Vilniaus universitetas Filosofijos fakultetas

Filosofijos institutas

Gytis Pranskūnas

Filosofijos studijų programa

Kursinis darbas

**Prasmės problema didžiųjų kalbos modelių veikime: filosofinis tyrimas.**

Darbo vadovas: Docentas dr. Mindaugas Gilaitis

Vilnius 2025

Turinys

[Įvadas 3](#_Toc201421290)

[Wittgensteinas 5](#_Toc201421291)

[Reikšmė kaip vartojimas. 5](#_Toc201421292)

[Kalbos žaidimai. 7](#_Toc201421293)

[Šeimyninis panašumas. 9](#_Toc201421294)

[LLM veikimo principai 11](#_Toc201421295)

[Mokymosi duomenys. 11](#_Toc201421296)

[Dėmesio mechanizmas. 13](#_Toc201421297)

[Semantinių ryšių modeliavimas. 14](#_Toc201421298)

[Išvados 15](#_Toc201421299)

[Literatūra 16](#_Toc201421300)

# Įvadas

Technologijų, ypač dirbtinio intelekto (*angl.* Artificial Intelligence, DI) pažanga per pastaruosius metus pravėrė naujas duris ir filosofijoje. Tokių didžiųjų kalbos modelių (*angl.* large language models, LLM) kaip ChatGPT ar BERT veikimo principai ir jų sąryšiai yra migloti, sunkiai įkandami. Šie modeliai, remdamiesi milžiniškais kalbos duomenų kiekiais, geba generuoti tekstą, kuris neretai atrodo prasmingas, tinkamas kontekstui ir funkciškai panašus į žmogaus kalbą. Filosofija atsako į klausimus juos panaikindama (FT §109). Kadangi LLM‘ai naudojasi kalba, o Wittgensteino tyrimo objektas ir buvo kalba – ar šių informacinių įrankių veikimas susijęs su vitgenšteiniška prasmės kaip vartojimo koncepcija? Kiek pagrindo apskritai yra lyginti Wittgensteino mąstymą su LLM? Kiek filosofinė tradicija duoda pagrindo nagrinėti sąsajas tarp Wittgensteino ir informacinių sistemų? Tad prasmės problema iš esmės kyla iš klausimo ar galime sieti Wittgensteino prasmės kaip vartojimo koncepciją su LLM‘ų veikimo principais, kuriuos aprašau šiame darbe – kaip jie yra mokomi, kaip veikia dvi bendros LLM rūšys, kaip suvokia žodžių santykius, kaip prasmingai vartoja žodžius. Mano manymu, yra nemažai pagrindo teigti, kad LLM‘ų veikimo principuose randame Wittgensteino koncepcijų, kurios iš esmės nurodo į prasmės kaip vartojimo koncepciją. Viena vertus, atrodo, kad šių sistemų kalbos vartojimas yra prasmingas dėl to kaip plačiai ir greitai išplito jų naudojimas visose gyvenimo sferose. O tai veda prie klausimo apie šių technologijų dalyvavimą kalbos žaidimuose.

Kokia šio tyrimo prasmė ir motyvacija, ką jis atskleidžia? Šis tyrimas yra paruošiamasis darbas, kadangi filosofijoje dar nėra plačiai tyrinėjami LLM‘ai, ypač sąsajoje su LLM‘ais, analizuojamas jų veikimas, ko pasekoje neturime žodyno ir schemų, kuriomis galėtume operuoti ir tikslingai svarstyti. Tad šio darbo racija ir noras – pabandyti pasėti mintis, sužaisti kalbos žaidimą, kas vėliau galbūt praverstų ir atvestų į tikresnį žinojimą. Nesieksiu atsakyti, ar DI/LLM „supranta”, „mąsto” ar „suvokia“ reikšmę tokiu būdu, kaip tai daro žmonės. Tad nepretenduodamas į neginčijamų atsakymų pateikimą, verčiau pavadinčiau šį darbą filosofiniu tyrimu.

