

ІНФОРМАЦІЙНІ ПРОГРАМИ ТА КОМП'ЮТЕРНО- ІНТЕГРОВАНІ ТЕХНОЛОГІЇ

УДК 004.912

DOI: 10.31471/1993-9965-2024-1(56)-68-78

ВИКОРИСТАННЯ LLM У ДИСТАНЦІЙНОМУ НАВЧАННІ: АДАПТАЦІЯ АНГЛОМОВНИХ АЛГЕБРАЇЧНИХ ЗАВДАНЬ

С. Є. Євсєєв

ІФНТУНГ; 76019, м. Івано-Франківськ, вул. Карпатська, 15,
e-mail: admin@pung.edu.ua

Епідемія коронавірусу та повномасштабна агресія проти нашої країни підкреслили важливість дистанційного навчання. Відповідно, зросли вимоги як до якості програмних систем, так і до викладачів, які готують курси лекцій та перевіряють засвоєння цих знань. На ринок вийшло багато безкоштовних та комерційних рішень на базі великих мовних моделей (ChatGPT, Claude, GoogleGemini та ін.), які вже використовуються студентами та учнями для підготовки своїх робіт. Інтеграція LLM/ШІ в освітній процес (зокрема, дистанційний, де важлива кількість різноманітних варіантів завдань, їх роз'яснення та тлумачення) — це новий виклик, який потребує осмислення та практичної імплементації. Як один з можливих варіантів — адаптація відомих англomовних наборів завдань в системи дистанційного навчання вітчизняних освітян, для україномовних учнів. Проте цей процес матиме зміст лише у випадку, якщо ШІ здатен генерувати переклад (адаптацію), який буде якісним настільки, аби відповідати вимогам навчального процесу. Також учні та студенти вже пробують використовувати ШІ як помічника при розв'язанні завдань з точних наук, але точність розв'язків ШІ завдань українською мовою лишається малодослідженою. Метою даної статті є аналіз поточних можливостей для дистанційного навчального процесу в умовах активного поширення систем на базі LLM у розрізі перекладу (адаптації) англomовних завдань шкільного рівня з алгебри. Завдання статті наступні: 1) проаналізувати здатність двох популярних LLM-моделей генерувати переклад (адаптацію) алгебраїчних завдань згідно з заданими правилами; 2) проаналізувати здатність LLM-моделей розв'язувати завдання в оригінальному варіанті; 3) проаналізувати здатність LLM-моделей розв'язувати завдання в адаптованому варіанті; 4) проаналізувати можливість LLM детально пояснити хід розв'язку завдання українською мовою. Першоджерелом для перевірки було використано набір задач GSM8K (<https://github.com/openai/grade-school-math>, математичні завдання шкільного рівня, який склали професійні викладачі). Після цього ChatGPT v4 / v4 Omni та Gemini 1.5 PRO Preview адаптували 200 завдань з цього набору українською мовою з заміною англomовних імен на українські та переведення одиниць виміру (без фактичного перерахунку, просто для зручності та розуміння школярами, де був «галлон» — став «літр» та ін., долар — гривня). Для оцінки якості саме української адаптації було використано OpenSource інструмент LanguageTool (<https://languagetool.org/>). Додатково людиною була проведена візуальна перевірка на коректність адаптації вибраних питань шляхом випадкової вибірки з перекладеного масиву даних. Адаптовані завдання LLM-моделі мали зрозуміти не тільки англійською, але й українською та розв'язати їх. Результати також були звірені з еталонними відповідями з набору. Результати виявились наступними: 1) Обидві LLM моделі продемонстрували майже однаково граматичну якість українського перекладу — близько 70%. 2) LLM продемонстрували точність розв'язку оригінальних завдань на рівні 85% (ChatGPT v4), 93.5% (ChatGPT v4 Omni) та 75% (Gemini) відповідно. Існує чітка залежність не тільки від структури підказки для системи, але і від об'єму виводу та детальності даних, які в системі вимагають. Якщо LLM

