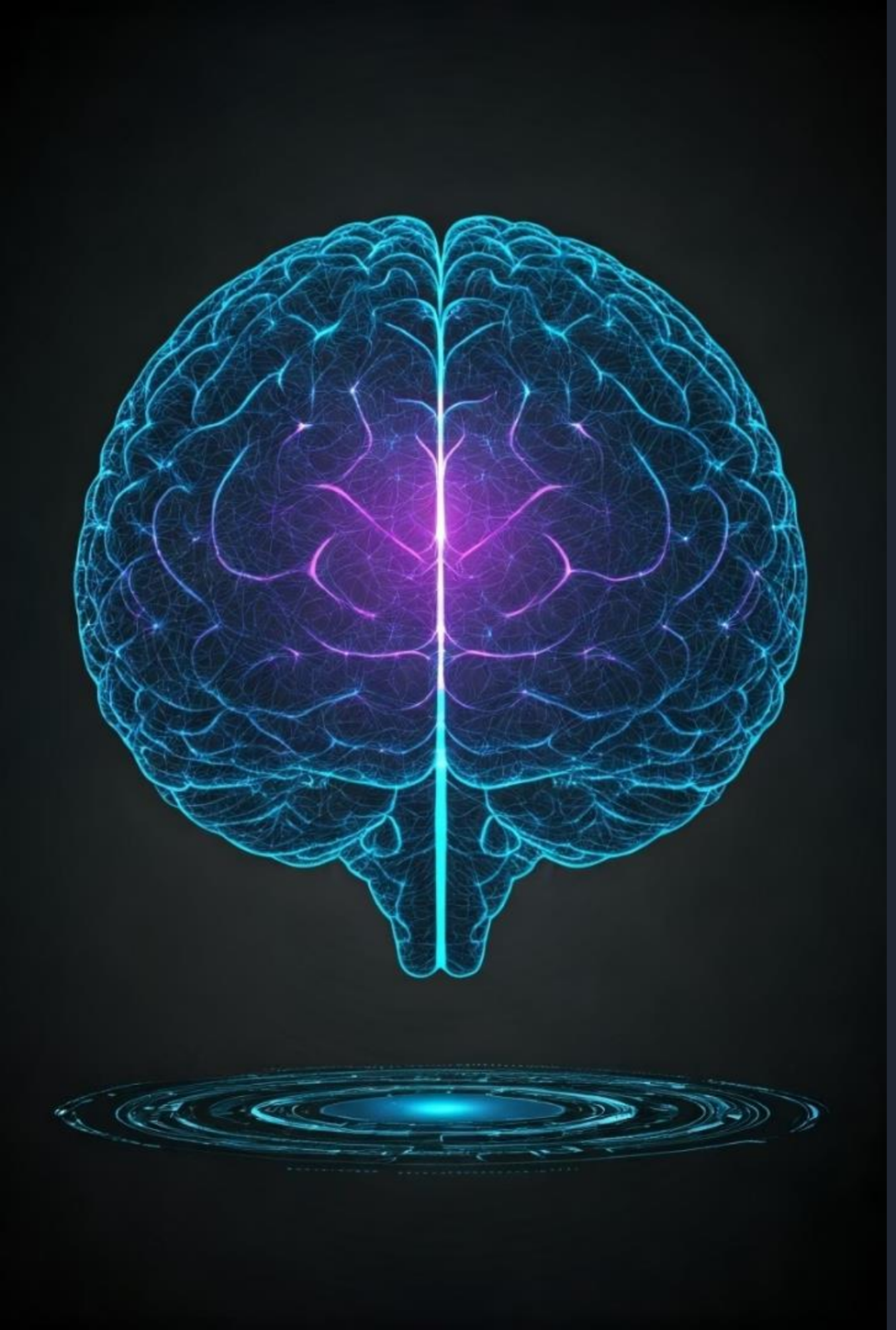


Agent Memory

Team Member

- CI 宋振維 Roy
- CI 許淇緣 CY
- CI 李婷琪 Dorbe
- DS 黃迪烯
- DS 何志偉
- CT 吳佳樺

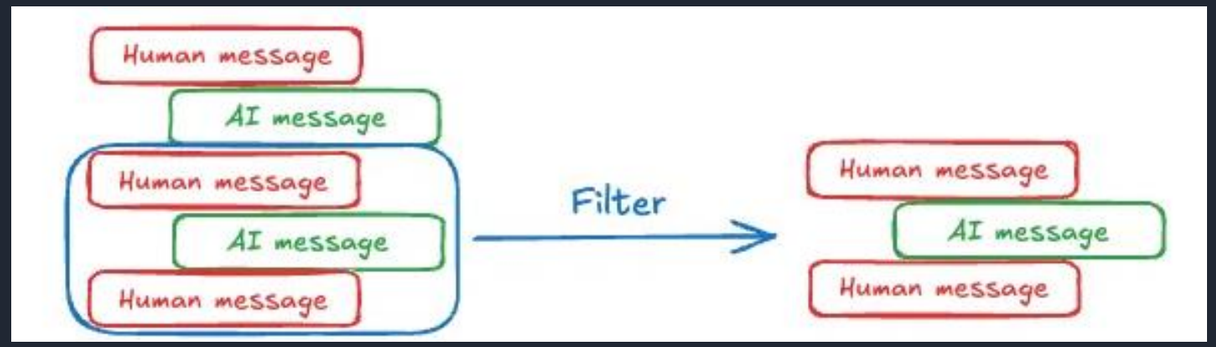
什麼是短期記憶？

對話線記憶

LLM會記住同一條「對話線」中之前的對話內容

