ЛОГІКА ТА МЕТОДОЛОГІЯ НАУКИ

DOI: 10.35423/2078-8142.2024.2.1.3
УДК 165:11

О. Л. Маєвський,
кандидат філософських наук,
молодший науковий співробітник
відділу логіки та методології науки
Інституту філософії імені Г. С. Сковороди НАН України,
м. Київ, Україна
e-mail: mayevsky@nas.gov.ua
ORCID: https://orcid.org/0000-0001-6063-6033

ЕПІСТЕМІЧНА ОБМЕЖЕНІСТЬ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Метою роботи є філософська експозиція концептуальних меж та відповідної їм епістемічної цінності застосувань діалогових великих мовних моделей (ВММ) архітектури Transformer як раціональних [спів-]агентів у контексті практичної пізнавально опосередкованої діяльності людини. Для цього у роботі виконуються такі дослідницькі завдання: 1) філософський огляд основоположних інтуїцій, загальної природи і функціональної механіки передових реалізацій ВММ; 2) визначення меж інструментальної епістемічної цінності ВММ як продуктів специфічного ланцюга процедур машинного навчання, що включає безпосередню експертну участь людей. Практичне значення дослідження полягає в інтуїтивно-наочній кристалізації гуманітарного розуміння ВММ через теоретико-філософську експлікацію їхньої природи й улаштування, що спирається на такі нові результати дослідження: ВММ представлено як функціоналістський механістичний проєкт статистичного

© Маєвський О. Л., 2024

моделювання мови і мовлення як моделі знання як смислової моделі дійсності – 1) модель 2) моделі 3) моделі дійсності; показано, що через значну дистанцію опосередкування дійсності внутрішньомодельні зв'язки втрачають свій фактологічний потенціал; продемонстровано, що ВММ є продуктом машинного навчання певній мовній поведінці з метою і цінностями, кардинально відмінними від мети і цінності людського пізнання, що робить принципово сумнівною виправданість їх застосування як компонентів будь-яких систем підтримки прийняття високоризикованих рішень.

Ключові слова: *велика мовна модель, епістемічна цінність, епістемічне значення, архітектура Transformer, філософія техніки, філософія штучного інтелекту, безпечність штучного інтелекту, доцільність штучного інтелекту.*

Усім відомі реалізації діалогових генеративних великих мовних моделей (ВММ) *справляють враження*. І у цьому – їхнє головне, конструктивно обумовлене завдання.

Принцип породження і суть ВММ – *ефективна імітація* (parroting). У цьому не так просто, але усе ж таки можна переконатись, самостійно розглянувши логічну механіку їх породження і наступної функціональної експлуатації на прикладі опису панівної на сьогоднішній день їх архітектури Transformer [6; 3]. (Докладніше про це див.: [7] та новіших роботах автора.)

В основі мовної компетентності будь-якої ВММ лежить оптимізований *словник* елементарних символів – «*токенів*» (деяких фрагментів слів – алфавіт плюс обмежений набір найчастотніших символічних сполучень), що не мають однозначно (частіше – жодним зрозумілим чином) інтерпретовного самостійного смислу і значення. Тобто генеративна ВММ сприймає, зберігає і відображає (моделює) виключно внутрішню імовірнісну логіку *символічних послідовностей* об'єктної мови у термінах власного словника. Ці моделі не

мають ні очевидної семантики, ні прагматики поза межами свого внутрішньо замкненого синтаксичного імперіалізму. Тому все, що може *нагадувати* смисл і значення у виконанні ВММ, виявляється (і оголошується) «*емерджентним*» – таким, що не може бути ефективно редукованим до її елементарних структур і властивостей.

Вирази мови у ВММ у загальному випадку – лише *послідовності* певних елементарних символьних «*токенів*». Генеративні ВММ – це однорангові моделі символічних послідовностей (*sequence-to-sequence models*) взагалі, у термінах словника, який не має обмежень на простоту і не мусить складатися з того, що ми загалом інтуїтивно розуміємо під «словами» як лексичними одиницями «звичайної» мови. Тобто, генеративний підхід Хомського тут знімається на користь редукції до чистоти елементарніших структур, без збереження якого там не було б оформлювального смислового зв'язку між тим, що надалі *повторно виникає* (*re-emerges*, виключно для нас і у нашому власному сприйнятті, не для ВММ) «емерджентно» (як слова, речення, тексти) на деякому *макрорівні*, і тим елементарним, із чого це останнє об'єктивно складене на *мікрорівні*.

