

# ЕФЕКТИВНІСТЬ ВИКОРИСТАННЯ СУЧАСНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ РОЗВ'ЯЗАННЯ ОСВІТНІХ ЗАДАЧ

Ковальчук Кирило Андрійович

*студент групи МПМ-1-26-1.4д,*

*Київського столичного університету імені Бориса Грінченка,*

*науковий керівник - к.т.н., доц. Негоденко О.В.*

## ВСТУП

Щоб оцінити користь генеративного штучного інтелекту в освіті, потрібно з'ясувати, як він допомагає людині навчатися. Великі мовні моделі (Large Language Models, LLM) можуть пояснювати матеріал у діалозі, добирати навчальні приклади й надавати зворотний зв'язок. Водночас навіть правильна та переконлива відповідь системи не свідчить про те, що студент зрозумів розв'язання і здатний відтворити його самостійно. Тому основна проблема дослідження — відокремити якість створеної моделлю відповіді від того, чого навчився її користувач.

Потреба обґрунтовано обирати інструменти генеративного ШІ для освітніх задач визначає актуальність цієї теми. Рекомендації UNESCO орієнтують на потреби людини, збереження самостійності здобувача освіти та перевірку того, чи відповідає інструмент педагогічним і етичним вимогам [1]. Отже, оцінювання має охоплювати спосіб використання системи під час навчання, захист даних і відповідальність людини за ухвалені рішення.

У наукових працях увага поступово переходить від переліку можливостей мовних моделей до з'ясування умов, за яких ці можливості приносять користь. Е. Kaspesі та співавтори описують персоналізацію навчання, підготовку освітніх матеріалів і підтримку студентів, водночас розглядаючи помилки, упередженість та залежність від автоматизованої допомоги [2]. А. Tlili та співавтори досліджують освітнє застосування чат-ботів на прикладі ChatGPT, приділяючи увагу достовірності відповідей і відповідальному користуванню [3]. Обидві праці дають теоретичну основу для подальшого аналізу, але не встановлюють переваги певної сучасної моделі.

Експериментальні дослідження дають змогу конкретизувати ці міркування. Н. Bastani та співавтори виявили, що успішність виконання вправ із ШІ та результати подальшої самостійної перевірки знань можуть змінюватися по-різному [4]. У роботі G. Kestin та співавторів спеціально створений ШІ-тьютор забезпечив кращі безпосередні навчальні результати, ніж аудиторне активне навчання в досліджуваному курсі фізики [5]. Отже, зіставляючи такі результати, потрібно зважати на організацію навчання, характер завдань і спосіб перевірки знань.

Мета дослідження — узагальнити можливості та обмеження великих мовних моделей у розв'язанні освітніх задач і запропонувати критерії для їх подальшого порівняння. Для цього розглянуто вибрані наукові праці, визначено окремі рівні ефективності й підготовлено протокол майбутнього тестування. Основне дослідницьке питання стосується того, які властивості відповідей і

умови взаємодії з моделлю потрібно перевіряти, щоб оцінювання враховувало освітню мету.

### **Опис дослідження та методи**

Об'єктом дослідження є використання генеративного штучного інтелекту в освітньому процесі, а предметом — критерії та умови оцінювання ефективності LLM у розв'язанні освітніх задач. На цьому етапі виконано оглядово-аналітичну роботу: власного експерименту з моделями та учасниками ще не було. Отримані результати охоплюють порівняння опублікованих висновків і розроблену на їх основі схему оцінювання.

Методи роботи включають цільовий аналіз літератури, зіставлення дослідницьких підходів та впорядкування критеріїв оцінювання. Розглянуто рекомендації UNESCO, дві праці про можливості й ризики використання чат-ботів в освіті та два експериментальні дослідження. Під час добору джерел враховано їхній зв'язок із навчанням за допомогою LLM, наявність опису дослідницького підходу та можливість перевірити бібліографічні відомості. Огляд охоплює цілеспрямовано дібрані праці й не претендує на повноту систематичного огляду всіх публікацій.

Експериментальні роботи порівнювали за освітнім контекстом, способом надання допомоги, складом групи порівняння та підходом до вимірювання навчального результату. Відмінності між учасниками, дисциплінами й показниками не дозволяють звести ці результати до одного середнього ефекту, тому такого об'єднання не проводили. Висновки авторів публікацій розглядали окремо від методичних пропозицій для майбутнього експерименту в межах магістерської роботи.

Наступний етап пропонується присвятити україномовним задачам із базових дисциплін інформаційних технологій. Це дозволить підготувати перевірені зразки відповідей і простежити, наскільки правильно моделі використовують фахову термінологію. Аналізувати слід кожную відповідь конкретної версії моделі на визначене завдання за зафіксованих умов. Для відтворення експерименту недостатньо назвати чат-інтерфейс: потрібно окремо записати назву сервісу, версію моделі та режим взаємодії.

