Ventanas de contexto, módulos de memoria y lo que significan para la educación

Autor: Zsolt Lengyel, iTStudy Hungary Training and Research Centre

Revisora: Mária Hartyányi, iTStudy Hungary Training and Research Centre

Objetivo de aprendizaje

Muchas sorpresas del uso cotidiano de la IA (el modelo "olvida" una instrucción, mezcla dos temas o sabe algo inesperado sobre nosotros) se explican por dos mecanismos: la ventana de contexto y las funciones de memoria. Este material explica qué pueden y qué no pueden retener los sistemas de IA durante una conversación y entre conversaciones. Al terminarlo, los docentes serán capaces de distinguir el contexto temporal de la conversación de las funciones de memoria a más largo plazo, explicar cómo influyen en la continuidad, la personalización y la privacidad, y aplicar rutinas prácticas para que las sesiones de trabajo largas sean fiables y los datos del alumnado estén seguros.

Introducción: el modelo no recuerda como un ser humano

Al chatear con un modelo de lenguaje, parece que recuerda toda la conversación. En realidad, el propio modelo no cambia durante la conversación. Antes de cada respuesta, el sistema reúne una entrada (la conversación hasta el momento, los archivos adjuntos, las instrucciones en segundo plano) y el modelo calcula la siguiente respuesta solo a partir de ella.

Una buena analogía es el protagonista de las películas *Memento* o *50 primeras citas*: alguien que no puede retener recuerdos nuevos, pero conserva sus conocimientos básicos. En un modelo de lenguaje, estos son lo que aprendió en el entrenamiento; todo lo demás hay que ponérselo delante una y otra vez.

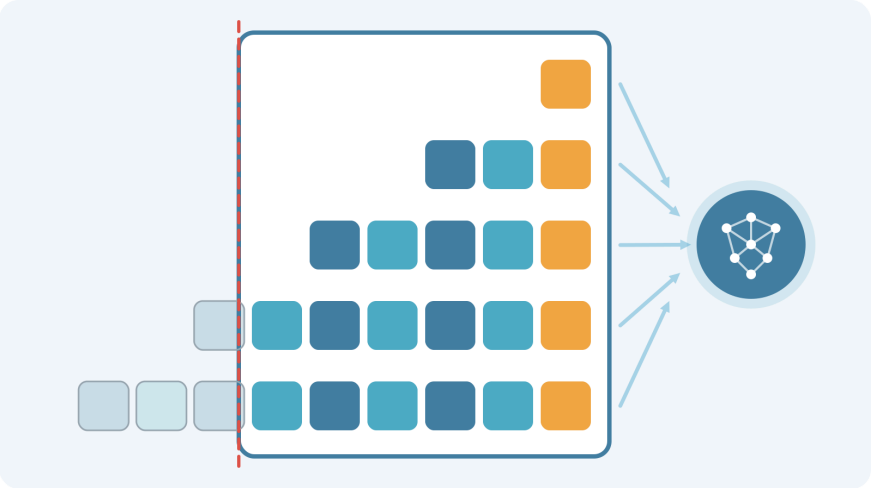


50 primeras citas (Fuente: https://nlc.hu/)

Cómo se construye una conversación: el contexto

En una interfaz de chat, la entrada crece con cada turno. En la primera ronda, el modelo recibe solo nuestro primer mensaje (M1) y responde (R1). En la segunda ronda, el sistema envía M1 + R1 + M2; en la tercera, M1 + R1 + M2 + R2

+ M3, y así sucesivamente. Todo lo que el modelo “ve” en un momento dado se denomina **contexto**.



Con cada turno la entrada crece; lo que ya no cabe en la ventana de contexto de tamaño fijo se descarta o se comprime

