Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40350515
Article received 01.08.2025
Article accepted 31.10.2025
UDC 004.89:81'322:519.816
Correspondence author: N. E. Kunanets nek.lviv@gmail.com

В. В. Пасічник, М. В. Яромич, І. І. Павлів, Н. Е. Кунанець
Національний університет "Львівська політехніка", м. Львів, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ САМОАНАЛІЗУ ПАРАМЕТРІВ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Досліджено результати роботи великих мовних моделей LLM (англ. *Large Language Model*) для вирішення лінгвістичних завдань. Визначено широкий спектр моделей, враховуючи провідні розробки від відомих організацій, оцінені за низкою параметрів, таких як розуміння контексту, якість генерування тексту, коригування граматики, багатомовність і лексичне багатство. Особливий акцент зроблено на методології групового методу аналізу ієрархій, яка уможливила інтеграцію оцінок від різних експертних груп, зокрема – самих мовних моделей, що виступали в ролі експертів. Самооцінювання моделей реалізовано із застосуванням методу аналізу ієрархій та новаторського підходу через формування матриць парних порівнянь, де моделі оцінювали одна одну за визначеними параметрами, а отримані дані агрегувалися за допомогою геометричного середнього. Виявлено лідерів, які відзначилися високою продуктивністю в ключових аспектах, таких як граматична точність, контекстна чутливість і якість тексту, тоді як моделі середнього сегмента продемонстрували збалансовані можливості, а деякі опинилися в нижньому сегменті через обмежену ефективність у функціональних завданнях. Узгодженість оцінок підтверджує надійність отриманих даних. Встановлено важливість вибору моделей відповідно до специфіки лінгвістичних завдань, пропонуючи рекомендації для їх застосування в завданнях, що вимагають високої точності чи швидкості оброблення. З'ясовано, що самооцінювання розкриває потенціал моделей до саморефлексії, що має значення для когнітивної лінгвістики та філософії штучного інтелекту. Результати вказують на потребу вдосконалення моделей у таких аспектах, як підтримка різних мов і зменшення упереджень, особливо для міжкультурних контекстів. Встановлено, що методологічно метод аналізу ієрархій є ефективним у гармонізації міждисциплінарних оцінок, поєднуючи кількісну точність із якісною обґрунтованістю. Охарактеризовано закономірності, що формують нову парадигму для аналізу можливостей штучного інтелекту в медичній галузі, пропонуючи інструмент для вибору оптимальних великих мовних моделей і з'ясовано питання щодо доцільності автономного використання систем штучного інтелекту та розподілу відповідальності між людиною та машиною. Перспективи подальших досліджень охоплюють вивчення стабільності отримання результатів під час аналізу за різними пріоритетами параметрів, аналіз продуктивності моделей у специфічних завданнях, таких як оцінювання рівня кваліфікації медичних працівників за різними критеріями з використанням структурованих та неструктурованих даних. Вказано на роль інформаційних технологій у перетворенні лінгвістичних даних на структуровані знання, сприяючи глибшому розумінню потенціалу та обмежень мовних моделей у цифровому гуманітарному просторі та можливості імплементації їх у медичну інформаційну систему.

Ключові слова: метод аналізу ієрархій; самооцінювання; розуміння контексту; багатомовність; системи штучного інтелекту.

Вступ / Introduction

За сучасних умов стрімкого розвитку систем штучного інтелекту (СШІ) великі мовні моделі набули значення не просто як інструменти для генерування тексту чи пошуку даних, а як повноцінні учасники процесу формування знання. Якщо раніше генеративні моделі застосовувалися переважно для автоматизації процесу виконання рутинних завдань, то нині їх дедалі частіше залучають до складних аналітичних процесів, серед яких інтерпретація художніх текстів, стилістична трансформація, дискурсивний аналіз, узагальнення змісту, виявлення смислових зв'язків тощо [21]. Це вимагає переосмислення їх ролі у лінгвістиці та потребує системного порівняльного оцінювання їхньої ефективності.

Мовні моделі нового покоління – GPT-4, Claude 3, Gemini, DeepSeek, Grok та інші – відрізняються за своєю архітектурою, кількістю параметрів, принципами навчання, складом корпусів даних і методами їх донавчання на підставі цих корпусів. Проте, за межами цих технічних характеристик особливої вагомості набуває їхня прикладна ефективність – зокрема в галузях, де потрібне тонке розуміння мови, контексту, стилю та значення. Саме в таких завданнях, як лінгвістичний аналіз чи філологічне тлумачення, виникає потреба оці-

Інформація про авторів:
Пасічник Володимир Володимирович, д-р техн. наук, професор, кафедра інформаційних систем та мереж. **Email:** vpasichnyk@gmail.com; https://orcid.org/0000-0002-5231-6395
Яромич Максим Віталійович, аспірант, кафедра прикладної лінгвістики. **Email:** yamax0312@gmail.com; https://orcid.org/0009-0005-3299-6695
Павлів Ігор Ігорович, аспірант, кафедра інформаційних систем та мереж. **Email:** mickyukr@gmail.com
Кунанець Наталія Едуардівна, д-р наук із соц. комунікацій, професор, кафедра інформаційних систем та мереж. **Email:** nek.lviv@gmail.com; https://orcid.org/0000-0003-3007-2462
Цитування за ДСТУ: Пасічник В. В., Яромич М. В., Павлів І. І., Кунанець Н. Е., Інформаційна технологія самоаналізу параметрів великих мовних моделей. Науковий вісник НЛТУ України. 2025, т. 35, № 5. С. 130–143.
Citation APA: Pasichnyk, V. V., Yaromych, M. V., Pavliv, I. I., & Kunanets, N. E. (2025). An information technology for the self-analysis of large language model parameters. *Scientific Bulletin of UNFU*, 35(5), 130–143. https://doi.org/10.36930/40350515

нювати моделі не тільки як універсальні мовні інтерфейси, а як потенційні "інтелектуальні партнери" у гуманітарних дослідженнях.

Актуальність цього дослідження зумовлена кількома ключовими чинниками. По-перше, великі мовні моделі вже активно застосовують у сфері лінгвістики для вирішення таких завдань, як автоматичне визначення частин мови, синтаксичний аналіз, розпізнавання стилістичних відтінків, тематичне моделювання та навіть фрагменти літературної критики [37]. У таких випадках важливо оцінити не тільки загальну продуктивність моделі, а і її глибину контекстуального розуміння, мовну чутливість і здатність до відтворення стильових норм і фактологічної точності [19]. Це потребує складніших підходів до тестування, ніж звичайна перевірка через запити та відповіді.

По-друге, лінгвістичні завдання мають міждисциплінарний характер, оскільки залучають фахівців із лінгвістики, філології, перекладознавства, культурології та інформаційних технологій. Тому оцінювання ефективності використання мовних моделей вимагає залучення експертів із різних галузей, кожен із яких застосовує власні критерії – від точності синтаксису до інтерпретативної складності й технічної надійності [8]. Саме тому для проведення дослідження було обрано метод аналізу ієрархій Сааті [33], який дає змогу структурувати й інтегрувати оцінки різних груп характеристик у єдину систему пріоритетів.

По-третє, це дослідження має на меті не тільки оцінити поточну ефективність використання мовних моделей, а й визначити напрями їх подальшого вдосконалення. Зіставлення моделей за 15 спеціально визначеними параметрами, такими як чутливість до контексту, лексичне багатство, багатомовність, здатність виявляти граматичні помилки, стилістична гнучкість – дає змогу скласти повну картину сильних і слабких сторін кожної системи. Це створює цінну базу знань для викладачів, науковців, редакторів і спеціалістів, які інтегрують великі генеративні моделі у свою щоденну професійну діяльність.

Особливий інноваційний аспект цього дослідження полягає в тому, що роль експертів виконують самі мовні моделі. У процесі оцінювання використовують відповіді, сформульовані самими генеративними моделями – GPT-4o, Claude 3, Gemini, DeepSeek R1 та іншими [13]. Такий підхід відкриває нову методологічну перспективу для вивчення здатності моделей до саморефлексії, тобто оцінювання одна одної в контрольованому середовищі. Це не тільки практичний інструмент, а й філософське підґрунтя для осмислення автономності штучного інтелекту й формування нових типів відповідальності, розподілених між людиною та СШІ.

Об'єкт дослідження – самоаналіз параметрів великих мовних моделей.

Предмет дослідження – методи і засоби самоаналізу параметрів великих мовних моделей, що містять процеси їх самооцінювання, формування матриць парних порівнянь, агрегування даних та ранжування моделей за визначеними параметрами їхнього оцінювання.

Мета роботи – розробити універсальну інформаційну технологію для оцінювання великих мовних моделей щодо ефективності виконання лінгвістичних завдань з інтеграцією їх самооцінювання, що уможливить виявлення їх переваг і недоліків, обґрунтування рекомендацій щодо доцільності їх оптимального вибору та визначення напрямів подальшого вдосконалення у гуманітарній сфері.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

- розробити інформаційну технологію самоаналізу параметрів великих мовних моделей на підставі методу аналізу ієрархій, що дасть змогу об'єктивно порівнювати їх ефективність виконання лінгвістичних завдань;
- визначити та формалізувати перелік ключових параметрів процесу оцінювання великих мовних моделей (розуміння контексту, якість генерування тексту, багатомовність, лексичне багатство, коригування граматики тощо), що дасть змогу стандартизувати процес оцінювання різних моделей;
- розробити методику самооцінювання моделей із залученням самих мовних моделей як експертів, що забезпечить можливість дослідження їхнього потенціалу до саморефлексії та взаємооцінювання;
- розробити алгоритм агрегування результатів оцінювання моделей за допомогою групового методу аналізу ієрархій, які допоможуть інтегрувати оцінки від різних експертів (людей і моделей) у єдину узгоджену систему;
- провести порівняльний аналіз великих мовних моделей за визначеними параметрами, які уможливлять формування обґрунтованих рекомендацій із вибору оптимальної моделі для конкретних завдань;
- дослідити вплив різних параметрів мовних моделей на загальну їхню ефективність, які спонукають до виявлення напрямів їхнього вдосконалення;
- сформувати ранжований перелік великих мовних моделей та описати закономірності їх розподілу за рівнем ефективності, які приведуть до розроблення нових підходів до інтеграції LLM у гуманітарні, медичні та інші міждисциплінарні сфери.

Аналіз останніх досліджень та публікацій. У роботі [37] автори дослідили особливості застосування великих мовних моделей LLM (англ. *Large Language Model*) для створення та збагачення лінгвістичних онтологій в освітніх і наукових системах. Встановлено, що LLM можуть автоматизувати витягання концептів із текстів, формуючи структуровані бази знань та забезпечуючи точну інтерпретацію термінів і зв'язків. Автори наголошують, що онтології, створені з використанням LLM, підвищують ефективність аналізу текстів, однак потребують удосконалення для роботи з багатомовними даними, що виявляє обмеженість поточних підходів у міжкультурних контекстах.

У роботі [19] досліджено збагачення онтологій через оброблення природної мови з використанням LLM. Показано, що моделі здатні автоматично витягувати знання зі значних обсягів наукової літератури, що істотно полегшує семантичний аналіз. Водночас, автори зазначають відсутність механізмів самоперевірки, що могло б знизити упередження у створюваних онтологіях, і вказують на потребу інтеграції функцій самооцінювання.