# Wittgensteinas

## Reikšmė kaip vartojimas.

Vienas esminių Wittgensteino vėlyvosios filosofijos teiginių – tai, kad žodžio reikšmė yra jo vartojimas kalboje. (Wittgenstein 1997, FT §43). „Nors ankstyvasis Wittgensteinas kartu su Russellu ir Frege pritarė tam tikrai šio atvaizdo [t.y. sakinio reikšmė yra objektas, į kurį nurodo] versijai, vėlyvasis Wittgensteinas pirmasis ėmėsi ją nuosekliai kritikuoti.“ (Glock 1996, p. 376) Reikšmė, Wittgensteino manymu, nėra tai, kas slypi *už* žodžio, paslėpta sąmonėje ar nurodo į objektą apie kurį kalba, greičiau tai atsiskleidžia per praktinį kalbos vartojimą tam tikrame kontekste. Norint suprasti kalbą, reikia stebėti, *kaip* ir *kur* ji veikia, o ne ieškoti jos loginės struktūros ar esminių dėsnių: „negalvok, o žiūrėk!” (FT §66). Prasmė atsiranda iš kalbinio veiksmo. Schroederis (2006) bei interpretuotojai pabrėžia, kad Wittgensteinas nekūrė naujos reikšmės teorijos, o veikiau bandė išsklaidyti filosofines painiavas, kylančias iš netinkamo kalbos supratimo (p. 170). Vadinasi, reikšmės analizė jam nėra teorinis apibrėžimas, bet praktinis tyrimas, kaip žodžiai veikia mūsų gyvenimo formose, ir kaip mūsų nesusikalbėjimai dažnai kyla ne dėl faktų stokos, o dėl netinkamo žodžių vartojimo supratimo. Pvz., (ibid., p. 171) kai bakalauro studentas vartoja sudėtingą akademinį žodyną nesuprasdamas jo tikros paskirties, klausytojas gali tai atpažinti: studentas nesugeba atskirti, kuo jo vartojamas terminas skiriasi nuo kitų, panašių. Tai reiškia, kad jis neturi veiksmingo vartojimo suvokimo – jo žinojimas yra paviršinis, nesuvoktas.

Reikšmė nėra atskirta nuo kalbos vartotojo. Schroederio pastebėjimu, vaikai neišmoksta „žodžių”, bet išmoksta, kaip jais naudotis – t. y. perima kalbos žaidimų taisykles, jų kalbos kompetencija formuojasi per praktikos įgūdžius, o ne per teorinį mokymąsi (ibid., p. 169). Būtent todėl vaikai gali pradėti vartoti žodį „ačiū”, „pienas“ ir kt. žinodami tik kelis pavyzdžius. Jie geba atskirti, kada galima pasakyti vieną ar kitą žodį, kam jis skirtas, kokiose situacijose jis priimtinas (plg. FT §7) .

Liu (2021) pripažįsta, jog Wittgensteino kalbos filosofija yra taip glaudžiai susijusi su semantiniais kompiuteriniais tinklais „kad nebegalime užmerkti akių prieš jos įsikūnijimą dirbtiniame intelekte“ (p. 427). Kaip pavyzdį pagrįsti ji pamini kaip J. F. Sowas savo konceptualiųjų grafų duomenų bazėms išradimą, kuris daro skirtumą tarp leksinės ir semantinės struktūrų, priskyrė Wittgensteino prasmės kaip naudojimo koncepcijai: Sową išmokė, jog kalbos dviprasmiškumas ir sudėtingumas kyla iš jos vartojimo naujuose kontekstuose (ibid.). Molino ir Tagliabue (2023, p. 1) tvirtinimu, M. Masterman – britų lingvistė, filosofė ir Wittgensteino mokinė – mokytojo koncepciją pritaikė mašininiam vertimui, orientuotą ne į sintaksę, o į reikšmę ir 1955m. įkūrė Kembridžo kalbos tyrimų skyrių (*angl*. Cambridge Language Research Unit). Liu (ibid.) apibendrina filosofės poziciją teigdama, kad „rakto į natūralios kalbos supratimą, taigi ir į tinkamą jos [už]kodavimą kompiuteryje, reikia ieškoti semantiniuose tinkluose, kurie vieninteliai sugeba susidoroti su žodžių reikšmių daugialypiškumu ir neapibrėžtumu.“ (p. 430). Nors Wittgensteino prasmės kaip vartojimo koncepcija neimplikuoja tiesioginio jos pritaikymo tokiose praktinėse sferose kaip kompiuterių mokslas ir jo tobulinimas, tačiau matome, jog prasmės priklausymas nuo konteksto padarė įtaką Sowui, o Masterman buvo viena pirmųjų pasiekusi rimtų rezultatų kompiuterio kalbos apdorojime.