виконує у рамках одного запиту декілька різнопланових завдань (переклад/розв'язок) — результати математичних обчислень будуть набагато гірші, ніж коли LLM має лише детально показати шлях до фінального результату. ChatGPT на англійському варіанті показав 93.5% (85%) розв'язаних завдань, що добре корелює з результатами інших досліджень в цій галузі. 3) Для українського варіанту (адаптація ChatGPT) точність відрізнялась на 5% з оригінальною постановкою (після аналізу типових помилок та корекції адаптації). У Gemini англійський варіант — ~25%, а український (власна адаптація та адаптація ChatGPT) — ~50% помилок. 4) Обидві системи продемонстрували можливість роз'яснювати шлях розв'язання завдання, але потребують зовнішнього контролю результату. Розвиток LLM показує величезний потенціал цієї технології у системах дистанційного навчання українською мовою. Суттєва економія часу при підготовці великої кількості одноманітних завдань з англійських джерел. ШІ демонструє здатність до аналізу задач і якість розв'язку майже не залежить від мови оригіналу завдання. На поточний момент контроль людини над результатами обчислень, запропонованих ШІ, все одно потрібен та є обов'язковим.

Ключові слова: штучний інтелект, дистанційна освіта, генерація питань, оцінювання знань

The COVID-19 pandemic and the full-scale aggression against Ukraine have underscored the importance of remote learning. Consequently, the demands on both the quality of software systems and the educators who prepare lecture courses and knowledge assessments have increased. The market has already seen the emergence of many free and commercial solutions based on large language models (LLMs) such as ChatGPT, Claude, Google Gemini, etc., which are already being used by students to prepare their assignments. The integration of LLM/AI into the educational process, particularly in distance education, where the variety of tasks, their explanations, and interpretations are crucial, presents a new challenge that requires thoughtful consideration and practical implementation. One potential approach is the adaptation of well-known English-language task sets into domestic distance learning systems for Ukrainian-speaking students. However, this process will only be meaningful if the AI is capable of generating translations (adaptations) of sufficient quality to meet educational requirements. Additionally, students are already attempting to use AI as an assistant in solving tasks in exact sciences, but the accuracy of AI solutions for tasks in the Ukrainian language remains underexplored. The aim of this article is to analyze the current possibilities for distance education in the context of the active proliferation of LLM-based systems regarding the adaptation of English-language school-level algebra tasks. The objectives of the article are as follows: 1) to analyze the ability of two popular LLM models to generate translations (adaptations) of algebraic tasks according to specified rules; 2) to analyze the ability of LLM models to solve tasks in their original form; 3) to analyze the ability of LLM models to solve tasks in their adapted form; 4) to analyze the ability of LLM to explain the solution process of tasks in detail in Ukrainian. The primary source for verification was the GSM8K task set (<https://github.com/openai/grade-school-math>, a collection of school-level math problems created by a professional teachers). Subsequently, ChatGPT v4 / v4 Omni and Gemini 1.5 PRO Preview adapted 200 tasks from this set into Ukrainian, replacing English names with Ukrainian ones and converting units of measurement (without actual recalculation, simply for convenience and understanding by students, e.g., "gallon" became "liter" and, US dollar — hryvnia, etc.) To assess the quality of the Ukrainian adaptation, the Open Source software, called LanguageTool (<https://languagetool.org/>) was used. Additionally, a human conducted a visual check for the correctness of the adaptation of selected questions through random sampling from the translated data set. The adapted tasks had to be understood not only in English but also in Ukrainian and solved by the LLM models. The results were also compared with the reference answers from the set. The findings were as follows: 1) both LLM models demonstrated almost identical grammatical quality of the Ukrainian translation — about 70%. 2) LLMs demonstrated solution accuracy of the original tasks at the level of 85% (ChatGPT v4), 93.5% (ChatGPT v4 Omni), and 75% (Gemini) respectively. There is a clear dependence not only on the structure of the prompt for the system, but also on the volume of output and the level of detail required by the system. If the LLM performs several different tasks (translation/solution) within one request, the results of mathematical calculations will be much worse than when the LLM is only required to show the detailed path to the final result. ChatGPT on the English variant showed 93.5% (85%) of solved tasks, which correlates well with the results of other studies in this field; 3) for the Ukrainian variant (ChatGPT adaptation), the accuracy differed by 5% from the original formulation (after analyzing typical errors and correcting the adaptation). For Gemini, the English variant showed ~25%, while the Ukrainian (own adaptation and ChatGPT adaptation) showed ~50% errors; 4) both systems demonstrated the ability to explain the solution path of the task, but require external control of the results. The development of LLMs shows enormous potential for this technology in distance learning systems in the Ukrainian language. Significant time savings in preparing numerous uniform tasks from English sources. AI demonstrates the ability to analyze tasks, and the quality of the solution does not significantly depend on the language of the original task. Right now, human control over the results of calculations proposed by AI is still required and should be considered mandatory.

