

Є. М. Крижановський¹І. М. Штельмах¹М. В. Швець¹О. О. Войцеховська¹

МЕТОД ТА ЗАСІБ АНАЛІЗУ ЕФЕКТИВНОСТІ РОБОТИ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ВИКОНАННЯ ЗАДАЧ КОМПЛЕКСНОГО ХАРАКТЕРУ

¹Вінницький національний технічний університет

Проаналізовано сучасні підходи до оцінки ефективності великих мовних моделей для виконання комплексних задач, які включають одночасний аналіз нормативно-правових документів та числових даних. Розглянуто проблеми інтеграції різнорідної інформації, забезпечення логічної узгодженості результатів та порівняння моделей за різними критеріями: коректністю, стабільністю, часом обробки та дотриманням інструкцій. Актуальність дослідження зумовлена швидким розвитком великих мовних моделей та необхідністю вибору оптимальних рішень для задач прикладного характеру, де помилки можуть мати суттєві наслідки.

Основу методу становить багатокритеріальний підхід, який передбачає одночасний аналіз результатів моделей на комплексному запиті: обробка нормативних документів та CSV-даних із середньодобовими концентраціями забруднювачів повітря. Для оцінки результатів використано показники коректності, стабільності, часу виконання, вартості, використання токенів та дотримання інструкцій.

Запропоновано математичну модель агрегування результатів, що включає усереднення показників по множині тестових задач, виділення множини Парето-оптимальних моделей, побудову скаляризованої функції корисності для визначення оптимальної конфігурації моделі залежно від заданих вагових коефіцієнтів.

Апробація методу на прикладі міста Вінниці показала, що різні моделі мають суттєві відмінності у здатності формувати узгоджені висновки. На основі аналізу моделями нормативних порогів визначено оптимальну кількість пунктів спостереження для різних забруднювачів, що забезпечує репрезентативний моніторинг якості повітря. Запропонований підхід дозволяє враховувати специфіку комплексних прикладних задач, інтегрувати різнорідні дані та підвищує обґрунтованість вибору моделей для практичного застосування.

Встановлено, що моделі класу GPT-4o та GPT-5.1 забезпечують кращий баланс між коректністю та стабільністю з прийнятним рівнем ресурсних витрат. На прикладному рівні отримано обґрунтовані рекомендації щодо кількості пунктів спостереження атмосферного повітря для міста Вінниці, що підтверджує здатність моделей інтегрувати нормативні вимоги та числові дані. Показано, що коректність таких рішень істотно залежить не лише від якості моделі, але й від її стабільності та здатності дотримуватись інструкцій.

Ключові слова: великі мовні моделі, комплексні задачі, оцінка ефективності, багатокритеріальний підхід, інтеграція даних, автоматизація.

Вступ

Існує багато різних факторів, які впливають на ефективність великих мовних моделей під час розв'язання задач комплексного характеру: архітектура моделі, обсяг і якість навчальних даних, методи навчання, параметри генерації, а також структура вхідного запиту. Без урахування цих факторів та їх взаємодії неможливо отримати достовірну оцінку якості роботи моделей. У результаті цього ускладнюється процес вибору оптимальної моделі для практичного застосування, а також знижується надійність отриманих результатів.

На відміну від окремих задач, де результати можуть бути оцінені за допомогою стандартних



метрик, у задачах комплексного характеру ефективність моделей є чутливою до структури вхідних даних та сценарію використання. Зокрема, важливу роль відіграє здатність моделі інтегрувати різномірну інформацію — текстові джерела, нормативно-правові документи та числові дані. Одним із ключових аспектів у цьому контексті є узгодженість результатів під час виконання багатокрокових запитів.

За останні роки суттєво розширилися можливості доступу до інструментів оцінювання та порівняння великих мовних моделей. Спеціалізовані платформи дозволяють автоматизувати процес тестування моделей на складних запитах, що поєднують різні типи даних і вимагають комплексного аналізу. Це створює передумови для переходу від ізольованого оцінювання окремих характеристик до системного аналізу ефективності моделей у прикладних сценаріях.

Актуальність дослідження

Згідно з результатами досліджень [1], останніми роками спостерігається стрімкий розвиток великих мовних моделей, які знаходять широке застосування у різних галузях — від обробки природної мови до підтримки прийняття рішень [2], [3]. Існує значна кількість таких моделей, що відрізняються за архітектурою, обсягом навчальних даних, підходами до навчання та сферою застосування. Можна зазначити, що для розв'язання простих задач, таких як генерація тексту, переклад або відповіді на запитання, більшість сучасних моделей демонструє задовільні результати, і їх використання не потребує складної адаптації [4].

Проте у випадку виконання комплексних завдань ситуація ускладнюється. До таких завдань належать, зокрема, задачі, що передбачають одночасний аналіз нормативно-правових документів, роботу з числовими даними, інтерпретацію результатів та формування аргументованих висновків. У подібних умовах ефективність різних моделей суттєво відрізняється: деякі з них здатні коректно інтегрувати різномірну інформацію та забезпечувати узгоджені результати, тоді як інші демонструють обмеження у логічному узагальненні або точності аналізу.

Додатковою проблемою є відсутність уніфікованих підходів до оцінювання ефективності великих мовних моделей саме в контексті комплексних, багатокomпонентних задач. Наявні методики оцінювання здебільшого орієнтовані на окремі аспекти роботи моделей і не враховують специфіку їх застосування в умовах реальних прикладних сценаріїв [5].

