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](#_Toc200656848)

[Wittgensteinas 4](#_Toc200656849)

[Reikšmė kaip vartojimas. 4](#_Toc200656850)

[Kalbos žaidimai. 5](#_Toc200656851)

[Šeimyninis panašumas. 6](#_Toc200656852)

[LLM veikimo principai 7](#_Toc200656853)

[Mokymosi duomenys. 7](#_Toc200656854)

[Dėmesio mechanizmas. 7](#_Toc200656855)

[Semantinių ryšių modeliavimas. 7](#_Toc200656856)

[GPT ir BERT algoritmai 8](#_Toc200656857)

[Sąsaja 9](#_Toc200656858)

[Kaip jų veikimas parodo „reikšmę kaip vartojimą”?  9](#_Toc200656859)

[Ar LLM‘ai dalyvauja kalbos žaidimuose? 9](#_Toc200656860)

[Kaip pasirodo šeimyninių panašumų koncepcija LLM‘ų veikime? 9](#_Toc200656861)

[Išvados 10](#_Toc200656862)

[Literatūra 11](#_Toc200656863)

# Įvadas

Wittgensteinas siūlė atsisakyti idėjos, kad žodžių reikšmės yra fiksuotos, vidinės ar loginės – ir vietoj to kalbėti apie reikšmę kaip vartojimą tam tikroje kalbinėje praktikoje. Tokios sąvokos kaip kalbos žaidimai, gyvenimo forma ir šeiminiai panašumai leidžia matyti, kaip kalba yra veikia, kurios prasmė priklauso nuo konteksto, funkcijos ir socialinės aplinkos.

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. O prie supratimo gryninimo neprisideda ekonominių interesų turintys individai, ypač iš to besipelnančios įmonės. Wittgensteino manymu, filosofija atsako į klausimus juos panaikindama (FT §109) . Todėl norėčiau būtent tam ir pasiryžti. Šio darbo tikslas parodyti ar prasmės kaip vartojimo koncepcija randama šiuolaikinių LLM, tokių kaip GPT ar BERT, veikime, o jei randama, tai koks yra santykis tarp jo ir prasmės kaip vartojimo sampratos, kurią siūlo vėlyvasis Wittgensteinas. Kokiu pagrindu ir kaip galime įmatyti vitgenšteinišką prasmės sampratą LLM‘ų veikime? Šio klausimo pagrįstumas kyla stebint šių technologijų tobulėjimo spartą ir įtikinančius rezultatus praktikoje – kadangi žmonės susišneka tarpusavyje, paliepia vienas kitam, tai į tokia padėtį papuolus LLM‘ams, žodžių prasmės kaip vartojimo supratimas tampa labai įtikinantis siekiant žinojimo. Susišnekėjimas ir, tam tikra prasme, bendradarbiavimas veda prie klausimo apie šių technologijų dalyvavimą kalbos žaidimuose bei tai, kaip pasirodo šeimyninių panašumų (*angl.* family resemblence) koncepcija LLM‘ų „supratimo” galimybėse. Šiame darbe nesiekiama atsakyti, ar šie dirbtinio intelekto modeliai „supranta”, „mąsto” ar „suvokia reikšmę” tokiu būdu, kaip tai daro žmonės. Mano manymu, tokie svarstymai neturint suvokimo struktūros, žodyno ar kalbos žaidimo taisyklių šiai temai yra pernelyg ankstyvi, spekuliaciniai. Vietoj to, tyrimas sutelktas į kalbos vartojimą ir tai, ar šie modeliai gali būti filosofiškai svarbūs, nagrinėjant reikšmės vartojimo kontekstus. Prieš žengiant prie šiuolaikinių technologijų analizės, būtina išsamiai panagrinėti paties Wittgensteino kalbos supratimą.

