KLAUSIMAI

**Ar yra pagrindo teigti, jog Wittgenstein’o prasmės kaip vartojimo koncepcija yra susijusi su Generatyvių didžiųjų kalbos modelių (*angl*. LLM) veikimu?**

Kuria prasme vitgenšteiniška prasmės koncepcija padėda suprasti “dėmėsio” mechanizmą (arba “word2vec” algoritmą)?

Kokiais būdais Wittgenstein’o konteksto svarbos supratimas paveikė Natūralios kalbos apdorojimo (NLP) ir LLM’ų vystymąsį?

Wittgenstein’as kritikavo referentinę prasmės teoriją, pagal kurią žodžiai atitinka objektus. Kaip galėtume apibrėžti LLM’ų kalbos vartojimą ir jų santykį su kalba, kurios vartojimo išmokime nėra virš-lingvistinio santykio su tikrove – vien tik lingvistinis pagrindas?

Galime susikalbėti su LLM’ais. Ar yra pagrindo manyti, jog “žmogiškoji” (šiuo atveju, vitgenšteiniškoji) p rasmės koncepcija skiriasi nuo LLM’o prasmės koncepcijos?

Kaip kalbos žaidimų koncepcija atsiskleidžia LLM‘ų mokymo procese, veikimo skirtingose intelekto taikymo srityse (pvz. matematikos uždavinių sprendimas, testų išlaikymas ir kt.)?

PAVADINIMAS

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

**🔍 Pagrindinės kryptys tavo mintyse:**

1. **Kalbos reikšmė pagal Wittgensteiną:**
   * Kalbos žodžių reikšmė kyla iš **jų naudojimo** kontekste.
   * Žodžių reikšmės nėra „esencinės“, jos nėra fiksuotos, bet priklauso nuo „žaidimo“, kuriame dalyvaujama (language games).
   * **Šeimyninis panašumas**: ne visi objektai/reiškiniai turi vieną bendrą savybę, bet panašumų tinklas juos jungia.
2. **LLM (GPT, BERT ir pan.) veikimo principai:**
   * Jie mokomi iš **milžiniško kiekio kalbos duomenų**, suprasdami **statistinius ryšius** tarp žodžių.
   * Dėmesio mechanizmas (attention) leidžia modeliams **atsižvelgti į kontekstą**, kad „suprastų“, ką reiškia žodis.
   * LLM’ai veikia tarsi „neformaliai“ mokosi kalbos žaidimų be aiškios semantinės teorijos – bet praktikoje veikia puikiai.
3. **Filosofo požiūris:**
   * Kaip šių modelių veikimas **įkūnija Wittgensteinišką požiūrį** į kalbą?
   * Ar tai **paneigia, patvirtina, ar papildomai paaiškina** jo mintis?
   * Galima net klausti: **ar LLM’ai dalyvauja kalbos žaidimuose?** Ar jie tik imituoja žaidimą?

?

* Ar galima sakyti, kad LLM'ai „supranta“ žodžių reikšmę, jei tai priklauso nuo konteksto ir vartojimo?
* Ar „šeimyninis panašumas“ gali būti paaiškinamas semantiškai – ar LLM'ai tai „atpažįsta“?
* Ar tai, kad modeliai gali daryti semantines analogijas (king : queen :: man : woman), rodo kažkokią naują reikšmės sampratą?
* Ar Wittgensteinas būtų sakęs, kad LLM’as supranta, ar tik imituoja supratimą?

**🧩 Galima struktūros pradžia tavo darbui:**

**1. Įvadas:**

* Tema ir jos aktualumas (AI ir filosofija).
  + „Šiame darbe nesiekiama atsakyti, ar dirbtinio intelekto modeliai „mąsto“ ar „suvokia reikšmę“ tokiu būdu, kaip tai daro žmonės. Vietoj to, tyrimas sutelktas į kalbos vartojimo struktūras ir tai, ar šie modeliai gali būti filosofiškai svarbūs, nagrinėjant reikšmės vartojimo kontekstus.“
* Kodėl pasirinkai Wittgensteiną?
  + Tai kartu ir apsprendžia nuostata temos atžvilgiu, tam tikrą imitaciją ar nusistatymą teorijų kaip objektų atžvilgiu, filosofijos kaip *atsakančios* į klausimus (greičiau ji “jas panaikina ir taip jas išsprendžia” kaip kad pasakytų Wittgensteinas)
* Tikslas: parodyti, kaip LLM'ai praktikoje atitinka Wittgensteino kalbos vartojimo teoriją. 🡪 Tikslas: patyrinėti santykį tarp Wittgensteino minčių santykio su generatyviniais (didžiaisiais) kalbos modeliais