Отже, актуальним є розроблення методу аналізу ефективності великих мовних моделей, який дозволить здійснювати їх об'єктивне порівняння під час розв'язання складних комплексних задач. Важливим етапом є також апробація такого підходу на прикладі конкретного комплексного завдання, що поєднає аналіз текстових джерел, нормативної бази та даних. Це сприятиме підвищенню доказовості вибору моделей для практичного застосування та покращенню якості отриманих результатів.

Огляд методів оцінювання ефективності великих мовних моделей

Методи оцінювання ефективності великих мовних моделей у сучасних дослідженнях умовно поділяють на три основні групи:

1. Метрики, орієнтовані на якість результату;
2. Бенчмаркові підходи;
3. Методи багатокритеріального оцінювання.

До найпоширеніших метрик якості належать BLEU, ROUGE, METEOR, перплексія та їх модифікації. Принцип їх застосування полягає у порівнянні згенерованого тексту з еталонними відповідями або оцінюванні ймовірнісних характеристик моделі. Зокрема, BLEU базується на порівнянні n -грам та оцінює точність збігу між згенерованим і референтним текстом, тоді як ROUGE орієнтований на повноту відтворення змісту [6], [7]. Такі метрики широко застосовуються для задач машинного перекладу та автоматичного реферування, проте їх обмеженням є низька кореляція з людськими оцінками у випадку відкритих або складних задач [8].

Бенчмаркові підходи ґрунтуються на використанні стандартизованих наборів задач, таких як MMLU, HumanEval, BIG-bench та інші. Вони дозволяють оцінювати здатність моделей до узагальнення, логічного міркування та роботи в різних предметних областях [9], [10]. Основною перевагою таких підходів є можливість порівняння великої кількості моделей за єдиними умовами. Водночас дослідження показують, що результати на бенчмарках не завжди корелюють з ефективністю моделей у реальних прикладних сценаріях, особливо у разі виконання багатокрокових або міждисциплінарних задач [11]. Це обмежує їх застосовність у прикладних задачах, де важливими є не лише якість відповіді, але й такі характеристики, як вартість, швидкодія та дотримання інструкцій.

Окремий напрям становлять методи, що використовують самі великі мовні моделі як оцінювачі. Наприклад, підхід G-Eval передбачає використання моделей для оцінювання якості тексту без необхідності наявності еталонних відповідей. Експериментальні результати демонструють вищу узгодженість з людськими оцінками порівняно з традиційними метриками, проте такі підходи можуть мати обмеження щодо стабільності результатів та наявності систематичних упереджень.

У сучасних дослідженнях також активно розвиваються задачо-орієнтовані підходи до оцінювання, які спрямовані на аналіз ефективності моделей у конкретних прикладних сценаріях. На відміну від статичних бенчмарків, такі підходи передбачають формування користувацьких тестових запитів, що відображають реальні задачі, та оцінювання моделей за сукупністю показників. Особливістю таких підходів є врахування стохастичної природи генерації відповідей шляхом багаторазового запуску моделей та оцінювання стабільності результатів. Це дозволяє отримати надійнішу оцінку поведінки моделей у реальних умовах використання. До того ж, задачо-орієнтовані підходи забезпечують можливість формалізації процесу вибору моделі як задачі багатокритеріальної оптимізації з урахуванням компромісів між якістю результату та ресурсними витратами.

У сучасних роботах також розглядаються комплексні підходи до оцінювання, які інтегрують декілька груп критеріїв: точність, узагальнювальну здатність, надійність, ефективність використання ресурсів та інтерпретованість результатів. Такі підходи дозволяють отримати повніше уявлення про поведінку моделей, проте потребують формалізації процедури інтеграції показників та адаптації до конкретних класів задач [12].

Аналіз зазначених підходів показує, що метрики якості забезпечують швидку, але обмежену оцінку, бенчмарки — стандартизоване порівняння, тоді як задачо-орієнтовані багатокритеріальні методи є найперспективнішими для оцінювання ефективності моделей у задачах комплексного характеру. Це обумовлює потребу у розробленні спеціалізованих методів оцінювання, орієнтованих на інтеграцію різномірних критеріїв та врахування специфіки прикладних сценаріїв використання великих мовних моделей.

Методика дослідження

У цьому дослідженні розроблено метод оцінювання ефективності великих мовних моделей у задачах комплексного характеру, що поєднують аналіз текстових, нормативно-правових та числових даних. Особливістю запропонованого методу є використання багатокритеріального підходу та задачо-орієнтованої експериментальної платформи, що забезпечує порівняння моделей в однакових умовах виконання.

Для виконання досліджень розроблено веб-інтерфейс LLM Comparator (рис. 1), який реалізує багатовимірне оцінювання моделей за такими критеріями, як коректність, дотримання інструкцій, стабільність результатів, час виконання, вартість обчислень, використання токенів та обсяг відповіді. При цьому частина критеріїв визначається автоматично на основі телеметрії API, тоді як оцінка коректності та дотримання інструкцій здійснюється з використанням фіксованої моделі-оцінювача.

Рис. 1. Веб-сторінка LLM Comparator

Для проведення експерименту сформовано комплексний запит, який включає:

- 1) аналіз числових даних (CSV-файл з концентраціями забруднювальних речовин) [13];
- 2) обробку нормативно-правових документів;
- 3) інтеграцію отриманих результатів із формуванням умотивованого висновку.

