

Finestre di contesto, moduli di memoria e cosa significano per la didattica

Autore: Zsolt Lengyel, iTStudy Hungary Training and Research Centre

Revisore: Mária Hartyányi, iTStudy Hungary Training and Research Centre

Obiettivo di apprendimento

Quando utilizziamo l'IA, può capitare che il sistema “dimentichi” un'istruzione, confonda due argomenti oppure ricordi informazioni su di noi che non pensavamo conoscesse. Questi comportamenti dipendono principalmente da due meccanismi: la finestra di contesto e le funzioni di memoria. Questo materiale spiega quali informazioni i sistemi di IA possono conservare durante una conversazione e tra conversazioni diverse. Al termine, i docenti sapranno distinguere le informazioni temporanee di una chat da quelle conservate attraverso la memoria, comprendere come questi meccanismi influenzino la continuità delle conversazioni, la personalizzazione delle risposte e la privacy, e adottare alcune semplici strategie per gestire attività lunghe e proteggere i dati degli studenti.

Introduzione: il modello non ricorda come un essere umano

Quando conversiamo con un modello linguistico, abbiamo l'impressione che ricordi tutto ciò che ci siamo detti. In realtà, il modello non modifica le proprie conoscenze durante la conversazione. Prima di ogni risposta, il sistema raccoglie le informazioni disponibili in quel momento, come i messaggi precedenti, i file allegati e le istruzioni attive, e le utilizza per generare una nuova risposta.

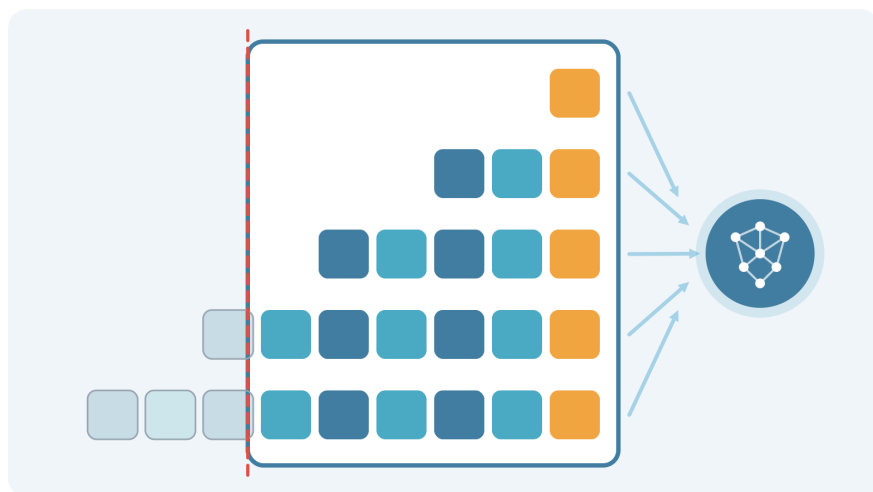
Un paragone utile è quello con i protagonisti dei film *Memento* o *50 volte il primo bacio*: persone che conservano le proprie conoscenze generali, ma hanno difficoltà a formare nuovi ricordi. Per un modello linguistico, le conoscenze di base sono quelle acquisite durante l'addestramento, mentre le informazioni relative alla conversazione devono essergli rese disponibili ogni volta.



50 volte il primo bacio (Fonte: <https://nlc.hu/>)

Come si costruisce una conversazione: il contesto

In una chat, le informazioni disponibili aumentano a ogni nuovo messaggio. All'inizio, il modello riceve il nostro primo messaggio (M1) e genera una risposta (A1). Quando scriviamo un secondo messaggio (M2), il sistema può fornire al modello anche lo scambio precedente: M1 + A1 + M2. Al terzo turno vengono aggiunti anche A2 e M3, e così via. L'insieme delle informazioni che il modello riceve e può utilizzare in un determinato momento si chiama **contesto**.