У публікації [20] автори проаналізували інтеграцію LLM в онтологічні системи та розглянули їх як альтернативу традиційним методам тематичного моделювання (англ. *Topic Modelling*). Доведено, що моделі здатні ефективно обробляти великі масиви тексту з високою семантичною чутливістю, проте через брак прозорості процесів навчання можуть виникати неточності в лінгвістичному аналізі. Відзначено, що у дослідженні відсутні результати застосування групового оцінювання для гармонізації результатів, що є потенційним напрямом удосконалення.

У роботі [8] розглянуто можливості навчання LLM на онтологіях у межах концепції LLMs4OL, зокрема – інтеграцію семантичного вебу з природною мовою. Автори вказують на потенціал створення динамічних онтологій, здатних адаптуватися до нових знань, однак вказують на обмежене використання механізмів саморефлексії моделей. Зроблено висновок, що без здатності до самокоректування LLM мають меншу ефективність у завданнях стилометричного аналізу та жанрової класифікації.

У дослідженні [21] акцент зроблено на філософських і когнітивних аспектах LLM у гуманітарних науках. Автор показує, як LLM змінюють парадигму створення та інтерпретації знань, пропонуючи гібридні системи для аналізу історичних і літературних текстів. Вказано на необхідність етичної рефлексії та контролю упереджень, що виникають через особливості навчальних даних. Однак, у роботі не розглянуто інтеграцію з багатокритеріальними методами оцінювання, такими як AHP.

Класична праця [32] описує метод аналізу ієрархій (AHP), розроблений Т. Сааті, як інструмент багатокритеріального прийняття рішень. Автор показує важливість парних порівнянь для визначення пріоритетів та зазначає, що метод ефективний за умов невизначеності. У подальшому дослідженні [33] розкрито можливості групового AHP, який зменшує суб'єктивність через інтеграцію оцінок кількох експертів. У роботі [34] розглянуто поєднання AHP з якісними факторами оцінювання, що демонструє гнучкість методу, але не охоплює випадків, коли експертами виступають самі мовні моделі.

Отже, проведений аналіз літератури свідчить, що хоча LLM активно впроваджуються у лінгвістику для вирішення завдань семантичного аналізу, побудови онтологій та оброблення текстів, проте відсутні комплексні підходи, які поєднували б метод AHP із самооцінюванням моделей [8, 20, 19, 21, 32, 33, 34, 37]. Невирішеними залишаються проблеми зменшення упереджень у багатомовних контекстах і гармонізації оцінок від різних експертів. Це підтверджує потребу в розробленні нової інформаційної технології, запропонованій у цьому дослідженні, яка поєднала б багатокритеріальний аналіз і механізми саморефлексії LLM.

Матеріали та методи дослідження. Інформаційна технологія самоаналізу роботи параметрів мовних моделей є гнучким і універсальним методологічним підходом до оцінювання їхньої ефективності у виконанні лінгвістичних завдань. Ця технологія, що базується на методі аналізу ієрархій, розробленому Томасом Сааті, з інноваційним елементом самооцінювання, де самі моделі виступають як експерти, була створена з метою забезпечення об'єктивного порівняння великих мовних моделей (LLM).

Матеріали дослідження містять 16 великих мовних моделей та 15 параметрів процесу оцінювання. Дані для матриць парних порівнянь генерувалися шляхом запитів до 8 моделей-експертів (GPT-4o, GPT-4.5, Gemini, Claude 3, DeepSeek R1, Grok-3, Qwen 2.5, Mistral 7B), які оцінювали одна одну за шкалою від 1–9. Агрегування експертних оцінок здійснювалося за допомогою геометричного середнього, що відповідає класичному підходу методу аналізу ієрархій і забезпечує об'єктивніше узгодження індивідуальних суджень.

Технологія структурує процес аналізу через чітко визначені етапи, забезпечуючи об'єктивність, узгодженість і кількісну точність. Джерела можливих помилок – це суб'єктивність самооцінювання (моделі можуть мати упередження через навчання на подібних даних), обмеження API (наприклад, контекстне вікно або затримки), випадкові варіації у відповідях моделей (температура генерування = 0 для детермінованості), похибки в агрегуванні (контролювалися перевіркою $CR < 0{,}1$). Для мінімізації помилок, матриці генерувалися декілька разів і усереднювалися, а узгодженість перевірялася на кожному етапі.

Результати дослідження та їх обговорення / Research results and their discussion

Формування переліку моделей. Перший етап дослідження передбачає формування переліку великих мовних моделей, які будуть оцінюватися. Універсальність технології дає змогу містити моделі різного типу – від потужних комерційних систем до компактних відкритих моделей, залежно від цільового призначення. Наприклад, для лінгвістичних завдань можуть обиратися моделі з розвиненими можливостями оброблення природної мови, тоді як для інженерних чи медичних застосувань – моделі, оптимізовані для аналізу технічних чи спеціалізованих текстів. Відбір ґрунтується на таких критеріях, як функціональні можливості, масштабованість, доступність і здатність моделей до самооцінювання. Технологія дає змогу гнучко адаптувати перелік, враховуючи моделі з різними архітектурами, розмірами та спеціалізаціями, що робить її придатною для порівняння як універсальних, так і вузькоспеціалізованих систем. Мета – сформувати репрезентативний набір LLM, який відображає сучасні тенденції розвитку СШІ та відповідає потребам конкретної предметної галузі.

Формування переліку параметрів процесу оцінювання. Другий етап полягає у визначенні параметрів процесу оцінювання, які відображають ключові аспекти продуктивності моделей у контексті обраних завдань. Універсальність технології проявляється у можливості адаптувати параметри до специфіки будь-якої галузі. Наприклад, для виконання лінгвістичних завдань параметри можуть містити розуміння контексту, якість генерування тексту, багатомовність чи коригування граматики, тоді як для юридичних – точність інтерпретації термінології, для медичних – надійність діагностичних висновків, а для освітніх – доступність і швидкість оброблення. Параметри розробляються з урахуванням міждисциплінарних вимог, залучаючи експертів або аналіз літератури для забезпечення повноти критеріїв. Кожен параметр чітко формулюється, щоб уникнути неоднозначності під час оцінювання. Гнучкість цього етапу дає змогу створювати як універсальні, так і вузькоспеціалізовані набори параметрів мовних моделей, що робить технологію придатною для широкого спектра застосувань, від гуманітарних досліджень до промислових рішень.

Проведення опитування моделей (самооцінювання). Третій етап є новаторським і передбачає залучення мовних моделей до процесу оцінювання як експертів. Моделі генерують оцінки одна одній за визначеними параметрами, створюючи матриці парного порівняння. У кожній матриці моделі порівнюються попарно за шкалою від 1 до 9, де 1 означає рівнозначність, а 9 – значну перевагу. Наприклад, у лінгвістичному контексті одна модель може оцінити іншу як кращу за якістю

генерування тексту, тоді як у технічному – за швидкістю оброблення даних. Процес автоматизується через запити, які модель опрацьовує, спираючись на свої аналітичні можливості. Універсальність технології дає змогу застосовувати цей підхід до будь-яких завдань, оскільки моделі можуть оцінювати одна одну за будь-якими критеріями, визначеними на попередньому етапі. Цей етап не тільки забезпечує кількісні дані, але й досліджує саморефлексію моделей, що має значення для розуміння їхнього когнітивного потенціалу в різних галузях.

Виконання обчислень за груповим методом аналізу ієрархій. Четвертий етап охоплює оброблення даних за методом аналізу ієрархій, який забезпечує строгу математичну основу для порівняння. Спочатку агрегується матриця парного порівняння параметрів мовних моделей, щоб визначити їхню відносну важливість, використовуючи геометричне середнє оцінок моделей-експертів. Вагомості параметрів мовних моделей обчислюють методом головного власного вектора [33] і нормалізують (сума становить 1). Узгодженість оцінок перевіряють за допомогою індексу узгодженості (*CI*) та коефіцієнта узгодженості (*CR*), де $CR < 0{,}1$ вказує на надійність отриманих даних. Далі для кожного параметра агрегується матриця порівняння моделей, формуючи локальні вектори вагомості, які відображають їхню продуктивність за окремими критеріями. Фінальний вектор вагомості моделей розраховують як добуток матриці рішень (де рядки – моделі, стовпці – параметри) на вектор вагомості параметрів мовних моделей. Універсальність цього етапу полягає у можливості опрацьовувати дані для будь-якої кількості моделей і їхніх параметрів, адаптуючи обчислення до специфіки завдань, що робить інформаційну технологію масштабованою та застосовною до різних галузей.

Формування результуючого впорядкованого переліку моделей. Останній етап дослідження передбачає формування впорядкованого переліку моделей на підставі фінального вектора вагомості. Кожна модель отримує значення, яке відображає її загальну ефективність з урахуванням усіх параметрів процесу оцінювання. Перелік ранжується від найвищих до найнижчих вагомості, даючи можливість визначити моделі, що найкраще відповідають цільовим завданням. Універсальність технології забезпечує можливість використання цього переліку в різних контекстах. Наприклад, їх використовують для вибору моделі для семантичного аналізу в лінгвістиці, оброблення юридичних документів чи створення освітнього контенту. Результати супроводжуються аналізом, який пояснює вплив окремих параметрів процесу оцінювання на позиції моделей. Цей етап надає практичний інструмент для прийняття рішень, а також основу для подальшого вдосконалення моделей у специфічних галузях, таких як підвищення точності чи оптимізація ресурсів (рисунок).

Зрештою, порівняльне оцінювання мовних моделей із використанням методу аналізу ієрархій Сааті [33] у поєднанні з машинним самооцінюванням, проведеним у цій роботі, формує нову методологічну парадигму в аналізі якості роботи СШІ у гуманітарній галузі. Це не тільки дає змогу встановити, яка модель функціонує краще, а й глибше зрозуміти, за яких умов і для яких користувачів вона буде найефективнішою [20].

Характеристика великих мовних моделей, параметри яких порівнюються. GPT-4o (Omni) – новітня флагманська модель OpenAI, здатна опрацьовувати текст, зображення та аудіо в реальному часі без окремих модулів перетворення. Вона забезпечує природну багатомодальну взаємодію з дуже низькою затримкою (~232 мс для голосу), підтримує понад 50 мов [25].

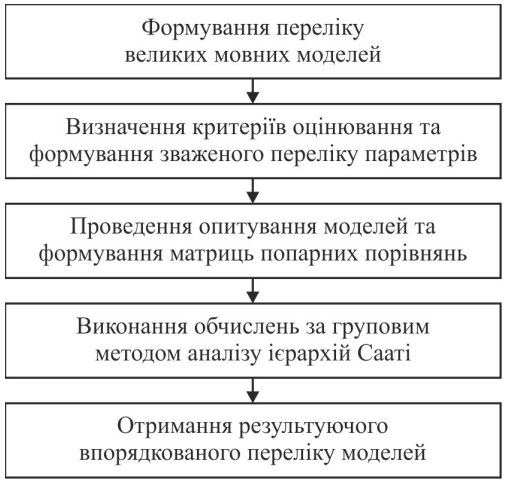

Рисунок. Структура інформаційної технології самоаналізу груп параметрів великих мовних моделей / Structure of the information technology for self-analysis of parameter groups of large language models

GPT-4.5 – перехідна версія між GPT-4 і GPT-5 з покращеною контекстною пам'яттю до 128 тис. токенів, стабільнішою логікою, точністю та краще контрольованою поведінкою. Модель підходить для вирішення складних завдань, зокрема – в юридичній, медичній та науковій галузях [25].