meaning is use” highlighted that in many cases what is said is significantly less/different than what is meant - semantics is important, but so is pragmatics. (tagliabue p. 3)

* Problema, jog kalbos kaip prasmės koncepcija iš esmės nepasako nieko apie pačią kalbą, jos vartojimą. Kokia yra kalba? Matome, kad tai iš tiesų vaisinga koncepcija, tačiau tuo pačiu svarbu pastebėti, kad toks kalbos mokymas AI yra vaisingas, tačiau jis siekia lubas, o tuo pačiu toks kalbos mokymas reikalauja labai daug išteklių – duomenų centrų, vaizdo plokščių, atminties atsargų, duomenų. Tad nors toks kalbos supratimas yra vaisingas, tačiau jis nėra beribis. Tad kyla alternatyvių prasmės teorijų klausimas.

## Kalbos žaidimai.

Kalbos žaidimų sąvoka (*angl.* language-games) yra neatskiriama nuo reikšmės kaip vartojimo sąvokos. Šie du elementai nėra atsitiktinai susiję – kalbos žaidimai yra terpė, kurioje vyksta kalbos vartojimas, o kartu kuriama reikšmė, kadangi supratimas priklauso nuo ištaros funkcijos tam tikrame kontekste, tad reikšmė priklauso nuo vartojimo kažkokiame kontekste (Obermeier 1983, p. 341) Jei reikšmė yra tai, kaip žodis vartojamas, tai neišvengiamai kyla klausimas: kokiuose kontekstuose ir pagal kokias taisykles tie žodžiai vartojami? „Jie [leksikografai] tiria, kaip žodis vartojamas, ypač išsilavinusių ir iškalbingų žmonių, ir bando tą vartojimą perteikti apibrėžimu ar parafrazėmis, o didesniame žodyne - iliustruoti keliomis citatomis. Žodynas teisingai pateikia žodžio reikšmę tada ir tik tada, kai tiksliai atspindi priimtą vartoseną.“ (2006, p. 169). Kalba yra „įpinta“ į veikimą ir neatsiejama nuo jo (FT §7). Vartojimas čia apima ne tik gramatinius sakinius, bet ir jų paskirtį: įsakymas, prašymas, aprašymas, pasakojimas, užuomina – visa tai yra skirtingi kalbos žaidimai (FT §23). Kiekvienas jų reikalauja skirtingo vartojimo būdo, o žodžių reikšmės skiriasi ne dėl to, kad žodžiai būtų iš esmės kitokie, bet todėl, kad jie įtraukti į skirtingus žaidimus. Tokiu būdu reikšmė kyla iš funkcinės, ne loginės sistemos, o žodžių supratimas tampa priklausomas nuo mūsų gebėjimo orientuotis konkrečiose vartojimo situacijose. Kalbos žaidimų daugiskaita pabrėžia, kad nėra universalaus kalbos naudojimo būdo – vietoj to egzistuoja įvairios praktikos, kuriose žodžiai įgauna skirtingas prasmes priklausomai nuo veikimo konteksto. Vadinasi kalbos žaidimai neturi griežtų ribų ar išbaigtų taisyklių, bet jungiasi per panašumus. Ši kalbos fragmentacija Wittgensteinui leidžia atmesti tiek loginę-semantiškai griežtą kalbos teoriją, tiek bet kokį bandymą kalbą „išgryninti” iki vienos reikšmės sampratos. Kitaip tariant, nėra vieno meta žaidimo (Penco 2024, p. 284)