ВММ експлуатують ту обставину, що якщо у мові щось *виглядає* як дещо, то воно і *означає* це дещо – принаймні в контексті усіх можливих інтерпретацій самої цієї мови, де самі ці інтерпретації достатньо лише розглядати як деякий невизначений функціонал. ВММ задумані і є принципово агностичними щодо своїх інтерпретацій і прагматики. І тому для них і слово, і речення, і текст (разом з усіма непростими процесами їх породження), власне, нічим не відрізняються від алгоритму написання деякої послідовності елементарних словникових складових. Якщо для людини мовлення – це *відповідальний осмислений акт*, то для ВММ це – просто зумовлена, *детерміністична дія з приблизного відтворення (відобра-*

ження) раніше почутого (із привнесеними елементами невизначеності – трюк для різноманітності).

Діалогові ВММ відтворюють «підслухані» на етапі навчання «діалоги» (деяку підмножину навчального корпусу символічних послідовностей), керуючись цільовою («об'єктивною») функцією задоволення оператора ВММ [5; 2]. Власне, діалогові ВММ – і це слід розуміти і пам'ятати кристально чітко – не мають за свою конструктивну мету нічого, крім проходження (виключно на етапі навчання) деякого доволі умовного, ресурсно обмеженого й принципово неповного епістемічного «тесту Тюринга» на ефективне уведення в оману деякої «приймальної комісії», у підсумку очолюваної конкретною людиною, рішенням якої навчання ВММ припиняється і вона уводиться до експлуатації [4].

Отже, завданням ВММ насправді є успішне уведення нас у *самооману*. Зокрема і щодо *епістемічного значення* і *цінності* маніфестованих ними виразів. Останнє являє собою як одне з потенційно найбільш корисних застосувань ВММ, так і, водночас, одне з найнебезпечніших – тоді, коли очікувана нами прагматична цінність інтерпретованих нами виразів ВММ значно перевищує виявлене чи припущене нами їх епістемічне значення.

У цій статті ми спробуємо дати загальний філософський нарис деяких логічних і прагматичних факторів принципової епістемічної обмеженості ВММ для побудови адекватних інтуїцій у частині інструментальних очікувань від них.

ВММ архітектури Transformer як така (без зовнішніх гетерогенних органічних розширень, про які йтиметься далі) є наближеною до відтворення навчального корпусу даних суперфункцією над «словами» (токенами) у послідовності – обробником слів (word processor). В оману щодо нього нас уводить, передусім, його складна поведінка, яку нам складно (і

практично неможливо) наївно передбачити як продукт деякої механіки.

Глибоко укорінене у філософії й гуманітаристиці взагалі переконання у привілейованій локалізації «розумності» в людині і подібних до неї (з різним ступенем абстрактності) її богах, при спостереженні мовленнєвої поведінки, *настільки* (як деякий поріг у деякому емпіричному тесті) подібної до людської, призводить до *абдуктивного* метависновку про наявність у ВММ відповідних такій поведінці атрибутів.

Гіпотетична *пояснювальна атрибуція* ВММ інтелектуальних здатностей, що традиційно визнаються властивими людині (як у гуманітарних науках, так і у повсякденному знанні), призводить, у такий спосіб, до метависновку про те, що емпірично спостережувана мовленнєва поведінка ВММ *певною мірою* (не абсолютно) *не суперечить* (або послідовно відповідає, is consistent with) зазначеній гіпотезі.

Водночас, індуктивні і дедуктивні метависновки щодо ВММ стикаються з майже нездоланними технічними труднощами своєї реалізації через надто великі розміри і складність ВММ як об'єкта логічного аналізу, наслідком чого також стає й практична неможливість ефективної інтерпретації внутрішніх логічних структур ВММ та їхньої функціональної динаміки на макрорівні, у термінах прогнозованої поведінки.