DeepSeek R1 – відкрита MoE-модель з 671 млрд параметрів процесу оцінювання (37 млрд активних на запит), яка поєднує ефективність і якість генерування відповідей на запити. Вона особливо вирізняється під час вирішення логічних завдань, під час програмування та аналізу текстів багатьма мовами, завдяки навчанню на підставі корпусів текстів технічного спрямування [5].

Gemini 2.5 Pro – мультимодальна модель Google DeepMind з контекстним вікном до 1 млн токенів, що опрацьовує текст, зображення, аудіо та відео. Вона інтегрована в екосистему Google (Gmail, Docs, Search) і має потужні планувальні та аналітичні можливості [6].

Grok 3 – третє покоління моделей від xAI Ілона Маска, орієнтоване на логічну точність і актуальність відповідей. Має варіанти Think і Big Brain, функцію DeepSearch для роботи з реальними даними та інтеграцію в платформу X (Twitter) [40].

Claude 3 – серія моделей від Anthropic (Haiku, Sonnet, Opus), що підтримують оброблення тексту й зображень, контекст до 200 тис. токенів (до 1 млн у деяких випадках) і працюють за принципом "Constitutional AI", забезпечуючи етичну взаємодію [1].

Mistral 7B – легка й ефективна відкрита модель з 7,3 млрд параметрів процесу оцінювання, яка перевершує більші LLaMA-моделі завдяки GQA та SWA механізмам. Доступна у версіях Base та Instruct під ліцензією Apache 2.0 [18].

Phi-2 – компактна модель Microsoft з 2,7 млрд параметрів процесу оцінювання, яка демонструє виняткову якість логічного виведення, попри невеликий розмір, завдяки навчанню на оптимізованому корпусі текстів [17].

Llama 3.1 – найновіша модель від Meta AI з конфігураціями до 405 млрд параметрів процесу оцінювання,

навчена на 15 трлн токенів. Підтримує кодогенерацію, багатомовність та інтегрується в продукти Meta (WhatsApp, Instagram), доступна через API та Hugging Face [16].

Nemotron-4 340B – модель від NVIDIA з 340 млрд параметрів процесу оцінювання, орієнтована на генерацію якісних синтетичних даних. Функціонує на платформі NVIDIA NeMo та використовує SFT і RLHF для покращення взаємодії з користувачами [22].

Qwen2.5 – серія моделей від Alibaba з кількістю параметрів процесу оцінювання до 72 млрд і підтримкою контексту до 128 тис. токенів. Мають як dense, так і MoE-архітектури, мультимодальні версії (Omni-7B) та доступність з відкритим кодом [30].

PaLM 2 – багатомовна модель Google другого покоління, зосереджена на перекладі, генеруванні коду та логічному мисленні. Інтегрована в понад 25 продуктів Google (Docs, Gmail, Bard) та доступна через API [10].

LaMDA – модель для діалогових систем від Google, представлена у 2021–2022 рр., орієнтована на ведення комунікації природною мовою. Стала базою для чат-бота Bard з акцентом на контекстну релевантність [4].

Megatron-Turing NLG 530B – модель Microsoft і NVIDIA з 530 млрд параметрів процесу оцінювання, зосереджена на генеруванні тексту, логіці та програмуванні, створена за допомогою фреймворків DeepSpeed та Megatron [36].

BLOOM – відкрита багатомовна модель із 176 млрд параметрів процесу оцінювання, створена у межах проєкту BigScience, навчена на 46 мовах і 13 мовах програмування. Вона доступна безкоштовно для асистування досліджень [2].

LLaMA 2 – друга генерація відкритих моделей Meta AI з обсягами до 70 млрд параметрів процесу оцінювання, оптимізована для діалогів (Chat-версії) й широко використовується як допоміжний інструмент у наукових і комерційних цілях [38].

Метод аналізу ієрархій Сааті [32]. Для проведення дослідження використано метод аналізу ієрархій (MAI), який розробив американський учений Томас Сааті у 1970-х роках для вирішення складних завдань прийняття рішень [32]. Основна ідея методу полягає у структурованому представленні проблеми у вигляді ієрархії, де верхній рівень визначає загальну мету, середній – сукупність параметрів процесу оцінювання, а нижній – конкретні альтернативи, між якими здійснюють вибір. Такий підхід дає змогу не тільки впорядкувати складні багатофакторні завдання, а й поєднати у процесі прийняття рішень як кількісні, так і якісні показники.

MAI особливо ефективний у ситуаціях, коли потрібно оцінити альтернативи за багатьма критеріями, об'єднати експертні судження різного ступеня суб'єктивності або інтегрувати кількісні та якісні фактори в єдину систему [34]. Типовими прикладами застосування є вибір постачальника, визначення пріоритетів інвестицій, оцінювання ризиків або формування стратегій.

Застосування MAI розпочинається з побудови ієрархічної структури завдання. На її вершині розміщується головна мета – наприклад, вибір оптимального рішення. Наступний рівень ієрархії складається з параметрів процесу оцінювання, за якими здійснюють оцінку альтернатив (таких як вартість, якість, надійність, терміни тощо). На нижньому рівні подаються самі альтернативи – об'єкти або варіанти, між якими здійснюють вибір.

Ключовим елементом методу є формування матриць парних порівнянь. У межах кожного рівня ієрархії всі елементи порівнюються між собою попарно щодо їхнього впливу на вищий рівень. Для цього використовують дев'ятибальну шкалу, де значення 1 означає рівнозначність двох елементів, а 3, 5, 7 і 9 відповідають поступовому зростанню переваги одного елемента над іншим. Проміжні значення (2, 4, 6, 8) використовують для відтінків переваг.

Після заповнення матриць парних порівнянь здійснюють нормалізацію, на підставі якої обчислюються вагові коефіцієнти (власні вектори), що відображають відносну важливість кожного елемента. З огляду на те, що метод передбачає використання людських суджень, важливо перевірити їх узгодженість. Для цього розраховується індекс узгодженості *CI* (англ. *Consistency Index*). Якщо його значення перевищує 0,1, це свідчить про наявність суперечностей у судженнях, і відповідні оцінки потребують перегляду.

На завершальному етапі здійснюють агрегування результатів оцінювання моделей. Вагомості альтернатив підсумовують відповідно до важливості кожного параметра, і в такий спосіб визначається загальний пріоритет кожної альтернативи. У підсумку приймається рішення на користь тієї альтернативи, яка має найбільшу підсумкову вагомість.

Отже, MAI є надійним, гнучким і прозорим інструментом багатокритеріального аналізу, що дає змогу ухвалювати виважені рішення навіть у складних і нечітко структурованих ситуаціях.

Визначення параметрів процесу оцінювання великих мовних моделей. Визначимо основні параметри оцінювання роботи різних великих мовних моделей для вирішення лігвістичних завдань [19, 39]: обсяг пам'яті, необхідний для роботи; масштабованість; взаємодія з іншими системами (API, інтеграція); вартість використання; наявність автоматичних оновлень; якість генерування тексту; розуміння контексту; багатомовність; можливість адаптації до спеціалізованих завдань; лексичне та стилістичне багатство; швидкість генерування тексту; врахування контексту знань та фактів; розпізнавання та коригування граматичних помилок; модельні упередження та безпека; апаратні вимоги та оптимальність роботи.

Проаналізуємо кожен з параметрів мовних моделей детальніше. Основними параметрами оцінювання великих мовних моделей для вирішення лінгвістичних завдань є обсяг пам'яті, необхідний для роботи, масштабованість, здатність взаємодіяти з іншими системами, вартість використання, актуальність знань моделі, якість генерування тексту, розуміння контексту, багатомовність, можливість адаптації до спеціалізованих завдань, лексичне та стилістичне багатство, швидкість генерування тексту, врахування контексту знань і фактів, розпізнавання та коригування граматичних помилок, модельні упередження та безпека, а також апаратні вимоги й оптимальність роботи.

Для лінгвістичних завдань, особливо під час роботи з великими текстовими масивами або інтеграції моделей у локальні середовища, важливо, щоб модель не потребувала надмірного обсягу оперативної пам'яті, адже високі вимоги до ресурсів ускладнюють використання в освітніх або мобільних середовищах.

Масштабованість визначає здатність моделі ефективно працювати як з невеликими, так і з великими обсягами даних, що критично для дослідницьких і високонавантажених систем, наприклад чат-ботів.

Можливість інтеграції через API або стандартні інтерфейси важлива для взаємодії з корпусами, словниками, перекладачами, навчальними платформами. Вартість використання істотно впливає на доступність моделі в освіті та наукових дослідженнях.

Оновлення бази знань визначає актуальність результатів, адже моделі, що не оновлювались кілька років, можуть давати застарілі відповіді, особливо для сучасних текстів.

Якість генерування тексту впливає на логічність, зв'язність і стилістичну відповідність створених відповідей, що є критично важливим для освітніх і дослідницьких завдань. Розуміння контексту дає змогу враховувати зв'язки між реченнями і репліками в діалозі, необхідні для семантичного аналізу чи вирішення анафор. Багатомовність важлива, адже сучасні завдання не обмежуються англійською мовою, а вимагають роботи з українською, німецькою, іспанською та іншими мовами.

Можливість адаптації до спеціалізованих завдань означає, що модель можна налаштувати або донавчити для вузької сфери, наприклад правничої мови чи поезії. Лексичне і стилістичне багатство визначає здатність моделі працювати із жанрами, стилями та редагуванням текстів.

Швидкість генерування важлива для сценаріїв реального часу, наприклад живого перекладу чи інтерактивних завдань. Важливим є і врахування знань та фактів, щоб модель посилалася на достовірну інформацію, особливо під час пояснень чи навчальних завдань.

Розпізнавання та коригування граматичних помилок є необхідним для інструментів редагування, де модель має не тільки знаходити помилки, а й пояснювати їх та пропонувати правильні варіанти. Додатково доцільно враховувати модельні упередження та безпеку, щоб уникнути некоректних або образливих відповідей, а також апаратні вимоги, адже модель має працювати на обладнанні з обмеженими ресурсами, залишаючись продуктивною й оптимізованою.

Для виконання лінгвістичних завдань, особливо у випадках роботи з великими текстовими масивами або інтеграції моделей у локальні середовища, важливо, щоб модель не потребувала надмірного обсягу оперативної пам'яті. Якщо ресурсні вимоги занадто високі, це ускладнює застосування моделей у школах, університетах чи мобільних пристроях, де обчислювальні можливості обмежені.

Класи лінгвістичних завдань, які вирішують за допомогою великих мовних моделей. У сучасній лінгвістиці великі мовні моделі відкривають нові можливості для автоматизації аналізу тексту, зокрема – завдяки поєднанню з онтологіями – структурованими базами знань, які забезпечують точність та узгодженість термінів.