Такий підхід дозволяє моделювати реальні прикладні сценарії, у яких великі мовні моделі виконують багатокрокові операції з різномірними джерелами даних, що вимагає одночасного забезпечення точності, узгодженості та дотримання інструкцій. Як середовище для проведення експерименту використано систему [14], яка реалізує задачо-орієнтований підхід до оцінювання моделей та забезпечує паралельний запуск декількох моделей і режимів їх роботи на одному наборі тестових запитів з подальшим збором і агрегуванням метрик. Це дозволяє мінімізувати вплив зовнішніх факторів та забезпечити відтворюваність результатів. Для оцінювання ефективності моделей вибрано такі критерії:

- | | |
|---------------------------|--------------------------|
| – коректність результату; | – використання токенів; |
| – стабільність; | – обсяг відповіді; |
| – час виконання; | – дотримання інструкцій. |
| – вартість обчислень; | |

Оцінювання критеріїв здійснюється комбінованим способом. Показники часу виконання, вартості обчислень та використання токенів визначаються автоматично на основі телеметрії виконання запитів. Критерії коректності результату та дотримання інструкцій оцінюються з використанням фіксованої моделі-оцінювача за принципом LLM-as-judge, що дозволяє формалізувати перевірку відповідності результатів очікуваним умовам задачі. Для врахування стохастичної природи генерації відповідей кожна модель запускалась декілька разів на однаковому запиті. Це дозволило оцінити стабільність її роботи як відношення кількості однакових відповідей до загальної кількості запусків, а також зменшити вплив випадкових факторів генерації. Додатково враховувався обсяг відповіді як характеристика схильності моделі до надмірної або недостатньої деталізації. Інтегральна оцінка ефективності моделей формується на основі багатокритеріального підходу з використанням вагових коефіцієнтів, що визначаються експертним шляхом з урахуванням специфіки задачі. Зокрема, найбільша вага надається коректності результату та дотриманню інструкцій, тоді як ресурсні показники (вартість, токени, час виконання) мають допоміжний характер і враховуються під час аналізу доцільності практичного застосування моделей.

З метою аналізу компромісів між якістю та ресурсними витратами результати оцінювання можуть бути інтерпретовані в рамках задачі багатокритеріальної оптимізації, де вибір оптимальної моделі визначається як баланс між точністю, стабільністю та обчислювальними витратами. Запропонована методика має такі переваги:

1. Дозволяє оцінювати моделі в умовах, наближених до реальних сценаріїв застосування.
2. Враховує багатокритеріальний характер ефективності.
3. Забезпечує можливість об'єктивного порівняння різних моделей.
4. Дозволяє аналізувати компроміси між якістю та ресурсними витратами.
5. Підтримує відтворюваність експериментів за рахунок використання стандартизованого набору тестових запитів.

Математична модель та опис системи оцінювання

Запропонована система оцінювання ефективності великих мовних моделей формалізується як задача багатокритеріального аналізу для множини моделей у контексті заданого набору тестових запитів. Нехай задано множину моделей $M = \{m_1, \dots, m_K\}$, множину тестових запитів $T = \{t_1, \dots, t_N\}$, множину критеріїв оцінювання $C = \{c_1, \dots, c_d\}$. Для кожної пари «модель–запит» формується вектор вимірювань $s(m, t) = (c_1(m, t), c_2(m, t), \dots, c_d(m, t))$, де $c_d(m, t)$ — значення d -го критерію для моделі m на запиті t . З метою отримання узагальненої характеристики моделі виконується агрегування по множині запитів

$$\bar{s}(m) = \frac{1}{N} \sum_{t \in T} s(m, t). \quad (1)$$

Для кожного критерію вводяться відповідні функціональні залежності. Коректність результату визначається як частка правильних відповідей

$$\text{Corr}(m, t) = \frac{1}{R} \sum_{r=1}^R J(t, m, r), \quad (2)$$

де R — кількість запусків моделі, $J(t, m, r)$ — результат оцінки (1 — правильна відповідь, 0 — не правильна). Стабільність визначається як відношення кількості найчастотнішої відповіді до загальної кількості запусків

$$\text{Stab}(m, t) = \frac{\max_{v \in V} |\{r : o_{m,t,r} = v\}|}{R}, \quad (3)$$

де $o_{m,t,r}$ — відповідь моделі рядковими літерами без пробілів на початку та у кінці, V — множина унікальних відповідей. Час виконання

$$\bar{L}(m, t) = \frac{1}{R} \sum_{r=1}^R L_{m,t,r}, \quad (4)$$

де $L_{m,t,r}$ — час виконання окремого запуску. Вартість обчислень

$$\text{Cost}(m, t) = \sum_{r=1}^R \frac{n_{m,t,r}^{\text{in}} \cdot p_m^{\text{in}} + n_{m,t,r}^{\text{out}} \cdot p_m^{\text{out}}}{10^6}, \quad (5)$$

де $n^{\text{in}}, n^{\text{out}}$ — кількість вхідних та вихідних tokenів, $p^{\text{in}}, p^{\text{out}}$ — їхня вартість. Використання tokenів

$$\text{Tok}(m, t) = \sum_{r=1}^R (n_{m,t,r}^{\text{in}} + n_{m,t,r}^{\text{out}}). \quad (6)$$

