

Kas yra LLM? Kaip kalbos modeliai generuoja tekstą – ir kodėl jie haliucinuoja

Autorius: Zsolt Lengyel, iTStudy Hungary Training and Research Centre

Recenzentė: Mária Hartyányi, iTStudy Hungary Training and Research Centre

Mokymosi tikslas

Pokalbių robotai, tokie kaip ChatGPT, Claude ar Gemini, atsako sklandžiai ir užtikrinta kalba – net tada, kai klysta. Šios medžiagos tikslas – konkrečiais pavyzdžiais paaiškinti, kas yra didysis kalbos modelis (LLM), kaip jis kuria tekstinius atsakymus ir kodėl gali pateikti netikslią, išgalvotą ar klaidinančią informaciją. Baigę šį mokymosi turinį, mokytojai gebės apibrėžti „žetono“ (angl. token) sąvoką, paaiškinti kito žetono prognozavimo principą, atpažinti tipines haliucinacijų formas ir įspėjamuosius ženklus bei taikyti praktines DI rezultatų tikrinimo strategijas ir priemones.

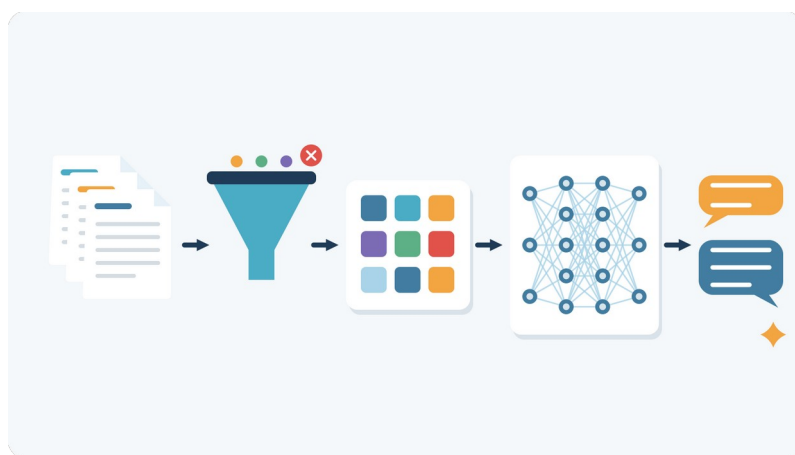
Įvadas

2022 m. lapkričio 30 d. buvo išleista pirmoji vieša ChatGPT versija, per penkias dienas užsiregistravo milijonas žmonių, o nuo tada didieji kalbos modeliai tapo kasdieniais žmonių darbo įrankiais. Jų atsakymai gerai struktūruoti ir užtikrinti. Būtent todėl jų klaidos pavojingos: klaidingas atsakymas atrodo lygiai taip pat kaip teisingas.

Mokytojams nebūtina išmanyti didžiųjų kalbos modelių matematikos, tačiau bendras supratimas, kaip veikia LLM, daug ką paaiškina: kodėl jis taip gerai tvarkosi su kalbomis, kodėl kartais ką nors išsigalvoja ir kodėl jo pateiktų rezultatų tikrinimas yra profesinė atsakomybė.

Kaip sukuriamas didysis kalbos modelis

Didysis kalbos modelis – tai DI sistema, apmokyta naudojant milžinišką kiekį teksto; pagrindinis jos gebėjimas – interpretuoti ir generuoti kalbą. „Didžiuoju“ jis vadinamas todėl, kad turi šimtus milijardų parametrų – reguliuojamų skaičių, kuriuose saugoma tai, ką modelis išmoko. Jo kūrimą sudaro keturi pagrindiniai etapai:



Nuo interneto tekstų filtravimo per tokenizavimą ir mokymą iki pokalbių asistento: pagrindiniai LLM kūrimo etapai

Duomenų rinkimas ir filtravimas

Kūrėjai atsisiunčia didelę viešojo interneto dalį ir pašalina šlamštą, dublikatus, žemos kokybės puslapius bei asmens duomenis – lieka dešimtys terabaitų teksto. Modelis gali „žinoti“ tik tai, kas buvo šiuose išfiltruotuose duomenyse, – ir tik iki tam tikros datos (tai vadinamoji *žinių riba*, angl. knowledge cutoff).