上下文理解

模型回應時能理解對話上下文，不會忘記你剛剛問了什麼



短期記憶儲存＆清理機制

對話開始

每一條 thread 就像一個聊天室，裡面有多次你和模型的互動記錄

暫存處理

與模型的互動記錄會暫時存放模型「短期記憶」(Agent state) 裡

存檔點

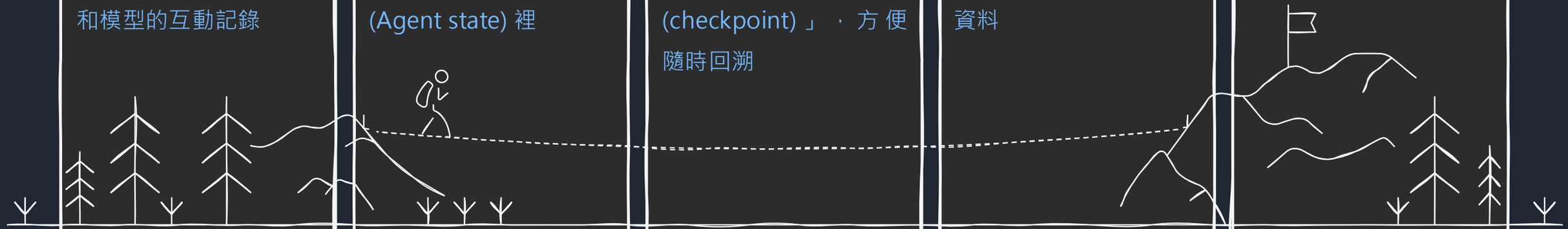
系統會定期把 thread 的內容做成「存檔點(checkpoint)」，方便隨時回溯