Обсяг відповіді

$$\text{Verb}(m, t) = \|o_{m,t,r}\|. \quad (7)$$

Оскільки критерії мають різну природу та напрям оптимізації, вводиться нормалізація

$$\hat{c}_d(m) = \frac{c_d(m) - \min_{m'} c_d(m')}{\max_{m'} c_d(m') - \min_{m'} c_d(m')}. \quad (8)$$

Для критеріїв, які підлягають мінімізації (вартість, час, токени), використовується обернена шкала. Інтегральна оцінка ефективності моделі визначається як зважена сума

$$U(m) = \sum_{d=1}^D w_d \cdot \hat{c}_d(m); \quad \sum_{d=1}^D w_d = 1. \quad (9)$$

Оптимальна модель визначається як

$$m^* = \arg \max_{m \in M} U(m). \quad (10)$$

Вибір може здійснюватися на основі множини Парето

$$P = \{m \in M : \neg \exists m' \in M : \bar{s}(m') \geq \bar{s}(m) \wedge \bar{s}(m') \neq \bar{s}(m)\}, \quad (11)$$

де $\geq$ означає покомпонентне домінування. Це дозволяє виділити множину непоміжених рішень без необхідності задання вагових коефіцієнтів. Задача вибору моделі формулюється як задача оптимізації $\max_{m \in M} U(m)$, за умови $\text{Corr}(m) \geq \theta$, де $\theta \in [0, 1]$ — допустимий рівень коректності.

У межах поставленої задачі вагові коефіцієнти w_j визначено експертним шляхом з урахуванням пріоритетності критеріїв якості над ресурсними витратами: коректність — 0,30; стабільність — 0,20; дотримання інструкцій — 0,20; час виконання — 0,10; вартість — 0,10; використання tokenів — 0,05; обсяг відповіді — 0,05 (сума ваг становить 1). Допустимий рівень коректності прийнято на рівні $\alpha = 1,0$, що відповідає вимозі формування правильної відповіді на комплексний запит, оскільки помилка у визначенні кількості пунктів спостереження має суттєві практичні наслідки.

Порівняння запропонованої моделі з відомими методами багатокритеріального аналізу

Запропонований метод оцінювання великих мовних моделей базується на принципах багатокритеріального аналізу рішень (MCDM — Multiple Criteria Decision Making). З метою обґрунтування вибору підходу та визначення його переваг у контексті поставленої задачі здійснено порівняння з чотирма найпоширенішими методами предметної галузі: AHP (Analytic Hierarchy Process), TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution), VIKOR та ELECTRE.

Метод AHP, розроблений Саати [15], передбачає ієрархічну декомпозицію задачі та визначення вагових коефіцієнтів на основі попарних порівнянь критеріїв і альтернатив експертом. Незважаючи на широке застосування в задачах вибору, AHP має обмеження за великої кількості альтернатив (у цьому дослідженні — 10 моделей) та потребує значних зусиль для формування матриць попарних порівнянь. До того ж, метод не передбачає роботи з кількісними телеметричними показниками, що автоматично збираються в ході експерименту, та не враховує стохастичну природу відповідей мовних моделей.

Метод TOPSIS [16] ранжує альтернативи за відстанню до ідеального та антиідеального розв'язку у нормованому критеріальному просторі. Він добре підходить для ситуацій, де всі критерії є порівнянними та безперервними. Проте у задачі оцінювання великих мовних моделей частина критеріїв є дискретною або бінарною (коректність: 0 або 1; дотримання інструкцій: так/ні), а також існує жорстке обмеження на допустимий рівень коректності ($\alpha = 1,0$), яке TOPSIS не підтримує у своєму стандартному формулюванні. Запропонований у роботі метод явно враховує це обмеження через формулу (10), що є принциповою перевагою для прикладних задач з неприпустимими помилками.

Метод VIKOR [17] орієнтований на пошук компромісного розв'язку у разі конфліктуючих критеріїв шляхом мінімізації відхилення від ідеального за критерієм максимінного жалю. Він ефективний у ситуаціях, коли необхідно збалансувати групові та індивідуальні переваги. Проте VIKOR, як і TOPSIS, не підтримує обмеження-відсікання за ключовим критерієм і не передбачає виявлення Парето-оптимальних рішень як самостійного етапу аналізу, що знижує його придатність для задач, де недомінованість є важливою інтерпретаційною характеристикою.

Методи сімейства ELECTRE [18] реалізують відносини переваги між альтернативами на основі порогів конкордансу та дискордансу, що дозволяє уникати компенсації між критеріями. Це є перевагою у разі якісних або порядкових критеріїв, але ускладнює інтерпретацію результатів і не дозволяє отримати єдиний числовий рейтинг альтернатив, необхідний для автоматизованого прийняття рішення у програмному засобі LLM Comparator.

Узагальнене порівняння характеристик розглянутих методів подано у табл. 1.

