

Konteksto langai ir atminties moduliai: reikšmė švietimui

Autorius: Zsolt Lengyel, ITStudy Hungary Training and Research Centre

Recenzentė: Mária Hartyányi, ITStudy Hungary Training and Research Centre

Mokymosi tikslas

Daugelį netikėtų DI sistemos veikimo ypatumų – ankstesnių nurodymų nepaisymas, skirtingų temų supainiojimas ar anksčiau pateiktos informacijos apie naudotoją panaudojimas – galima paaiškinti dviem mechanizmais: konteksto langu ir atminties funkcijomis. Šios mokomosios medžiagos tikslas – paaiškinti, kokią informaciją DI sistemos gali išlaikyti pokalbio metu ir išsaugoti vėlesniems pokalbiams. Išnagrinėję medžiagą, mokytojai gebės atskirti laikiną pokalbio kontekstą nuo ilgalaikę atmintį imituojančių funkcijų, paaiškinti jų poveikį darbo tęstinumui, atsakymų individualizavimui ir privatumui. Jie taip pat gebės taikyti praktinius metodus, padedančius užtikrinti ilgų darbo sesijų patikimumą ir mokinių duomenų apsaugą.

Įvadas: kalbos modelio ir žmogaus atminties skirtumai

Bendraujant su kalbos modeliu gali susidaryti įspūdis, kad jis prisimena visą pokalbį. Tačiau pats modelis pokalbio metu nesikeičia. Prieš generuojant kiekvieną atsakymą, sistema suformuoja įvestį iš ankstesnio pokalbio turinio, pridėtų failų ir sisteminių nurodymų. Kitą atsakymą modelis generuoja remdamasis šia įvestimi.

Šį veikimo principą galima iliustruoti filmų „Memento“ ar „50 pirmų pasimatymų“ veikėjų analogija: žmogus negali išsaugoti naujų prisiminimų, tačiau išlaiko pagrindines žinias. Analogiškai kalbos modelio bendrosios žinios įgyjamos mokymo metu, o kita atsakymui reikalinga informacija turi būti pateikiama pakartotinai.

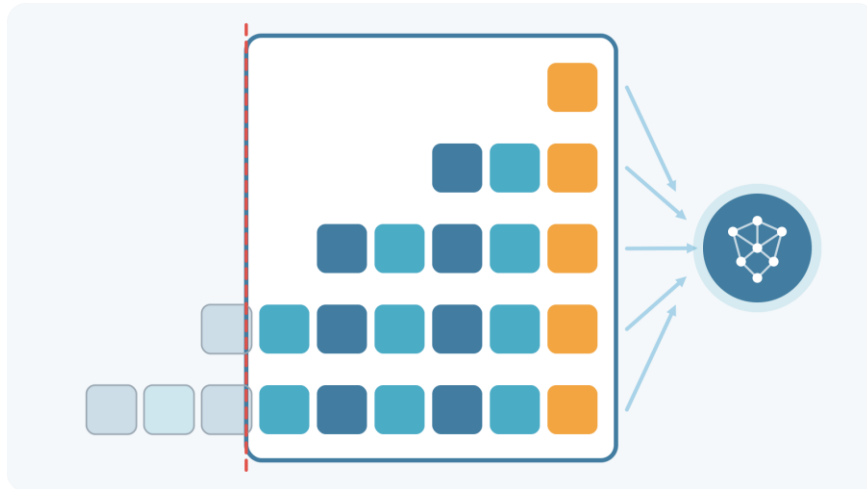


„50 pirmų pasimatymų“ (šaltinis: <https://nlc.hu/>)

Pokalbio konteksto formavimas

Pokalbių sąsajoje įvesties apimtis didėja su kiekvienu apsikeitimu žinutėmis. Pirmajame etape modelis gauna pirmąją naudotojo žinutę (Ž1) ir pateikia atsakymą (A1). Antrajame etape sistema perduoda seką Ž1 + A1 + Ž2, trečiajame

– $\tilde{Z}_1 + A_1 + \tilde{Z}_2 + A_2 + \tilde{Z}_3$ ir t. t. Visa konkrečiu momentu modeliui pateikiama informacija vadinama **kontekstu**.



