



¿Cómo recuerda la IA?

La verdad logística tras sus respuestas

Una guía educativa sobre la Ingeniería de Contexto

Una situación cotidiana: Organizando un viaje

Semana 1

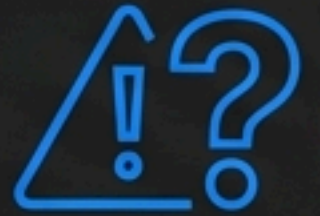
Usted indica: Presupuesto
2.000€ | Destino: Playa |
Fecha: Agosto.

Semana 2

Usted actualiza:
Presupuesto 3.000€ |
Fecha: Septiembre.

Semana 3

La IA propone:
Hotel de lujo...
¡en Agosto!



¿Por qué recuerda la playa, pero olvida el cambio de fecha si se lo dijimos la semana pasada?

La Ventana de Contexto: El escritorio de la IA

La IA no aprende en tiempo real. Utiliza un espacio de trabajo limitado.



La regla del escritorio: Todo lo que esté sobre el escritorio es lo que la IA puede ver para responder. Lo que no cabe, no existe en ese instante.

El espacio se mide en tokens. Un límite de 200.000 tokens equivale a unas 500 páginas físicas sobre el escritorio. (Anthropic, 2024)

Las 4 cosas que solemos llamar memoria (y no lo son)



1. Información de Entrenamiento

El conocimiento general estático con el que se fabricó el modelo (con fecha de corte).



2. Conversación Actual

El hilo de mensajes que se están intercambiando en este momento exacto.



3. Datos del Usuario

Información externa y persistente, como las preferencias de su perfil.



4. Información Recuperada (RAG)

Datos que el sistema busca en bases de datos externas para complementar la respuesta actual.

El proceso logístico: Guardar, Recuperar, Comprender

1. GUARDAR (Almacenamiento)

Un acto técnico. La aplicación guarda su historial en un disco duro externo.



2. RECUPERAR (Logística)

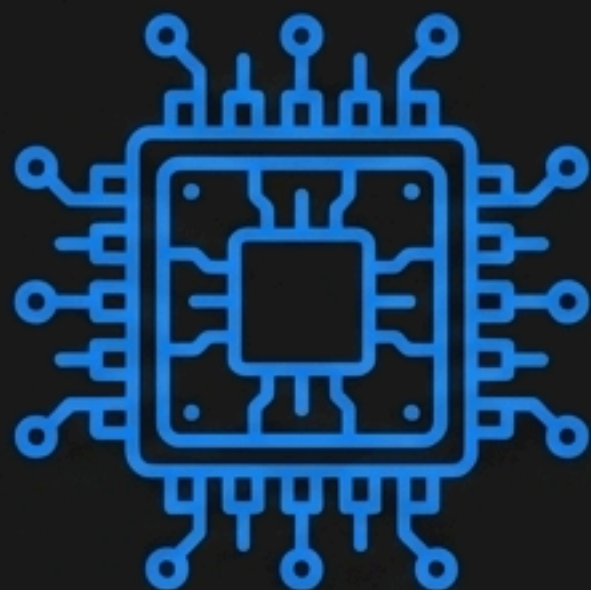
El sistema busca activamente esos datos y los lleva al escritorio de la IA.



3. COMPRENDER (Razonamiento)

La IA solo puede procesar la información una vez que está dentro de su ventana de contexto.

La IA gestionada como un Sistema Operativo

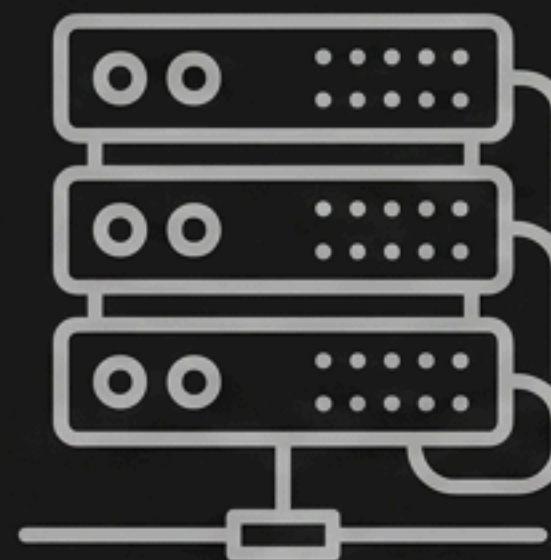


Memoria Principal (Ventana de Contexto)

El espacio de trabajo actual,
rápido pero muy limitado.



Flujo de datos



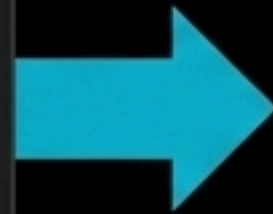
Almacenamiento Externo (Base de Datos)

Espacio masivo y lento, situado
fuera de la ventana.