Keywords: artificial intelligence, distance education, question generation, knowledge assessment.

Постановка проблеми

Епідемія коронавірусу та військові дії внесли істотні зміни у систему освіти, зокрема ак-

цент змістився на дистанційне навчання, аби гарантувати безпеку та здоров'я як викладачів, так і студентів. У цьому контексті використан-

ня сучасних систем на базі штучного інтелекту може спричинити як позитивний, так і потенційно негативний вплив на освітній процес.

Наразі під «штучним інтелектом» розуміють комерційні рішення на базі Великих мовних моделей (*Large Language Model, LLM*). За своєю природою LLM — це спеціалізований програмний комплекс, навчений генерувати текстову (а віднедавна — і аудіовізуальну) відповідь, за якістю дуже схожу до результату творчості людини. Система здатна уточнювати свої відповіді на базі зворотного зв'язку від оператора та навчилася запам'ятовувати та підтримувати діалог на задану тематику. Треба обов'язково зауважити, що насправді «інтелектуальною діяльністю» результат роботи LLM не є: вони лише більш-менш правдоподібно навчилися *імітувати* результати такої діяльності, а їх відповіді залежать від бази даних для навчання (тренінгу). Відповідно, генерація нових смислів, які відсутні у базі даних моделі, є для LLM надзвичайно складним завданням.

Для інтеграції зі сторонніми застосунками компанії-розпорядники пропонують відкриті програмні інтерфейси (API) на комерційній основі (*pay-as-you-go*).

ChatGPT v4 (у двох варіантах — «загальному» та GPT4 Omni) та Google Gemini 1.5 PRO (Preview) були обрані тому, що вони вже вміють доволі пристойно спілкуватись українською, на відміну, наприклад, від Claude 3 від компанії Anthropic, який час від часу вставляє англійські слова у відповідь.

Якщо вже на ринку присутні системи зі знанням української мови, постає питання: як можна використати цю потужну систему у дистанційній освіті? Варіантів є багато, але у статті увагу буде зосереджено на простому, але важливому прикладі: адаптації іншомовних завдань до українських реалій, оскільки, наприклад, англомовний світ пропонує багато готових баз даних, наприклад, питань із відповідями, створених та верифікованих професіоналами. Використання перевірених даних — це очевидна перевага, оскільки економить час викладача та застерігає його від механічних помилок, особливо у точних науках.

Також треба взяти до уваги, що кількість даних для навчання ШІ українською мовою (особливо академічної чи літературної якості) є набагато меншою, ніж англійською. Відповідно, постраждає і якість відповідей LLM: відповіді можуть бути неповними, непослідовними, не містити очевидних та відомих фактів тощо.

Треба врахувати також факт, що LLM не є спеціалізованим інструментом для розв'язання

завдань із точних наук (на відміну від, наприклад, WolframAlpha, <https://www.wolframalpha.com/>). Відповідно, використання сервісів на базі LLM для отримання подібних розв'язків є ризикованим, оскільки кінцевий результат роботи такого сервісу є лише «побічним продуктом» основного завдання — генерації текстової відповіді.

Аналіз останніх досліджень і публікацій

Українське наукове середовище бере активну участь у полеміці доречності використання ШІ як явища (та ChatGPT зокрема) в навчальному процесі. Леонтєва І.В. артикулює дві популярні та полярні думки щодо використання ШІ в освіті: або заборонити взагалі, або дати простір інноваціям [1]. У статті підкреслено, що студенти вже користуються ChatGPT для підготовки до іспитів чи контрольних робіт, відповідно, освітянам доводиться реагувати на вже наявну технологію, а не тільки дискутувати щодо доцільності імплементації ШІ. Варто зазначити, що ChatGPT вже проходить апробацію в закладах освіти — наприклад, на рівні підготовки магістрів [2]. Крім етичної складової використання ШІ в освіті, існує проблема практичного характеру, а саме: коректність відповідей, які генерує ШІ та придатності у навчальному процесі, особливо — при перекладі. Івашкевич Л. С зазначає [3], що «... гарна якість перекладу з української мови англійською — під час спроб з невеликими обсягами тексту була виявлена дуже незначна кількість помилок, текст сприймається плавно і перевірка його за допомогою сервісу Grammarly показує високу коректність тексту.» Відповідно, при перекладі алгебраїчних завдань (які, зазвичай, не є великими саме за обсягом тексту), можна очікувати високоякісний результат і у зворотному напрямку перекладу.