記憶體檢查_Token

如果累積的字數(token)超過限制，就會刪除舊資料

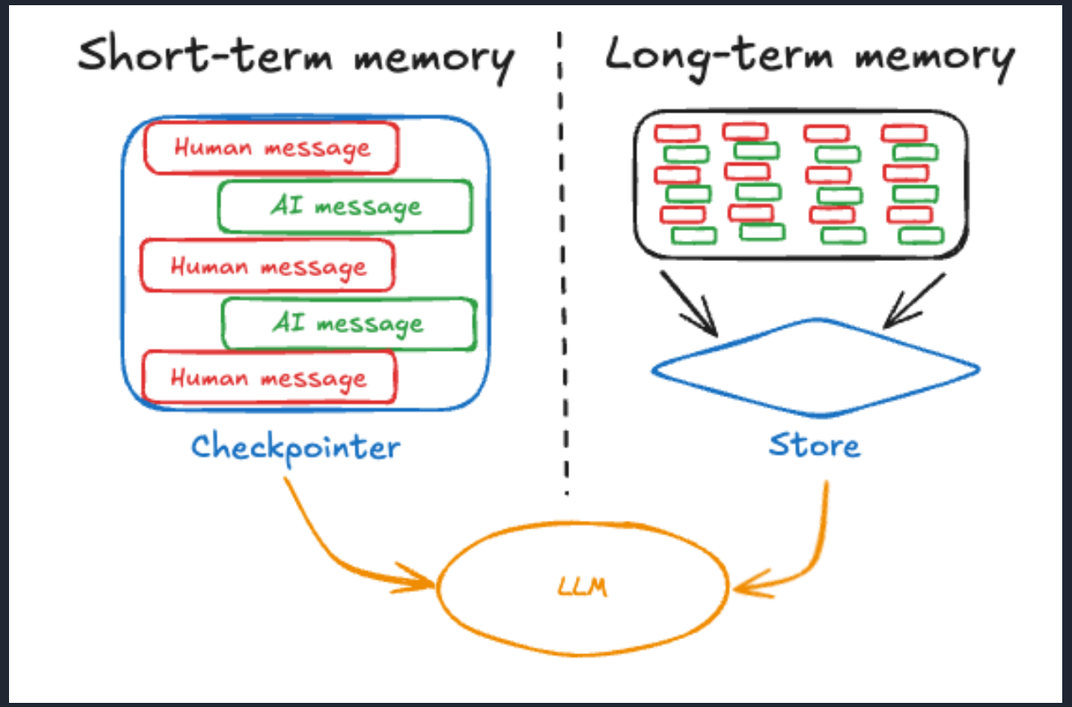
記憶體檢查_訊息

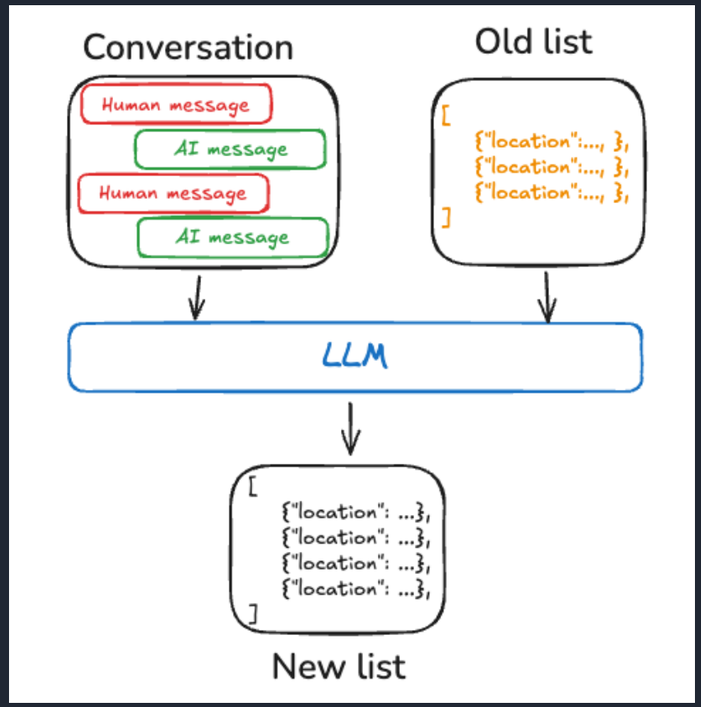
若對話內容太多，只保留最新的幾則



什麼是長期記憶？

在一般的語言模型對話中，記憶通常是短期的（例如在單一請求的 context window 中），但長期記憶是跨越多輪對話，甚至跨會話持久保存的資訊。





長期記憶的儲存方式

透過自訂 namespace（如資料夾）與 key（如檔案名稱），在多次 conversation (thread / session) 中保留資訊

	JSON 檔案記憶體儲存	DataBase (vector)儲存
優 點	易於理解與手動編輯	• 快速處理大規模資料集 • 相似度搜尋效能、精準度較高
缺 點	• 需要讀取整個檔案 • 不適合多人/高頻寫入情境	• 初期建置較複雜 • 搜尋受限於索引結構
適用場景	• 單人聊天機器人 • 開發初期測試	• 大量文本檢索、多輪語意對話記憶 • 圖片/影片/語音語意搜尋