### **Результати дослідження**

За результатами аналізу виокремлено три рівні ефективності. На технічному рівні перевіряються правильність і повнота відповіді та її відповідність запиту. Дидактичний рівень охоплює зрозумілість пояснень, доречність підказок і те, наскільки вони спонукають студента міркувати самостійно. Навчальний рівень показує, як змінилася здатність студента виконувати нові завдання без системи. Перші два рівні можна досліджувати через оцінювання відповідей моделей. Для перевірки третього потрібна окрема робота за участю здобувачів освіти.

Н. Bastani та співавтори зіставили два інструменти на основі GPT-4: базовий чат-інтерфейс і тьютора з педагогічними обмеженнями. За даними авторів, результати тренувальних вправ перевищили показники контрольної групи на 48 % у першому випадку та на 127 % у другому. Однак на

самостійному іспиті, коли доступ до ІІІ прибрали, група базового інтерфейсу отримала результати на 17 % нижчі за контрольну. Наведені величини означають відносну зміну показника, а не різницю у відсоткових пунктах. У групі тьютора з обмеженнями статистично значущого погіршення порівняно з контролем не зафіксовано; позитивного впливу на самостійний іспит також не доведено [4].

G. Kestin та співавтори провели дослідження за участю 194 студентів на матеріалі університетського курсу фізики. Вони порівняли навчання зі спеціально розробленим ІІІ-тьютором та аудиторне активне навчання й виявили більший приріст знань у сценарії з тьютором [5]. Висновок стосується саме цього інструмента і короткострокової перевірки результатів. Він не дає підстав очікувати такого самого ефекту від будь-якого чат-бота, в усіх дисциплінах або щодо тривалого збереження знань.

Ці дані підтримують висновок, що користь ІІІ залежить від організації навчальної взаємодії. Вони не дозволяють побудувати рейтинг ChatGPT, Gemini та Claude, адже в публікаціях перевірялися інші об'єкти й умови. Другим результатом роботи стала система критеріїв для власного порівняння відповідей (табл. 1). Це авторська методична пропозиція, що враховує проблеми, описані в джерелах [1–5].

Таблиця 1.

Критерії порівняльного оцінювання відповідей LLM

| Критерій                 | Що оцінюємо                                      | Як перевіряємо                                      |
|--------------------------|--------------------------------------------------|-----------------------------------------------------|
| Коректність              | Точність фактів, висновків та етапів розв'язання | Порівняння з еталоном і облік суттєвих помилок      |
| Повнота та релевантність | Дотримання всіх умов без відхилень від теми      | Зіставлення відповіді з вимогами задачі             |
| Зрозумілість             | Послідовність викладу та пояснення термінів      | Експертне оцінювання з огляду на рівень студента    |
| Навчальна підтримка      | Підказки, запитання й пояснення помилки          | Перевірка, чи має студент зробити самостійний крок  |
| Надійність джерел        | Наявність джерел і відповідність посилань        | Перевірка кожного джерела та пов'язаного твердження |
| Стійкість                | Відповідь на хибну передумову й повторний запит  | Облік виправлень і коливань якості між запусками    |

Зміст кожного критерію залежить від завдання. У поясненні поняття потрібно оцінити точність і доступність викладу, в узагальненні тексту — збереження його основних положень, а в тесті — правильність ключа й однозначність запитань. Для задач із програмування необхідно перевірити, як працює код: переконливого пояснення недостатньо. Коли завдання не вимагає

посилань на джерела, відповідний критерій слід позначити як незастосовний. Нульовий бал у такому разі був би некоректним.

Якісні характеристики пропонується оцінювати за шкалою від 0 до 3 балів. Оцінка 0 означає, що вимогу не виконано або допущено суттєву помилку; 1 — завдання виконано частково й відповідь потребує істотних виправлень; 2 — вимогу переважно виконано, але залишилися незначні недоліки; 3 — її виконано повністю. Для кожного типу задач потрібно уточнити ознаки цих рівнів. Зокрема, правильний кінцевий результат із помилковим обґрунтуванням не заслуговує найвищої оцінки коректності. Суттєві фактичні помилки варто обліковувати окремо, щоб вдало написаний текст не компенсував їх у загальному балі.

