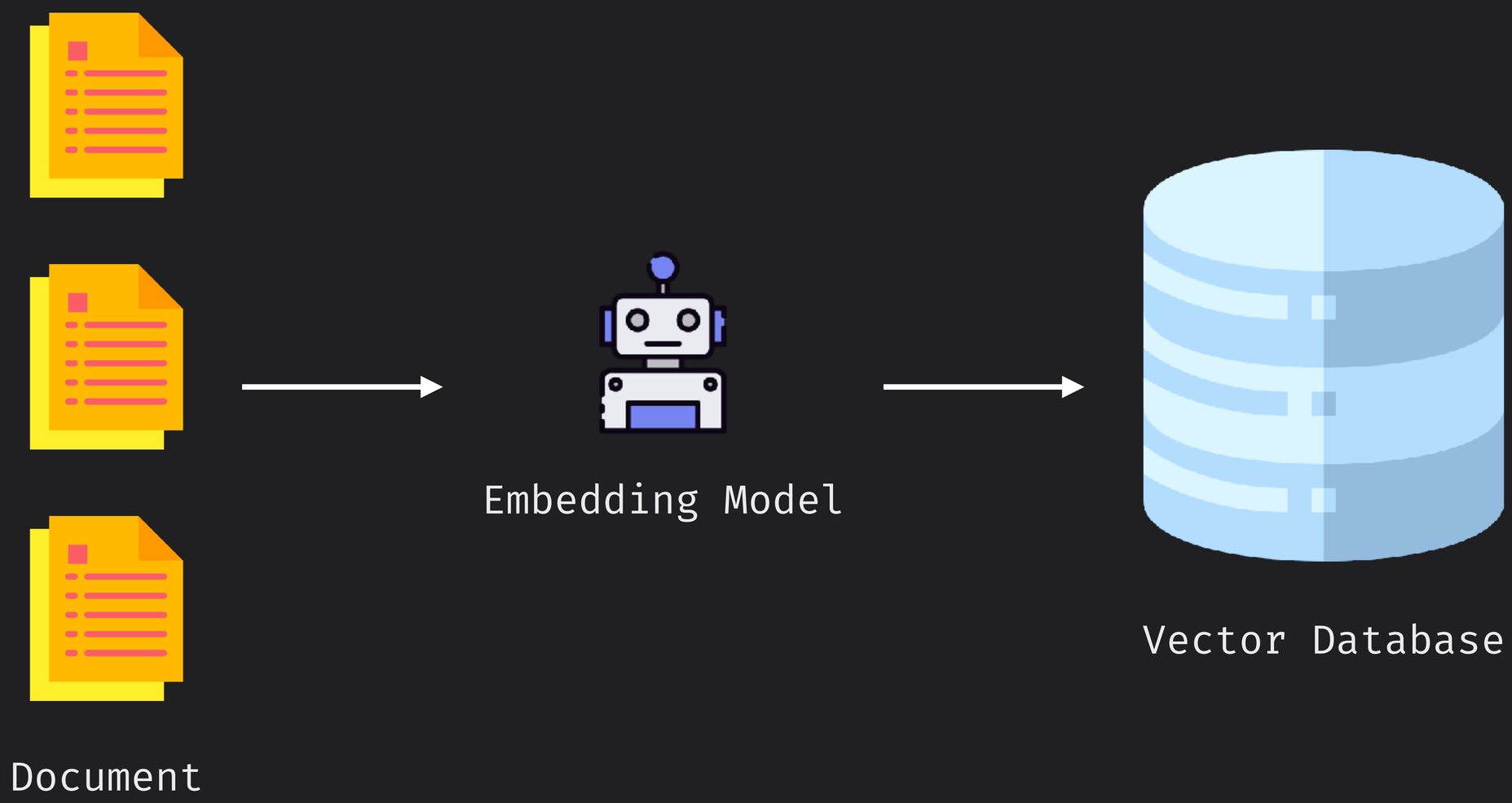
Documents

26 Documents

1 ร้านอาหาร / Document

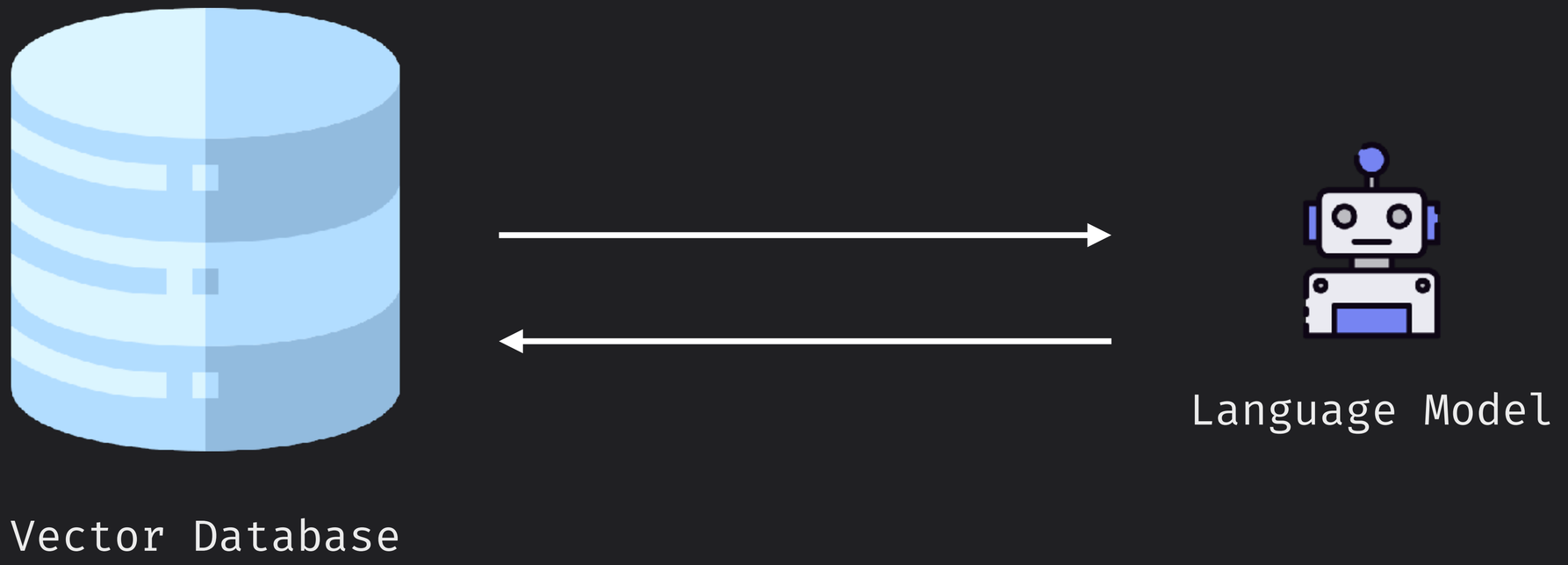