Wittgensteinas aprašė statybininkų kalbą, kurią sudaro 4 žodžiai, kuria kalbėtojas A paliepia padėjėjui B (FT §2). Kaip kalbos žaidimų taikymą dirbtinio intelekto srityje Penco pamini SHRDLU žaidimą, vieną pirmųjų *įgyvendintų* kalbos žaidimų praktiškai: „Winograd‘o SHRDLU - programa, skirta dialogui su simuliuotu robotu, su kuriuo galima bendrauti žaislų pasaulyje, sudarytame iš įvairių formų ir spalvų dėžių, kubų ir piramidžių. Žaidimas buvo labai paprastas: duoti nurodymus, kaip perkelti tuos blokelius, užduoti klausimus arba suteikti pavadinimus naujoms tų blokelių kompozicijoms.“ (ibid., p. 283). Jis taip pat pastebi, kad panašumas tarp Winogrado ir Wittgensteino idėjų yra kalbos kaip skirtingų kalbos žaidimų rinkinio (p. 284). Kita vertus, Bojin (2008) aptardamas Wittgensteino ir jo amžininko svarbaus mokslininko DI srityje Turingo įtampą mašinų mąstymo klausimu teigia, jog pastarojo apibrėžtos mašinos (ir to pasekoje visas dabartinis DI) tegali pateikti jau iš anksto apibrėžtas išvestis (p. 4), tačiau konceptai kalbos žaidimuose neturi „prasminio apibrėžtumo, todėl žaidimas [su jais] yra nuolat kintantis ir banguojantis procesas, kuris vis labiau artėja prie apibrėžtumo, bet niekada jo nepasiekia“ (p. 5). Kitaip tariant, dabartiniai LLM dalyvauja įvairiuose kalbos žaidimuose (juokelių, programų, teisinių kalbų rašymas ir t.t.), tačiau jų išvestys, nors ir netiesiogiai bei sunkiai apibrėžtai, vis dėlto *yra* iš anksto apibrėžti jiems pateikiamų duomenų, kuriais jie yra mokomi. Tuo tarpu, matome, jog kalbos žaidimai iš tiesų nėra tiesiog teorija, tačiau realizuojamas procesas (SHRDLU) pragmatiškai.

## Šeimyninis panašumas.

Šeimyninio panašumo (*angl.* family resemblance) sąvoka yra būtinas kalbos žaidimų ir reikšmės kaip vartojimo sampratos papildymas. Jei reikšmė kyla iš vartojimo, o vartojimas vyksta įvairiuose kalbos žaidimuose, tai neišvengiamai kyla klausimas: kas sieja šiuos žaidimus tarpusavyje? Wittgensteinas šeimyninių panašumų idėją pristato per žaidimų pavyzdį: „Pavyzdžiui, panagrinėkime procesus, kuriuos vaidiname „žaidimais“, turiu mintyje šachmatų žaidimą, lošimą kortomis, žaidimą su kamuoliu, sporto žaidimus ir t.t. Kas bendra jiems visiems? – Nesakyk: „Jiems *turi būti* kažkas bendra, kitaip jie nesivadintų „žaidimais“ – bet *įsižiūrėk*, ar jie turi ką nors bendra“ ir toliau paaiškina, jog įsižiūrėję pamatysime „ne tai, kas jiems *visiems* bendra, bet panašumus, giminingumus, ir jų yra labai daug“ (FT §66). Kitaip tariant, daugelis „aiškių“ ir „apibrėžtų“ sąvokų iš tikrųjų neturi vieno bendro esminio bruožo, vietoj to jos yra viena kitą užklojančių panašumų tinkle (ibid.). Pasak Glocko, šeimyninių panašumų sąvoka yra svarbi pasisakant prieš „požiūrį, kad visiems sąvokos atvejams turi būti kažkas bendro, kas paaiškina, kodėl jie priskiriami šiai sąvokai“ ir „kad vienintelis tinkamas arba teisėtas žodžio paaiškinimas yra analitinis apibrėžimas, kuriame nustatomos būtinos ir pakankamos jo taikymo sąlygos“ (1996, p. 120). Kitaip tariant, dažnai neryškus vaizdas yra kaip tik tai, ko mums reikia (FT §71).

Tas pats žodis gali funkcionuoti skirtinguose kontekstuose išlaikydamas tapatumą. Pavyzdžiui, žodis „žinojimas” filosofijos („žinau, kad nieko nežinau“), mokslo („sužinojome, kad nieko negali būti greitesnio už šviesos greitį“) ir kasdienybės („žinau kada atvažiuos troleibusas“) žaidimuose vartojamas skirtingai, nors visi jie susiję su informacijos įgijimo, patikrinimo, abejojimo, spėjimo aspektais. Tuo matome, jog nėra absoliutaus, vienatinio konteksto, kuriame kaip kalbos vartotojas galėčiau turėti užbaigtą apibrėžimą: „Aprašymo išsamumas priklauso nuo konkrečios aplinkybės, kurioje esu, poreikių ir pagrindinių prielaidų.“ (Penco 2024, p. 283) Kalbos mokymasis yra galimas be išankstinio visų taisyklių žinojimo. Vaikai neišmoksta apibrėžimų – jie perima vartojimo modelius, kurie formuojasi per šeiminius panašumus su jau žinomais atvejais (plg. FT §73, Schroeder 2008 p. 169) arba rodo į objektus, priklijuoja jiems vardą ar „kartoja mokytojo pasakytus žodžius“ (FT §7)