Таким чином, у нас немає інструментів для перевірки висновків ВММ як таких, на основі аналізу їхньої мікромеханіки. А отже, логічна послідовність і пов'язана з нею цінність таких висновків завжди лишатиметься під питанням доречності їхньої імовірнісної інтерпретації (яка також є лише інтерпретацією і видимою формою таких висновків, а не імовірністю як такою).

ВММ як такі не мають безпосереднього доступу до фактів світу, їх моделі (world model) і зворотного зв'язку з ними – це *моделі мови*, про світ і не тільки. Тому ВММ на етапі екс-

плуатації (висновування), взагалі кажучи, «бачить» і «чує» лише себе: у її контексті немає інструментів для розрізнення між собою і світом, між собою і своїм співбесідником.

Діалогові ВММ можуть містити маркування для позначення (кодування) співбесідника у контексті, що дає можливість розпізнавати його внесок у контекст за місцем і асоціаціями як певний патерн, притаманний тренувальним діалогам ВММ взагалі. Однак при цьому ці діалоги все одно залишаються для моделі лише текстами про діалоги, зображеннями (описами, дескрипціями) діалогів як деяких символічних послідовностей, що містять деякий узагальнений патерн такої «діалогічності».

У контексті від «співбесідника» залишається присутнім лише і виключно те, що було виражено текстом вводу у його виконанні. ВММ (без зовнішніх розширень) не може будувати припущень чи узагальнень щодо «співбесідника», виходячи з ширшого контексту дійсності, зазвичай доступного людині через її включеність у життя. ВММ має виключно ту інформацію, яка є частиною її контексту, незалежно від способу і порядку її отримання.

Авторегресивний характер генерації вихідного контексту ВММ також призводить до того, що якщо до контексту і включаються фрагменти (такі, як зовнішній ввід), які не були згенеровані самою ВММ, то ці фрагменти для ВММ не розпізнаються як «чужі». ВММ типу Transformer у режимі генерації має справу з контекстом як з деякою знеособленою даністю свого внутрішнього, феноменального «досвіду» у цілому – не як з історією: для моделі тут, для кожного слова, існує лише пара «все, що до, разом» (внутрішньо пов’язане і так зважене механізмом «уваги») і «наступне слово» (як розподіл за словником).

Тому, якщо продукт генерації ВММ і має певну епістемічну *цінність*, то вона полягатиме, передусім, у формулю-

ванні текстів, які могли б бути нами проінтерпретовані як такі, що мають *гіпотетичне значення* у зв'язку з нашим (зовнішнім для ВММ) вводом фрагментів контексту. При цьому інтерпретація продуктів генерації як осмислених гіпотез (як і можливість її) лишаються метазавданням відносно самої ВММ, яке покладається споживачем таких продуктів (нами) на самого себе.

Простою мовою взаємовідносини з ВММ можна було б описати як намагання зварити «кашу із сокири» за допомогою «системи ОБС» (нар., тж. «одна баба сказала…»). Тобто, практично, ми *отримуємо у «відповідь» на втручання* (яким і є наш зовнішній ввід) у генерацію *деяку проєкцію цієї відповіді* у латентному внутрішньому просторі ВММ плюс асоційовану за мірою «близькості» наступну генерацію, зважену механізмом «уваги» з урахуванням нашого вводу й усього попереднього досвіду ВММ.

Міра структурної подібності нашого вводу до узагальненої внутрішньої картини (представлення) ВММ дає можливість сподіватися, що наше втручання (яким є кожен наш «запит», prompt) не призведе до суттєвих коливань у загальному «смисловому» розподілі у внутрішньому векторному просторі моделі. Цей розподіл також *імпліцитно враховує* й патерни того, що можна назвати наближеним граматичним виводом (grammar induction) у результаті машинного навчання. Тому те, що є достатньо схожим на граматично правильні форми, може також підтримувати й логічно правильні форми у граматично правильних виразах.