Tokenizavimas

Modelis mato ne raides ar žodžius, o skaičius. **Žetonai** yra žodžių fragmentai, ištisi žodžiai arba skyrybos ženklai, kurių kiekvienas turi identifikacinį numerį. Tipinio modelio žodyną sudaro apie šimtą tūkstančių žetonų. Ilgas lietuviškas žodis dažnai suskaidomas į kelis žetonus – tai viena iš priežasčių, kodėl LLM stebėtinai prastai suskaičiuoja, kiek kartų tam tikra raidė pasikartoja žodyje.

Išankstinis mokymas

Modelis gauna žetonų seką ir bando numatyti kitą žetoną, tada patikrina, ar atspėjo, ir šiek tiek pakoreguoja savo parametrus. Tai kartojama trilijonus kartų. Rezultatas – labai sudėtinga „teksto tęsimo mašina“, bet dar ne asistentas.

Tolesnis mokymas

Po to modelis mokomas naudojant kruopščiai parašytus klausimų ir atsakymų pavyzdžius, kad išmokytų atsakyti naudingai, mandagiai ir tinkamu formatu – bei suprastų, kada reikia atsisakyti atsakyti. Galiausiai žmonės arba vertinimo modeliai reitinguoja alternatyvius atsakymus, todėl modelis išmoksta, kurie atsakymai labiau pageidaujami.

Atsakymo generavimas: po vieną žetoną

Kai išsiunčiame žinutę, modelis kiekvienam galimam kitam žetonui apskaičiuoja, kokia tikimybė, kad jis eis po iki šiol parašyto teksto. Jis pasirenka vieną žetoną, prideda jį prie teksto ir kartoja procesą – žetonas po žetono. Tai, ką ekrane matome kaip „rašymą“, yra šių žetonų atkeliavimas. Paimkime dirbtuvėse girdimo sakinio pradžią:

„Prieš pjaudami plieninį vamzdį kampiniu šlifukuokliu, visada dėvėkite ...“

Tęsiniai, pavyzdžiui, *apsauginius akinius, pirštines ar apsaugines ausines*, yra labai tikėtini, nes panašūs sakiniai daug kartų pasitaikė mokymo duomenyse, o tęsinys *šlepetes* – itin mažai tikėtinas. Modelis neieško taisyklės – jis tęsia tekstą statistiškai tikėtiniu būdu.



Kito žetono prognozavimas: modelis įvertina galimus tęsinius ir vieną iš jų prideda prie teksto

Verta įsidėmėti tris pasekmes:

Atsakymai skiriasi

Modelis ne visada pasirenka vienintelį labiausiai tikėtiną žetoną, todėl į tą patį klausimą galima gauti skirtingus atsakymus.

Modeliui „mąstyti“ reikia žetonų

Paprašytas pateikti tik rezultatą, modelis kartą atsakė, kad trys mašinos, kurioms atskirai tam pačiam darbui atlikti reikia 12, 8 ir 6 valandų, kartu jį atliktų per 3,43 valandas. Kai jam buvo leista spręsti žingsnis po žingsnio, jis gavo teisingą atsakymą – 2 valandas 40 minučių. „Mąstantys“ modeliai tokią samprotavimo grandinę sugeneruoja prieš pateikdami atsakymą.

Modelis neprisimena taip, kaip žmogus

Kiekvienas atsakymas generuojamas iš teksto, kurį modelis gauna tuo metu (žr. M1-T1-MLC5); naujausia informacija gaunama tik per įrankius, pavyzdžiui, paiešką internete, kodo vykdymą ar pridėtus failus.

Tokenizavimas – tai teksto skaidymas į mažus vienetus ir unikalaus identifikacinio numerio priskyrimas kiekvienam iš jų. Pavyzdžiui, GPT-4 modelio „žodyną“ – visų galimų žetonų sąrašą – sudaro 100 277 elementai. Taigi, kai mūsų žinutė pasiekia modelį, ji jau yra skaičių seka – pavyzdžiui, [15339, 1917, 264, 3918].

Atsakymas generuojamas tuo pačiu principu, tik atvirkščia tvarka: modelis grąžina žetonus vieną po kito, o paprasta programa juos vėl paverčia skaitomu tekstu. Ekrane tai matome kaip nenutrūkstamą rašymą, tačiau iš tikrųjų žetonai atkeliauja vienas po kito.