**IŠ KO/KAIP KYLA PRASMĖS PROBLEMA LLM‘Ų VEIKIME?**

# 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). Schroederis bei interpretuotojai pabrėžia, kad Wittgensteinas nekūrė naujos reikšmės teorijos, o veikiau bandė išsklaidyti filosofines painiavas, kylančias iš netinkamo kalbos supratimo (2006, 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) . Prasmė atsiranda iš kalbinio veiksmo.

Liu pripažįsta, jog Wittgensteino kalbos filosofija yra taip glaudžiai susijusi su semantiniais kompiuteriniais tinklais „kad nebegalime užmerkti akių prieš jos įsikūnijimą dirbtiniame intelekte“ (Liu 2021, p. 427). Kaip pavyzdį pagrįsti ji pamini, jog 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 (Molino et Tagliabue 2023, p. 1) tvirtinimu, Margaret 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 apibendrina Masterman poziciją, teigdama, kad „rakto į natūralios kalbos supratimą, taigi ir į tinkamą jos kodavimą kompiuteryje, reikia ieškoti semantiniuose tinkluose, kurie vieninteliai sugeba susidoroti su žodžių reikšmių daugialypiškumu ir neapibrėžtumu.“ (2021, 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, kuris nebūtų simbolinis, statiškas, bet semantinis.

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) aptardami Wittgensteino ir jo amžininko svarbaus mokslininko DI srityje Turingo įtampa 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) gaimybę: „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. 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, nors išvada, jog jie dalyvauja kartu su mumis kalbos žaidimuose yra intuityviai nemaloni ar atžari. Galbūt galėtume teigti tiek, kad LLM‘ai demonstruoja kaip kalba veikia per tarpusavio ryšius ir vartojimą.

## Dėmesio mechanizmas.

Dėmesio mechanizmas (*angl*. attention mechanism) pasirodė 2017 metais labai tinkamo pavadinimo straipsnyje „Attention is all you need“. Neuroniniai tinklai besinaudoją Dėmesiu vadinami transformeriais. Tai ne tik kalbos modeliai. Transformeriai naudojami šnekos atpažinime, nuotraukų generavime, kalbos vertime. Geriausiai Dėmesį apibūdina kalbinis pavyzdys: „jis ištraukė picą iš orkaitės. Ji labai skani“. Sakinys neatrodo problematiškas, tačiau kalbos modeliui nėra taip lengva nustatyti į ką nurodo įvardis „ji“ – gali nurodyti tiek į picą, tiek į orkaitę. Žinome, kad jokia orkaitė nėra skani, taip pat, kaip jokia pica nėra metalinė ir negali iškepti maisto. Dėmesio pagalba kalbos modelis apskaičiuoja ir nustato žodžių tarpusavio santykį.

## Semantinių ryšių modeliavimas.

Ne ką mažiau svarbus GKM veikimo principas – semantinis žemėlapis arba ryšių modeliavimas (Mikolov et al. 2013). Kaip minėta, teksto dalys įrašomos kaip skaitinės reikšmės, tiksliau, vektorių forma, pagal kuriuos koordinačių sistemoje žodžiui priskiriama vieta, kuri žymi žodžio semantiką arba prasmę. Kaip paaiškės, reikšmė yra nulinė be kitų žodžių. Erdvės, kurioje vietą užima žodis, dimensijos priklauso nuo modelio parametrų skaičiaus, tad kuo jų daugiau, tuo erdvė yra semantiškai gilesnė – žodžiai įgauna vis turtingesnę semantiką, nes vektoriai talpina daugiau skaitinių reikšmių. Tad panašūs žodžiai yra netoli vienas kito erdvėje. Pvz. žodis „karalius“ yra arti žodžio „vyras“, o „karalienė“ arti „moteris“. Atėmus „karaliaus“ vektorių iš „vyro“ vektoriaus ir pridėjus „moters“ vektorių gauname žodį „karalienė“ (Vylomova 2016). Tai įgalina paprastus samprotavimus remiantis linijine algebra (Molino et Tagliabue 2023, 2) Kalbos modeliai turi *n* skaičių parametrų, kurie sukuria *n*-dimensijų erdvę. Erdvė atstovauja kaip reikšmės žemėlapis, kuriame tarsi yra susikūrę miestai ir miesteliai – panašūs semantine reikšme žodžiai yra netoli vienas kito. 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“ (ibid., 1)