3 Sources

Document Preparation

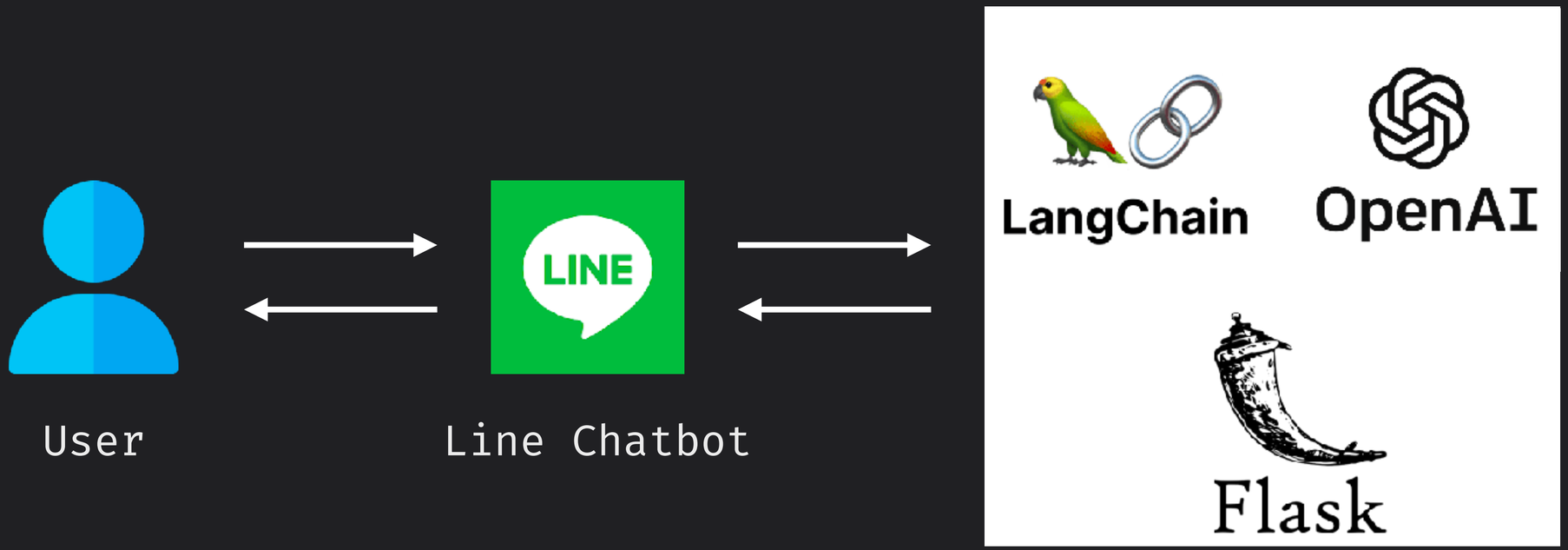


Retrieval Augmented Generation (RAG)