**2. Wittgensteino teorija trumpai:**

* Reikšmė kaip vartojimas. 🡪 pradinė ir pirminė sąvoka, kuri duoda starta betkokiam mąstymui ar prielaidavimui apie šių dviejų dalykų ryšį. Jį papildo vėlesni dėmenys (kalbos žaidimai, šieminynis panašumas).
* Kalbos žaidimai. 🡪 specializuoti kalbos modeliai, geresnis suvokimas kontekste, kalbos žaidimai kaip semantinės erdvės vietos, kampučiai, besiskiriantys savo apimtimi, gilmumu (pvz. Skiriasi paliepimų kalbos žaidimas nuo kvantinės mechanikos kalbos žaidimo, fizikos ar filosofijos kalbos žaidimų).
* Šeimyninis panašumas. ---🡪 Semantinė erdvė ir jos analizė, savybės tarsi atskleidžia kaip žodžiai atskirai sukuria tam tikrus ryšius, jie išreiškia tam tikrą semantinę aritmetiką, kuri tarsi leidžia spėti žodžius or something like it.

**3. LLM veikimo principai:**

* Mokymosi duomenys.
  + Learning (FT 2)
* Dėmesio mechanizmas.
* Semantinių ryšių modeliavimas.
* GPT vs Bert? Tiesiog parodymui
  + Pirmiausia verta trumpai pažvelgti, kaip veikia transformerių architektūra „iš viršaus“: transformerių modeliai, tokie kaip GPT ar BERT, remiasi didelės apimties sekų (žodžių ar simbolių) duomenimis ir mokosi atpažinti konteksto priklausomybes per daugiasluoksnius uždėjimus (angl. „self-attention“) mechanizmus. Jie neprogramuojami konkrečiomis griežtomis kalbos taisyklėmis, o „moko“ žodžių reikšmes per jų vienu metu pamatomą kontekstą – kas, griežtai tariant, yra savaime „vartybos“ pavyzdys: žodžio prasmė kyla iš jo 20/28 santykio su kitais žodžiais toje pačioje sakinio ar pastraipos „žaidimo erdvėje“. Jei pažvelgsime į tai filosofiškai, matysime, kad transformerių „supratimas“ yra formos, o ne esmės lygmens: jie nekuria vidinės semantinės esybės, bet demonstruoja, kaip kalba veikia per tarpusavio ryšius ir vartojimą – o tai atitinka Wittgensteino idėją, kad „žodžio reikšmė yra jo vartojimas kalboje“.
  + Dabar, norint lyginti du konkrečius modelius (pavyzdžiui, GPT, kuris yra autoregresinis, ir BERT, kuris naudoja maskavimo strategiją), galime atkreipti dėmesį, kaip jų priešmokymai (angl. „pretraining“) skiriasi: GPT serija augina teksto sekas po vieną simbolį, stengdamasi nuspėti kitą žodį, ir taip treniruojasi „būti tolesniu sakinio tęsiniu“; tuo tarpu BERT ištekdavo ištisus žodžius (ar jų porcijas), juos „užmaskuoja“ ir bando juos sėkmingai „atspėti“, skatindamas modelį labiau „suprasti“ tiek kairę, tiek dešinę konteksto pusę. Filosofiškai tai galima interpretuoti kaip du skirtingus būdus „žaisti kalbos žaidimą“: vienas – tarsi klausimas „kas toliau?“, kitas – „ką skamba šio konteksto dalys?“. Abi strategijos rodo, kad „reikšmė“ LLM čia yra tarpusavio ryšių funkcija, o ne loginė struktūra, ir jos tarpusavio palyginimas gali padėti atskleisti, kaip skirtingi „žaidimo taisyklių“ pasirinkimai formuoja skirtingas „supratimo“ niuansų versijas – nors nė vienas modelis savaime neturi žmogaus patirties ar pasaulio įsijautimo, jų veikimo skirtumai parodo, kokius kalbos vartojimo aspektus jie akcentuoja