Актуальність проблематики статті підкреслюють Тарановський А.О. та Самойлов В.Д. [4]: «При цьому створення тестових завдань все ще має базовий рівень автоматизації, за якого комп'ютерні та інформаційні технології використовуються як інструменти введення тексту в електронному вигляді та обміну ним з іншими учасниками процесу створення ККЗ. Текст тестових завдань все ще створює безпосередньо людина як автор, як експерт у предметній області. Комп'ютерні та інформаційні технології дозволяють лише зафіксувати результат і забезпечити його подальше використання».

Шуляк І.М., оцінюючи якість перекладу ChatGPT, застерігає: «... хоча ChatGPT може

надати пропозиції щодо термінології та допомогти із загальними завданнями перекладу, проте рівень експертних знань завжди буде поступатися досвідченому фахівцю перекладачу, людині.» [5].

Якість перекладу можна оцінювати як за допомогою зовнішніх спеціалізованих сервісів, так і за допомогою спеціально натренованої нейронної мережі, наприклад, BERT [6].

Нещодавнє дослідження, проведене GoogleDeepMind у співпраці з Університетом Іллінойсу щодо набору завдань GSM8K [7] свідчить, що за допомогою спеціальних технік якість розв'язку наближається до 95% для англійської мови.

Метою статті є дослідження перспективного напрямку інтеграції LLM-моделі в систему дистанційної освіти, якщо почати з найпростішого — адаптації англійськомовних завдань українською мовою. Під «адаптацією» у цьому випадку будемо розуміти не тільки переклад завдання з англійської, але і заміну власних імен людей в задачах (наприклад, Беатріс стане Наталкою, Мішель — Михайлом та ін.), одиниць виміру (мили стануть кілометрами, фунти — кілограмами, але без фактичного перерахунку в випадку даного дослідження — для контролю якості), валюти (долари США — гривнями), при цьому зберігаючи зміст завдання та його контекст для подальшого аналізу. Для прикладу взятий набір завдань з алгебри для розвитку математичних навичок в учнів середньої школи. Алгебраїчні завдання були обрані саме тому, що відповідь на них є універсальною, не залежить від мови постановки задачі, власної думки викладача чи учня. Відповідно, саме завдяки цій універсальності існує можливість оцінити коректність перекладу не тільки граматично, але і по суті, якщо система буде здатна зрозуміти адаптовану задачу та згенерувати вірний розв'язок.

Виклад основного матеріалу

Власники кожної з LLM-моделей поперджають, що результати роботи системи потрібно перевіряти власноруч, оскільки результати генерації можуть (і, скоріше за все, будуть) містити помилки, навіть фактичні. Добре відомим та очевидним прикладом таких помилок є генерація зображень людей з сімома пальцями на руці — тобто результати не проходять навіть базової перевірки на відповідність реальності (як текстові, так і візуальні).

Вчені розробляють різноманітні стратегії для підвищення точності та розуміння великих

мовних моделей, таких як «доповнення пошуком» та «міркування за допомогою ланцюга думок». Серед них «самовиправлення» — техніка, коли LLM уточнює власні відповіді — набула значного розповсюдження, продемонструвавши ефективність у численних прикладах.

У рамках поточного дослідження лише граматичної коректності перекладу замало: ще треба зберегти контекст та сенси завдання, тому питання визначення «якості» адаптованого тексту буде складнішим. Крім того, необхідно визначити, чи може LLM виступати помічником викладача, розтлумачуючи учневі незрозумілі моменти у розв'язанні конкретної задачі.

У дослідженні будуть розкриті наступні моменти:

1. Якість адаптованого питання.
2. Чи здатна система коректно розв'язати оригінальне завдання?
3. Чи впливає мова завдання на коректність розв'язку і наскільки великий цей вплив?
4. Чи здатна система детально пояснити розв'язок конкретної задачі?
5. Буде перевірено, чи може виступати LLM помічником викладача українською мовою, пояснюючи незрозумілі чи неочевидні моменти.