การทำให้ Language Model เข้าถึงข้อมูลภายนอกได้ โดยปราศจากการ train



Workflow

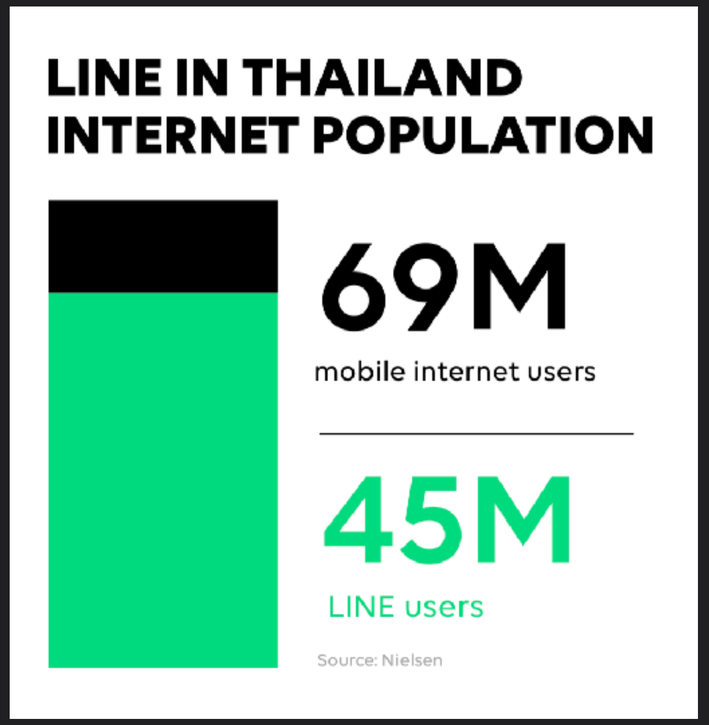


User

Line Chatbot

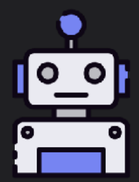
Chatbot API

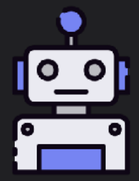
ทำไมต้องใช้ Line Chatbot



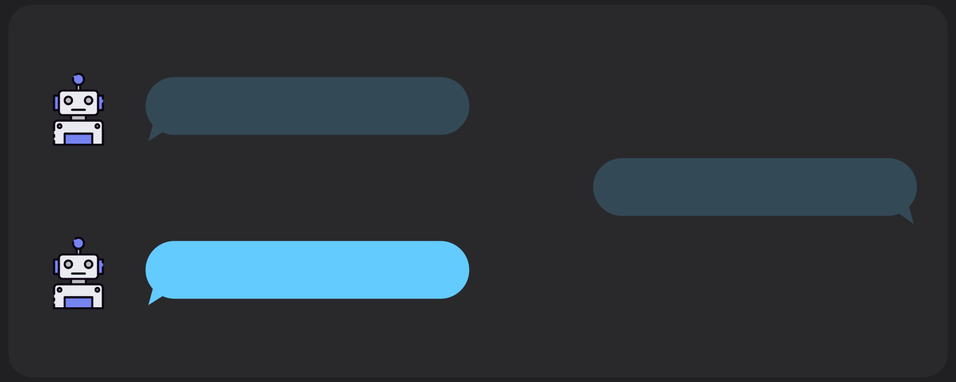
https://www.adxmatching.com/line.html

Chatbot จำข้อความก่อนหน้าไม่ได้

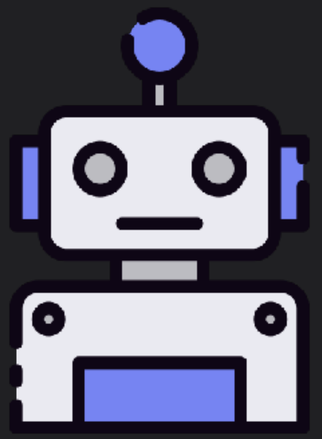




Chatbot จำข้อความก่อนหน้าไม่ได้

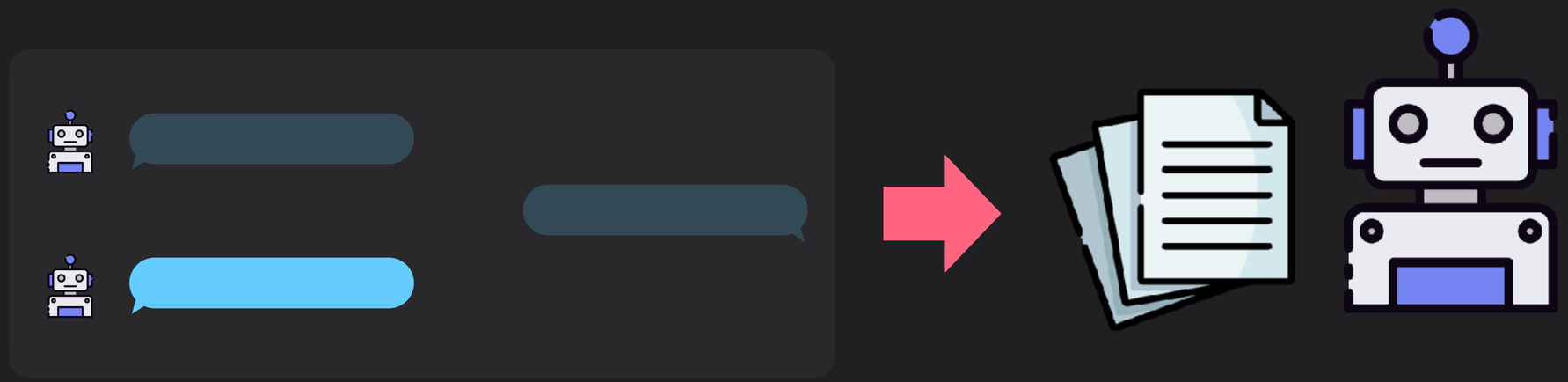


`ConversationBufferMemory`



ปัญหา: เมื่อประวัติการพูดคุยเยอะขึ้น จะทำให้ prompt ที่นำเข้าโมเดลนั้นมีขนาดเกิน max_tokens

Chatbot จำข้อความก่อนหน้าไม่ได้



`ConversationBufferSummaryMemory`

ปัญหา: ในกรณีที่ข้อความมีการเรียงเป็นลำดับ การสรุปความอาจจะทำให้ลำดับผิดเพี้ยนไปได้

Input เกิน max_token

เปลี่ยน prompt ให้มีความยาวที่สั้นลง

Input เกิน max_token

GPT-3.5

GPT-3.5 models can understand and generate natural language or code. Our most capable and cost effective model in the GPT-3.5 family is `gpt-3.5-turbo` which has been optimized for chat using the Chat Completions API but works well for traditional completions tasks as well.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
gpt-3.5-turbo-1106	**Updated GPT 3.5 Turbo** New The latest GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0613.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k	Currently points to gpt-3.5-turbo-0613.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-instruct	Similar capabilities as text-davinci-003 but compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-0613 Legacy	Snapshot of gpt-3.5-turbo from June 13th 2023. Will be deprecated on June 13, 2024.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k-0613 Legacy	Snapshot of gpt-3.5-16k-turbo from June 13th 2023. Will be deprecated on June 13, 2024.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo-0301 Legacy	Snapshot of gpt-3.5-turbo from March 1st 2023. Will be deprecated on June 13th 2024.	4,096 tokens	Up to Sep 2021

เปลี่ยน chat model

gpt-3.5-turbo-1106	**Updated GPT 3.5 Turbo** New The latest GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
gpt-3.5-turbo	Currently points to gpt-3.5-turbo-0613.	4,096 tokens	Up to Sep 2021
gpt-3.5-turbo-16k	Currently points to gpt-3.5-turbo-0613.	16,385 tokens	Up to Sep 2021