La aplicación mueve apuntes entre el archivo y el escritorio para simular memoria a largo plazo sin saturar el sistema. (Arquitectura MemGPT, Oct. 2023)

El viaje: Qué ocurrió realmente entre bastidores

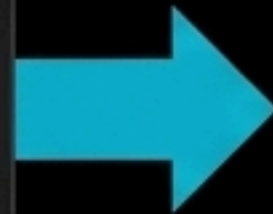
Semana 1: Presupuesto
2.000€ y agosto.



Acción técnica (Guardar):

La aplicación almacena este dato en su historial externo.

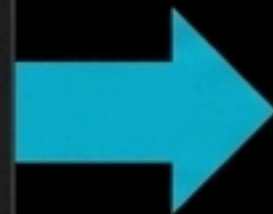
Semana 2:
Cambio a 3.000€ y
septiembre.



Acción técnica (Compactar):

La aplicación genera un resumen para ahorrar espacio, pero omite el cambio de fecha.

Semana 3:
Pregunta por hoteles.

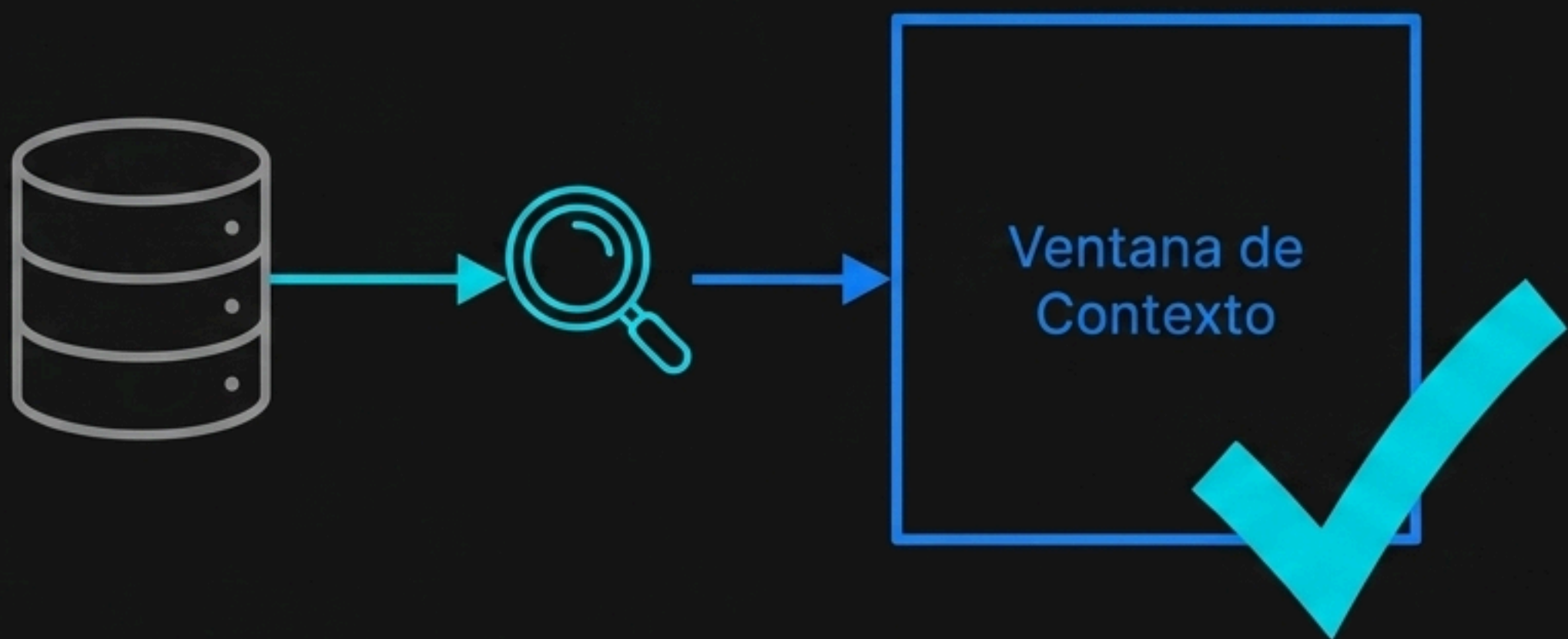


Acción técnica (Recuperar):

El sistema extrae el resumen desactualizado y lo pone en el escritorio.

El éxito depende de los apuntes, no de la IA

La IA no es más inteligente por recordar, simplemente ha **recibido la información correcta en el momento preciso.**



- **Recuperación Bajo demanda:** El sistema busca activamente en sus archivos solo cuando lo necesita.
- Entrega el apunte exacto (ej. Septiembre) al modelo justo antes de que este genere la respuesta.