Išbandykime!

Tokenizavimą galima išbandyti interaktyvioje svetainėje Tiktokerizer (<https://tiktokerizer.vercel.app/>). Įvedę tekstą pamatysime, kaip GPT-4o modelis jį suskaido į žetonus: kuris žodžio fragmentas taps atskiru žetonu, o kuris – ne.

Kodėl DI haliucinauja

Dažnai tenka išgirsti klaidinantį posakį „DI meluoja“. Iš tiesų pasitaiko, kad į klausimą gauname klaidingą atsakymą, tačiau taip nutinka dėl modelio mokymo būdo: mokymo pavyzdžiuose po klausimų eina užtikrinti atsakymai, todėl modelis išmoksta pateikti atsakymą primenantį tekstą net tada, kai neturi patikimų žinių.



Žmonių haliucinacijos kyla iš vaizduotės! (Šaltinis: ChatGPT)

Galbūt ir gerai, kad literatūroje įsivertino terminas ***haliucinacija***, nors pats procesas iš esmės skiriasi nuo žmogaus haliucinacijų: **DI neturi žmogiškos vaizduotės; jis nemato ir neįsivaizduoja to, ko nėra, o generuoja tikėtiną informaciją remdamasis išmoktais dėsningumais – informaciją, kuri kartais gali būti neteisinga.**

Kalbant apie didžiuosius kalbos modelius, haliucinacija – tai DI sugeneruotas turinys, kuris atrodo sklandus, nuoseklus ir tikėtinas, tačiau yra faktiškai neteisingas arba nepagrįstas patikimais įrodymais. Ji atsiranda todėl, kad modelis generuoja tekstą prognozuodamas statistiškai tikėtinas žodžių sekas, o ne tikrindamas savo teiginius pagal tikrovę.

OpenAI tyrėjai taip pat parodė, kad įprasti mokymo ir vertinimo metodai labiau skatina spėlioti nei pripažinti netikrumą (Kalai ir kt., 2025). Naujesni modeliai dažniau sako „nežinau“, tačiau problema neišnyko. Haliucinuotas atsakymas statistiškai dera su tikrais atsakymais – tas pats tonas, tas pats formatas, – tik jis nėra teisingas.

Tipiniai pavyzdžiai

Išgalvoti faktai ir asmenys.

Paklaustas apie mažai žinomą asmenį, pokalbių robotas kartais pateikia išsamią biografiją su profesija, darbais, metais, nors visa ji yra atsitiktinai sudurstyta.

Išgalvotos nuorodos ir teisės aktai.

Paklaustas apie saugos taisyklės teisinį pagrindą, modelis gali nurodyti nutarimą su numeriu, data ir punktu, pateiktais neprieikštingu formatu, – nors toks nutarimas neegzistuoja arba reglamentuoja visai ką kita. 2023 m. Niujorko teismas nubaudė advokatus, pateikusius procesinį dokumentą su ChatGPT išgalvotais teismų sprendimais (Mata v. Avianca).

Tikėtinios, bet klaidingos techninės vertės.

Varžto priveržimo momentas, kabelio skerspjūvis, gaminimo temperatūra ar leistinasis nuokrypis gali atrodyti teisingi, nes patenka į įprastą intervalą, – tačiau skirtis nuo gamintojo ar standarto nurodytos vertės. Skaičiavimuose klaidų pasitaiko lengvai!

Tylaus pakeitimo rizika.

Modelis gali palaikyti retus techninius terminus rašybos klaidomis ir pakeisti juos panašiais, dažniau vartojamais žodžiais. Pavyzdžiui, nežinomą augalo, gyvūno ar mašinos detalės pavadinimą jis gali automatiškai pakeisti žinomu žodžiu, kuris atrodo ar skamba panašiai. Gautas atsakymas vis tiek gali atrodyti sklandus ir įtikinamas, nors pasuka visiškai kita ir klaidinga linkme.

Praleidimai ir šališkumas.

Paklaustas, kaip nupjauti metalinį vamzdį, DI gali teisingai pasiūlyti kampinį šlifluoklį su pjovimo disku – ir nieko neužsiminti apie akių, rankų ir klausos apsaugą. Modeliai taip pat gali atkartoti stereotipus iš savo mokymo duomenų.