https://platform.openai.com/docs/models/gpt-3-5

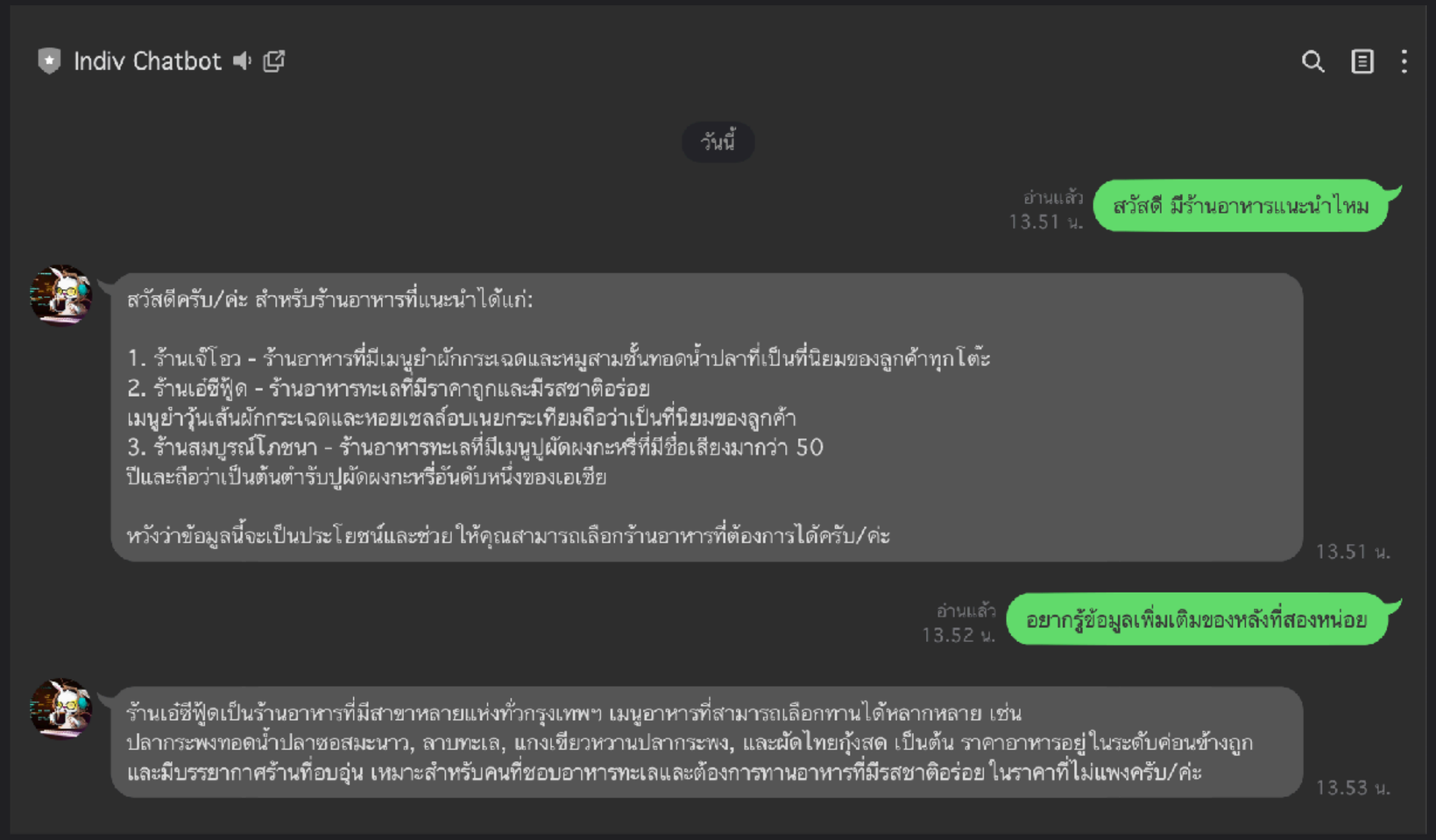
Hallucination

Thai Documents

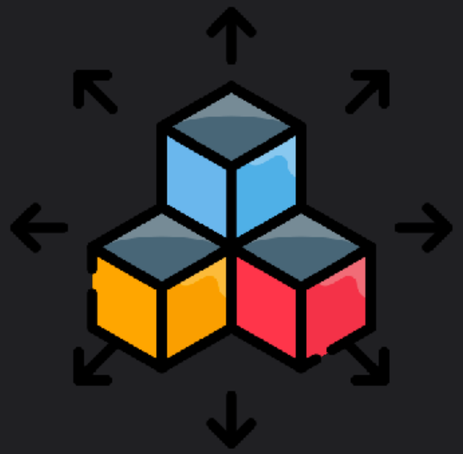
English Documents

เนื่องจาก ConversationBufferSummaryMemory สรุปความเป็นภาษาอังกฤษ

Final Product



Future Plan



ทำให้ข้อมูล update แบบ Real-time

เพิ่ม Document

Optimize ความเร็วของ Chatbot