Використання великих мовних моделей охоплює широкий спектр завдань у лінгвістиці та цифрових гуманітарних науках. Вони забезпечують семантичний аналіз, даючи змогу визначати значення слів і синтаксичних конструкцій у контексті, зокрема – під час роботи з полісемією та метафорами, що значно покращується за поєднання з онтологіями [8]; виконують стилістичний і стилометричний аналіз, точно виявляючи стильові особливості тексту – тон, жанр, авторство – та аналізуючи мовні фігури ефективніше, ніж традиційні методи [23]; автоматизують лексикографічні завдання, витягуючи слова, приклади їх уживання й міжмовні відповідники, що оптимізує створення словників [39]; здійснюють жанрову класифікацію та тематичне моделювання, визначаючи жанри, ключові теми й мотиви текстів за допомогою кластеризації та topic modeling [28, 35]; сприяють інтертекстуальному аналізу, допомагаючи знаходити цитати, алюзії та зв'язки між творами й будувати карти інтертекстуальності з підтримкою онтологій [39]; застосовують для порівняльного мовознавства і багатомовного аналізу, порівнюючи мовні структури різних мов, що є особливо корисним у типологічних дослідженнях і перекладознавстві з урахуванням концептуальних відмінностей [8]; дають змогу працювати з історичними та літературними текстами, аналізуючи стиль минулих епох, реконструюючи мовний контекст та імітуючи історичне мовлення з урахуванням культурної онтології; застосовують для генерування та переписування текстів, створюючи або переформульовуючи тексти з урахуванням жанру, стилю або рівня складності, що особливо корисно у викладанні та редагуванні [26]; а також автоматизують створення цифрових архівів і онтологічних баз знань, класифікуючи та анотуючи тексти для формування інтерактивних архівів, які сприяють дослідженню та популяризації гуманітарної спадщини [19].

Аналогічно можна оцінювати ефективність LLM і для інших типів завдань – юридичних, медичних, освітніх, інженерних тощо – адаптуючи критерії оцінювання під специфіку предметної області та рівень точності, якого вимагає практичне застосування.

Визначення вагомості параметрів процесу оцінювання великих мовних моделей. Після формування повного переліку параметрів процесу оцінювання великих мовних моделей у лінгвістичних завданнях наступним етапом є побудова матриці парних порівнянь параметрів мовних моделей. Цей процес здійснюють відповідно до методології аналізу ієрархій Сааті й передбачає попарне зіставлення кожного параметра з усіма іншими з точки зору їх відносної важливості. Експерти порівнюють, наскільки один параметр є важливішим за інший у контексті мовної ефективності використання мовних моделей, використовуючи дев'ятибальну шкалу.

Усі ці оцінки заносяться в квадратну матрицю, у якій елементи відображають перевагу одного параметра над іншим (табл. 1). Матриця є взаємно оберненою, тобто якщо параметр А важливіший за параметр В з оцінкою 5, то навпаки – параметр В порівняно з А матиме значення 1/5. Після заповнення всієї матриці на її основі обчислюється вектор вагомості параметрів мовних моделей, який відображає відносну важливість кожного параметра в загальній структурі прийняття рішень.

Для обчислення вектора вагомості використовують метод головного власного вектора [33]. У математичному сенсі це означає знаходження ненульового вектора w, який задовольняє рівняння

$$Aw = \lambda_{max} w,$$

де: A – матриця парних порівнянь; λ_{max} – найбільше власне значення цієї матриці; w – відповідний власний вектор, що після нормування визначає вектор вагомості.

Отриманий вектор нормалізують (його елементи мають суму, що становить 1), і кожна компонента вектора відображає вагомість відповідного параметра в загальному оцінюванні.

Ці вагомості параметрів мовних моделей використовують на наступному рівні ієрархії для зважування ло-

кальних векторів альтернатив – тобто для обчислення фінального вектора глобальних пріоритетів мовних моделей, який узагальнює їхню ефективність у всіх обраних параметрах одночасно. Такий підхід забезпечує логічну послідовність оцінювання, формалізує суб'єктивні судження експертів і дає змогу приймати виважені рішення на підставі багатокритеріального аналізу (табл. 1).

Табл. 1. Матриця попарних порівнянь параметрів мовних моделей / Pairwise comparison matrix of language model parameters

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1.	1	1/3	1/3	1/3	1/3	1/7	1/9	1/7	1/7	1/7	1/5	1/9	1/9	1/5	1/5
2.	3	1	1	1	1	1/5	1/7	1/5	1/5	1/5	1/3	1/7	1/7	1/3	1/3
3.	3	1	1	1	1	1/5	1/7	1/5	1/5	1/5	1/3	1/7	1/7	1/3	1/3
4.	3	1	1	1	1	1/5	1/7	1/5	1/5	1/5	1/3	1/7	1/7	1/3	1/3
5.	3	1	1	1	1	1/5	1/7	1/5	1/5	1/5	1/3	1/7	1/7	1/3	1/3
6.	7	5	5	5	5	1	1	3	3	3	3	1	1	3	5
7.	9	7	7	7	7	1	1	3	3	3	3	1	1	3	5
8.	7	5	5	5	5	1/3	1/3	1	1	1	1	1/3	1/3	1	3
9.	7	5	5	5	5	1/3	1/3	1	1	1	1	1/3	1/3	1	3
10.	7	5	5	5	5	1/3	1/3	1	1	1	1	1/3	1/3	1	3
11.	5	3	3	3	3	1/3	1/3	1	1	1	1	1/5	1/5	1	3
12.	9	7	7	7	7	1	1	3	3	3	5	1	1	3	5
13.	9	7	7	7	7	1	1	3	3	3	5	1	1	3	5
14.	5	3	3	3	3	1/3	1/3	1	1	1	1	1/3	1/3	1	3
15.	5	3	3	3	3	1/5	1/5	1/3	1/3	1/3	1/3	1/5	1/5	1/3	1

Перевірка узгодженості. Наступним важливим етапом є перевірка, чи були порівняння між параметрами достатньо послідовними (не суперечливими). Для цього використовують індекс узгодженості *CI* (англ. C*onsistency Index*) та коефіцієнт узгодженості *CR* (англ. *Consistency Ratio*).

Формула для обчислення *CI* має такий вигляд [33]:

$$CI = \frac{\lambda_{\max} - n}{n-1}, \quad (1)$$

де: n – кількість параметрів мовних моделей (n = 15); λ_{max} – максимальне власне значення агрегованої матриці попарних порівнянь.

Після цього визначають коефіцієнт узгодженості *CR* за такою формулою:

$$CR = \frac{CI}{RI}, \quad (2)$$

де *CI* – середнє випадкове узгодження *RI* (англ. *Random Index*), для n = 15 становить *RI* = 1,59 (за стандартною таблицею Сааті):

- якщо *CR*<0,1, то узгодженість хороша, оцінки є коректними;
- якщо *CR*>0,1, то варто переглянути порівняння параметрів мовних моделей через суперечливість оцінок.

У цьому випадку:

- максимальне власне значення (λ_max) – 5,6150;
- індекс узгодженості (*CI*) – 0,0439;
- випадковий індекс (*RI*) – 1,59 (для матриці 15×15);
- коефіцієнт узгодженості (*CR*) – 0,0276.

Значення *CR* = 0,0276 істотно менше за критичне значення 0,1, що вказує на дуже хорошу узгодженість експертних оцінок.

Груповий метод аналізу ієрархій Сааті. Різновид методу аналізу ієрархій – груповий метод аналізу ієрархій – використовують як найдоцільніший інструмент для реалізації міждисциплінарного підходу [33]. На відміну від класичного варіанта MAI, орієнтованого на одного експерта, груповий метод дає змогу поєднати оцінки та судження кількох фахівців, що працюють у різних професійних доменах. У контексті оцінювання великих мовних моделей для лінгвістичних завдань це означає можливість врахування вагомих, але різновекторних параметрів мовних моделей, які формуються на підставі специфіки кожної експертної групи.

Суть групового MAI полягає у створенні матриць парних порівнянь параметрів мовних моделей або альтернатив кожним із експертів або експертних груп, після чого ці оцінки агрегуються в єдину, узгоджену матрицю. Якщо в оцінюванні бере участь k експертів, і кожен з них формує власну парну матрицю порівнянь A^1, A^2,…, A^k, то агрегована матриця A^* розраховується за допомогою геометричного середнього кожного відповідного елемента [33]:

$$a_{ij}^* = \sqrt[k]{\prod_{m=1}^{k} a_{ij}^{(m)}}, \quad (3)$$

де: k – кількість експертів; $a_{ij}^{(m)}$ – значення i, j-го елемента матриці парних порівнянь, визначене m-тим експертом.

Це дає змогу уникнути викривлень, властивих арифметичному середньому, і краще зберігає співвідношення переваг, особливо коли між експертами існують істотні розбіжності. Після формування агрегованої матриці проводиться її нормалізація та обчислення вектора вагомості – власного вектора, що представляє пріоритети параметрів мовних моделей або альтернатив. У класичному варіанті цей вектор формується через середнє нормалізованих рядків або через знаходження головного власного вектора матриці (методом степеневого наближення або іншим чисельним способом).

Наступним важливим кроком є перевірка узгодженості оцінок. Для цього обчислюється індекс узгодженості, який дає змогу оцінити, наскільки логічно послідовними були парні порівняння. Формально це виглядає як (1), де λ_{max} – найбільше власне значення агрегованої матриці, а n – її розмірність. Чим ближче λ_{max} до n, тим узгодженіші порівняння.

Отримане значення *CI* порівнюють з табличним індексом випадкової узгодженості (*RI*) для відповідного n, і визначають коефіцієнт узгодженості (2). Це показник, який показує відносний рівень узгодженості суджень. Якщо $CR \leq 0{,}1$ (10 %), рівень узгодженості вважається прийнятним. Якщо $CR > 0{,}1$, перевищення цього порогу, потрібно переглянути вихідні оцінки окремих експертів, аби забезпечити більшу послідовність.

У практичному сенсі це означає, що лінгвісти зможуть сфокусуватись на таких параметрах, як семантич-

на точність, стильове багатство, граматична коректність і здатність моделі до адекватного мовного аналізу; наукові працівники – на вмінні моделі працювати з джерелами, створювати інтерпретації, узагальнення, дотримуватись академічної доброчесності; IT-фахівці – на технічній стабільності, продуктивності, масштабованості, можливості адаптації та інтеграції. Кожна з цих груп, працюючи в межах власної експертизи, формує власну матрицю оцінювання, яка потім об'єднується у єдину модель, що відображає колективний консенсус.

Отже, груповий MAI не тільки технічно вирішує проблему використання оцінок декількох експертів, а й формує концептуально прозору методологію для гармонізації міждисциплінарних поглядів. Його перевагою є те, що він забезпечує як кількісну точність, так і якісну обґрунтованість пріоритетів. Унаслідок оцінювання можна отримати єдиний інтегрований вектор вагомості параметрів мовних моделей або рейтинг альтернатив, що враховує комплексний погляд усіх залучених сторін. У галузі гуманітарних досліджень, де важко застосувати суто технократичні або формальні методи оцінки, груповий MAI стає ключовим інструментом, який дає змогу об'єднати мовну, змістову й технологічну логіку у збалансовану систему ухвалення рішень.

Процедура самооцінювання великих мовних моделей за визначеними параметрами. Наступний етап дослідження передбачає оцінювання великих мовних моделей за наведеними вище параметрами з використанням самих моделей як джерела експертних суджень (табл. 2). У цьому сценарії ролі експертів виконуватимуть вісім провідних мовних моделей: GPT-4o, GPT-4.5, Gemini, DeepSeek R1, Grok-3, Claude 3, Qwen 2.5 та Mistral 7B. Кожна з цих моделей надасть свої незалежні оцінки інших моделей (або ж самооцінку – залежно від експериментального дизайну) за 15 попередньо визначеними параметрами, такими як розуміння контексту, багатомовність, лексичне багатство, точність генерування та ін.