Penco užsimena kaip šeimyninių panašumų tam tikra eksplikacija yra „statistiniai telkiniai“ pagal kuriuos žodžiai gali būti apibrėžti kitų žodžių „telkiniais“ (2024, p. 286), o Masterman naudojosi koncepcija kaip strategija atskleidžiančią žodžių schemas (ibid.) Tuo tarpu Bojin (2008) pasitelkia šeimyninį panašumą ir Wittgensteino mintis apie mąstančias mašinas ginčyti tokio diskurso (pradinę) galimybę: „Norint aptarti galinčios mąstyti mašinos idėją, pirmiausia reikėtų susitarti dėl dalykų, kurie yra bendri veiklai, kurią vadiname „mąstymu“ ar net „protu““ (p. 5). Smith ir Cook (2023) siekdami atsakyti į klausimą ar dirbtinio intelekto (DI) generuojamos nuotraukos yra menas pasitelkia Wittgensteino šeimyninių panašumų idėją pagrįsti teigiamam atsakymui.

Taigi, šia dalimi norėjau atskleisti kaip Wittgensteino idėjos jau yra naudojamos pagrįsti bendrai DI srities žinojimui bei naujoms idėjoms ir prieigoms generuoti. Darosi aišku, kad Wittgensteino idėjos turi santykį su DI, kas išryškėjo, jog santykis yra padrikas ir yra daugiau šeimyninių panašumų tarp Wittgensteino idėjų ir DI. Kol kas literatūroje yra mažai tiesioginio Wittgensteino idėjų, šiuo atveju šeimyninio panašumo, kalbos žaidimų ir prasmės kaip vartojimo santykio artikuliavimo su LLM veikimu. Panašu, kad santykis yra, tačiau klausimas kaip šios idėjos atsiskleidžia konkretesniame ir labiau apibrėžtoje ir fokusuotoje srityje kaip LLM ir jų veikimas. Juk Wittgensteino analizės objektas buvo kalba, todėl kyla klausimas kaip kalbą įkūnijančių kompiuterio programų veikimo principai yra susiję su Wittgensteino minėtomis idėjomis. Verta paminėti, kad šeimyninių panašumų ir kalbos žaidimų idėjas apibrėžia ir apibendrina reikšmės kaip vartojimo idėja, todėl pirmiausia, nagrinėjant LLM veikimo principų santykį, fundamentaliai bus nagrinėjamas santykis su prasmės kaip vartojimo koncepcija. Iš čia ir kyla prasmės problema LLM‘ų veikime.

# LLM veikimo principai

## Mokymosi duomenys.

DI srityje LLM yra lokalizuoti mašininio mokymosi (*angl.* Machine Learning, ML) srityje, kuri naudojasi algoritmais ir duomenimis siekiant „imituoti“ žmonių mokymosi būdą, palaipsniui didinant tikslumą (IBM 2021). Iš esmės, ML tikslai yra nuspėti (*angl*. predict) ir klasifikuoti duomenis: „remdamasis tam tikrais įvesties duomenimis, kurie gali būti pažymėti arba nepažymėti, algoritmas parengia įvertį apie duomenų modelį“ (ibid.). Tarkime, kad turime duomenų rinkinį susijusį su širdies ligų simptomais (krūtinės skausmas, pasunkėjęs kvėpavimas, nuovargis ir t.t.) ir juos turinčiais žmonėmis bei atsakymais, ar jiems buvo diagnozuota liga. Remiantis duomenimis tinklas išgauna funkciją, į kurią įstačius reikalaujamų kintamųjų reikšmes bus galima nuspėti kurie žmonės turi širdies ligą.