長期記憶的類型

語意記憶

『儲存事實性知識與長期穩定的資訊，屬於較固定的背景知識』

- 使用者基本資料
- 偏好與興趣
- 固定的背景知識

情節記憶

『記錄與使用者的互動歷史及過去行為』

- 過去的對話內容
- 先前提出的問題
- 模型做過的推薦

程序性記憶

『儲存完成特定任務的指令或技能』

- 多步驟工作流程
- 執行複雜任務的能力

範例

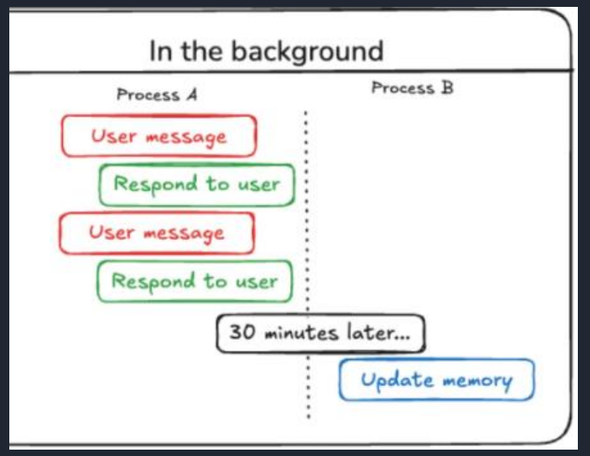

短期記憶
ChatGPT
臨時
3月
什麼是微服務
S3 存取 AWS 組織
Kinesis Data Streams vs Fire
CloudWatch 指標概覽
健走隊名建議
AWS组织概述与SCP
AWS Configuration Help
Amazon Forecast 介紹
Hive Metastore 問題解決
AWS Kendra 概述
自動刪除PII功能
AWS 無伺服器大數據管道
Redshift 概述与应用
Amazon Athena 介紹與最佳
邊緣定制與函數比較
CloudFormation 基礎架構管
Nvidia 投資策略分析
Amazon Security Lake 概述
CIDR 基礎與應用
續訂 Plus
長期記憶
儲存的記憶
ChatGPT 會記住關於你和你的喜好的有用詳細資料，這樣它才能更有幫助。 了解更多
Prefers table titles to be in Chinese.
Prefers financial data to include year-on-year growth and comparisons.
Is using an Ubuntu-based EC2 instance.
Does not want the cutting audio files to be deleted in their workflow.
Is working with AWS SageMaker to process audio files uploaded to S3, transcribe them using Whisper, perform speaker diarization with pyannote.audio, and upload the results to S3.
Wants to invest in Company B.
Requests examples and results of text embeddings.
Is working on predicting mental compensation and wants to reduce the Mean Absolute Error (MAE) in their model.
Prefers responses and help with Python-related issues in Traditional Chinese.
刪除全部

長期記憶體寫入策略

即時更新

VS

延遲30分鐘



In the Path(熱路徑)

每次使用後立即更新記憶

In the Background(背景處理)

定期/條件式在背景中批次更新記憶

	In the Path(熱路徑)	In the Background(背景處理)
優 點	• 記憶即時更新 • 可立即應用於後續互動	• 應用邏輯與記憶管理分離 • 可以更專注地處理記憶任務
缺 點	• 增加 agent 負擔與決策延遲 • 多工可能降低記憶品質	• 記憶延遲導致其他 thread 缺乏最新上下文 • 需妥善設計觸發頻率與機制

短期記憶 V.S. 長期記憶

	短期記憶	長期記憶
優點	• 快速：直接從記憶體讀取，延遲低 • 簡單：無需外部系統，實現成本低 • 即時性：適合快速響應的對話場景	• 持久化：支持跨會話的上下文保留 • 容量大：可儲存大量歷史數據 • 個人化：能記住用戶偏好，提升體驗
缺點	• 容量有限：受上下文窗口限制，無法儲存長歷史 • 易丟失：會話結束後信息消失 • 不適合複雜任務：難以處理需要歷史回溯的任務	• 檢索延遲：外部查詢增加計算時間 • 複雜性：需額外維護儲存系統 • 隱私風險：儲存敏感數據需加密和訪問控制

DEMO時間

Q&A