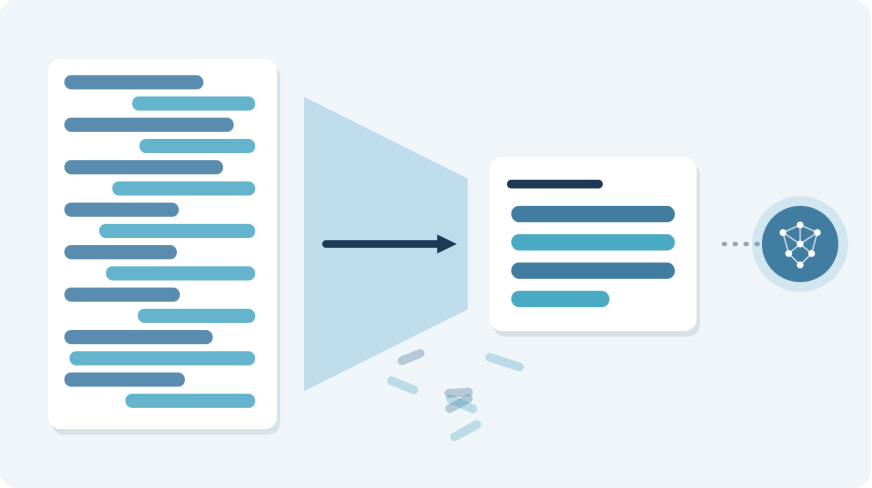
El contexto se mide en **tokens**: pequeñas unidades de texto, a menudo fragmentos de palabras. El número máximo de tokens que un modelo puede manejar a la vez es la **ventana de contexto**. Es grande, pero finita: al redactar este material, los modelos punteros trabajan con ventanas desde unos cientos de miles de tokens (p. ej., unos 400 000 en los modelos de la clase GPT-5) hasta alrededor de un millón (p. ej., Gemini 3.1 Pro), lo que equivale a varios libros extensos. El límite de una aplicación de chat puede ser inferior al máximo técnico del modelo y suele depender de la suscripción. En muchos idiomas distintos del inglés, una palabra ocupa de media más tokens.

¿Por qué es importante?

- **Límite:** cuando una conversación se alarga demasiado, el sistema elimina las partes más antiguas o las resume.
- **Coste y velocidad:** cuanto más larga es la entrada, más cálculo requiere; las respuestas pueden ser más lentas y agotar antes los límites de uso.
- **Calidad:** más contexto puede ayudar, pero también aporta ruido. Las investigaciones muestran que los modelos aprovechan mejor la información del principio y del final de un contexto largo que la del medio, que puede perderse. El material antiguo irrelevante hace más probables los malentendidos.

Compactación: cuando la conversación se comprime

Cuando una conversación alcanza el límite de contexto, muchas aplicaciones **compactan** automáticamente los mensajes antiguos: los sustituyen por un resumen. Lo esencial se mantiene y los detalles se difuminan; así siempre hay sitio para información nueva. No podemos controlar cuándo ocurre, y el inconveniente es que pueden perderse detalles finos o la redacción exacta anterior. Las conversaciones muy largas también pueden ralentizar el navegador.



Compactación: una conversación larga se condensa en un breve resumen, y algunos detalles se pierden

Consejos prácticos para sesiones de trabajo largas

- Cuando cambie de tema, abra un chat nuevo. Haga lo mismo cuando un tema se haya alargado mucho.
- Antes de cambiar, pida: "Resume las decisiones, los requisitos y las preguntas abiertas de esta conversación de forma que pueda pegarlo en un chat nuevo". Revise el resumen: rara vez lo contendrá todo.
- Adjunte los documentos largos en lugar de pegarlos en el mensaje; algunas aplicaciones lo hacen automáticamente cuando se pega un texto muy largo.

Tres capas: conocimiento general, memoria y contexto actual

Resulta útil pensar en lo que "sabe" un asistente de IA en tres capas:



El conocimiento general del entrenamiento, una pequeña memoria editable y el contexto actual alimentan juntos la respuesta

- El **conocimiento general** procede del entrenamiento. Es amplio, pero no está actualizado: solo llega hasta la fecha de corte de conocimiento del modelo. Para temas recientes, el modelo tiene que recurrir a la búsqueda web.

- La **memoria** es un conjunto relativamente pequeño de información sobre el usuario que la aplicación pasa al modelo en segundo plano al inicio de cada conversación nueva. Todo chat nuevo empieza con un contexto vacío, salvo lo que aporta el módulo de memoria.
- El **contexto actual** es lo que ponemos ahora delante del modelo: la conversación, un archivo, un ejemplo, un mensaje de error, el formato esperado. Los modelos rinden mejor con lo que les mostramos de forma concreta: un buen ejemplo suele valer más que una larga explicación.