Įliustruoti mokymui norėčiau trumpai parodyti mokymosi ir veikimo principų pavyzdžius konkrečiuose GPT (*angl.* Generative Pre-trained Transformer) ir BERT (*angl.* Bidirectional Encoder Representations from Transformers) kalbos modelių algoritmuose. GPT modeliai augina tekstą po vieną simbolį, stengdamasi nuspėti kitą žodį, ir taip treniruojasi būti tolesnio sakinio tesėju. Stebėtina, jog algoritmai besiremiantys tokia strategija „kuriuose kiekvienas dokumento žodis nuspėjamas pagal prieš tai buvusius žodžius, yra tikrai veiksmingi mokantis kaip kurti sklandų, paprastai gramatiškai ir tematiškai nuoseklų tekstą.“ (Molino & Tagliabue 2023, p. 3). Tuo tarpu BERT sakinyje paslepia atsitiktinius žodžius ir iš likusių bando nuspėti pirmuosius. Kitaip tariant, matome jog iš esmės minėti algoritmai „spėlioja“ žodžius, kol po kurio laiko geba vartoti kalbą ir susišnekėti su žmonėmis.

Kas išryškėja, mano manymu, yra Wittgensteino prasmės kaip vartojimo koncepcija minėtų LLM‘ų veikime. BERT ir GPT algoritmai parodo kaip kalbos vartojimo pagrindas yra tam tikras kontekstų išmokimas, dalyvavimas įvairiausiose kalbos žaidimuose. LLM‘ai demonstruoja kaip kalba veikia per tarpusavio ryšius ir vartojimą. Žinoma, jų prieiga prie jų yra visiškai kitokia nei mūsų, tačiau mūsų egzistavimo ir buvimo būdai taip pat yra visiškai kitokie. Kita vertus, kyla klausimas ar iš tiesų toks dalyvavimas kalbos žaidimuose ir prasmės atsiskleidimas iš tiesų yra atitinka tai, ką turėjo omenyje Wittgensteinas? Bojin teigimu, „Svarbu atkreipti dėmesį į tai, kad kalbos žaidimai yra įsišakniję gyvenimo formose, t. y. kalbos žaidimai yra kažkas, kas vyksta tam tikroje gyvenimo formoje“ (p. 2). Kadangi LLM mokosi iš žmonių duomenų ir dalyvavimo kalbos žaidimuose, tai išeina, kad jie tarsi patenka į mūsų gyvenimo formas ir dalyvauja juose kartu su mumis dar ir dėl to, jog galime su jais bendrauti, „[kalbos žaidimams kaip praktikai] būdinga tam tikra organiška kokybė ... kurios šaknys glūdi dalyvaujančiųjų dinamiškume“ (ibid.) Tačiau galima būtų nesunkiai prieštarauti tokiai išvadai, kadangi LLM‘ai nieko nekuria *per se* ir jų dalyvavimas kalbos žaidimuose yra labai sąlyginis. Kita vertus, mes labai panašiai išmokstame dalyvauti kalbos žaidimuose, o tam tikrų žodžių reikšmes suprantame per praktiką ir vartojimą, kaip mokinys supranta mokytojo kalbą, kartais prašydamas pakartoti. Nėra akivaizdu kodėl Wittgensteino prasmės kaip vartojimas neatsiskleidžia LLM‘ų veikime ir mokymesi, kadangi jų kalbos ir prasmės mokymasis yra panašus į mūsų, o su šių įrankių vis platesne integracija, mūsų dalinimasis gyvenimo formomis ir bendras dalyvavimas taps vis platesnis.

Galiausiai, sunku užtikrintai priskirti LLM‘ams dalyvavimą kalbos žaidimams. Wittgensteinas apibrėždamas statybininkų kalbą pabrėžia, jog „viena pusė sušunka žodžius, kita atitinkamai veikia“ (FT §7). Kalbos žaidimai yra neatsiejami nuo veikimo ir praktikos. Tik kalbos vartojimas dar neįtraukia kalbos žaidimo ir iš pirmo žvilgsnio galime atmesti galimybę LLM‘am dalyvauti juose, tačiau čia matau sunkumų. Pirma, minėti LLM‘ai yra porūšis DI – dabar galime generuoti vaizdo įrašus, nuotraukas, muziką. Nors išeiname iš kalbos ribų, tačiau paneigti DI dalyvavimą ir praktiką būtų sunku, kadangi tai turi praktinių, pvz. ekonominių, padarinių mūsų gyvenimams. O kita vertus, informatikoje jau yra paplitęs LLM‘ų informacinių sistemų ir programų rašymas vartotojo paliepimu. Kas išryškėja yra tai – LLM‘ai (ir DI) yra ta pusė, kuri „atitinkamai veikia“ ir neigti jų dalyvavimą kalbos žaidimuose būtų neigti mūsų pačių dalyvavimą kalbos žaidimuose.