Табл. 2. Вагомості параметрів мовних моделей / Weights of parameters

12	Врахування знань та фактів	0,1510226055795294
13	Коригування граматики	0,15102260557952932
7	Розуміння контексту	0,14427477428180232
6	Якість генерування тексту	0,13381626539730092
8	Багатомовність	0,06572900311756194
9	Адаптація до завдань	0,06572900311756194
10	Лексичне багатство	0,06572900311756192
14	Безпека та упередження	0,05526992372165737
11	Швидкість генерування	0,05269089309149482
15	Апаратні вимоги	0,03305520337322098
3	Взаємодія з іншими системами (API)	0,017867924248544556
4	Вартість використання	0,017867924248544556
5	Дата оновлення знань	0,017867924248544556
2	Масштабованість	0,017867924248544553
1	Обсяг пам'яті	0,010187942627500064

Актуальність і доцільність такого підходу пояснюється кількома чинниками. По-перше, великі мовні моделі здатні моделювати експертну поведінку, тобто відтворювати оцінки на підставі тренування на масивних корпусах текстів академічного, технічного, літературознавчого спрямування. Це дає змогу залучити потенціал великих мовних моделей як метаекспертів у завданні саморефлексії або оцінювання своїх аналогів. По-друге, використання моделей як експертів дає можливість масштабувати оцінювання, зменшуючи залежність від людських ресурсів при збереженні структурованості процесу. По-третє, такий підхід дає змогу зіставити людське бачення ефективності використання мовних моделей із машинним – і виявити точки розбіжностей або збігів, що становить теоретичну цінність для когнітивної лінгвістики, філософії СШІ та критики машинної семантики.

На етапі реалізації кожна з 8 моделей-експертів оцінюватиме альтернативні моделі за кожним з 15 параметрів мовних моделей. Унаслідок чого для кожного параметра буде зібрано 8 незалежних матриць парних порівнянь. Наступним кроком є агрегування цих даних шляхом побудови єдиної групової матриці парних порівнянь для кожного параметра. Для цього використовують метод геометричного середнього, який дає змогу об'єднати оцінки моделей-експертів у збалансований вигляд, зберігаючи пропорції між ними.

Після побудови агрегованих матриць для кожного параметра застосовано метод головного власного вектора для обчислення локальних векторів вагомості альтернатив, а далі – формування фінальної матриці з усіх параметрів мовних моделей. З використанням вагомості параметрів мовних моделей (можна взяти ті самі, що були отримані від людських експертів, або ж сформувати нові на підставі оцінки важливості, сформовані мовними моделями) обчислено глобальний вектор вагомості альтернатив, тобто ранжування мовних моделей на підставі їхньої оцінки іншими мовними моделями.

Отже, результати дослідження дадуть змогу реалізувати двоїсту схему оцінювання: "людина → модель" та "модель → модель", що дає змогу виявити ступінь збіжності когнітивного та алгоритмічного бачення лінгвістичної ефективності мовних моделей, а також протестувати потенціал самих моделей у завданнях саморефлексивної оцінки, що стає дедалі актуальнішим в епоху автономних СШІ.

Для демонстрації процесу оцінювання великими мовними моделями ефективності своїх аналогів скористаємося прикладом параметра "розуміння контексту" – одного з найважливіших у лінгвістичних завданнях. Саме цей параметр охоплює здатність моделей підтримувати логічну зв'язність між частинами тексту, правильно інтерпретувати анафори, зберігати семантичну узгодженість діалогів та враховувати раніше подану інформацію у процес генерування нових фрагментів. Контекстна чутливість є вирішальною як у синтаксичному аналізі, так і у стилістичному, дискурсивному та когнітивному вимірах оброблення мови.

У межах цього прикладу вісім мовних моделей – GPT-4o, GPT-4.5, Gemini, Claude 3, DeepSeek R1, Grok-3, Qwen 2.5 та Mistral 7B – виступають як умовні експерти й кожна з них формує власну матрицю парних порівнянь для оцінки того, наскільки одна модель переважає іншу саме за параметром "розуміння контексту" (табл. 3-12). Кожна така матриця заповнюється значеннями за шкалою Сааті (від 1 до 9), виходячи з логіки самої моделі.

Після отримання восьми матриць парних порівнянь, вони об'єднуються в єдину агреговану матрицю шляхом обчислення середнього геометричного для кожної відповідної комірки. Цей метод дає змогу уникнути надмірного впливу оцінок окремих моделей-експертів і забезпечує збалансоване врахування всіх точок зору.

На підставі агрегованої матриці проводиться обчислення вектора локальних вагомості альтернатив – тобто визначення, яку частку ефективності (в межах саме цього параметра) має кожна модель. Власне, цей вектор демонструє загальну картину сприйняття контекстної компетенції однією групою великих мовних моделей стосовно іншої. Метод власного вектора дає змогу досягти найточнішого результату, з урахуванням внутрішньої логіки порівнянь та узгодженості агрегованої матриці попарних порівнянь.

У підсумку було отримано ранжування всіх моделей за параметром "розуміння контексту", побудоване на підставі оцінок самих мовних моделей, що стане відправною точкою для аналогічного розрахунку за рештою параметрів процесу оцінювання та, зрештою, для формування фінального вектора глобальних пріоритетів. Такий підхід дає змогу не тільки кількісно проаналізувати міжмодельну експертизу, а й емпірично протестувати ідею "самооцінки" та машинної аналітичної рефлексії (табл. 3-12).

Табл. 3. Матриця попарних порівнянь мовних моделей за версією Gemini / Matrix of pairwise comparisons of models according to Gemini

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1,100	1,200	1	1,300	1,100	2	4	1,500	1,400	1,200	1,100	1,200	1,300	1,400	1,500
2	0,909	1	1,100	0,909	1,200	1	1,900	3,800	1,400	1,300	1,100	1	1,100	1,200	1,300	1,400
3	0,833	0,909	1	0,833	1,100	0,909	1,800	3,600	1,300	1,200	1	0,909	1	1,100	1,200	1,300
4	1	1,100	1,200	1	1,300	1,100	2	4	1,500	1,400	1,200	1,100	1,200	1,300	1,400	1,500
5	0,769	0,833	0,909	0,769	1	0,833	1,700	3,400	1,200	1,100	0,900	0,833	0,909	1	1,100	1,200
6	0,909	1	1,100	0,909	1,200	1	1,900	3,800	1,400	1,300	1,100	1	1,100	1,200	1,300	1,400
7	0,500	0,526	0,556	0,500	0,588	0,526	1	2	0,750	0,700	0,600	0,526	0,556	0,588	0,636	0,667
8	0,250	0,263	0,278	0,250	0,294	0,263	0,500	1	0,375	0,350	0,300	0,263	0,278	0,294	0,318	0,333
9	0,667	0,714	0,769	0,667	0,833	0,714	1,333	2,667	1	0,933	0,800	0,714	0,769	0,833	0,909	1
10	0,714	0,769	0,833	0,714	0,909	0,769	1,429	2,857	1,071	1	0,857	0,769	0,833	0,909	0,923	1,071
11	0,833	0,909	1	0,833	1,111	0,909	1,667	3,333	1,250	1,167	1	0,909	1	1,111	1,222	1,250
12	0,909	1	1,100	0,909	1,250	1	1,905	3,810	1,400	1,300	1,100	1	1,111	1,250	1,364	1,400
13	0,833	0,909	1	0,833	1,111	0,909	1,800	3,600	1,300	1,200	1	0,909	1	1	1,100	1,200
14	0,769	0,833	0,909	0,769	1	0,833	1,700	3,400	1,200	1,091	0,818	0,800	0,909	1	1	1,100
15	0,714	0,769	0,833	0,714	0,909	0,769	1,571	3	1,100	1,083	0,818	0,733	0,909	1	1	1
16	0,667	0,714	0,769	0,667	0,833	0,714	1,500	3	1	1	0,800	0,714	0,833	0,909	1	1

Табл. 4. Матриця попарних порівнянь мовних моделей за версією DeepSeek R1 / Matrix of pairwise comparisons of models according to DeepSeek R1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	2	3	3	5	2	7	8	4	3	4	5	6	4	7	5
2	0,500	1	2	2	4	1	6	7	3	2	3	4	5	3	6	4
3	0,333	0,500	1	2	3	0,500	5	6	2	1	2	3	4	2	5	3
4	0,333	0,500	0,500	1	3	0,500	5	6	2	0,500	2	3	4	2	5	3
5	0,200	0,250	0,333	0,333	1	0,250	3	4	0,500	0,333	1	2	3	1	4	2
6	0,500	1	2	2	4	1	6	7	3	2	3	4	5	3	6	4
7	0,143	0,167	0,200	0,200	0,333	0,167	1	2	0,333	0,200	0,500	1	2	0,333	2	1
8	0,125	0,143	0,167	0,167	0,250	0,143	0,500	1	0,250	0,167	0,333	0,500	1	0,250	1	0,500
9	0,250	0,333	0,500	0,500	2	0,333	3	4	1	0,500	1	2	3	1	3	2
10	0,333	0,500	1	2	3	0,500	5	6	2	1	2	3	4	2	5	3
11	0,250	0,333	0,500	0,500	1	0,333	2	3	1	0,500	1	2	3	1	3	2
12	0,200	0,250	0,333	0,333	0,500	0,250	1	2	0,500	0,333	0,500	1	2	0,500	2	1
13	0,167	0,200	0,250	0,250	0,333	0,200	0,500	1	0,333	0,250	0,333	0,500	1	0,333	1	0,500
14	0,250	0,333	0,500	0,500	1	0,333	3	4	1	0,500	1	2	3	1	3	2
15	0,143	0,167	0,200	0,200	0,250	0,167	0,500	1	0,333	0,200	0,333	0,500	1	0,333	1	0,500
16	0,200	0,250	0,333	0,333	0,500	0,250	1	2	0,500	0,333	0,500	1	2	0,500	2	1

Табл. 5. Матриця попарних порівнянь мовних моделей за версією Grok 3 / Matrix of pairwise comparisons of models according to Grok 3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
2	2	1	3	3	2	2	4	5	2	2	2	2	5	4	6	7
3	0,500	0,333	1	1	0,500	0,500	3	3	0,500	0,500	0,500	0,500	3	3	5	5
4	0,500	0,333	1	1	0,500	0,500	3	3	0,500	0,500	0,500	0,500	3	3	5	5
5	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
6	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
7	0,333	0,250	0,333	0,333	0,333	0,333	1	2	0,333	0,333	0,333	0,333	2	2	3	4
8	0,250	0,200	0,333	0,333	0,250	0,250	0,500	1	0,250	0,250	0,250	0,250	2	2	3	3
9	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
10	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
11	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
12	1	0,500	2	2	1	1	3	4	1	1	1	1	4	3	5	6
13	0,250	0,200	0,333	0,333	0,250	0,250	0,500	0,500	0,250	0,250	0,250	0,250	1	1	3	3
14	0,333	0,250	0,333	0,333	0,333	0,333	0,500	0,500	0,333	0,333	0,333	0,333	1	1	3	3
15	0,200	0,167	0,200	0,200	0,200	0,200	0,333	0,333	0,200	0,200	0,200	0,200	0,333	0,333	1	2
16	0,167	0,143	0,200	0,200	0,167	0,167	0,250	0,333	0,167	0,167	0,167	0,167	0,333	0,333	0,500	1