Haliucinuotas atsakymas gali atrodyti oficialus ir išsamus – kol kas nors nepatikrina, koks šaltinis už jo slypi

Haliucinacijų tikimybė didesnė, kai tema yra reta, vietinė ar nauja (nacionaliniai teisės aktai, įmonės procedūros, praėjusio mėnesio pokyčiai), kai reikia tikslių skaičių, pavadinimų ar citatų ir kai klausime slypi klaidinga prielaida.

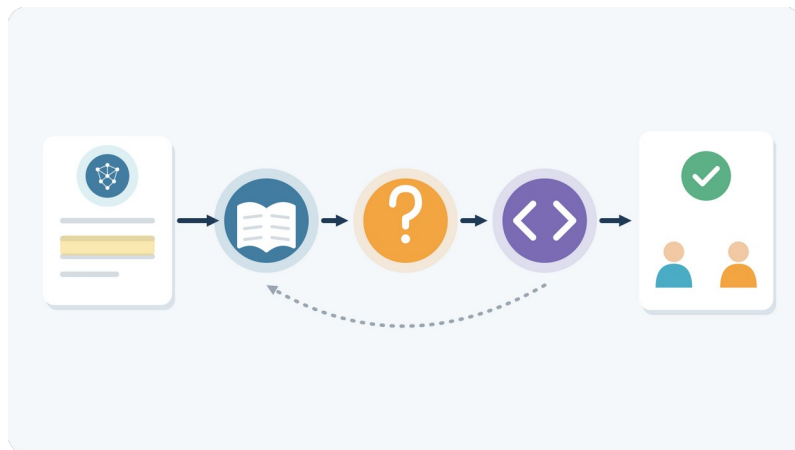
Užduokite ChatGPT šį klausimą:

Sugrupuok pagrindines sritis ir užduočių tipus pagal tai, kiek patikimi paprastai būna ChatGPT atsakymai. Sudaryk lentelę, kurioje būtų nurodytas apytikslis patikimumo intervalas, pagrindinė klaidos rizika ir reikiamas tikrinimo lygis. Procentus pateik kaip išmatuotus rezultatus tik tada, kai juos pagrindžia patikimi tyrimai; kitu atveju aiškiai pažymėk, kad tai įverčiai.

Atsakyme pateikta lentelė suteikia svarbių gairių, kaip sąmoningai naudoti DI. Pavyzdžiui, iš jos matyti, kad programavimo užduočių atveju haliucinacijų tikimybė yra santykinai maža (5–15 %), tačiau, kai kalbama apie naujausią informaciją, o prieigos prie interneto šaltinių nėra, ji gali siekti net 60 %.

Klaidų aptikimas ir rezultatų tikrinimas

Sklandumas ir užtikrintumas nėra teisingumo požymiai. Reikia klausti ne „ar tai skamba teisingai?“, o „iš kur žinau, kad tai teisinga?“ Padeda šios strategijos:



Patikrinkite šaltinį, užduokite papildomų klausimų ir naudokite įrankius – ir tik tada naudokite turinį pamokoje

Tikrinkite šaltinius!

Paprašykite nurodyti šaltinius ir juos atsiverskite: ar dokumentas egzistuoja ir ar jame iš tiesų teigiama tai, ką teigia atsakymas? Palyginkite su patikimais šaltiniais: gamintojo instrukcija, standartu, vadovu, mokymo programa.

Užduokite papildomų klausimų!

„Dėl kurių atsakymo dalių nesi tikras?“, „Iš kur gauta ši vertė?“, „Kas pasikeistų, jei...?“ Užduokite tą patį klausimą dar kartą naujame pokalbyje arba kitame modelyje: nenuoseklūs atsakymai yra įspėjamasis ženklas.

Liepkite jam naudoti įrankius!

Naujausiems faktams paprašykite atlikti paiešką internete, o matematiniams ir techniniams skaičiavimams – „naudoti kodą“. Tokiu atveju DI parašys programą uždaviniui išspręsti ir ją paleis, kad pateiktų atsakymą, – tai gerokai sumažina klaidų tikimybę.

Remkitės savo dokumentais!

Pridėkite reikiamą skyrių ar techninių duomenų lapą ir nurodykite: „Atsakyk tik remdamasis pridėtu dokumentu; jei atsakymo jame nėra, taip ir pasakyk.“

Patikrinkite prieš naudodami pamokoje!