Su kiekvienu apsikeitimu žinutėmis įvesties apimtis didėja; į ribotos talpos konteksto langą nebetelpanti informacija pašalinama arba suglaudinama

Konteksto apimtis matuojama **žetonais** (angl. tokens) – nedideliais teksto vienetais, dažnai atitinkančiais žodžių dalis. Didžiausias žetonų skaičius, kurį modelis gali apdoroti vienu metu, apibrėžia **konteksto lango talpą**. Ji gali būti didelė, tačiau yra ribota: šios medžiagos rengimo metu pažangiausių modelių konteksto langai apima nuo kelių šimtų tūkstančių žetonų (pvz., apie 400 000 GPT-5 klasės modeliuose) iki maždaug milijono (pvz., „Gemini 3.1 Pro“). Tokia apimtis atitinka kelių didelės apimties knygų tekstą. Pokalbių programoje taikoma riba gali būti mažesnė už techninę modelio konteksto lango talpą ir dažnai priklauso nuo prenumeratos. Daugelyje kalbų vienam žodžiui užkoduoti vidutiniškai reikia daugiau žetonų nei anglų kalboje.

Konteksto lango praktinė reikšmė

- **Talpos riba:** pokalbiui viršijus konteksto lango talpą, sistema pašalina seniausią jo dalį arba ją apibendrina.
- **Sąnaudos ir veikimo greitis:** didesnei įvesties apimčiai apdoroti reikia daugiau skaičiavimų, todėl atsakymų generavimas gali sulėtėti, o naudojimo limitai – būti pasiekti greičiau.
- **Atsakymų kokybė:** didesnis kontekstas gali būti naudingas, tačiau jame gali daugėti ir nereikšmingos informacijos. Tyrimai parodė, kad modeliai patikimiau panaudoja ilgo konteksto pradžioje ir pabaigoje esančią informaciją nei pateiktą jo viduryje. Todėl dalis informacijos gali likti nepanaudota. Ankstesnė, su dabartine užduotimi nesusijusi medžiaga didina klaidingo interpretavimo tikimybę.

Konteksto suglaudinimas

Pokalbiui pasiekus konteksto lango ribą, daugelis programų automatiškai **suglaudina** senesnes žinutes, pakeisdamos jas santrauka. Taip išlaikoma pagrindinė informacija ir atlaisvinama vietos naujam turiniui, tačiau dalis detalių prarandama. Naudotojas negali kontroliuoti, kada šis procesas įvyksta. Esminis suglaudinimo trūkumas – galimas smulkių detalių ir tikslų ankstesnių

formuluočių praradimas. Be to, itin ilgi pokalbiai gali sulėtinti naršyklės veikimą.



Suglaudinant kontekstą, ilgas pokalbis pakeičiamas trumpa santrauka, todėl dalis detalių gali būti prarasta

Ilgų darbo sesijų organizavimo rekomendacijos

- Keisdami temą, pradėkite naują pokalbį. Tą patį rekomenduojama daryti, kai pokalbis viena tema tampa itin ilgas.
- Prieš pradėdami naują pokalbį, pateikite užklausą: „Apibendrink šio pokalbio sprendimus, reikalavimus ir neišspręstus klausimus taip, kad santrauką galėčiau perkelti į naują pokalbį.“ Patikrinkite, ar santraukoje išliko visa tolesniam darbui svarbi informacija.
- Ilgus dokumentus pridėkite kaip failus, užuot įklijavę jų tekstą į žinutę. Kai kurios programos, įklijavus itin ilgą tekstą, automatiškai pateikia jį kaip failą.