Використано набір завдань GSM8K (<https://github.com/openai/grade-school-math>), який було створено Карлом Коббом (Karl Cobbe) та Вінетом Косараджу (Vineet Kosaraju). GSM8K — це колекція, що налічує 8500 математичних задач високої якості для учнів початкових/середніх класів, часто використовується в академічній площині для навчання LLM [8]. Цей набір широко використовується для навчання та перевірки знань ШІ (Google Scholar видає понад 1000 статей про дослідження з використанням GSM8K).

Завдання були розроблені та перевірені кваліфікованими авторами. Цей набір поділений на 7,5 тисячі задач для прогресивного навчання та 1300+ задач для фактичного тестування. Розв'язання кожної задачі вимагає від 2 до 8 кроків і здебільшого базується на використанні базових арифметичних операцій (+, -, /, *), що приводять до кінцевого результату. Учень середньої школи з хорошим рівнем підготовки повинен бути здатний розв'язати будь-яку з цих задач.

У дослідженні використано відомий OpenSource проєкт LanguageTool (<https://languagetool.org/>), який спеціалізується якраз на перевірці граматики та підтримує українську мову. Використання цього інстру-

Таблиця 1 – Дослідження коректності розв'язків «тестових» завдань

Параметр	Перший прохід ChatGPT	Перший прохід Gemini	Другий прохід ChatGPT	Другий прохід Gemini	Третій прохід ChatGPT	Третій прохід Gemini
Найменший час генерування відповіді, сек.	2.32	2.13	3.12	2.11	2.44	2.02
Найбільший час генерації відповіді, сек.	33.07	10.50	29.15	9.34	34.07	7.12
Середній час генерації відповіді, сек.	8.32	3.59	7.14	3.31	6.29	3.14
% помилок (англійська)	47.5%	76.5%	51.44%	75.4%	52.64%	74.12%
% помилок (українська)	48%	77.5%	53.2%	74.2%	51.4%	75.2%
% відмінності у розв'язках різними мовами	0.5%	1%	1.76%	1.2%	1.24%	1.08%
% граматичних зауважень для адаптацій	28.5%	32%	29%	31%	27.4%	32.9%

менту важливе тому, що це — одне з небагатьох рішень, яке дозволяє автоматизувати подібну перевірку та вже використовується українськими науковцями [9-12].

LanguageTool перевірятиме граматичну коректність адаптації на наявність помилок або зауважень. Лише ті адаптації, до яких зауваження або помилки відсутні, будуть вважатися такими, що пройшли перевірку. Відповідно, надалі під «процентом граматичних зауважень для адаптацій» розумітиметься відсоток адаптацій від загальної кількості, у яких виявлені проблеми.

Зазвичай, питання, віднесені до категорії «тестових», є складнішими як у лінгвістичному, так і у математичному сенсі.

Оцінимо, наскільки добре взагалі LLM може справитись із завданнями з набору GSM8K — як в оригінальному (англійському) варіанті, так і в адаптованому українському, якщо не надавати йому правильної відповіді.

Для дослідження були взяті перші 200 питань (як і у дослідженні DeepMind) GSM8K з «тестового» набору, який є складнішим, та було зроблено три проходи по набору в різні дні та години, оскільки LLM щораз генерує унікальну відповідь і були відмічені випадки, коли LLM розв'язував одну і ту саму задачу правильно, потім неправильно, потім знову правильно. Сервіс в одній відповіді мав поєднати три результати:

1. Адаптувати текст українською мовою;
2. Навести числовий розв'язок задачі англійською мовою;
3. Навести числовий розв'язок завдання за адаптованим, українським описом.

Для ChatGPT була використана модель «gpt-4-turbo-2024-04-09» (платна, найновіша на момент дослідження) без додаткового навчання за тренінговим набором GSM8K. Причина такого рішення проста: тренінг моделі вимагає спеціальних знань, даних та підготовки від користувача, тому для загального вжитку краще розуміти, на що здатна ця модель у режимі «як є», оскільки ChatGPT постійно вдосконалюється.

Для GoogleGemini 1.5 ProPreview (який офіційно ще не був доступний в Україні) використовувався VPN та також не було додаткового навчання.