Por qué falló el presupuesto: El efecto Perdido en el medio

Tener un escritorio enorme no garantiza que la IA use todo correctamente.



Presupuesto 3.000€

- Los modelos priorizan la información situada al principio y al final del texto.
- El rendimiento cae drásticamente si el dato clave queda enterrado en la mitad de una conversación muy larga.

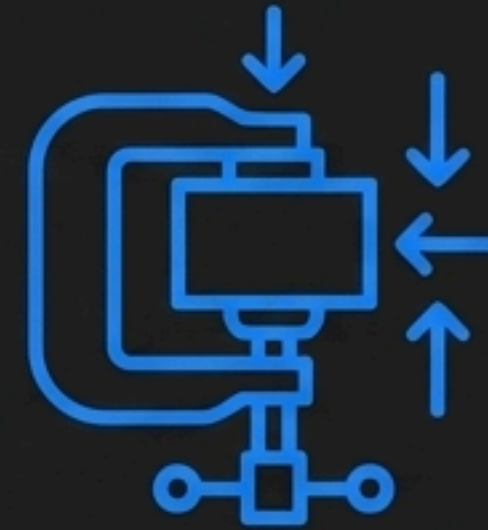
Estudio: Lost in the Middle (Liu et al., Jul. 2023)

Estrategias para gestionar el límite



Eliminación

Se descartan proactivamente las partes de la conversación menos relevantes (saludos, búsquedas antiguas) para liberar espacio vital en el escritorio.

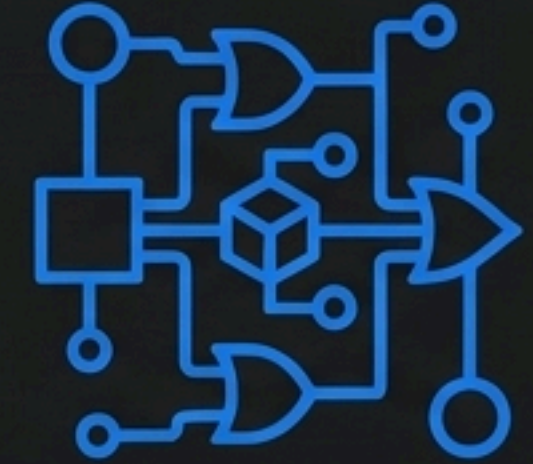


Resumen / Compactación

El sistema toma los mensajes antiguos, extrae únicamente los detalles críticos y crea un resumen muy corto que sustituye al historial original.

Guardado no significa Comprendido

Que la aplicación haya registrado su cambio de planes en su base de datos no garantiza que la IA lo entienda.



Context Rot

La Confusión (Context Rot):

A medida que la charla crece y el número de tokens inunda la ventana, el modelo empieza a perder precisión y a confundir conceptos.

Solo razonará correctamente si la logística pone el dato específico ante sus ojos en el momento exacto.

Fallos logísticos más comunes

Olvidar una condición

El dato clave se salió de la ventana de contexto por ser demasiado antiguo.

Recuperación equivocada

El buscador interno se confundió con palabras similares al extraer datos del archivo.

Información no válida

El sistema conservó un resumen inicial e ignoró una corrección posterior.

Mezclar tareas

Datos de un escenario anterior (ej. viaje a la montaña) se filtran por error en la tarea actual.

Riesgos de privacidad cruzada

Dado que las aplicaciones guardan historiales externos a largo plazo para simular memoria, se generan nuevas vulnerabilidades.



Las IAs pueden inferir información altamente privada cruzando pequeños datos inofensivos recolectados a lo largo de múltiples sesiones.

Marco de Gestión de Riesgos: NIST AI 600-1 (Julio 2024)

Proteja su información: Verifique su aplicación

- ☒ ¿Qué datos específicos se están guardando de mis sesiones?
- ☒ ¿Dónde se almacena exactamente este historial externo?
- ☒ ¿Durante cuánto tiempo se conservan mis preferencias?
- ☒ ¿Cómo puedo corregir un dato que la aplicación recuerda mal?
- ☒ ¿Cómo puedo eliminar mi historial de forma permanente?

La IA no tiene memoria biográfica. Su aparente continuidad es solo el resultado de una logística incansable: guardar, resumir y recuperar datos. Es una herramienta potente, pero inherentemente susceptible a la pérdida de contexto.

Anthropic. Building Effective AI Agents (Dic. 2024); Contextual Retrieval (Sep. 2024); Effective Context Engineering (Sep. 2025).
Packer et al. MemGPT: Towards LLMs as Operating Systems (Oct. 2023).
Liu et al. Lost in the Middle: How Language Models Use Long Contexts (Jul. 2023).
NIST AI 600-1: Generative AI Profile (Jul. 2024).
Lewis et al. Retrieval-Augmented Generation (May. 2020) | Niu et al. RAGTruth (Dic. 2023).