## Dėmesio mechanizmas.

Svarbi aptariamų LLM‘ų veikimo dalis yra dėmesio mechanizmas (*angl*. attention mechanism), kuris pasirodė 2017 metais labai tinkamo pavadinimo straipsnyje „Attention is all you need“ (*liet.* Tereikia dėmesio). LLM‘ai dėl to vadinami transformeriais, jog naudojasi minėtu mechanizmu. Geriausiai mechanizmą apibūdina dažnai pasitelkiamas pavyzdys: „jis ištraukė picą iš orkaitės. Ji labai skani“. Sakinys neatrodo problematiškas, tačiau LLM‘ui nėra taip lengva nustatyti ką nurodo įvardis „ji“ – šiame kontekste tai tiek pica, tiek orkaitė. Žinoma, jokia orkaitė nėra skani, taip pat, kaip jokia pica nėra metalinė. Mechanizmo pagalba kalbos modelis vis dėlto nustato žodžių tarpusavio santykį ir teisingai apskaičiuoja, jog šiuo atveju pica yra skani. Tad mechanizmas nusako tarpusavio žodžių ryšius ir jų *pasitaikymą* kartu viename ar kitame kontekste – „skani“ dažniau apibūdina picą, nei orkaitę.

Dėmesio mechanizmas atrodo kaip prasmės kaip vartojimo praktinė realizacija. Žodžiai turi prasmę tik tam tikrame kontekste, kuris padeda susieti žodžius tarpusavyje ir įtikinamai dalyvauti kalbos žaidimuose. Tuo pačiu ryškėja labiau koncentruota koncepcijos paskirtis – ne suprasti žodžių reikšmę tiesiogiai – *x* reiškia *y* – bet susieti žodžius tarpusavyje. Tad galime pradėti kalbėti ne tik apie pavienių žodžių reikšmes vienuose ar kituose kontekstuose, bet apie bendresnę sakinio prasmę, kuri priklauso nuo tinkamo žodžių susiejimo. Tuo pačiu galėtume netgi teigti, kad dėmesio mechanizmas apibrėžia kontekstą, jei kontekstą apibrėžtume kaip žodžių tarpusavio santykis ir išreiškiama mintis tam tikru sakiniu ar jų rinkiniu. Kita vertus, ar tikrai galime kalbėti apie dėmesio mechanizmą kaip tam tikro sakinio prasmės apibrėžimą? Mano manymu, apie tai sunku kalbėti, ką turime omenyje vartodami žodžius „prasmė“ ir „apibrėžimas“? Pastarasis tebūnie „išsakymas ar eksplikavimas, kuris duoda tam tikrą paskatą veiksmui ar idėją, mintį“ (o kam įpareigoja sušukimas „Šaunu!“ (plg. FT §19)). „Prasmės“ atveju intuityviai norisi kalbėti apie „mąstymą“, „protą“, kas veda prie Tiuringo testo, Searle kinų kambario ir daug problemų, kurios yra anapus šio darbo ribų.

## Semantinių ryšių modeliavimas.

Ne ką mažiau svarbus LLM veikimo principas – semantinis žemėlapis arba ryšių modeliavimas (Mikolov et al. 2013). Kaip teigia Molino ir Tagliabue, „buvo pasiūlyta daug įvairių algoritmų, skirtų žodžiams atvaizduoti kaip vektoriams daugiamatėje erdvėje, paprastai taikant statistinius metodus, kuriais stebimas žodžių sutapimas su kitais žodžiais arba kitais kontekstų reprezentavimais konkrečiame korpuse“ (2023, p. 1) Teksto dalys LLM įrašomos kaip skaitinės reikšmės, vektorių forma, pagal kuriuos koordinačių sistemoje žodžiui priskiriama vieta, žyminti žodžio semantiką. Panašūs žodžiai yra netoli vienas kito, pvz. „karalius“ yra arti „vyras“, o „karalienė“ – arti „moteris“. Atėmus „karaliaus“ vektorių iš „vyro“ ir pridėjus „moters“ gauname žodį „karalienė“ (Vylomova 2016). Tai įgalina paprastus samprotavimus remiantis linijine algebra (ibid., p. 2).