Форми внутрішньої логічної послідовності (послідовність, consistency; зв'язність, coherence) так само імпліцитно враховуються під час навчання моделі на зразках, які ці форми містять. Найімовірніше, йтиметься про найбільш репрезентовані у навчальному корпусі такі форми. Й очікувати на здатність ВММ до використання некласичних логік, немоно-

тонних міркувань і абсурдистських установок (без спеціального донавчання моделі для таких спеціальних кейсів) тут не доводиться.

Факти ж для ВММ, таким чином, також виявляються як аналітичні структурні дескрипції (чогось) лише у нашій, зовнішній інтерпретації їх функції у тому, що ми сприймаємо як логічно правильні форми суджень у граматично правильному мовленні ВММ. При цьому «факти» для самої ВММ (якщо її продовжувати послідовно розглядати як осучаснений різновид «китайської кімнати», Chinese room argument, by John Searle) – це *факти мови*, які не мають іншого значення поза внутрішнім когерентизмом ВММ: їх може верифікувати, валідувати чи спростувати виключно зовнішній (наш) довільний (arbitrary) ввід до контексту.

У результаті, загальна епістемічна цінність продуктів генерації ВММ визначається передусім їх граматичною правильністю і внутрішньою логічною послідовністю. І це відповідає першому етапу машинного навчання ВММ. Далі ж змістовна *епістемічна цінність* та *епістемічне значення* продуктів генерації ВММ *визначаються її користувачем* (людиною чи машиною) шляхом так чи інакше осмисленого останнім втручання в процес генерації на основі власного оцінювання і встановлення значення її проміжних результатів.

Фундаментальний внутрішній (і, до того ж, стохастичний) когерентизм та знеособленість ВММ у цілому покладають *обов'язок інтерпретації* їх виводу *і наступної перевірки* (верифікації та валідації) змісту такої інтерпретації на зовнішнього щодо самої ВММ спостерігача (користувача). У цьому розумінні ВММ як знеособлені генератори зовнішньо схожих на судження символічних послідовностей є повністю *безвідповідальними.* Нині таку відповідальність покладають, меншою мірою, на конструкторів і вчителів ВММ (тут зміст відповідальності має політичний характер і реалізується через

RLHF), і найбільшою, що цілком справедливо, – на їхніх кінцевих користувачів (які відповідають за наслідки, зазвичай, практично, своїм життям).

Дискусії навколо обґрунтованості атрибуції ВММ властивих людських інтелектуальних здатностей не завершуються і набувають вигляду протистояння між ідеєю і практикою, між неповнотою теорії і недоліками реалізації, між алармізмом прихильників *AI safety* та оптимізмом адептів *AGI* (artificial general intelligence). При цьому відверто крайні позиції поділяє мало хто, і, загалом, нині лінія поділу тут пролягає між фундаментальними інтересами дослідників та практичними інтересами компаній, які вкладають надзвичайні матеріальні ресурси у розвиток і корисне втілення ВММ на основі експлуатації існуючих принципових технологій, однією з яких є архітектура Transformer.

Діалектика співзалежних взаємовідносин між дослідниками і інженерами тут є неусувною: успішні ВММ стали *практично* можливими виключно *завдяки залученню* до їх конструювання і навчання *значних апаратних, матеріальних і людських ресурсів.* Адже, справді, без гігантських обчислювальних кластерів і, буквально, армій людей, які оформлювали і розмічали не менш гігантські навчальні набори даних вручну, теорія не набула б тієї валідації, яку вона має нині. ВММ – ще й досі надзвичайно ресурсозатратні (принаймні, на етапі навчання) технології, у яких розмірні характеристики доступних ресурсів мають визначальне значення. При цьому одним з «темних боків» ВММ саме є доволі об'ємне пряме використання людських ресурсів як на етапах їх навчання з підкріпленням (RLHF), так і в процесах їх приведення у відповідність до тих чи інших людських очікувань (human alignment, political correctness).