A ogni turno l'input cresce; ciò che non entra più nella finestra di contesto, di dimensione fissa, viene escluso o compresso

Il contesto si misura in **token**: piccole unità di testo, spesso frammenti di parole. Il numero massimo di token che un modello può gestire contemporaneamente è la **finestra di contesto**. È ampia ma finita: al momento della stesura, i modelli di punta lavorano con finestre che vanno da alcune centinaia di migliaia di token (ad es. circa 400.000 per i modelli della classe GPT-5) fino a circa un milione (ad es. Gemini 3.1 Pro), l'equivalente di diversi libri lunghi. Il limite disponibile in un'applicazione di chat può essere inferiore al massimo tecnico del modello e spesso dipende dall'abbonamento. In molte lingue diverse dall'inglese una parola richiede in media più token.

Perché è importante?

- **Limite:** quando una conversazione diventa troppo lunga, il sistema ne taglia le parti più vecchie oppure le riassume.
- **Costi e velocità:** più lungo è l'input, più calcolo serve; le risposte possono rallentare e consumare più in fretta i limiti di utilizzo.
- **Qualità:** più contesto può aiutare, ma porta anche rumore. La ricerca ha dimostrato che i modelli usano in modo più affidabile le informazioni che si trovano all'inizio e alla fine di un contesto lungo, mentre quelle nel mezzo rischiano di andare perse. Il materiale vecchio e non pertinente rende più probabili i fraintendimenti.

Compattazione: quando la conversazione viene compressa

Quando una conversazione raggiunge il limite di contesto, molte applicazioni **compattano** automaticamente i messaggi più vecchi, sostituendoli con un riassunto. L'essenza rimane, i dettagli sfumano – e così c'è sempre spazio per nuove informazioni. Non possiamo controllare quando ciò avviene, e lo svantaggio è che dettagli sottili o formulazioni precise precedenti possono andare persi. Le conversazioni molto lunghe possono anche rallentare il browser.



Compattazione: una lunga conversazione viene condensata in un breve riassunto – e alcuni dettagli si perdono

Consigli pratici per le sessioni di lavoro lunghe

- Quando cambi argomento, apri una nuova chat. Fai lo stesso quando un argomento è diventato molto lungo.
- Prima di cambiare chat, chiedi: "Riassumi le decisioni, i requisiti e le questioni aperte di questa conversazione in una forma che posso incollare in una nuova chat." Controlla il riassunto: raramente conterrà tutto.
- Allega i documenti lunghi invece di incollarli nel messaggio; alcune applicazioni lo fanno automaticamente quando si incolla un testo molto lungo.

Tre livelli: conoscenza generale, memoria e contesto attuale

È utile pensare a ciò che un assistente di IA "sa" come organizzato su tre livelli:



La conoscenza generale acquisita con l'addestramento, una piccola memoria modificabile e il contesto attuale confluiscono tutti nella risposta

- La **conoscenza generale** deriva dall'addestramento. È ampia ma non aggiornata: arriva solo fino alla data limite delle conoscenze del modello. Per gli argomenti recenti, il modello deve ricorrere alla ricerca web.
- La **memoria** è un insieme relativamente piccolo di informazioni sull'utente che l'applicazione trasmette al modello in background all'inizio delle nuove conversazioni. Ogni nuova chat parte con un contesto vuoto, a parte ciò che fornisce il modulo di memoria.
- Il **contesto attuale** è ciò che mettiamo davanti al modello in quel momento: la conversazione, un file, un esempio, un messaggio di errore, il formato atteso. I modelli lavorano meglio a partire da ciò che mostriamo loro concretamente: un buon esempio vale spesso più di una lunga spiegazione.

Le funzioni di memoria in pratica (l'esempio di ChatGPT)

La maggior parte dei principali assistenti IA (ChatGPT, Claude, Gemini) offre ormai funzioni simili a una memoria. In ChatGPT, per esempio, attualmente comprendono:

- **Memorie salvate** – informazioni che gli abbiamo chiesto esplicitamente di ricordare o che ha salvato automaticamente. Sono conservate separatamente dalle chat: eliminare una chat non elimina una memoria salvata.
- **Riferimento alla cronologia delle chat** – se attivato, l'assistente può attingere alle conversazioni passate per personalizzare le risposte.
- **Riepilogo della memoria e fonti della memoria** – una panoramica di ciò che l'assistente può usare per la personalizzazione e, sotto alcune risposte, quali chat o file precedenti hanno influito sulla risposta. Entrambi possono essere corretti o eliminati.
- **Chat temporanea** – una conversazione che non crea né aggiorna memorie.
- **Memoria limitata al progetto** – in un Progetto, il contesto può essere limitato alle chat e ai file di quel progetto (vedi il materiale sulla gestione di compiti complessi).