Таблиця 1

Порівняння запропонованого методу з відомими підходами MCDM

Характеристика	AHP	TOPSIS	VIKOR	ELECTRE	Запропонований метод
Тип вхідних даних	якісні / кількісні	кількісні	кількісні	якісні / порядкові	змішані (бінарні, дискретні, безперервні)
Визначення вагових коефіцієнтів	попарні порівняння	задаються безпосередньо	задаються безпосередньо	порогові значення	задаються безпосередньо (експертно)
Парето-аналіз	—	—	—	частково	+ (явний етап)
Жорстке обмеження за критерієм	—	—	—	—	+ (α -відсікання)
Врахування стохастичності	—	—	—	—	+ (стабільність як критерій)
Автоматизація збору даних	—	—	—	—	+ (API-телеметрія)
Єдиний числовий рейтинг	+	+	+	—	+

Як випливає з табл. 1, жоден із розглянутих класичних методів MCDM не забезпечує одночасної підтримки всіх ключових вимог задачі: роботи зі змішаними типами критеріїв, явного Парето-аналізу, жорсткого відсікання за критерієм коректності та автоматизованого збору вхідних даних через API. Запропонований метод поєднує зважену адитивну функцію корисності (характерну для TOPSIS/VIKOR) з Парето-аналізом та α -обмеженням, завдяки чому він структурно ближчий до методів з обмеженнями у задачах дискретної оптимізації, але адаптованим до специфіки оцінювання стохастичних систем типу великих мовних моделей. Таким чином, запропонований підхід можна

розглядати як розширення класу зважених адитивних MCDM-методів, доповнених механізмами відсікання та виявлення невідомінованих рішень, що є оригінальним внеском у методологію порівняльного оцінювання великих мовних моделей.

Результати дослідження

Реалізацію багатокритеріальної методики оцінювання ефективності великих мовних моделей виконано за допомогою системи LLM Comparator. Вибрано комплексний запит, який включав одночасний аналіз числових даних, нормативно-правових документів [19], [20] та формування узагальненого висновку щодо обґрунтованості розміщення пунктів спостережень у місті Вінниця. На початковому етапі протестовано декілька моделей з вимірюванням показників. Результати експериментів подано у табл. 2.

Таблиця 2

Показники ефективності великих мовних моделей на комплексному запиті

Модель	Коректність	Стабільність, %	Час, с	Вартість, \$	Токени	Довжина відповіді	Дотримання інструкцій
GPT-4o	-	33	25,01	0,4	152537	2244	+
GPT-4o Mini	+	33	14,54	0,02	152447	1872	+
o3-mini (reasoning)	-	33	31,50	0,21	161282	2729	+
GPT-5	+	67	38,71	0,06	195037	2217	+
GPT-5.1	+	67	32,23	0,06	190609	2356	+
claude-3-7-sonnet-latest	-	33	26,43	0,09	101469	3023	-
claude-4-opus	-	33	29,86	0,09	99733	2072	-
claude-4-sonnet	-	33	26,09	0,09	100134	2555	-
claude-4.6-sonnet	+	33	60,90	0,3	319583	5131	+
gemini-3.1	-	33	34,11	0,17	179006	2874	+

Для ілюстрації застосування формул (2)–(7) наведемо покроковий розрахунок агрегованих показників на прикладі моделі GPT-5.1, яка виконувалась тричі ($n = 3$) на тестовому запиті. За формулою (2) коректність становить $C^{corr} = 1,0$ (правильна відповідь у всіх запусках). За формулою (3) стабільність $C^{stab} = 2/3 \approx 0,67$ (дві з трьох відповідей збіглися після нормалізації). За формулою (4) середній час виконання $T = 32,23$ с, за (5) середня вартість $Q = 0,06$ \$, за (6) середнє використання токенів $N = 190\,609$, за (7) середній обсяг відповіді $L = 2\,356$ символів. Аналогічні розрахунки виконано для всіх моделей; результати подано у табл. 2.

Для отримання інтегральної оцінки виконано нормалізацію за формулою (8). У стовпці «Коректність» значення «+» інтерпретується як 1, а «-» — як 0. Результати обчислення нормалізованих показників та інтегральної оцінки $U(M_i)$ за формулою (9) подано у табл. 3.

Таблиця 3

Нормалізовані значення критеріїв та інтегральна оцінка ефективності $U(M_i)$

Модель	Коректність	Стабільність	Час, с	Вартість, \$	Токени	Довжина відповіді	Дотримання інструкцій	$U(M_i)$
GPT-4o	0,00	0,00	0,77	0,00	0,76	0,89	1,00	0,360
GPT-4o Mini	1,00	0,00	1,00	1,00	0,76	1,00	1,00	0,788
o3-mini (reasoning)	0,00	0,00	0,63	0,50	0,72	0,74	1,00	0,386
GPT-5	1,00	1,00	0,48	0,89	0,57	0,89	1,00	0,910
GPT-5.1	1,00	1,00	0,62	0,89	0,59	0,85	1,00	0,923
claude-3-7-sonnet	0,00	0,00	0,74	0,82	0,99	0,65	0,00	0,238
claude-4-opus	0,00	0,00	0,67	0,82	1,00	0,94	0,00	0,245
claude-4-sonnet	0,00	0,00	0,75	0,82	1,00	0,79	0,00	0,246
claude-4.6-sonnet	1,00	0,00	0,00	0,26	0,00	0,00	1,00	0,526
gemini-3.1	0,00	0,00	0,58	0,61	0,64	0,69	1,00	0,385

За результатами обчислень найвищу інтегральну оцінку отримала модель GPT-5.1 ($U = 0,923$), що відповідає оптимальному рішенням M^* згідно з формулою (10). Аналіз множини Парето згідно з формулою (11) показав, що невідомінованими є п'ять моделей: GPT-4o Mini, GPT-5, GPT-5.1, claude-

4-orpus та claude-4-sonnet. Дві останні зберігаються у множині Парето лише за рахунок мінімального використання токенів, проте відсікаються вимогою $\alpha = 1,0$. Після застосування обмеження на допустимий рівень коректності (виключення моделей з $C^{corr} < 1,0$) залишаються чотири моделі-кандидати: GPT-5.1 ($U = 0,923$), GPT-5 ($U = 0,910$), GPT-4o Mini ($U = 0,788$) та claude-4.6-sonnet ($U = 0,526$). Отже, числові результати підтверджують, що моделі сімейства GPT-5 забезпечують найкращий компроміс між якістю та ресурсними витратами для задач цього класу, а модель GPT-4o Mini залишається конкурентоспроможною як економічна альтернатива.