На цьому фоні можна зрозуміти, чому в цілому влучна і небезпідставна характеристика ВММ як *stochastic parrot*

(«стохастичний імітатор») [1] не знаходить свого схвалення (але не розуміння) у доволі широких колах спеціалістів, залучених до практичної інженерно-технічної роботи у цій сфері.

Натомість, незважаючи на фундаментальну обмеженість ВММ, з різним успіхом ведуться розробки методик та системних інтеграцій, які дали б можливысть пом'якшити чи замаскувати (але не усунути) ефекти цієї обмеженості. І ці зусилля, у підсумку, дають результати.

Так, методики «конструювання підказок» (prompt engineering) на практиці засновуються на імпліцитному визнанні того, що за зміст відповіді ВММ несе відповідальність той, хто запитує (be careful what you wish for!). І якщо певним чином «правильно спитати», то ВММ з вищою імовірністю надасть відповідь, що може краще відповідати очікуванням.

Практично, ці методики зводяться до ручних і автоматизованих методів втручання у контекст генерації для наповнення його необхідною для надання відповіді інформацією. Ця інформація далі, через механізм «уваги», впливає на зміст наступної за нею генерації, демонструючи, таким чином, «емерджентний» ефект «навчання з контексту» (in-context learning).

«Навчання» із самого контексту в режимі експлуатації ВММ насправді не відбувається – внутрішній латентний векторний простір ВММ змін не зазнає. Тим не менше, на основі емпіричних спостережень за поведінкою ВММ як «чорного ящика», практики «конструювання підказок» розглядають методики уведення до контексту прикладів бажаного міркування чи результату (few-shot learning), інструкцій керування структурою і послідовністю кроків і операцій міркування (COT, chain-of-thought; TOT, tree-of-thought), цитат відомостей із зовнішніх пошукових джерел (RAG, retrieval-augmented generation), відповідей від зовнішніх спеціалізованих інформаційних систем (LangChain).

Створюючи комплексні програмні застосунки, у яких ВММ є лише одним з компонентів, розробники так можуть обмежувати фактичне використання можливостей моделі, відокремлюючи її компетентності у частині *розуміння мови* від її часто проблематичних «емерджентних» компетентностей у частині *розуміння предмету мовлення*.

Зокрема, дефіцит обчислювальної повноти (за Тюрингом) у ВММ може компенсуватися шляхом перекладу запиту (частини запиту) з «природної мови» на деяку об'єктну мову програмування, виконання отриманого програмного коду у відповідному зовнішньому середовищі і вставки виводу результатів такого виконання на відповідне місце у контексті генерації.

Аналогічно, контекст також може наповнюватись вибіркою анотацій веб-сторінок, віднайдених за скерованим до зовнішніх пошукових систем запитом на основі відповідної частини контексту.

Нарешті, умовно-циклічне (залежно від поточного стану виводу генерації) автоматизоване часткове наповнення контексту зовнішнім вводом із виводу самої ВММ (синтетичними даними) або інтегрованих з нею зовнішніх її гетерогенних розширень називають *агентом* (agent). Принципово, немає теоретичних чи технічних обмежень для того, аби такий «агент» міг вважатися «інтелектуальним» настільки, наскільки таким міг би вважатися весь інтегрований комплекс систем, включно з ВММ, разом узятий.

Та, більше того, взагалі немає принципових обмежень і на підключення до ВММ будь-яких систем зовнішнього вводу («сенсорів», sensors) чи виводу («актуаторів», actuators). І тоді усю видимість процесу і результатів міркування ВММ у якості системи управління (control system) можна було б спостерігати як вже не виключно мовленнєву історію її носія у світі.

Наразі такими носіями інформаційних сигналів управління з боку ВММ добровільно, хоч і не завжди усвідомлено, виступають самі їх користувачі, які наділяють продукти генерації ВММ епістемічною цінністю і значенням у своїх оцінках і діях.

Приклади негативного і навіть шкідливого соціоекономічного впливу від некритичної оцінки можливостей і фундаментальних обмежень сучасних ВММ є відомими: проблема «галюцинацій» (hallucinations, вигаданих правдоподібних фактів) ВММ зачепила усі сфери, де їх намагались застосовувати (від судової практики і публічних консультацій й до освіти і медицини).