**4. Palyginimas:**

* Ar LLM‘ai dalyvauja kalbos žaidimuose?
  + Kai kalbos modelis generuoja atsakymą, jis ne **interpretuoja pasaulį**, bet **dalyvauja kalbinėje veikloje**, kurios forma yra panaši į žmonių kalbos žaidimą (pokalbis, paaiškinimas, klausimas ir pan.). Čia galima kelti klausimus:
    - Ar LLM sukuria reikšmingą *panašumą* į kalbos žaidimą?
    - Kokią reikšmės formą galima jam priskirti, jei ji remiasi tik vartojimu be intencijos?
    - Ar galima kalbėti apie *pragmatiškai veikiančią* reikšmę be sąmoningo reikšmės suvokimo?
    - *„Žodžio reikšmė yra jo vartojimas kalboje.“*  
      *(Philosophical Investigations, §43)*
* Kaip jų veikimas parodo „reikšmę kaip vartojimą“?
  + Tu jau minėjai šį momentą. Tai labai geras aspektas – galima parodyti, kad net be tikslių apibrėžimų, kalbos modeliai „mokosi“ struktūrų, kurios atitinka **panašumo tinklus** tarp žodžių ar sąvokų. Pvz.:
    - „karalienė“ yra panaši į „karalių“ tam tikrame semantiniame lauke (gender, power, monarchy).
    - LLM modeliai išmoko šių panašumų ne kaip taisyklių, bet kaip **panašumo per vartojimą**. Trumpai galima užsiminti apie ankstyvus simbolinės logikos bandymus atgaivinti ar prikelti robotus tarsi zombius, kurie remiasi taisyklėmis, kad taptų protingi,bet būna zr dumb.
    - Tai galima interpretuoti kaip *funkcinį analogą* šeiminių panašumų logikai, kurią aprašo Wittgensteinas.
* Kur kyla ribos (LLM vs žmogus)?

**5. Kritinė refleksija:**

* Ką tai sako apie kalbos prigimtį?
* Ar LLM’ai padeda suprasti Wittgensteiną geriau ar jį paneigia? O kaip atvirkščiai?

**6. Išvados:**

* Ką galima teigti apie kalbą, reikšmę ir DI?
* Ka galima teigti apie filosofijos statusą, kaip praktinio užsiėmimo ir praktines pasėkmes turinčio mokslo.

**Kalba kaip gyvenimo forma**

Galiausiai Wittgensteinas įveda sąvoką *forms of life* – gyvenimo formos, kurios yra kontekstas, kuriame atsiranda kalbos žaidimai. Kalba egzistuoja tik tam tikrame socialiniame ir kultūriniame audinyje. Tai reiškia, kad kalba nėra nepriklausomas reiškinys – ji susijusi su žmonių veikla, tikėjimais, tradicijomis ir institucijomis.

*“To imagine a language is to imagine a form of life.”*  
(*Philosophical Investigations*, §19)

**=== intersting thought kurios galbut dar neitrauksiu į kursinį.**

**Įvadas (juodraštis)**

Dirbtinio intelekto pažanga per pastaruosius dešimtmečius atvėrė naujas galimybes ne tik technologijose, bet ir filosofijoje. Šiame darbe tyrinėjama sąsaja tarp kalbos vartojimo sampratos, kurią siūlo vėlyvasis Ludwigas Wittgensteinas, ir šiuolaikinių didžiųjų kalbos modelių (*large language models*, LLM), tokių kaip GPT ar BERT, veikimo principų. Šie modeliai, remdamiesi milžiniškais kalbos duomenų kiekiais, geba generuoti tekstą, kuris neretai atrodo prasmingas, tinkamas kontekstui ir funkciškai panašus į žmogaus kalbą.

~~Tačiau ar šių modelių gebėjimas naudoti kalbą reiškia, kad jie~~ *~~supranta~~* ~~reikšmes?~~ Ar jie dalyvauja kalbos žaidimuose, kaip tai suprato Wittgensteinas~~? Ir ar jų generuojama kalba grindžiama tam tikra „gyvenimo forma“, ar tai tik~~ statistinis ~~paviršius be sąsajos su realia praktika?~~

Šie klausimai veda į platesnį filosofinį tyrimą apie reikšmės pobūdį, kalbos vartojimą ir semantinių struktūrų atpažinimą. Vėlyvojo Wittgensteino mintys apie reikšmę kaip vartojimą, kalbos žaidimus ir šeimyninius panašumus gali pasiūlyti konceptualų pagrindą, padedantį suprasti, ką reiškia kalbinis „veikimas“ tiek žmogaus, tiek dirbtinio intelekto atveju. Kitaip tariant, šiame darbe siekiama išnagrinėti, ar šiuolaikiniai kalbos modeliai veikia taip, tarsi jie būtų įtraukti į tam tikrus kalbos žaidimus, ir ar jų kalbinė kompetencija rodo gilesnį reikšmės apdorojimą, ar tik jos paviršinę imitaciją.