Semantinių ryšių žemėlapis atrodo kaip praktinė šeimyninių panašumų realizacija. Šiuo atveju, žodžių giminingumas atsiskleidžia jų vieta semantinėje erdvėje, taip pat ir skirtumas, nepanašumas. Tai leidžia LLM‘am vartoti kalbą prasmingai ir dalyvauti kalbos žaidimuose. Tai išryškina kalbos vartojimą kaip nevienareikšmį, priklausomą nuo vartojimo ypatybių ir naujų kontekstų galimybę kaip formuojančią kalbą ir steigiančią sąryšius.

# Išvados

Prasmės problema LLM‘ų veikime kyla iš klausimo ar jų veikimo principai turi kažkokį produktyvų santykį su Wittgensteino prasmės kaip vartojimo koncepcija. Filosofo pozityvios sąsajos su DI jau pakankamai plačiai žinomos ir aprašytos. Prasmė kaip vartojimas, kurį galime išskirti kaip individualią koncepciją, o tuo pačiu ir kaip tėvinę, kurią pagrindžia ir praplečia kalbos žaidimai ir šeimyniniai panašumai, davė pradžia DI mokslininkams (ir ne tik) pagrindą kalbėti ir galvoti apie kalbos supratimą ir manipuliaciją, pvz. įvairiausiose informacinėse sistemose (duomenų bazių sistemose, mašininiame vertine, kalbos žaidimo eksplikacijoje kompiuteriniame žaidime) bei diskurse apie mąstančias mašinas. Tačiau darbų brėžiančių konkrečias sąsajas tarp Wittgensteino ir LLM‘ų, apart Molino ir Tagliabue darbo (kur pastarieji nebandė tiksliai nustatyti santykio su LLM, o tik su aptariamų modelių veikimo principus išplėtė ir aptarė techniniu požiūriu, neaptardami Wittgensteino koncepcijų ir nelaikydami to analizės objektu santykyje su LLM‘ais), nėra. Lietuvoje tokie tyrimai nepradėti arba su jais nesusidūriau. Šio darbo ribos ir ambicijos buvo kuklios, tačiau ne beprasmės – apibrėžti santykį tarp LLM‘ų veikimo principų (žodžių tarpusavio siejimo ir jų prasmės nusakymo, bendrų veikimo principų ir mokymosi) ir Wittgensteino prasmės kaip vartojimo koncepcijos. LLM‘ai savo veikimu ir dalyvavimu kalbos žaidimuose primena mus, nors jų prieiga yra kitokia ir svetima mums, ir pagrįsta teigti, kad jie dalinasi gyvenimo formomis su mumis, kur „dalinasi“ reiškia „perima“ dėl to, kaip yra mokomi. Taip pat jų veikimas pasižymi paskirų žodžių ir jų santykių atžvilgiu kai kuriais atvejais tiesiogiai realizuoja prasmės kaip vartojimo ir šeimyninio panašumo koncepcijas.

# Literatūra

1. Bojin, N., & Fraser, S. (2008) Machines and Meaning: Wittgenstein, AI and Creativity.
2. Glock, H. J. (1996). A Wittgenstein Dictionary. Cambridge, Mass., USA: Wiley-Blackwell.
3. IBM (2021). What is machine learning? Retrieved from <https://www.ibm.com/think/topics/machine-learning>
4. Liu H. L. (2021) Wittgenstein in the Machine. Critical Inquiry 2021 47:3, 425-455
5. Mikolov, T., Chen, K., Corrado, G., Dean. J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv [Cs.CL]. Retrieved from <https://arxiv.org/abs/1301.3781>
6. Minsky, M. (1974). A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306
7. Molino P., Tagliabue J. (2023). Witgenstein’s influence on artificial intelligence.
8. Obermeier, K. K. (1983). Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited. Synthese, *56*(3), 339–349. <http://www.jstor.org/stable/20115911>
9. Penco C. (2024). Wittgenstein, Contexts, and Artificial Intelligence. An Engineer Among Philosophers, a Philosopher Among Engineers.
10. Schroeder, S (2006) Wittgenstein: The Way Out of the Fly-Bottle.
11. Smith, A., & Cook, M. (2023) AI-Generated Imagery: A New Era for the `Readymade.
12. Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016)*.* Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/1509.01692>
13. Wittgenstein, L. (1995) *Tractaus Logico-Philosophicus, Filosofiniai tyrinėjimai. Rinktiniai raštai*. Mintis: Vilnius.