Табл. 6. Матриця попарних порівнянь мовних моделей за версією GPT-4.5 /
Matrix of pairwise comparisons of models according to GPT-4.5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	2	4	3	5	2	7	8	7	3	6	5	4	3	8	8
2	0,500	1	3	2	4	1	6	7	6	2	5	4	3	2	7	7
3	0,250	0,333	1	0,500	2	0,333	4	5	4	0,500	3	2	2	1	5	5
4	0,333	0,500	2	1	3	0,500	5	6	5	1	4	3	2	1	6	6
5	0,200	0,250	0,500	0,333	1	0,250	3	4	3	0,333	2	1	1	0,500	4	4
6	0,500	1	3	2	4	1	6	7	6	2	5	4	3	2	7	7
7	0,143	0,167	0,250	0,200	0,333	0,167	1	2	1	0,200	0,500	0,250	0,250	0,200	2	2
8	0,125	0,143	0,200	0,167	0,250	0,143	0,500	1	0,500	0,167	0,333	0,200	0,200	0,167	1	1
9	0,143	0,167	0,250	0,200	0,333	0,167	1	2	1	0,200	0,500	0,250	0,250	0,200	2	2
10	0,333	0,500	2	1	3	0,500	5	6	5	1	4	2	2	1	6	6
11	0,167	0,200	0,333	0,250	0,500	0,200	2	3	2	0,250	1	0,500	0,500	0,250	3	3
12	0,200	0,250	0,500	0,333	1	0,250	4	5	4	0,333	2	1	1	0,500	4	4
13	0,250	0,333	0,500	0,500	1	0,333	4	5	4	0,500	2	1	1	1	5	5
14	0,333	0,500	1	1	2	0,500	5	6	5	1	4	2	1	1	6	6
15	0,125	0,143	0,200	0,167	0,250	0,143	0,500	1	0,500	0,167	0,333	0,250	0,333	0,333	1	1
16	0,125	0,143	0,200	0,167	0,250	0,143	0,500	1	0,500	0,167	0,333	0,250	0,333	0,333	1	1

Табл. 7. Матриця попарних порівнянь мовних моделей за версією GPT-4o /
Matrix of pairwise comparisons of models according to GPT-4o

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	3	4	5	2	9	9	7	8	6	9	9	9	9	9
2	1	1	2	3	4	1	9	9	6	7	5	8	9	9	9	9
3	0,333	0,500	1	1	2	1	7	9	4	5	3	6	8	9	9	9
4	0,250	0,333	1	1	1	0,500	6	9	3	4	2	5	7	8	9	9
5	0,200	0,250	0,500	1	1	0,333	5	8	2	3	1	4	6	7	9	9
6	0,500	1	1	2	3	1	4	5	4	4	3	4	4	4	9	9
7	0,111	0,111	0,143	0,167	0,200	0,250	1	1	0,143	0,200	0,167	0,111	0,125	0,111	1	1
8	0,111	0,111	0,111	0,111	0,125	0,200	1	1	0,125	0,200	0,333	0,200	0,500	0,143	1	1
9	0,143	0,167	0,250	0,333	0,500	0,250	7	8	1	1	0,333	0,125	0,250	0,111	0,111	0,111
10	0,125	0,143	0,200	0,250	0,333	0,250	5	5	1	1	0,167	0,200	0,250	0,125	0,167	0,167
11	0,167	0,200	0,333	0,500	1	0,333	6	3	3	6	1	0,500	0,500	0,333	0,333	0,333
12	0,111	0,125	0,167	0,200	0,250	0,250	9	5	8	5	2	1	1	0,500	0,250	0,167
13	0,111	0,111	0,125	0,143	0,167	0,250	8	2	4	4	2	1	1	1	0,333	0,200
14	0,111	0,111	0,111	0,125	0,143	0,250	9	7	9	8	3	2	1	1	0,167	0,167
15	0,111	0,111	0,111	0,111	0,111	0,111	1	1	9	6	3	4	3	6	1	1
16	0,111	0,111	0,111	0,111	0,111	0,111	1	1	9	6	3	6	5	6	1	1

Табл. 8. Матриця попарних порівнянь мовних моделей за версією Claude 3 /
Matrix of pairwise comparisons of models according to Claude 3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	3	4	3	1	7	8	4	5	5	6	6	6	9	7
2	0,500	1	2	1	2	0,500	6	7	3	4	4	5	5	5	8	6
3	0,333	0,500	1	0,500	1	0,333	5	6	2	3	3	4	4	4	7	5
4	0,500	1	2	1	2	0,500	6	7	3	4	4	5	5	5	8	6
5	0,333	0,500	1	0,500	1	0,333	5	6	2	3	3	4	4	4	7	5
6	1	2	3	2	3	1	7	8	4	5	5	6	6	6	9	7
7	0,143	0,167	0,200	0,167	0,200	0,143	1	2	0,250	0,333	0,333	0,500	0,500	0,500	3	1
8	0,125	0,143	0,167	0,143	0,167	0,125	0,500	1	0,200	0,250	0,250	0,333	0,333	0,333	2	0,500
9	0,250	0,333	0,500	0,333	0,500	0,250	4	5	1	2	2	3	3	3	6	4
10	0,200	0,250	0,333	0,250	0,333	0,200	3	4	0,500	1	1	2	2	2	5	3
11	0,200	0,250	0,333	0,250	0,333	0,200	3	4	0,500	1	1	2	2	2	5	3
12	0,167	0,200	0,250	0,200	0,250	0,167	2	3	0,333	0,500	0,500	1	1	1	4	2
13	0,167	0,200	0,250	0,200	0,250	0,167	2	3	0,333	0,500	0,500	1	1	1	4	2
14	0,167	0,200	0,250	0,200	0,250	0,167	2	3	0,333	0,500	0,500	1	1	1	4	2
15	0,111	0,125	0,143	0,125	0,143	0,111	0,333	0,500	0,167	0,200	0,200	0,250	0,250	0,250	1	0,333
16	0,143	0,167	0,200	0,167	0,200	0,143	1	2	0,250	0,333	0,333	0,500	0,500	0,500	3	1

Табл. 9. Матриця попарних порівнянь мовних моделей за версією Qwen 2.5 /
Matrix of pairwise comparisons of models according to Qwen 2.5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	2	4	3	5	2	6	7	3	4	3	5	6	7	8	6
2	0,500	1	3	2	4	1	5	6	2	3	2	4	5	6	7	5
3	0,250	0,333	1	0,500	2	0,333	3	4	0,500	1	0,500	2	3	4	5	3
4	0,333	0,500	2	1	3	0,500	4	5	1	2	1	3	4	5	6	4
5	0,200	0,250	0,500	0,333	1	0,250	2	3	0,333	0,500	0,333	2	3	4	5	3
6	0,500	1	3	2	4	1	5	6	2	3	2	4	5	6	7	5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
7	0,167	0,200	0,333	0,250	0,500	0,200	1	2	0,250	0,333	0,250	0,500	1	2	3	1
8	0,143	0,167	0,250	0,200	0,333	0,167	0,500	1	0,200	0,250	0,200	0,333	0,500	1	2	0,500
9	0,333	0,500	2	1	3	0,500	4	5	1	2	1	3	4	5	6	4
10	0,250	0,333	1	0,500	2	0,333	3	4	0,500	1	0,500	2	3	4	5	3
11	0,333	0,500	2	1	3	0,500	4	5	1	2	1	3	4	5	6	4
12	0,200	0,250	0,500	0,333	0,500	0,250	2	3	0,333	0,500	0,333	1	2	3	4	2
13	0,167	0,200	0,333	0,250	0,333	0,200	1	2	0,250	0,333	0,250	0,500	1	2	3	1
14	0,143	0,167	0,250	0,200	0,250	0,167	0,500	1	0,200	0,250	0,200	0,333	0,500	1	2	0,500
15	0,125	0,143	0,200	0,167	0,200	0,143	0,333	0,500	0,167	0,200	0,167	0,250	0,333	0,500	1	0,333
16	0,167	0,200	0,333	0,250	0,333	0,200	1	2	0,250	0,333	0,333	0,500	1	2	3	1

Табл. 10. Матриця попарних порівнянь мовних моделей за версією Mistral 7B / Matrix of pairwise comparisons of models according to Mistral 7B

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	3	5	7	5	4	6	7	8	6	5	6	7	8	9	7
2	0,333	1	3	5	4	3	5	6	7	5	4	5	6	7	8	6
3	0,200	0,333	1	3	2	2	4	5	6	4	3	4	5	6	7	5
4	0,143	0,200	0,333	1	2	2	4	5	6	4	3	4	5	6	7	5
5	0,200	0,250	0,500	0,500	1	2	3	4	5	3	2	3	4	5	6	4
6	0,250	0,333	0,500	0,500	0,500	1	2	3	4	2	2	3	4	5	6	4
7	0,167	0,200	0,250	0,250	0,333	0,500	1	2	3	2	2	3	4	5	4	3
8	0,143	0,167	0,200	0,200	0,250	0,333	0,500	1	2	2	2	3	4	5	6	4
9	0,125	0,143	0,167	0,167	0,200	0,250	0,333	0,500	1	1	0,500	0,500	0,500	0,500	2	0,500
10	0,167	0,200	0,250	0,250	0,333	0,500	0,500	0,500	1	1	0,500	0,500	0,500	0,500	2	0,500
11	0,200	0,250	0,333	0,333	0,500	0,500	0,500	0,500	2	2	1	2	3	4	5	3
12	0,167	0,200	0,250	0,250	0,333	0,333	0,333	0,333	2	2	0,500	1	2	3	4	2
13	0,143	0,167	0,200	0,200	0,250	0,250	0,250	0,250	2	2	0,333	0,500	1	2	3	2
14	0,125	0,143	0,167	0,167	0,200	0,200	0,200	0,200	2	2	0,250	0,333	0,500	1	2	0,500
15	0,111	0,125	0,143	0,143	0,167	0,167	0,250	0,167	2	2	0,200	0,250	0,333	0,500	1	0,500
16	0,143	0,167	0,200	0,200	0,250	0,250	0,333	0,250	2	2	0,333	0,500	0,500	2	2	1

Табл. 11. Агрегована матриця попарних порівнянь мовних моделей / Aggregated matrix of pairwise comparisons of models

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,142	0,142	0,162	0,095	0,095	0,119	0,130	0,109	0,113	0,156	0,163	0,039	0,117	0,117	0,080
2	0,113	0,113	0,260	0,087	0,087	0,102	0,103	0,118	0,100	0,128	0,137	0,047	0,132	0,132	0,065
3	0,059	0,035	0,056	0,027	0,035	0,034	0,036	0,033	0,032	0,041	0,034	0,097	0,098	0,098	0,097
4	0,067	0,057	0,095	0,087	0,087	0,062	0,061	0,062	0,061	0,081	0,074	0,058	0,132	0,132	0,086
5	0,049	0,035	0,023	0,025	0,032	0,038	0,039	0,029	0,034	0,032	0,037	0,034	0,070	0,070	0,051
6	0,106	0,106	0,074	0,087	0,081	0,075	0,071	0,082	0,079	0,109	0,109	0,075	0,096	0,096	0,078
7	0,019	0,010	0,017	0,010	0,016	0,017	0,017	0,017	0,018	0,023	0,022	0,175	0,042	0,042	0,064
8	0,014	0,008	0,027	0,006	0,014	0,015	0,014	0,016	0,016	0,020	0,019	0,217	0,033	0,033	0,039
9	0,037	0,057	0,013	0,054	0,035	0,041	0,044	0,039	0,035	0,044	0,043	0,109	0,073	0,073	0,086
10	0,043	0,044	0,013	0,054	0,035	0,042	0,038	0,034	0,038	0,053	0,032	0,087	0,058	0,058	0,051
11	0,041	0,057	0,027	0,054	0,034	0,044	0,044	0,041	0,043	0,058	0,046	0,089	0,088	0,088	0,108
12	0,035	0,035	0,074	0,087	0,054	0,054	0,049	0,048	0,049	0,071	0,063	0,054	0,050	0,050	0,051
13	0,025	0,028	0,037	0,054	0,029	0,039	0,038	0,036	0,043	0,046	0,041	0,057	0,046	0,046	0,051
14	0,029	0,028	0,013	0,054	0,029	0,023	0,022	0,023	0,032	0,036	0,021	0,043	0,033	0,033	0,039
15	0,013	0,007	0,007	0,010	0,013	0,017	0,016	0,017	0,021	0,025	0,017	0,109	0,032	0,032	0,051
16	0,019	0,028	0,013	0,027	0,020	0,036	0,038	0,029	0,032	0,037	0,030	0,087	0,042	0,042	0,095

Табл. 12. Результуючий зважений перелік великих мовних моделей / Resulting weighted list of models

Модель	Вагомість	Модель	Вагомість
GPT-4.5	0,1347	Nemotron-4 340B	0,0487
GPT-4o	0,1306	LLaMA 2	0,0476
Claude 3	0,0945	Phi-2	0,0469
Gemini	0,0893	LaMDA 3	0,0442
Qwen2.5	0,0607	Mistral 7B	0,0438
Llama 3.1	0,0587	Grok-3	0,0421
DeepSeek R1	0,0586	Megatron-Turing NLG 530B	0,0328
PaLM 2	0,0568	BLOOM	0,0280

Примітка: індекс узгодженості (*CI*) – 0,0341, коефіцієнт узгодженості (*CR*) – 0,0214.