Las funciones de memoria en la práctica (con ChatGPT como ejemplo)

La mayoría de los grandes asistentes (ChatGPT, Claude, Gemini) ofrecen ya funciones similares a una memoria. En ChatGPT, por ejemplo, actualmente son estas:

- **Memoria guardada**: datos que le pedimos expresamente que recordara o que guardó automáticamente. Se guardan aparte de los chats: borrar un chat no borra una memoria guardada.
- **Consultar el historial de chats**: si está activada, el asistente puede basarse en conversaciones anteriores para personalizar las respuestas.
- **Resumen y fuentes de la memoria**: una visión general de lo que el asistente puede usar para personalizar y, bajo algunas respuestas, qué chats o archivos anteriores influyeron en ellas. Ambos pueden corregirse o eliminarse.
- **Chat temporal**: una conversación que no crea ni actualiza memorias.
- **Memoria solo del proyecto**: en un Proyecto, el contexto puede limitarse a sus propios chats y archivos (véase el material sobre la gestión de tareas complejas).

Importante
La memoria no es una base de datos. El asistente elige lo que considera importante; no retiene todos los detalles, lo almacenado puede quedar desfasado o ser erróneo, y puede aplicarse en situaciones que no esperábamos. Revise con regularidad el resumen de su memoria.

Qué significa esto para la educación

Continuidad

Como el contexto a corto plazo se pierde entre chats, la continuidad hay que diseñarla. Para cada unidad didáctica, mantenga un breve “paquete de contexto”: los objetivos de la unidad, el nivel del grupo (sin datos personales) y el estilo y formato preferidos de los materiales. Péguelo al inicio de los chats nuevos o guárdelo como instrucciones y archivos en un Proyecto. Es más fiable que esperar que la memoria recuerde los detalles.

Personalización

La memoria puede ahorrar tiempo: el asistente puede recordar que usted imparte, por ejemplo, electromecánica de vehículos a alumnos de 16 años y que prefiere tablas con solucionario. Pero la personalización también puede

estrechar las respuestas o arrastrar suposiciones erróneas. Si necesita una respuesta neutral o quiere ver cómo se comporta la herramienta con un estudiante, use un chat temporal.

Privacidad y protección de datos



La memoria puede revisarse, desactivarse y borrarse, pero los datos más seguros son los que nunca introducimos

- Nunca introduzca datos personales del alumnado (nombres, notas, datos de salud o familiares, fotos) en herramientas de IA de uso general. Use ejemplos anonimizados o ficticios.
- Cumpla el RGPD y las normas de protección de datos de su centro y, si existen, use las cuentas institucionales de IA del centro.
- Revise la configuración: la memoria y la consulta del historial de chats pueden desactivarse, y cada memoria puede borrarse. Los proveedores pueden guardar registros de las memorias borradas durante un tiempo limitado (en ChatGPT, hasta 30 días).
- Cuidado al proyectar la pantalla en clase: la personalización basada en la memoria o los títulos de chats de la barra lateral pueden revelar información que no quería mostrar.
- Enseñe al alumnado la misma higiene digital: qué recuerda una IA de ellos, cómo comprobarlo y cómo borrarlo.

Ideas clave para docentes

- El modelo calcula cada respuesta a partir de la entrada actual; no recuerda como un ser humano.
- La ventana de contexto es finita: los chats largos se vuelven más lentos y ruidosos y, al final, se comprimen.
- Tema nuevo, chat nuevo. Traslade las decisiones importantes con un resumen revisado o un paquete de contexto.
- La memoria es pequeña, selectiva y editable; revísela y desactívela cuando no sea adecuada.
- Muestre al modelo ejemplos y archivos concretos: funciona mejor a partir de lo que tiene delante.
- Proteja los datos del alumnado: anonimice, utilice cuentas institucionales y nunca confíe a la memoria información sensible.

Referencias

- Liu, N. F. et al. (2024): Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157–173.
- OpenAI Help Center (2026): Memory in ChatGPT; Projects in ChatGPT; ChatGPT release notes.
- Lengyel, Zs. (2026): Introduction to the concepts of artificial intelligence. Presentación de la formación VITA.

Enlaces

- La memoria en ChatGPT: help.openai.com/en/articles/8590148-memory-faq
- Proyectos en ChatGPT: help.openai.com/en/articles/10169521-projects-in-chatgpt
- Demostración de un tokenizador: tiktokenizer.vercel.app

Derechos de autor © 2026 Consorcio VITA

Esta obra puede compartirse y adaptarse bajo los términos de la Licencia Internacional Creative Commons Atribución/Reconocimiento-CompartirIgual 4.0 (CC BY-SA 4.0).

Financiado por la Unión Europea. Las opiniones expresadas son únicamente las de los autores y no reflejan necesariamente las de la Unión Europea ni las de la Agencia Nacional.