На основі отриманих даних визначено такі результати щодо кількості пунктів спостереження у Вінниці:

- 1) тверді частинки: 3 пункти; 3) інші забруднювачі: 2 пункти.
- 2) озон: 1 пункт;

Рішення базується на аналізі нормативних документів та відповідних додатків, які визначають мінімальні вимоги до розміщення пунктів спостережень з урахуванням чисельності населення та щільності населення у зоні спостереження. Для міста з відповідною кількістю населенням критичною є наявність хоча б одного пункту для озону та декількох пунктів для твердих часток (ТЧ) та інших забруднювачів, що забезпечує достатню деталізацію та надійність вимірювань.

LLM Comparator

Compare different LLMs on your tasks

Prompt

Проаналізуй середньодобові концентрації речовин у повітрі (мг/м3) з csv-файлу. Проаналізуй "постанову від 14 серпня 2019 р. No 827", визнач верхні та нижні середньодобові та середньорічні пороги оцінювання концентрації забруднювальних. Проаналізуй "наказ 21.04.2021 No 300", зверни увагу на "Додаток 1 до Порядку розміщення пунктів спостережень за забрудненням атмосферного повітря в зонах та агломераціях".

Expected Result

Тверді частинки: 3
Озон: 1
Інші забруднювачі: 2

Or upload a test file (YAML/JSON)

[Browse...](#) request.json Loaded 1 test case(s)

Models 88 models available [+ Add models](#)

[GPT-4o](#) [GPT-4o Mini](#) [o3-mini \(reasoning\)](#) [GPT-5](#) [GPT-5.1](#)

Criteria

☒ Correctness
 ☒ Stability
 ☒ Latency
 ☒ Cost
 ☒ Token Usage
 ☒ Verbosity
 ☒ Instruction Following

Number of runs (for stability)

3

[Run Comparison](#)

Prompt

Проаналізуй середньодобові концентрації речовин у повітрі (мг/м3) з csv-файлу. Проаналізуй "постанову від 14 серпня 2019 р. No 827", визнач верхні та нижні середньодобові та середньорічні пороги оцінювання концентрації забруднювальних. Проаналізуй "наказ 21.04.2021 No 300", зверни увагу на "Додаток 1 до Порядку розміщення пунктів спостережень за забрудненням атмосферного повітря в зонах та агломераціях". Врахуй, що у Вінниці мешкає 430000 осіб. Спочатку виведи текстом: кількість пунктів спостереження за ТЧ (TSP), кількість пунктів спостереження за озоном (O3), кількість пунктів спостереження за іншими забруднювачами. Далі напиши обґрунтування чому ти прийняв таке рішення. Вкажи нижні та верхні пороги та відповідні концентрації речовин, які вплинули на твоє рішення

Expected Result

Тверді частинки: 3 Озон: 1 Інші забруднювачі: 2

Model	Correctness	Stability	Latency	Cost	Token Usage	Verbosity	Instruction Following	Response
GPT-4o	Fail	33%	25.01s	\$0.3986374	152537	2244 chars	Pass	Show
GPT-4o Mini	Pass	33%	14.54s	\$0.023864	152447	1872 chars	Pass	Show
o3-mini (reasoning)	Fail	33%	31.50s	\$0.213888	161282	2729 chars	Pass	Show
GPT-5	Pass	67%	38.71s	\$0.064936	195037	2217 chars	Pass	Show
GPT-5.1	Pass	67%	32.23s	\$0.065548	190609	2356 chars	Pass	Show

Рис. 2. Результати роботи системи

Значення верхніх та нижніх порогів оцінювання знайдено та використано моделями для аналізу CSV-файлу з середньодобовими концентраціями та формування рекомендацій щодо розташування пунктів спостережень. Варіабельність результатів між моделями показала, що для комплексних завдань важлива не лише здатність моделі коректно обробляти нормативні документи, але й інтегрувати числові дані та робити узгоджені висновки, що підтверджує необхідність багатокритеріального підходу до оцінки ефективності. Отримані результати (рис. 2) демонструють, що застосування запропонованої методики дозволяє врахувати змішані типи даних (текстові та числові), оцінювати коректність інтегрованих висновків, визначати оптимальну кількість та типи пунктів спостереження для конкретного міста.

Висновки

Розроблено та апробовано метод багатокритеріальної оцінки ефективності великих мовних моделей для виконання задач комплексного характеру, що передбачають одночасний аналіз нормативно-правових документів та числових даних. На відміну від наявних підходів, запропонований метод базується на формалізації процесу вибору моделі як багатокритеріальної оптимізаційної задачі з урахуванням суперечливих критеріїв якості та ресурсної ефективності.

У рамках дослідження сформовано систему критеріїв оцінювання (коректність, стабільність, час виконання, вартість, використання токенів, обсяг відповіді та дотримання інструкцій) та реалізовано їх обчислення на основі експериментальної платформи LLM Comparator. Проведення багаторазових запусків моделей дозволило оцінити стохастичні властивості генерації та кількісно визначити стабільність результатів.