Prieš žengiant prie šiuolaikinių technologijų analizės, būtina išsamiai panagrinėti paties Wittgensteino kalbos teoriją, jos pagrindines sąvokas ir jų tarpusavio ryšius.

Šiame darbe siekiama tyrinėti tam tikras sąsajas tarp Ludwig Wittgensteino vėlyvosios kalbos reikšmės teorijos ir šiuolaikinių didžiųjų kalbos modelių (LLM), tokių kaip GPT. Mano tikslas nėra pateikti galutinio atsakymo į klausimą, ar šie modeliai „supranta“ kalbą, ar gali būti laikomi kalbos žaidimų dalyviais, kaip tai suprastų Wittgensteinas. Atvirkščiai, mano intencija – atverti šias sąsajas kaip filosofinius klausimus, kuriuos verta tyrinėti, o ne spręsti.

LLM modeliai, nors neturintys sąmonės ar intencionalumo, geba generuoti kalbą, kuri atrodo tinkama kontekstui ir semantiškai taikli. Tai kelia klausimą: ar jų elgsena yra tik imitacija, ar čia veikia tam tikros gilesnės struktūros, kurios, nors ir ne žmoginės prigimties, vis dėlto įdomiai rezonuoja su Wittgensteiniškomis įžvalgomis apie reikšmę, kontekstą, kalbos vartojimą, šeiminius panašumus?

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 veikla, kurios prasmė priklauso nuo konteksto, funkcijos ir socialinės aplinkos**. Tuo pačiu leidžia geriau suvokti ir apsibrėžti generatyvinius kalbos modelius, kurie dažnai gali pasirodyti apsupti migla, išpūsti technokratų ar ekonomikų.

* The "fly in a bottle" is Wittgenstein's famous metaphor - philosophy creates problems by trapping us in conceptual confusions, like a fly buzzing around inside a bottle. The philosopher's job is to "show the fly the way out."

**AI yra tas butelys, iš kurio tarsi reikia išssivaduoti išsiaiškinant tą santykį tarp AI ir Wittgenstein, o tuo paciu tarp kalbos, tarp kulturinio ir populiaraus objekto, į kurį dar nežino atsakymų, kurio panaudojimas dar nėra savaime suprantamas.**

* Tai labai stipri mintis. Iš tikro, tu čia formuoji poziciją, kuri leidžia *filosofiškai* kalbėti apie kalbos modelių klausimus neprivalant įsipareigoti nei vienai ontologinei ar metafizinei pozicijai apie „sąmonę“, „mastymą“ ar „intencionalumą“. Tu darai tai, ką galima vadinti epistemologiniu perorientavimu į vartojimą, ir tai yra labai *Wittgensteiniška* dvasia.

**Vietoj to, kad siektume atsakyti į klausimą „Ar dirbtinis intelektas mąsto?“, arba „Ar LLM turi intencionalumą?“, šiame tyrime žiūrima į kalbos naudojimą. Toks požiūris nėra neutralus: jis remiasi vėlyvojo Wittgensteino filosofija, kurią galima suvokti kaip atsisakymą metafizinių spekuliacijų apie kognityvines būkles, ir vietoj to siūlantį kalbos reikšmę aiškinti per praktikas.  
 „Filosofiniai klausimai išsprendžiami ne atradimais, bet parodant, kaip kalba veikia.“ (Philosophical Investigations, §109)**

* Kalbėti apie tai, *kaip* LLM vartoja kalbą, ir kokiose kalbos žaidimų struktūrose jis funkcionuoja, yra **filosofiškai pagrįstas tyrimo objektas**, net jeigu (ar būtent todėl kad) negalime atsakyti į klausimą apie jo „vidinį protą“.

**Tada galim nesivolioti ir negundyti savo proto miglotai ar vargiai apibrėžtomis sąvokomis kaip “suvokimas”, “supratimas” ir t.t.. Kita vertus, tai *kažkokią* poziciją implikuoja, mano manymu, filosofijos kaip terapijos ar tiesai neįpareigojančios veiklos statusą, apie kurį pasisakė ir Wittgensteinas.**

Šis darbas yra bandymas pažvelgti į tai, kaip šios Wittgensteino sąvokos gali būti panaudotos ne tiek tam, kad paaiškintų, pateisintų ar nuteistų dirbtinį intelektą, kiek tam, kad giliau suprastume, *ką* iš tikrųjų šie modeliai daro ir *kaip* jų veikimą galima (ir reikia) mąstyti filosofiniu požiūriu.