## GPT ir BERT algoritmai

Galiausiai, norėčiau trumpai parodyti minėtų veikimo principų pavyzdžius konkrečiuose GPT (*angl.*Generative Pre-trained Transformer) ir BERT (*angl.* Bidirectional Encoder Representations from Transformers) transformerių kalbos modelių algoritmuose. GPT modeliai augina tekstą po vieną simbolį, stengdamasis 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 gramatinį ir paprastai temiškai nuoseklų tekstą.“ (ibid., 3) Tuo tarpu BERT pasiima sakinį, paslepia atsitiktinius žodžius ir iš likusių bando nuspėti pasleptus. Kitaip tariant, matome jog iš esmės minėti algoritmai „spėlioja“ žodžius, kol po kurio laiko įgusta vartoti kalbą taip, jog tai būtų naudinga, pvz. gali susišnekėti su žmonėmis. Galėtume netgi teigti, jog kalbos modeliai nekuria vidinės semantinės esybės, bet demonstruoja, kaip kalba veikia per tarpusavio ryšius ir vartojimą.

# Sąsaja

## 

Parasyti apie tai kas filosofijoje siaip plaukioja ir tik po to daryti savo apibendrinimus ar teiginius, savo mintis.

Pristatykite žinomas problemą nagrinėjančias pozicijas. Kelkite sau uždavinį rasti klausimų, į kuriuos pristatoma koncepcija nepajėgi atsakyti, ar situacijų, kurių filosofas negalėtų paaiškinti, ar klausimų, kurie nekyla žvelgiant iš šios koncepcijos perspektyvos.

WITTGES TIESIOGINIA IKONCEPTU PRISTATYMAI

## Kalbos zaidimai – kaip kiti autoriai mato kad tai susije su ai – mano klaussimas ko nepaaiskian (ka galbut as galeciau paaiskint toliau)

TD AI KONCEPTAI

* KAIP KIEKVIENAS PAPILDO ARBA ATSAKO I KLAUSMA, YRA SUSIJE SU WITGE, KOKS KLAUSIMAS VISDELTO KYLA?

## Kaip jų veikimas parodo „reikšmę kaip vartojimą”?

## Ar LLM‘ai dalyvauja kalbos žaidimuose?

## Kaip pasirodo šeimyninių panašumų koncepcija LLM‘ų veikime?

# Išvados

# Literatūra

1. Glock, H. J. (1996). A Wittgenstein Dictionary. Cambridge, Mass., USA: Wiley-Blackwell.
2. IBM (2021). What is machine learning? Retrieved from <https://www.ibm.com/think/topics/machine-learning>
3. 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>
4. Minsky, M. (1974). A Framework for Representing Knowledge. MIT-AI Laboratory Memo 306
5. Molino P., Tagliabue J. (2023). Witgenstein’s influence on artificial intelligence.
6. 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>
7. Penco C. (2024). Wittgenstein, Contexts, and Artificial Intelligence. An Engineer Among Philosophers, a Philosopher Among Engineers.
8. Schroeder, S (2006) Wittgenstein: The Way Out of the Fly-Bottle.
9. 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>
10. Wittgenstein, L. (1995) *Tractaus Logico-Philosophicus, Filosofiniai tyrinėjimai. Rinktiniai raštai*. Mintis: Vilnius.
11. Liu H. L. (2021) Wittgenstein in the Machine. Critical Inquiry 2021 47:3, 425-455
12. Bojin, N., & Fraser, S. (2008) Machines and Meaning: Wittgenstein, AI and Creativity.
13. Smith, A., & Cook, M. (2023) AI-Generated Imagery: A New Era for the `Readymade.