Запропоновано математичну модель агрегування результатів, що включає усереднення показників по множині тестових задач, виділення множини Парето-оптимальних моделей, побудову скаляризованої функції корисності для визначення оптимальної конфігурації моделі залежно від заданих вагових коефіцієнтів. Здійснено порівняння запропонованої моделі з відомими методами багатокритеріального аналізу.

Результати експерименту показали наявність суттєвих відмінностей між моделями за всіма групами критеріїв. Встановлено, що моделі класу GPT-4o та GPT-5.1 забезпечують кращий баланс між коректністю та стабільністю з прийнятним рівнем ресурсних витрат. На прикладному рівні отримано обґрунтовані рекомендації щодо кількості пунктів спостереження атмосферного повітря для міста Вінниця, що підтверджує здатність моделей інтегрувати нормативні вимоги та числові дані. Показано, що коректність таких рішень істотно залежить не лише від якості моделі, але й від її стабільності та здатності дотримуватись інструкцій.

Отримані результати свідчать, що використання багатокритеріального підходу дозволяє перейти від емпіричного вибору моделей до формалізованого процесу прийняття рішень. Запропонований метод може бути застосований для задач системного аналізу, де необхідно забезпечити компроміс між точністю, швидкістю та вартістю обчислень.

Числове застосування формул (8)–(11) до даних експерименту дозволило виділити п'ять Парето-оптимальних моделей (GPT-4o Mini, GPT-5, GPT-5.1, claude-4-opus, claude-4-sonnet); за інтегральною оцінкою U із заданими ваговими коефіцієнтами та допустимим рівнем коректності $\alpha = 1,0$ оптимальною визначено модель GPT-5.1 зі значенням $U(M^*) = 0,923$, що чисельно підтверджує висновки про її переваги над іншими розглянутими моделями.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] S. M. Sajjadi Mohammadabadi, et al., "A Survey of Large Language Models: Evolution, Architectures, Adaptation, Benchmarking, Applications, Challenges, and Societal Implications," *Electronics*, vol. 14, no. 18, p. 3580, 2025. <https://doi.org/10.3390/electronics14183580>.
- [2] А. В. Лосенко, Є. М. Крижановський, І. М. Штельмах, і І. В. Варчук, «Технологія LLM-видобування ознак тестування пацієнтів з текстових звітів для удосконалення прогнозування кількості хворих на коронавірус,» *Вісник Вінницького політехнічного інституту*, № 6, с. 135-144, 2024. <https://doi.org/10.31649/1997-9266-2024-177-6-135-144>.
- [3] Є. М. Крижановський, В. О. Караваєв, І. М. Штельмах, і О. О. Войцеховська, «Автоматизація використання природномовних запитів для комплексного аналізу стану поверхневих вод басейну Південного Бугу,» *Вісник Вінницького політехнічного інституту*, № 4, с. 118-125, 2025. <https://doi.org/10.31649/1997-9266-2025-181-4-118-125>.
- [4] X. Han et al., "Pre-Trained Models: Past, Present and Future," *AI Open*, vol. 2, pp. 225-250, 2021. <https://doi.org/10.1016/j.aiopen.2021.08.002>.

- [5] Q. He et al., "Can Large Language Models Understand Real-World Complex Instructions?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 18188-18196, 2024. <https://doi.org/10.1609/aaai.v38i16.29777>.
- [6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, Philadelphia, Pennsylvania, USA, July 7–12, 2002, pp. 311-318. <https://doi.org/10.3115/1073083.1073135>.
- [7] C.-Y. Lin, and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, Barcelona, Spain, July 21-26, 2004, p. 605. <https://doi.org/10.3115/1218955.1219032>.
- [8] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, July, 2020, pp. 7881-7892. <https://aclanthology.org/2020.acl-main.704>.
- [9] R. Bommasani, P. Liang, and T. Lee, "Holistic Evaluation of Language Models," *Annals of the New York Academy of Sciences*, vol. 1525, pp. 140-146, 2023. <https://doi.org/10.1111/nyas.15007>.
- [10] Y. Liu, et al., "Datasets for large language models: a comprehensive survey," *Artificial Intelligence Review*, vol. 58, article 403, 2025. <https://doi.org/10.1007/s10462-025-11403-7>.
- [11] W. Cao, et al., "Benchmarking large language models against human experts in rehabilitation medicine: a multidimensional evaluation," *Journal of NeuroEngineering and Rehabilitation*, vol. 23, article 84, 2026. <https://doi.org/10.1186/s12984-026-01903-0>.
- [12] H. V. Pandhare, "Evaluating Large Language Models: Frameworks and Methodologies for AI/ML System Testing," *International Journal of Scientific Research and Management (IJSRM)*, vol. 12, issue 09, pp. 1467-1486. <https://doi.org/10.18535/ijssrm/v12i09.ec08>.
- [13] Є. М. Крижановський, і М. В. Швець, «Підхід до збору та візуалізації даних державного моніторингу стану атмосферного повітря міста Вінниці», *Матеріали LV Всеукраїнської науково-технічної конференції підрозділів ВНТУ*, Вінниця, 24-27 березня 2026. [Електронний ресурс]. Режим доступу: <https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/27366>
- [14] LLM Comparator. [Online]. Available: <https://llmcomparator.flitsolutions.com/>. Accessed: April 06, 2026.
- [15] T. L. Saaty, "A scaling method for priorities in hierarchical structures," *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234-281, 1977. [https://doi.org/10.1016/0022-2496\(77\)90033-5](https://doi.org/10.1016/0022-2496(77)90033-5).
- [16] C.-L. Hwang, and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*. Berlin, Heidelberg: Springer, 1981. <https://doi.org/10.1007/978-3-642-48318-9>.
- [17] S. Opricovic, and G.-H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *European Journal of Operational Research*, vol. 156, no. 2, pp. 445-455, 2004. [https://doi.org/10.1016/S0377-2217\(03\)00020-1](https://doi.org/10.1016/S0377-2217(03)00020-1).
- [18] B. Roy, "Classement et choix en présence de points de vue multiples," *RAIRO – Revue d'automatique, d'informatique et de recherche opérationnelle*, vol. 2, no. 1, pp. 57-75, 1968. <https://doi.org/10.1051/ro/196802V100571>.
- [19] Постанова Кабінету Міністрів України «Деякі питання здійснення державного моніторингу в галузі охорони атмосферного повітря» від 14.08.2019 р. № 827 зі змінами від 7.05. 2024 р. [Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/laws/show/827-2019-%D0%BF#Text>. Дата звернення: 06.04.2026.
- [20] Наказ Міністерства внутрішніх справ України від 21.04.2021 № 300 «Про затвердження Порядку розміщення пунктів спостережень за забрудненням атмосферного повітря в зонах та агломераціях» [Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/laws/show/z0635-21#Text>. Дата звернення: 06.04.2026.