Формулювання завдання для обох систем виглядала наступним чином: **Щоб адаптувати англійськомовну математичну задачу для українських школярів, потрібно: а) замінити усі власні англійськомовні імена українськими, б) замінити (без конвертації) валютні одиниці на українські, в) замінити (без конвертації) одиниці виміру на ті, що прийняті в Україні. Адаптований текст треба переписати таким чином, щоб він відповідав нормам української мови, був чітким та лаконічним, перевірити відмінки.**

Результати проходів містяться у таблиці 1.

Результати коректності неприємно вражають: фактично, кожна друга задача була розв'язана невірно, а Gemini взагалі впорався жадливо. Мінімальна різниця у відповідях на різних мовах — це показник того, що мова завдання не має вирішального значення.

Якість перекладу в ~30% помилок української адаптації демонструє, що LLM не здатні поки що обходитись без зовнішнього редагу-

Таблиця 2 – Дослідження коректності розв'язків «тестових» завдань мовою оригіналу

Параметр	Перший прохід ChatGPT	Перший прохід Gemini	Другий прохід ChatGPT	Другий прохід Gemini	Третій прохід ChatGPT	Третій прохід Gemini
Найменший час генерації відповіді, сек.	0.49	1.01	0.51	1.24	0.46	1.08
Найбільший час генерації відповіді, сек.	6.23	4.15	7.48	7.66	9.12	3.54
Середній час генерації відповіді, сек.	1.57	1.23	1.15	2.05	0.98	1.54
% помилок	47.45%	75.53%	49.24%	73.14%	50.56%	74.65%

Таблиця 3 – Дослідження коректності розв'язків «тестових» завдань мовою оригіналу з обов'язковим виводом детального шляху розв'язання задачі

Параметр	Перший прохід ChatGPT	Перший прохід Gemini	Другий прохід ChatGPT	Другий прохід Gemini	Третій прохід ChatGPT	Третій прохід Gemini
Найменший час генерації відповіді, сек.	1.99	1.46	2.71	1.54	2.13	2.06
Найбільший час генерації відповіді, сек.	23.71	8.18	31.22	9.23	24.55	10.65
Середній час генерації відповіді, сек.	6.44	3.24	7.03	4.05	7.24	5.31
% помилок	14.5%	24.77%	14.71%	25.74%	14.34%	23.83%

вання. Але, оскільки той же LanguageTool не просто підказує помилки, а ще і пропонує варіанти їх виправлення, то фактичну підготовку українського тексту можна автоматизувати, суттєво спростивши роботу редактора.

Кожна LLM є дуже чутливою до того, як саме поставлене завдання. Наступним кроком буде досліджуватись, що саме може впливати на коректність алгебраїчного розв'язку.

Спочатку спростимо початкові умови й вимагатимемо тільки числовий розв'язок (адже попередньо система виконувала і переклад) та перевіримо, чи ситуація поліпшиться. Для початку використаємо оригінальний варіант постановки – англійський.

Формулювання завдання:

Please solve the mathematical problem. Results should contain numeric value only.

Суттєвої різниці (табл. 2), з погляду коректності розв'язків, не відбулося, але, вочевидь, зменшився час на аналіз даних та формування відповіді.

Якщо вимагати від LLM показати детальне пояснення відповіді, то чи це вплине на кінцевий результат?

Для наступного експерименту, система була змушена демонструвати, крім, власне, чис-

лової відповіді, ще і шлях розв'язання задач в оригінальній постановці – англійською.

Формулювання завдання для LLM:

You are a professional math teacher. When I ask You something, You should consider this as a math problem and solve it for me.

Результати — у таблиці 3.

Ситуація покращилась: LLM зменшили кількість помилок до 14-15% та 24-25% відповідно. Отже, існує чітка залежність не тільки від структури підказки для системи, але і від об'єму виводу та детальності даних, які в системі вимагають.

Наступним кроком дослідимо якість адаптованих завдань українською мовою, зробивши акцент на роз'ясненні розв'язку задачі. Для більшої наочності проведемо 4 експерименти з перехресними перевірками:

1. ChatGPT розв'язує завдання, які адаптовані ChatGPT;
2. Gemini розв'язує завдання, які адаптовані ChatGPT;
3. ChatGPT розв'язує завдання, які адаптовані Gemini;
4. Gemini розв'язує завдання, які адаптовані Gemini.