Як наслідок, усвідомлення неминучості прямого перетікання пов'язаних із заснованими на машинному навчанні системами штучного інтелекту (включно з ВВМ) епістемічних ризиків у ризики управління, Європейським парламентом було прийнято Закон про штучний інтелект (Artificial Intelligence Act), яким, зокрема, уводяться чіткі зобов'язання щодо використання систем штучного інтелекту у високоризикованих сферах, до яких (згідно з переліком) відносяться практично всі сфери прав людини і організації систем критичного забезпечення соціуму і держави.

Цей документ у цілому уможливлює промислове використання таких систем, як основи для *автоматизованого прийняття [управлінських] рішень* щодо громадян за умови наявності «систем управління ризиками» (risk management system) і «нагляду з боку людини» (human oversight). Передбачається, що громадяни матимуть право *подавати скарги* на системи штучного інтелекту та *отримувати пояснення щодо рішень*, заснованих на системах штучного інтелекту у високоризикованих сферах, які впливають на їхні права і хід життя.

Таким чином, у систем штучного інтелекту з'являється законний універсальний кредит довіри на ухвалення щодо

громадян екзистенційного значення рішень, які, фундаментально, ефективних пояснень мати не можуть. Адже там, де такі пояснення є можливими, діють правила не статистичного, а класичного логічного умовиводу. А використання статистичного профілювання, зазвичай, властиві при ухваленні дискреційних рішень, оскільки лише цей тип рішень дає можливість інкорпорувати ефективно непояснювані аргументи як компонент відповідального пояснення.

Іншими словами, якщо громадянин погоджується з висновком системи штучного інтелекту, то пояснення доводів на користь цього стає його особистою проблемою і відповідальністю. А якщо громадянин не погоджується – тоді тут висновок системи перекваліфіковується у дискреційний висновок людини-службовця, якій доведеться відповідально пояснювати, на яких дійсних підставах, крім висновку системи штучного інтелекту, була прийнята та чи інша кваліфікація, класифікація чи рішення щодо громадянина.

Така підміна суб'єкта відповідальності не здається надто безпечною, оскільки, у загальному випадку, відповідальність за рішення «самовпевненої» (overconfident) системи штучного інтелекту перекладається на громадянина, який тут вже не є повністю добровільним користувачем цієї системи. При цьому громадянинові, як і у випадку з ВММ, належить самостійно *судити про себе*, приймаючи рішення про те, чи погоджуватись з автоматичним висновком про себе, не маючи для ефективної аргументації ні інструментальних засобів, ні достатньої інформації.

У зв'язку з цим ми можемо лише сподіватися, що запропонована вище експозиція зразка «гуманітарно-технічного» розуміння великих мовних моделей типу Transformer, через інтуїтивно-наочну теоретико-філософську експлікацію їх природи, улаштування і можливостей, дасть змогу краще зрозуміти місце, роль, значення, ступінь необхідності і прийнятні

способи соціальної участі цих систем у процесах практичної пізнавально опосередкованої діяльності людини.

Адже показані нами фундаментальні засади епістемічної обмеженості сучасних діалогових великих мовних моделей (зокрема у ролі баз знань та генераторів міркувань) залишають принципові сумніви у надійності і безпечності застосування їх реалізацій у якості основних компонентів компетентних агентів щонайменше у високоризикованих сферах людського життя.

ЛІТЕРАТУРА ТА ПРИМІТКИ

1. Bender E. M., Gebru T., McMillan-Major A., Shmitchell Sh. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FaccT '21*. New-York : Association for Computing Machinery. P. 610–623. doi:10.1145/3442188.3445922.

2. Kaelbling L. P., Littman M. L., Moore A. W. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research. 1996. Vol 4.* P. 237–285. URL : https://arxiv.org/abs/cs/9605103

3. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature.* 2015. V. 521(7553). P. 436-444.