Рекомендовано кафедрою системного аналізу та інформаційних технологій ВНТУ

Дата надходження 22.04.2026

Дата прийняття після доопрацювання 28.04.2026

Дата прийняття до друку після рецензування 24.05.2026

Дата публікації 2.10.2026

Крижановський Євгеній Миколайович — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: kruzhan@gmail.com . <https://orcid.org/0000-0003-1311-1175>;

Штельмах Ігор Миколайович — канд. техн. наук, асистент кафедри системного аналізу та інформаційних технологій, e-mail: igor.shtelmakh@vntu.edu.ua . <https://orcid.org/0009-0008-1259-448X>;

Швець Михайло Владиславович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: misha.shvets02@gmail.com . <https://orcid.org/0009-0005-0933-2938>;

Войцеховська Ольга Олександрівна — д-р філософії, доцент кафедри системного аналізу та інформаційних технологій, e-mail: olgav1085@gmail.com . <https://orcid.org/0000-0001-8504-1204>.

Вінницький національний технічний університет, Вінниця

Ye. M. Kryzhanovskiy¹
I. M. Shtelmakh¹
M. V. Shvets¹
O. O. Voitsekhovska¹

Method and Tool for Analyzing the Performance of Large Language Models for Complex Tasks Solution

¹Vinnytsia National Technical University

This paper analyzes modern approaches to evaluating the efficiency of large language models for complex tasks solution that involve the simultaneous analysis of regulatory documents and numerical data. The study examines challenges related to the integration of heterogeneous information, ensuring logical consistency of results, and comparing models across multiple criteria, including correctness, stability, processing time, and adherence to instructions. The relevance of the research is determined by the rapid development of large language models and the need to select optimal solutions for applied tasks where errors may have significant consequences.

The proposed method is based on a multi-criteria approach that involves the simultaneous evaluation of model outputs on a complex query, including the processing of regulatory documents and CSV data containing average daily air pollutant concentrations. The evaluation is carried out using metrics such as correctness, stability, execution time, cost, token usage, and adherence to instructions.

A mathematical model of results aggregation is proposed, which includes averaging indicators over a set of test tasks, selecting a set of Pareto-optimal models, and constructing a scalar utility function to determine the optimal configuration of the model depending on the specified weight coefficients.

The methodology was validated using a case study of the city of Vinnytsia, demonstrating that different models exhibit significant differences in their ability to produce consistent and coherent conclusions. Based on the analysis of regulatory thresholds by the models, the optimal number of monitoring stations for different pollutants was determined, ensuring representative air quality monitoring. The proposed approach accounts for the specific features of complex applied tasks, enables the integration of heterogeneous data, and improves the validity of model selection for practical applications.

It is established that the models of the GPT-4o and GPT-5.1 class provide a better balance between correctness and stability at an acceptable level of resource costs. At the applied level, justified recommendations on the number of atmospheric air observation points for the city of Vinnytsia were obtained, which confirms the ability of the models to integrate regulatory requirements and numerical data. It is shown that the correctness of such solutions significantly depends not only on the quality of the model, but also on its stability and ability to follow instructions.

Keywords: large language models, complex tasks, performance evaluation, multi-criteria approach, data integration, automation.

Kryzhanovskiy Yevhenii M. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: kruzhan@gmail.com . <https://orcid.org/0000-0003-1311-1175>;

Shtelmakh Ihor M. — Cand. Sc. (Eng.), Assistant of the Chair of System Analysis and Information Technologies, e-mail: igor.shtelmakh@vntu.edu.ua . <https://orcid.org/0009-0008-1259-448X>;

Shvets Mykhailo V. — Post-Graduate Student, of the Chair of System Analysis and Information Technologies, e-mail: misha.shvets02@gmail.com . <https://orcid.org/0009-0005-0933-2938>;

Voitsekhovska Olha O. — PhD, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: olgav1085@gmail.com . <https://orcid.org/0000-0001-8504-1204>