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 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.

# 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, 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 interpretuotojai pabrėžia, kad Wittgensteinas nekūrė naujos reikšmės teorijos, o veikiau bandė išsklaidyti filosofines painiavas, kylančias iš netinkamo kalbos supratimo(2006, 170). Kaip rašo Wittgensteinas, „Filosofija yra kova prieš mūsų proto pakerėjimą mūsų kalbos priemonėmis” (FT §109). 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., (2006, 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 (Schroeder 2006, 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. Tai rodo, kad kalbos supratimas nėra teorinių žinių turėjimas, bet gebėjimas dalyvauti gyvenimo formose, veikti. Reikšmė gimsta ne iš apibrėžimų, bet iš vartojimo – iš to, kaip žodžiai veikia mūsų gyvenimo situacijose, kontekstuose.

## Kalbos žaidimai.

Kalbos žaidimų sąvoka (*angl.* language-games) yra neatskiriama nuo reikšmės kaip vartojimo sąvokos. Šie du elementai nėra atsitiktinai susiję – kalbos žaidimai yra būtent ta terpė, kurioje vyksta kalbos vartojimas, o kartu ir kuriama reikšmė, kadangi supratimas priklauso nuo ištaros funkcijos tam tikrame kontekste, tad reikšmė priklauso nuo vartojimo kažkokiame kontekste (Obermeier 1983, 341) Jei reikšmė yra tai, kaip žodis vartojamas kalboje, tai neišvengiamai kyla klausimas: kokiuose kontekstuose ir pagal kokias taisykles tie žodžiai vartojami? Būtent į šį klausimą atsako kalbos žaidimų samprata. Kalbos žaidimas yra tam tikra veikimo forma, kurioje kalbiniai ženklai įgyja prasmę, nes jie yra naudojami pagal tam tikras „žaidimo” taisykles – konvencijas, tikslus, socialinius lūkesčius

Kalba nėra atskira nuo mūsų kasdienybės: kiekvienas jos panaudojimo atvejis yra tam tikros socialinės veiklos dalis – 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, todėl ir žodžių reikšmės skiriasi ne dėl to, kad žodžiai būtų iš esmės kitokie, bet todėl, kad jie įtraukti į skirtingas žaidimų struktūras. Tokiu būdu reikšmė kyla iš funkcinės, ne loginės sistemos, o žodžių supratimas tampa priklausomas nuo mūsų gebėjimo orientuotis konkrečiose vartojimo situacijose.

Kalbos žaidimai yra būdas stebėti reikšmės susidarymą praktikoje. Žodžio prasmę įmanoma suprasti tik dalyvaujant tame žaidime, kuriame tas žodis naudojamas. Tai reiškia, kad tiek kalbos taisyklių mokymasis, tiek kalbinis supratimas yra įsitraukimas į socialinį veiksmą, kurio taisyklės dažnai nėra aiškiai išsakytos, bet yra perimamos per praktiką. „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.“ (Schroeder 2006, 169).

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, 284)

Kalbos žaidimai sukuria ir apibrėžia reikšmės galimybės erdvę – kada tam tikras žodis ar posakis turi prasmę, o kada jis netenka savo veikimo galios. Pvz., „argumentas” filosofijoje reiškia loginę struktūrą, o kasdienybėje – nesutarimą. Reikšmė šiuose žaidimuose skirtinga, bet ne klaidinga – jos prasmingos savo žaidimo kontekste. Apibendrinant galime tarti, jog be žaidimo nėra vartojimo, o be vartojimo – reikšmės.

## 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? Kodėl mes vis dar kalbame apie „kalbą” ar „reikšmę”, jei jos pasireiškia tokiais skirtingais būdais? Šeimyninis panašumas atskleidžia, kad ryšiai tarp kalbos formų yra ne loginiai ar esminiai, bet panašūs į šeimos narių bruožų panašumus.

Wittgensteinas šią 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.), kaip šeimos nariai gali turėti panašų nosies pavidalą, akių spalvą ar ūgį, bet nė vieno bruožo, kuris būtų bendras visiems. Tarsi žinome, kas yra „kėdė“ ar „žaidimas“, tačiau negalime apibrėžti logiškai kaip elemento, priklausančio tam tikra klasių hierarchijai ar kokia formalia taisykle (Minsky 1974, 51).

Svarbu tai, kad šeimyninis panašumas nėra kompromisas ar nepilnas apibrėžimas. Pasak Glocko (Glock 1996, 120), 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“. 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.“ (2024, 283) Kitaip tariant, šeimyniniai panašumai rodo, kad kalba yra praktinė – ji kyla iš mūsų gebėjimo atpažinti ir dalyvauti panašiose veiklos formose, kurios formuoja kalbą. Juk 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) arba rodo į objektus, priklijuoja jiems vardą ar „kartoja mokytojo pasakytus žodžius“ (FT §7)

# LLM veikimo principai

## Mokymosi duomenys.

Dirbtinio intelekto 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.

## Dėmesio mechanizmas.

Dėmesio mechanizmas (*angl*. Attention mechanism, Dėmesys) 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

## 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.