Pasitelkite savo – arba kolegos – profesines žinias ir patikrinkite kiekvieną faktą, skaičių ir saugos nurodymą, kuriais remsis mokiniai.

Penki patikrinimai prieš naudojant pamokoje

1. Kurie teiginiai yra faktai, skaičiai, pavadinimai ar nuorodos – ir ar kiekvieną iš jų patikrinau?
2. Ar yra kas nors, kas svarbu saugai? Ar tai atitinka gamintojo duomenis ir galiojančias taisykles?
3. Ko trūksta (apsaugos priemonių, išankstinių sąlygų, išimčių)?
4. Ar lygis, terminija ir kontekstas tinka mano mokiniams?
5. Ar pasirašyčiau po šia medžiaga savo vardu?

Pedagoginis ir etinis klausimas

Haliucinacija nėra vien techninis nepatogumas. Kai mokytojas išdalija DI sugeneruotą užduočių lapą, jame esančios klaidos įgyja mokytojo autoritetą, o mokiniai, kurie dar tik kaupia profesines žinias, mažiausiai pajėgūs jas pastebėti. Dirbtuvėse klaidinga vertė ar praleistas saugos žingsnis gali baigtis sugadinta įranga ar sužalojimu. Mokinių įsimintų klaidų sunku atsikratyti.

Todėl atsakomybė tenka mokytojui: DI gali parengti juodraštį, tačiau jį patikrina mokytojas. Ne mažiau svarbu to paties išmokyti ir mokinius – klausti, iš kur gauta informacija, lyginti ją su patikimais šaltiniais ir laikyti DI atspirties tašku, o ne galutiniu autoritetu. Tai yra DI ir informacinio raštingumo dalis, aprašyta DigComp 3.0.

Šiek tiek istorijos

Nuo rašto išradimo žmonija nuolat siekė išsaugoti sukauptas rašytines žinias, perduoti jas ateities kartoms ir padaryti jas prieinamas.

Raštas, atsiradęs apie 3200 m. pr. Kr. Mezopotamijoje, leido žinioms išlikti ilgiau nei žmogaus atmintis. Molinės lentelės su dantiraščiu buvo patvarios, tačiau jas galėjo perskaityti tik nedidelis išmokytų raštininkų sluoksnis. Papiruso ir pergamento ritiniai padarė rašytines žinias lengviau pernešamas ir leido sukurti dideles antikios bibliotekas, tačiau prieiga prie jų iš esmės liko ribota – ja naudojosi tik nedaugelis privileijuotųjų.

Viduramžiais ritinį palaipsniui pakeitė kodeksas – įrišta, iš puslapių sudaryta knygos forma. Vienuolynuose ranka perrašomos knygos buvo labai brangios, o jų gamyba užtrukdavo ilgai, todėl knygų ir jose sukaupytų žinių buvo mažai. XV a. viduryje esminį pokytį atnešė spaustuvės išradimas: tekstus tapo įmanoma dauginti dideliais kiekiais ir vis pigiau. Knygų plitimas, o vėliau viešosios bibliotekos, masinis švietimas ir augantis raštingumas nuolat didino skaičių žmonių, galinčių tiesiogiai naudotis sukauptomis žiniomis.

XX a. pabaigoje dar vieną didelį šuolį atnešė kompiuteriai ir internetas. Skaitmeninimas ir pasauliniai tinklai beveik akimirksniu padarė prieinamą milžinišką informacijos kiekį. Daugelis fizinių kliūčių išnyko: šiandien vienas išmanusis telefonas gali suteikti prieigą prie išteklių, kuriems kadaise reikėjo bibliotekų, kelionių ar institucinių privilegijų. Tačiau tai sukūrė naują iššūkį. Pagrindinė problema nebėra informacijos trūkumas – dabar svarbu orientuotis jos gausoje, ją suprasti ir atskirti patikimą informaciją nuo nepatikimo turinio.