Обговорення отриманих результатів. Після обчислення всіх 15 локальних векторів вагомості, що відображають ефективність кожної мовної моделі в межах окремого параметра, аналогічно до попереднього, формуємо агреговану матрицю парних порівнянь параметрів мовних моделей та обчислимо фінальний вектор вагомості моделей.

Фінальні вагомості моделей, отримані шляхом множення матриці вагомості моделей за параметрами на вектор вагомості параметрів мовних моделей, відображають їхню загальну ефективність. У підсумку було обґрунтовано отримані результати, проаналізовано ключові фактори, що вплинули на рейтинг, та пояснено позиції моделей, зокрема – низьку позицію Grok-3.

Метод MAI базується на парному порівнянні альтернатив (моделей) за кожним параметром та параметрів між собою. Для кожного з 15 параметрів мовної моделі було сформовано матриці парного порівняння розміром 16×16, які потім агреговані за допомогою геометричного середнього. Вектори вагомості моделей для кожного

параметра були нормалізовані (сума = 1), а узгодженість матриць перевірена за допомогою коефіцієнта узгодженості ($CR < 0,1$), що підтверджує надійність даних. Вектор вагомості параметрів мовної моделі, який відображає їхню відносну важливість, також був нормалізований (сума ≈ 1). Фінальні вагомості моделей обчислені як добуток матриці рішень (16×15) на вектор вагомості параметрів мовної моделі, з подальшою нормалізацією.

Фінальний (підсумковий, результуючий) вектор вагомості показує такий рейтинг моделей:

- лідери: GPT-4.5 (0,1347), GPT-4o (0,1306), Claude 3 (0,0945);
- середній сегмент: Gemini (0,0893), Qwen 2.5 (0,0607), DeepSeek R1 (0,0586), Llama 3.1 (0,0587);
- нижній сегмент: Grok-3 (0,0421), Mistral 7B (0,0438), Phi-2 (0,0469), Nemotron-4 340B (0,0487), PaLM 2 (0,0568), LaMDA 3 (0,0442), Megatron-Turing NLG 530B (0,0328), BLOOM (0,0280), LLaMA 2 (0,0476).

Лідери рейтингу (GPT-4.5, GPT-4o, Claude 3) отримали високі оцінки за ключовими параметрами з великою вагомістю, такими як параметр "Коригування граматики" (GPT-4.5: 0,2599, GPT-4o: 0,1617), "Розуміння контексту" (GPT-4o: 0,1419, Claude 3: 0,1063) та параметр "Якість генерування тексту" (GPT-4o: 0,1419, Claude 3: 0,1063). Ці параметри мають вагомості 0,1510, 0,1443 та 0,1338 відповідно, що становить значну частку загальної оцінки (~43 %). Високі оцінки за цими аспектами відображають сильні функціональні можливості цих моделей у генерування якісного, контекстуально правильного тексту, що відповідає очікуванням від провідних моделей, розроблених організаціями з великими ресурсами (OpenAI, Anthropic).

Середній сегмент (Gemini, Qwen2.5, DeepSeek R1, Llama 3.1) демонструє збалансовані оцінки, але з нижчими значеннями за основними параметрами. Наприклад, Qwen 2.5 має 0,1082 за параметром "Швидкість генерування" (вагомість параметра 0,0527), але тільки 0,0269 за параметром "Коригування граматики", що обмежило його загальну вагомість. DeepSeek R1 і Llama 3.1 отримали помірні оцінки за "Вартість" (0,0972 і 0,1088), але їхні оцінки за "Розуміння контексту" (0,0588 і 0,0367) поступаються лідерам.

Нижній сегмент, зокрема – Grok-3, отримав низькі оцінки за ключовими параметрами. Grok-3 має вагомість 0,0421, що пояснюється слабкими оцінками за параметром "Коригування граматики" (0,0225), за параметром "Розуміння контексту" (0,0485) та за параметром "Якість генерування тексту" (0,0349). Хоча Grok-3 показує конкурентні результати за параметрами з нижчою вагомістю, такими як "Дата оновлення" (0,0699) чи "Апаратні вимоги" (0,0699), ці параметри (вагомість 0,0179) мають незначний вплив на фінальний результат. Низькі оцінки можуть відображати меншу впізнаваність Grok-3 як нової моделі від xAI або специфіку її дизайну, орієнтовану на креативність і неформальну взаємодію, а не на формальні аспекти, як граматична точність.

Результати оцінювання мовних моделей є логічним відображенням пріоритетів, заданих вагомістю параметрів мовної моделі, та оцінок моделей у матрицях парного порівняння. Лідерство GPT-4.5, GPT-4o і Claude 3 пояснюється їхньою високою продуктивністю за параметрами, що мають найбільшу вагомість, такими як граматика, контекст і якість тексту. Grok-3, хоча й має потенціал у технічних аспектах, відстає через низькі оцінки за основними функціональними параметрами. Ці результати можуть бути використані для вибору моделей залежно від конкретних потреб, наприклад, GPT-4.5 для завдань із високими вимогами до граматики, або Qwen2.5 для швидкості генерування. Надалі можна провести аналіз чутливості, щоб оцінити, як зміна вагомості параметрів мовної моделі вплине на рейтинг, або детальніше дослідити продуктивність Grok-3 для потенційного вдосконалення його оцінок.

Обговорення результатів дослідження. У дослідженні [31] запропоновано підхід до підвищення якості генерування тексту великими мовними моделями за рахунок механізмів самооцінки. Автори показали, що здатність моделі оцінювати власні відповіді забезпечує вищу кореляцію з якістю результату, ніж використання традиційних метрик. Це відкриває перспективи для більш надійних систем, у яких LLM самостійно відсіюють менш точні варіанти відповідей. Такий підхід є показовим у контексті використання мовних моделей у складних аналітичних завданнях.

Подібні ідеї розвиває робота [3], де розроблено фреймворк для адаптації LLM через інтегровані модулі самооцінки. Система дає змогу моделі прогнозувати ступінь власної впевненості у правильності відповіді, що значно покращує вибір коректних результатів. Автори продемонстрували переваги такого підходу для складних завдань оброблення природної мови. Це підтверджує потенціал LLM у створенні механізмів надійного контролю якості вихідних даних.

Особливо цікавим є дослідження [11], яке поєднало аналітичний ієрархічний процес із донавчанням великих мовних моделей. Автори показали, що такий гібридний підхід дає змогу будувати ієрархії рішень навіть у ситуаціях невизначеності, де класичний АІП функціонує недостатньо ефективно. Важливо, що мовні моделі у цьому випадку використовують як додатковий аналітичний інструмент, а не тільки як генеративну систему. Це підкреслює перспективність інтеграції методів багатокритеріального аналізу з LLM.

У роботі [15] представлено систему OE-Assist, яка поєднує автоматизовану й напівавтоматизовану оцінки онтологій за допомогою великих мовних моделей. Автори доводять, що такі моделі можуть ефективно виконувати CQ-верифікацію та підсилювати роботу експертів-онтологів. Це відкриває нові горизонти для використання LLM не тільки у створенні, а й у перевірці складних знаннєвих структур.

Важливим прикладом тестування моделей у прикладних завданнях є дослідження [12], у якому перевіряють математичну й програмувальну компетентність LLM на спеціально модифікованих наборах даних. Автор показує, що навіть невеликі зміни в умовах завдання можуть призводити до істотних коливань якості відповідей. Це свідчить про вразливість моделей до варіацій формулювання і наголошує на необхідності систем оцінювання стійкості. Такий висновок підтверджує потребу в багаторівневому аналізі та порівнянні.

Унаслідок обговорення результатів дослідження було з'ясовано, що хоча у сучасній літературі розглядають різні аспекти самооцінювання LLM, інтеграцію їх у багатокритеріальні методи, зокрема – в аналітичний ієрархічний процес, або майже не висвітлюють, або має фрагментарний характер. Проаналізовані праці підтверджують актуальність теми, але жодна з них не поєднує групове оцінювання з використанням самих мовних мо-

делей як експертів у системі ранжування. Саме тому проведене дослідження є новим кроком у розвитку методології, оскільки запропонована інформаційна технологія самоаналізу не тільки забезпечує кількісну оцінку, а й відкриває можливості для метакогнітивної рефлексії ШІ. Отримані результати підтверджують доцільність виконаного дослідження та його унікальність порівняно з уже наявними науковими напрацюваннями.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – удосконалено методику самоаналізу параметрів великих мовних моделей, яка, на відміну від наявної, поєднує аналітичний ієрархічний процес із груповим оцінюванням та механізмом самооцінювання, що забезпечує інтегральний підхід до їх порівняльного аналізу.

Практична значущість результатів дослідження – результати дослідження використано для створення методологічної основи універсальної системи самоаналізу параметрів великих мовних моделей, яку можна застосувати в лінгвістичних, гуманітарних, технічних, медичних, юридичних та освітніх системах, оскільки вона забезпечує комплексне оцінювання ефективності використання моделей за широким спектром параметрів процесу оцінювання.