Trys informacijos sluoksniai: bendrosios žinios, atmintis ir dabartinis kontekstas

DI asistento naudojamą informaciją galima suskirstyti į tris sluoksnius:



Atsakymą lemia mokymo metu įgytos bendrosios žinios, ribotos apimties redaguojama atmintis ir dabartinis kontekstas

- **Bendrosios žinios** įgyjamos modelio mokymo metu. Jos apima daugelį sričių, tačiau nebūtinai atspindi naujausią informaciją: jų aktualumą riboja modelio žinių ribos data (angl. knowledge cut-off). Atsakydamas į klausimus apie naujausius įvykius ar pokyčius, modelis turi pasitelkti paiešką internete.
- **Atmintis** – palyginti nedidelės apimties informacijos apie naudotoją rinkinys, kurį programa perduoda modeliui naujo pokalbio pradžioje. Ankstesnio pokalbio kontekstas į naują pokalbį savaime neperkeliamas, išskyrus atminties modulio pateikiamą informaciją.
- **Dabartinis kontekstas** – konkrečiu momentu modeliui pateikiama informacija: pokalbio turinys, failas, pavyzdys, klaidos pranešimas ar pageidaujamas atsakymo formatas. Modeliai veiksmingiausiai apdoroja aiškiai pateiktą informaciją: vienas tinkamai parinktas pavyzdys dažnai yra naudingesnis už ilgą paaiškinimą.

Atminties funkcijų taikymas: „ChatGPT“ pavyzdys

Dauguma plačiai naudojamų DI asistentų („ChatGPT“, „Claude“, „Gemini“) siūlo atmintį imituojančias funkcijas. Pavyzdžiui, šios medžiagos rengimo metu „ChatGPT“ siūlomos šios funkcijos:

- **Išsaugoti prisiminimai** – informacija, kurią naudotojas aiškiai paprašė įsiminti arba kurią asistentas išsaugojo automatiškai. Ji saugoma atskirai nuo pokalbių, todėl ištrynus pokalbį išsaugotas prisiminimas nepašalinamas.
- **Rėmimasis pokalbių istorija** – funkcija, leidžianti asistentui naudoti ankstesnių pokalbių informaciją atsakymams individualizuoti.
- **Atminties santrauka ir atminties šaltiniai** – informacijos, kurią asistentas gali naudoti atsakymams individualizuoti, apžvalga ir prie kai kurių atsakymų pateikiami duomenys apie ankstesnius pokalbius ar failus, turėjusius įtakos atsakymui. Šią informaciją galima koreguoti arba ištrinti.
- **Laikinasis pokalbis** – pokalbis, kurio metu nekuriami ir neatnaujinami išsaugoti prisiminimai.
- **Tik projekto atmintis** – funkcija, leidžianti projekto kontekstą apriboti tik to projekto pokalbiais ir failais (žr. medžiagą apie sudėtingų užduočių valdymą).

Atminties funkcijų ribotumai

Atminties funkcija nėra lygiavertė duomenų bazei. Asistentas pats atrenka, kurią informaciją laikyti svarbia, todėl neišsaugo visų detalių. Išsaugota informacija gali pasenti, būti klaidinga arba būti panaudota naudotojui netikėtame kontekste. Dėl šių priežasčių atminties santrauką rekomenduojama reguliariai peržiūrėti.

Reikšmė švietimo praktikai

Tęstinumas

Kadangi trumpalaikis kontekstas neišsaugomas pereinant iš vieno pokalbio į kitą, darbo tęstinumą būtina numatyti iš anksto. Kiekvienai kurso temai parenkite trumpą konteksto aprašą, kuriame būtų nurodyti temos tikslai, mokinių grupės pasirengimo lygis (be asmens duomenų), pageidaujamas medžiagos stilius ir formatas. Šį aprašą pateikite naujų pokalbių pradžioje arba išsaugokite projekto nurodymuose ir failuose. Toks būdas patikimesnis nei prielaida, kad atminties funkcija išsaugos visas reikalingas detales.