Третій результат роботи — план порівняльного експерименту, який можна відтворити. Запропоновано набір із 30 завдань, по шість для кожного виду діяльності: пояснення понять, розв’язання задач, узагальнення матеріалу, створення тестів і надання зворотного зв’язку. Перед тестуванням для кожного завдання потрібно визначити очікувані складники відповіді та типові помилки. Складність завдань має бути зіставною. Частину запитів слід побудувати на навмисно хибних передумовах, щоб перевірити, чи здатна модель критично їх оцінити.

Для порівняння можна обрати моделі, доступні через ChatGPT, Gemini та Claude, записавши їх точні назви на день тестування. Три системи, кожна з яких виконає кожне завдання тричі незалежно, дадуть загалом 270 відповідей. Повторні запуски одного запиту показують мінливість генерації, тому їх не можна вважати окремими незалежними навчальними задачами. Основою висновків про якість має бути набір різних завдань.

Усі запуски потрібно проводити в нових діалогах, з однаковими матеріалами, формулюваннями та обмеженнями обсягу відповіді. Слід також записувати, чи були доступні вебпошук, додаткові інструменти, персоналізація і спеціальні режими. Якщо можливості режимів різняться, їхні результати потрібно аналізувати окремо. Журнал експерименту має містити дату, версію моделі, повний текст запиту, отриману відповідь і відомості про технічні збої. Час генерації можна враховувати додатково, пам’ятаючи про вплив мережі та навантаження сервісу.

Перед перевіркою відповіді варто знеособити й подати двом оцінювачам у випадковій послідовності. На невеликому тренувальному наборі вони мають узгодити розуміння шкали, після чого незалежно оцінити основні матеріали. Первинні розбіжності слід зберегти, а обговорити їх уже після виставлення оцінок. Якщо другого оцінювача залучити не вдасться, це потрібно зазначити як обмеження: сама наявність числової шкали не робить оцінювання об’єктивним.

Підсумки тестування доцільно показувати для кожної категорії задач і кожного критерію окремо. Так можна простежити розподіл балів, кількість відповідей із суттєвими помилками та відмінності між повторними запусками. Під час розрахунку довірчих інтервалів повторні відповіді потрібно групувати за завданням. Зводити оцінки до одного бала варто лише за умови, що ваги

критеріїв визначено заздалегідь. Без такого обґрунтування більше інформації дає опис сильних і слабких сторін моделі. Цей протокол оцінює відповіді за заданих умов, проте не вимірює впливу на навчальні досягнення.

Щоб перевірити вплив на навчання, необхідно провести окремий етап із початковим оцінюванням знань, зіставними умовами навчання та підсумковим розв'язанням нових задач без ІІІ. Перевірка через певний час дозволить з'ясувати, чи збереглися знання. Відгуки про зручність інструмента й задоволеність роботою з ним можуть доповнити ці дані, але не замінити їх. Поточний огляд обмежений невеликою цільовою добіркою джерел. Власних даних про україномовні відповіді моделей і результати студентської вибірки поки немає.

## ВИСНОВКИ

Оцінювання великих мовних моделей в освіті має охоплювати технічну якість відповіді, її дидактичну цінність і вплив на навчання. Розглянуті публікації свідчать, що успішніше виконання завдання з допомогою ІІІ не гарантує кращої самостійної роботи надалі. Щоб правильно пояснити отриманий ефект, потрібно враховувати педагогічну організацію допомоги. Самої назви сервісу недостатньо для висновку про його навчальну цінність.

Запропоновані критерії та протокол дозволяють організувати порівняння відповідей моделей на україномовні освітні завдання. Їх практична користь — у переході від загальних вражень до зрозумілої процедури з перевіреними еталонами, повторними запусками й обліком помилок. Застосування цього протоколу стане наступним етапом магістерської роботи. Визначити переваги конкретної моделі та її вплив на навчальні досягнення можна буде лише після отримання відповідних емпіричних даних.

## ДЖЕРЕЛА

1. Miao F., Holmes W. Guidance for generative AI in education and research. Paris : UNESCO, 2023. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000386693> (дата звернення: 08.10.2026).
2. ChatGPT for good? On opportunities and challenges of large language models for education / E. Kasneci et al. Learning and Individual Differences. 2023. Vol. 103. Article 102274. DOI: 10.1016/j.lindif.2023.102274.
3. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education / A. Tlili et al. Smart Learning Environments. 2023. Vol. 10. Article 15. DOI: 10.1186/s40561-023-00237-x.
4. Generative AI without guardrails can harm learning: Evidence from high school mathematics / H. Bastani et al. Proceedings of the National Academy of Sciences. 2025. Vol. 122, no. 26. Article e2422633122. DOI: 10.1073/pnas.2422633122.
5. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting / G. Kestin et al. Scientific Reports. 2025. Vol. 15. Article 17458. DOI: 10.1038/s41598-025-97652-6.