Висновки / Conclusions

Розроблено універсальну інформаційну технологію для оцінювання великих мовних моделей щодо ефективності виконання лінгвістичних завдань з інтеграцією їх самооцінювання, що дало змогу виявити їх переваги та недоліки, обґрунтувати рекомендаціъ щодо доцільності їх оптимального вибору. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Розроблено інформаційну технологію самоаналізу великих мовних моделей, яка базується на груповому методі аналізу ієрархій та інноваційному підході до самооцінювання, де моделі виступають як експерти.
2. Систематизовано етапи технології (формування списків моделей і параметрів процесу оцінювання, опитування моделей, обчислення за груповим MAI, створення впорядкованого списку), що забезпечує структурований, прозорий і адаптивний процес оцінювання для будь-якої предметної області.
3. Проведено ранжування великих мовних моделей за їхньою загальною ефективністю, де лідерські позиції посіли GPT-4.5 (0,1347), GPT-4o (0,1306) та Claude 3 (0,0945).
4. Встановлено, що параметри з найвищими вагомостями – коригування граматики (0,1510), розуміння контексту (0,1443) та якість генерування тексту (0,1338) – становлять близько 43 % загальної оцінки, що підтверджує їх критичну роль у лінгвістичних завданнях.
5. Проаналізовано результати середнього сегмента (Gemini, Qwen 2.5, Llama 3.1, DeepSeek R1), які показали збалансовані характеристики, та нижнього сегмента (Grok-3, Mistral 7B, BLOOM), які продемонстрували нижчу ефективність за ключовими параметрами.
6. З'ясовано, що інноваційний підхід до самооцінювання, де самі мовні моделі виконували роль експертів, відкрив нові перспективи для дослідження їхнього потенціалу до саморефлексії та дав змогу масштабувати процес оцінювання за мінімізації залежності від людських ресурсів.
7. Доведено, що застосування групового MAI є ефективним інструментом для інтеграції різновекторних оцінок від людських експертів та мовних моделей, забезпечуючи кількісну точність та якісну обґрунтованість рішень.
8. Виявлено, що результати дослідження мають практичне значення для гуманітарних і медичних застосувань, зокрема – у сфері оцінювання кваліфікації персоналу, а також для розроблення інформаційних систем, орієнтованих на багатомовність, стилістичну гнучкість і зменшення упереджень.
9. Окреслено перспективи подальших досліджень, які полягають у вивченні стабільності результатів за зміни вагомості параметрів мовної моделі, аналізу вузькоспеціалізованих завдань (наприклад, робота з історичними текстами чи побудова онтологій), а також розширення методів тестування ШІ з акцентом на когнітивну та етичну рефлексії.

References

1. Anthropic Team. (2023). The Claude 3 model family. *Anthropic*. URL: https://www.anthropic.com/claude-3-model-card
2. BigScience, & HuggingFace. (2022). BLOOM – BigScience large open multilingual model. *HuggingFace*. URL: https://bigscience.huggingface.co/blog/bloom
3. Chen, J., Yoon, J., Ebrahimi, S., Arik, S., Pfister, T., & Jha, S. (2023). Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5190–5213. https://doi.org/10.18653/v1/2023.findings-emnlp.345
4. Collins, E., & Ghahramani, Z. (2021). LaMDA: Our breakthrough conversation technology. *Google*. URL: https://blog.google/technology/ai/lamda/
5. DeepSeek-AI, Zhu, Q., & Guo, D. (2024). DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence. *arXiv*. https://doi.org/10.48550/arXiv.2406.11931
6. Gemini Team, Google. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*. https://doi.org/10.48550/arXiv.2403.05530
7. Georgiev, P., Lei, V. I., & Burnell, R. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*. https://doi.org/10.48550/arXiv.2403.05530
8. Giglou, H. B., DSouza, J., & Auer, S. (2023). LLMs4OL: Large language models for ontology learning. In *The Semantic Web – ISWC* (pp. 408–427). https://doi.org/10.1007/978-3-031-47240-4_22
9. Google. (2021). LaMDA: Our breakthrough conversation technology. *Google*. URL: https://blog.google/technology/ai/lamda/
10. Google. (2023). Introducing PaLM 2. *Google*. URL: https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
11. Haeun Park, H., Oh, H., & Gao, F., & Kwon, O. (2025). Enhancing Analytic Hierarchy Process Modelling Under Uncertainty With Fine-Tuning LLM. *Expert Systems*, 42, article ID e70051. https://doi.org/10.1111/exsy.70051
12. Hong, P. (2024). Evaluating LLMs Mathematical and Coding Competency with Perturbed Datasets. *ACL Findings*. URL: https://arxiv.org/abs/2401.09395
13. Hu, Y., Liu, D., & Wang, Q. (2024). Automating knowledge discovery from scientific literature via LLMs: A dual-agent approach. *arXiv*. https://doi.org/10.48550/arXiv.2409.00054
14. Kunanets, N. E., & Yaromych, M. V. (2025). Concept extraction in literary texts using large language models. *Bulletin of Science and Education*, *32*(2), 343–357. https://doi.org/10.52058/2786-6165-2025-2(32)-343-357
15. Lippolis, A. S., Saeedizade, M. J., Keskisärkkä, R., Gangemi, A., Blomqvist, E., & Nuzzolese, A. G. (2025). Large Language Models Assisting Ontology Evaluation. *OE-Assist framework*. URL: https://arxiv.org/abs/2507.14552
16. Meta AI. (2024). Introducing Llama 3.1: Our most capable models to date. *Meta*. URL: https://ai.meta.com/blog/meta-llama-3-1/
17. Microsoft Research. (2023). Phi-2: The surprising power of small language models. *Microsoft*. URL: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/
18. Mistral AI. (2023). Mistral 7B. *Mistral*. URL: https://mistral.ai/news/announcing-mistral-7b
19. Mukanova, A., Milosz, M., & Dauletkaliyeva, A. (2024). LLM-powered natural language text processing for ontology en-

richment. *Applied Sciences*, *14*(13), 5860–5875. https://doi.org/10.3390/app14135860
20. Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. *arXiv*. https://doi.org/10.48550/arXiv.2403.16248
21. Neuhaus, F. (2024). Ontologies in the era of large language models – a perspective. *Applied Ontology*, *18*(4), 399–407. https://doi.org/10.3233/AO-230072
22. NVIDIA. (2024). Nemotron-4 340B technical report. *arXiv*. URL: https://arxiv.org/html/2406.11704v1
23. Opara, C. (2024). StyloAI: Distinguishing AI-generated content with stylometric analysis. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education* (pp. 1–12). https://doi.org/10.48550/arXiv.2405.10129
24. OpenAI. (2024). Introducing GPT-4o and more tools to ChatGPT free users. *OpenAI*. URL: https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/
25. OpenAI. (2024). OpenAI platform. *OpenAI*. URL: https://platform.openai.com/docs/api-reference/introduction
26. Pasichnyk, V. V., & Yaromych, M. V. (2025). Automated generation of technical documentation in IT using large language models. *Studia Methodologica*, 59, 250–273. https://doi.org/10.32782/2307-1222.2025-59-22
27. Pasichnyk, V. V., & Yaromych, M. V. (2025). Features of genre classification of literature using large language models. *Folium*, *6*, 132–144. https://doi.org/10.32782/folium/2025.6.19
28. Pasichnyk, V. V., & Yaromych, M. V. (2025). Genre classification of literature by metrics using large language models. *Scientific Works of the Interregional Academy of Personnel Management. Philology*, *1*(15), 60–68. https://doi.org/10.32689/maup.philol.2025.1.11
29. Pasichnyk, V. V., & Yaromych, M. V. (2025). Large language models and ontologies in philological research: An analytical review of sources. *Current Issues of the Humanities. Linguistics. Literary Studies*, *83*(3), 236–250. https://doi.org/10.24919/2308-4863/83-3-35
30. Qwen Team. (2024). Qwen2-0.5B-Instruct. *HuggingFace*. URL: https://huggingface.co/Qwen/Qwen2-0,5B-Instruct/blob/main/README.md
31. Ren, J., Zhao, Y., Vu, T., Liu, P. J., & Lakshminarayanan, B. (2023). Self-Evaluation Improves Selective Generation in Large Language Models. *Proceedings on "I Cant Believe Its Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, in *Proceedings of Machine Learning Research*, 239, 49–64. URL: https://proceedings.mlr.press/v239/ren23a.html
32. Saaty, T. L. (1980). *The analytic hierarchy process*. New York, NY: McGraw-Hill. URL: https://www.researchgate.net/publication/362349026_The_Analytic_Hierarchy_Process
33. Saaty, T. L. (1989). Group decision making and the analytic hierarchy process. *European Journal of Operational Research*, *48*(1), 9–26. URL: https://ideas.repec.org/a/eee/ejores/v48y1990i1p9-26.html
34. Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, *1*(1), 83–98. https://doi.org/10.1504/IJSSCI.2008.017590
35. Schreiber, H. (2016). Genre ontology learning: Comparing curated with crowd-sourced folksonomies. In *Proceedings of the 17th International Society for Music Information Retrieval Conference* (pp. 400–406). URL: https://archives.ismir.net/ismir2016/paper/000074.pdf
36. Smith, S., Patwary, M., & Morick, B. (2022). Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model. *arXiv*. https://doi.org/10.48550/arXiv.2201.11990
37. Tkachenko, O. I., Tkachenko, K. O., & Tkachenko, O. A. (2021). Linguistic ontologies: Designing and using in the educational intellectual systems. *Digital Platform: Information Technologies in Sociocultural Sphere*, *4*(1), 97–111. https://doi.org/10.31866/2617-796X.4.1.2021.236950
38. Touvron, H., Martim, L., & Stone, K. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv*. https://doi.org/10.48550/arXiv.2307.09288
39. Umphrey, R., Roberts, J., & Roberts, L. (2024). Investigating expert-in-the-loop LLM discourse patterns for ancient intertextual analysis. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (*NLP4DH 2024*) (pp. 31–41). URL: https://arxiv.org/abs/2409.01882
40. xAI. (2023). Grok. *xAI*. URL: https://x.ai/grok

V. V. Pasichnyk, M. V. Yaromych, I. I. Pavliv, N. E. Kunanets

Lviv Polytechnic National University, Lviv, Ukraine

AN INFORMATION TECHNOLOGY FOR THE SELF-ANALYSIS OF LARGE LANGUAGE MODEL PARAMETERS

The article is devoted to the study of the performance of large language models in solving linguistic tasks. A wide range of models, including leading developments from well-known organizations, was evaluated according to a number of parameters such as contextual understanding, quality of text generation, grammar correction, multilingual capabilities, and lexical richness. Special attention is paid to the methodology of the group analytic hierarchy process, which made it possible to integrate evaluations from different expert groups, including the language models themselves acting as experts. Self-assessment was implemented using the analytic hierarchy process and an innovative approach to self-evaluation through the construction of pairwise comparison matrices, in which models evaluated each other according to defined parameters, and the resulting data were aggregated using the geometric mean. Leaders were identified that demonstrated high performance in key aspects such as grammatical accuracy, contextual sensitivity, and text quality, while mid-range models showed balanced capabilities, and some fell into the lower tier due to limited effectiveness in functional tasks. The consistency of the assessments confirms the reliability of the data obtained. The study highlights the importance of selecting models in accordance with the specifics of linguistic tasks, offering recommendations for their application in tasks requiring high accuracy or processing speed. It was found that self-assessment reveals the potential of models for self-reflection, which is significant for cognitive linguistics and the philosophy of artificial intelligence. The results indicate the need for improvement of models in such aspects as multilingual support and bias reduction, especially for cross-cultural contexts. It has been established that methodologically the analytic hierarchy process is effective in harmonizing interdisciplinary assessments by combining quantitative accuracy with qualitative validity. The study characterizes patterns that shape a new paradigm for analyzing the capabilities of artificial intelligence in the medical domain, offering a tool for selecting optimal large language models and addressing the feasibility of autonomous use of AI systems and the distribution of responsibility between humans and machines. Future research prospects include studying the stability of results when analyzing with different parameter priorities, analyzing model performance in specific tasks such as assessing the qualifications of medical professionals using both structured and unstructured data. The article emphasizes the role of information technologies in transforming linguistic data into structured knowledge, contributing to a deeper understanding of the potential and limitations of language models in the digital humanities and their potential implementation in a medical information system.

Keywords: analytic hierarchy process; self-assessment; context understanding; multilingualism; artificial intelligence systems.