Importante

La memoria non è un database. L'assistente seleziona ciò che ritiene importante; non conserva ogni dettaglio, le informazioni memorizzate possono diventare obsolete o essere errate e possono essere applicate in situazioni in cui non ce lo aspettavamo. Controlla regolarmente il riepilogo della tua memoria.

Che cosa significa per la didattica

Continuità

Poiché il contesto a breve termine si perde da una chat all'altra, la continuità va progettata. Per un'unità didattica, tieni un breve "pacchetto di contesto": gli obiettivi dell'unità, il livello del gruppo (senza dati personali), lo stile e il formato preferiti dei materiali. Incollalo all'inizio delle nuove chat oppure

salvalo come istruzioni e file in un Progetto. È più affidabile che sperare che la memoria ricordi i dettagli.

Personalizzazione

La memoria può far risparmiare tempo, perché permette all'assistente di tenere conto di alcune preferenze o informazioni già comunicate. Ad esempio, può ricordare che insegni mecatronica dell'autoveicolo a studenti di 16 anni e che preferisci ricevere materiali organizzati in tabelle con le soluzioni. Tuttavia, la personalizzazione può anche limitare le risposte o portare il modello a fare supposizioni non corrette. Quando desideri una risposta neutra, oppure vuoi verificare come l'IA risponderebbe a uno studente senza conoscere le tue preferenze, puoi utilizzare una chat temporanea.

Privacy e protezione dei dati



La memoria può essere consultata, disattivata ed eliminata – ma i dati più sicuri sono quelli che non inseriamo mai

- Non inserire mai dati personali degli studenti (nomi, voti, informazioni sanitarie o familiari, foto) in strumenti di IA di uso generale. Usa esempi anonimizzati o fittizi.
- Rispetta il GDPR e le regole sulla protezione dei dati della tua scuola e, dove disponibili, usa gli account istituzionali di IA della scuola.
- Controlla le impostazioni della memoria: la memoria e il riferimento alla cronologia delle chat possono essere disattivati e le singole memorie possono essere eliminate. I fornitori possono conservare i log delle memorie eliminate per un periodo limitato (nel caso di ChatGPT, fino a 30 giorni).
- Fai attenzione quando proietti lo schermo in classe: la personalizzazione basata sulla memoria o i titoli delle chat nella barra laterale possono rivelare informazioni che non intendevi mostrare.
- Insegna agli studenti la stessa igiene digitale: che cosa un'IA ricorda di loro, come verificarlo e come eliminarlo.

Indicazioni per i docenti

- Il modello calcola ogni risposta a partire dall'input attuale; non ricorda come un essere umano.
- La finestra di contesto è finita: le chat lunghe diventano più lente, più rumorose e alla fine vengono compresse.

- Nuovo argomento – nuova chat. Riporta le decisioni importanti con un riassunto verificato o un pacchetto di contesto.
- La memoria conserva soltanto alcune informazioni e può essere incompleta o non aggiornata. Controllala periodicamente e disattivala quando non è necessaria.
- Mostra al modello esempi e file concreti: lavora meglio a partire da ciò che ha davanti.
- Proteggi i dati degli studenti: anonimizza, usa gli account istituzionali e non affidarti mai alla memoria per le informazioni sensibili.

Bibliografia

- 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. Presentazione formativa VITA.

Link

- Memoria in ChatGPT: help.openai.com/en/articles/8590148-memory-faq
- Progetti in ChatGPT: help.openai.com/en/articles/10169521-projects-in-chatgpt
- Dimostrazione del tokenizzatore: tiktokenizer.vercel.app

Copyright © 2026 Consorzio VITA

Quest'opera è concessa in licenza secondo i termini della Creative Commons
Attribuzione-Condividi Allo Stesso Modo 4.0 Internazionale (CC BY-SA 4.0).



Finanziato dall'Unione europea. Le opinioni espresse sono esclusivamente quelle degli autori e non riflettono necessariamente quelle dell'Unione europea o dell'Agenzia nazionale.