Atsakymų individualizavimas

Atminties funkcija gali padėti taupyti laiką: asistentas gali išsaugoti informaciją, kad mokote šešiolikmečius automobilių mechatronikos ir teikiate pirmenybę lentelėms su teisingais atsakymais. Tačiau individualizuojant atsakymus gali susiaurėti jų turinys arba būti pakartotinai pritaikytos klaidingos prielaidos. Kai reikalingas neutralus atsakymas arba siekiama įvertinti, kaip priemonė atsakytų mokiniui, rekomenduojama naudoti laikinąjį pokalbį.

Privatumas ir duomenų apsauga



Atminties turinį galima peržiūrėti ir ištrinti, o funkciją – išjungti; patikimiausia duomenų apsauga užtikrinama jų apskritai nepateikiant sistemai

- Niekada nepateikite mokinių asmens duomenų – vardų, pažymių, informacijos apie sveikatą ar šeimą, nuotraukų – bendrosios paskirties DI priemonėms. Naudokite anonimizuosius arba išgalvotus pavyzdžius.
- Laikykitės BDAR reikalavimų ir mokyklos duomenų apsaugos taisyklių. Jei yra galimybė, naudokite mokyklos suteiktas institucines DI paskyras.
- Patikrinkite atminties nustatymus: atminties ir rėmimosi pokalbių istorija funkcijas galima išjungti, o atskirus išsaugotus prisiminimus – ištrinti. Paslaugų teikėjai ribotą laiką gali saugoti ištrintų prisiminimų žurnalus („ChatGPT“ atveju – iki 30 dienų).
- Demonstruodami savo ekraną pamokoje, įvertinkite galimo informacijos atskleidimo riziką: atmintimi grindžiamas atsakymų

individualizavimas ar šoninėje juostoje matomi pokalbių pavadinimai gali atskleisti informaciją, kurios neketinate rodyti.

- Ugdykite mokinių saugaus skaitmeninių priemonių naudojimo įpročius: mokykite juos nustatyti, kokią informaciją DI apie juos išsaugo, kaip ją peržiūrėti ir ištrinti.

Pagrindinės išvados mokytojams

- Modelis kiekvieną atsakymą generuoja remdamasis dabartine įvestimi; jo veikimas skiriasi nuo žmogaus atminties.
- Konteksto lango talpa yra ribota: ilgi pokalbiai gali sulėtinti sistemos veikimą, padidinti nereikšmingos informacijos kiekį ir galiausiai būti suglaudinti.
- Pradėdami nagrinėti naują temą, pradėkite naują pokalbį. Svarbius sprendimus perkelkite pateikdami patikrintą santrauką arba konteksto aprašą.
- Atminties apimtis ribota, joje saugoma informacija atrenkama ir gali būti redaguojama. Reguliariai ją peržiūrėkite, o prireikus išjunkite atminties funkciją.
- Pateikite modeliui konkrečius pavyzdžius ir failus: jis veiksmingiausiai dirba su tiesiogiai pateikta informacija.
- Saugokite mokinių duomenis: juos anonimizuokite, naudokite institucines paskyras ir niekada neišsaugokite jautrios informacijos DI atmintyje.

Literatūra

- 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. VITA mokymų pristatymas.

Nuorodos

- „ChatGPT“ atmintis: help.openai.com/en/articles/8590148-memory-faq
- „ChatGPT“ projektai: help.openai.com/en/articles/10169521-projects-in-chatgpt
- Teksto skaidymo į žetonus demonstracija: tiktokenizer.vercel.app

Autorių teisės © 2026 m. VITA konsorciumas

Šį kūrinį galima pakartotinai naudoti ir adaptuoti pagal „Creative Commons“ priskyrimo ir analogiško platinimo 4.0 tarptautinės licencijos (CC BY-SA 4.0) sąlygas.



Finansuoja Europos Sąjunga. Išsakytos nuomonės ir požiūriai priklauso tik autoriui (-iams) ir nebūtinai atspindi Europos Sąjungos ar nacionalinės agentūros nuomonę.