4. OpenAI. GPT-4 Technical Report. *arXiv:2303.08774 [cs.CL]*. https://doi.org/10.48550/arXiv.2303.08774

5. Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida. Training language models to follow instructions with human feedback. *arXiv:2203.02155 [cs.CL].* https://doi.org/10.48550/arXiv.2203.02155

6. Vaswani A., Shazeer N., Parmar N., Uszkoreit J. Attention is all you need. *Proceedings of Neural Information Processing Systems (NeurIPS).* 2017. https://doi.org/10.48550/arXiv.1706.03762

7. Маєвський О. Л. Функціональний успіх інтелектуальних автоматів. *Наукові записки НаУКМА. Філософія та релігієзнавство.* Т. 5. Київ, 2020. С. 15–25. https://doi.org/10.18523/2617-1678.2020.5.15-25; Маєвський О. Л. Кластерний аналіз і механіка досвіду. *Семіотичний аналіз явищ культури*. Київ, 2021.

С. 350–393. URL : https://web.archive.org/web/20220415072219/https://www.filosof.com.ua/Mentaltheorie/P8.pdf; Маєвський О. Л. Комунікативна раціональність сучасних інтелектуальних автоматів. *Комунікативні трансформації в сучасній науці.* Київ, 2022. С. 219–278. URL : https://www.filosof.com.ua/elektronna_biblioteka

REFERENCES

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, Sh. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21* (pp. 610-623). New York: Association for Computing Machinery. doi:10.1145/3442188.3445922

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research, 4,* 237-285. Retrieved from https://arxiv.org/abs/cs/9605103

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Long, O., Jeff, Wu, Xu J., Diogo, A., Carroll, L., Wainwright, P., & Ryan, L. (2022). Training language models to follow instructions with human feedback. In *arXiv:2203.02155 [cs.CL].* https://doi.org/10.48550/arXiv.2203.02155

Mayevsky, A. (2022). Communicative Rationality in Contemporary Intelligent Automata. In *Communicative Transformations in Contemporary Sciences* (p. 219-278). IF NANU Retrieved from https://www.filosof.com.ua/elektronna_biblioteka [In Ukrainian].

OpenAI (2023). GPT-4 Technical Report. In *arXiv:2303.08774 [cs.CL].* https://doi.org/10.48550/arXiv.2303.08774

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. … Polosukhin, I. (2017). Attention is all you need. In *Proceedings of Neural Information Processing Systems (NeurIPS).* https://doi.org/10.48550/arXiv.1706.03762

Alexander Mayevsky
Candidate of Philosophical Sciences, Junior Researcher at the Department of Logic and Methodology of Science, H. Skovoroda Institute of Philosophy, NAS of Ukraine; Kyiv; Ukraine; e-mail: mayevsky@nas.gov.ua; ORCID: https://orcid.org/0000-0001-6063-6033

Epistemic Limitations of Large Language Models

Abstract
The purpose of this paper is a philosophical exposition of the conceptual limitations and the corresponding epistemic value of applications of dialogic large language models (LLMs) based on the Transformer architecture as rational [co-]agents in the context of practical cognitively mediated human activity. To this end, the following research tasks are accomplished in the study: 1) a philosophical review of the fundamental intuitions, general nature and functional mechanics of advanced LLM implementations; 2) finding the limitations of the instrumental epistemic value of LLMs as products of a specific chain of machine learning procedures, which includes direct expert participation by humans. The practical significance of the study lies in the intuitive and compelling crystallization of the humanitarian understanding of LMMs through the theoretical and philosophical explication of their nature and design, which is grounded on the following new research results: LLMs are presented as a functionalist mechanistic project of statistical language and speech modeling as a model of knowledge as a semantic model of reality – 1) a model 2) of a model 3) of a model of reality; it is shown that due to the significant distance of reality mediation, intra-model connections lose in their factual capacity; it is also demonstrated that an LLM is a product of machine learning of a certain linguistic behavior with a purpose and values that are radically different from those of human cognition, which makes the justifiability of their employment as components of any high-risk decision support systems to be fundamentally questionable.

Keywords*: large language model; epistemic value; epistemic significance; Transformer architecture; philosophy of technology; philosophy of AI; AI safety; AI alignment.*