Didieji kalbos modeliai – dar vienas žingsnis demokratizuojant prieigą prie žinių. Internetas suteikė žmonėms prieigą prie informacijos; LLM žengia toliau – jie suteikia natūraliosios kalbos sąsają darbui su ta informacija. Naudotojams nebereikia tiksliai žinoti, kurią knygą ar duomenų bazę atsiversti, kokio techninio termino ieškoti ar net kaip suformuluoti tobulą paieškos užklausą. Jie gali klausti kasdiene kalba, užduoti papildomų klausimų, prašyti paprastesnių ar išsamesnių paaiškinimų, pavyzdžių, santraukų, palyginimų ar vertimų. Tą pačią temą galima paaiškinti pradedančiajam ar specialistui, skirtingu detalumo lygiu ir skirtingomis kalbomis.

Tai vienas svarbiausių demokratizuojančių LLM poveikių: **jie gali sumažinti ne tik kliūtis rasti informaciją, bet ir kalbines, technines bei su paieška susijusias kliūtis ją suprasti.** Ankstesnės technologijos palaipsniui padarė knygas, o vėliau ir informaciją, prieinamas vis daugiau žmonių. Kalbos modeliai prideda naują dimensiją: jie gali pritaikyti informaciją prie naudotojo klausimo, kalbos, konteksto ir žinių lygio.



Žmonijos žinių išsaugojimas: nuo molinių lentelių iki kalbos modelių (Šaltinis: ChatGPT)

Tai taip pat žymi svarbų skirtumą nuo ankstesnių žinių formų. Knygoje išsaugotas nekintamas tekstas; LLM į kiekvieną klausimą generuoja naują atsakymą, remdamasis mokymo metu išmoktais dėsningumais. Dėl to jis nepaprastai lankstus, bet ne neklystantis. Jis gali pateikti netikslius ar net visiškai klaidingus atsakymus, o teiginio kilmės ne visada įmanoma atsekti iki konkretaus šaltinio. **Kuo lengviau pasiekiamos žinios, tuo svarbesnis tampa gebėjimas įvertinti ir patikrinti gautus atsakymus.**

Taigi nuo molinės lentelės per spausdintą knygą ir internetą iki LLM matome ilgą procesą, kurio metu **žinios tapo prieinamos vis daugiau žmonių, o kliūčių vis mažėjo. Išskirtinis LLM indėlis – tai, kad kitas demokratizacijos etapas nebėra vien prieiga prie informacijos: vis labiau tai prieiga prie supratimo.**

Pagrindinės išvados mokytojams

- LLM generuoja tekstą žetonas po žetono, rinkdamasis tikėtinus tęsinius pagal mokymo duomenyse esančius dėsningumus.
- Jis neieško faktų: sklandūs, užtikrinti atsakymai gali būti klaidingi.
- Būkite ypač atidūs su skaičiais, nuorodomis, teisės aktais, retais terminais, vietinėmis ar naujomis temomis ir saugos nurodymais.
- Tikrinkite šaltinius, užduokite papildomų klausimų, naudokite paiešką ir kodą, pateikite savo šaltinius.
- Patikrinkite prieš naudodami pamokoje – ir mokykite mokinius taip pat tikrinti DI rezultatus.

Literatūra

- Kalai, A. T., Nachum, O., Vempala, S. S. & Zhang, E. (2025): Why Language Models Hallucinate. arXiv:2509.04664.
- Huang, L. ir kt. (2024): A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems. DOI: 10.1145/3703155.
- Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Cosgrove, J. & Cachia, R. (2025): DigComp 3.0 – European Digital Competence Framework. 5-asis leid., Publications Office of the European Union.
- Lengyel, Zs. (2026): Introduction to the concepts of artificial intelligence; Concrete applications for content generation. VITA mokymų pristatymai.

Nuorodos

- Tiktokerizer – pažiūrėkite, kaip tekstas suskaidomas į žetonus: tiktokenizer.vercel.app
- Interaktyvi trimatė LLM vizualizacija: bbbycroft.net/llm
- A. Karpathy: Deep Dive into LLMs like ChatGPT (vaizdo įrašas): youtube.com/watch?v=7xTGNNLPyMI

Autorių teisės © 2026 VITA konsorciumas

Šia medžiaga galima dalintis ir ją adaptuoti pagal Creative Commons Priskyrimo — Analogiško platinimo 4.0 tarptautinės licencijos (CC BY-SA 4.0) sąlygas.





Finansuojama Europos Sąjungos. Išsakytos nuomonės priklauso tik autoriui(-iams) ir nebūtinai atspindi Europos Sąjungos ar Nacionalinės agentūros nuomonę.