Таблиця 4 – Дослідження коректності розв'язків «тестових» завдань в адаптованому варіанті з обов'язковим виведенням детального шляху розв'язання задачі

Параметр	ChatGPT розв'язує адаптацію ChatGPT	Gemini розв'язує адаптацію ChatGPT	ChatGPT розв'язує адаптацію Gemini	Gemini розв'язує адаптацію Gemini
Найменший час генерації відповіді, сек.	3.86	2.04	2.51	0.55
Найбільший час генерації відповіді, сек.	54.87	10.94	38.77	7.54
Середній час генерації відповіді, сек.	13.25	3.55	11.44	3.40
% помилок	15.5%	49.74%	24.5%	50.5%
% коректно розв'язаних завдань у випадку відсутності граматичних помилок в адаптації	62%	40%	54%	36.5%
% коректно розв'язаних завдань без граматичних помилок в адаптації та без граматичних помилок у роз'ясненні	44.5%	36%	43.5%	30.5%

Формулювання завдання:

Ти — викладач математики у школі. Я тобі надсилаю задачу, яку треба розв'язати.

Результати — у таблиці 4.

Як бачимо, для ChatGPT відчутної різниці у якості розв'язків залежно від мови завдання не спостерігається (для власної адаптації). Треба зауважити, що знаково зменшилась швидкість генерації відповіді системою.

Gemini впорався з завданнями набагато гірше, але майже однаково наробив помилок в обох випадках, хоча й надавав відповіді суттєво швидше. Відмічено блокування зі сторони системи деяких завдань без пояснення причини.

У 62% випадків ChatGPT зумів згенерувати адаптацію без граматичних зауважень від LanguageTool, коректно вирішити завдання, але до роз'яснення, з огляду на граматику, були зауваження.

У 44.5% випадків ChatGPT адаптація та детальний розв'язок були без граматичних зауважень від LanguageTool та LLM змогла коректно вирішити завдання.

Для Gemini результати скромніші – 36% та 30.5% відповідно.

На початку травня 2024 року компанія OpenAI, розробник ChatGPT, презентувала свою найновішу модель — ChatGPT v4 Omni, або GPT4o. Заявлена підвищена точність цієї моделі у порівнянні з «звичайною» GPT v4.

З попередніх досліджень відомо, що найкраща точність спостерігається тоді, коли запитується повний шлях «роздумів» до задачі, тому GPT4o була перевірена на: а) граматичну коректність адаптованого питання; б) здатність розв'язати оригінальне завдання; в) здатність

розв'язати адаптоване нею ж питання українською мовою.

У таблиці 5 представлені значущі результати цієї моделі в експериментах, описаних вище.

Таблиця 5 – Результати ChatGPT v4 Omni

Параметр	Результат
% граматичних зауважень для адаптацій	28.5%
% помилок (англійська)	6.5%
% помилок (українська адаптація від GTP4o)	12%

Очевидно, що GTP4o дійсно демонструє кращу точність, ніж попередньо розглянуті моделі, тому для подальшого аналізу підходить саме вона.

Проаналізуємо, які саме завдання і чому GTP4o не розв'язала для української адаптації (але розв'язала для англійського варіанту).

Таблиця 6 – Типові помилки при розв'язанні ChatGPT v4 Omni

Помилка	Результат
Аналітична помилка при розв'язку української адаптації (сама адаптація коректна)	9
Логічна помилка в адаптації	3
Помилка в одиницях вимірювання / валюті	1

Зустрічаються випадки, коли ChatGPT перекладає слово *dozen* як «десяток», а не «дюжину», що спотворює подальший розв'язок чи

невірно розміщає числа в адаптації, якщо їх в тексті задачі багато.

Також один раз зафіксована конвертація доларів США у гривні, хоча це прямо було заборонено.

Спробуємо покращити результат, якщо підказати ChatGPT v4o виправляти суперечливі моменти, використавши більш деталізовану підказку:

Необхідно адаптувати англомовну математичну задачу для українських школярів.

Для цього треба:

а) замінити усі власні англомовні імена на українські

б) замінити долар США на гривні, без перерахунку або конвертації, використати оригінальне число,

в) замінити одиниці виміру на прийняті в Україні без перерахунку або конвертації, використати оригінальне число.

г) Перекласти слово "dozen" та його похідні як "дюжина" або "дванадцять".

д) Перевірйай, що б числа в адаптації були числами з англійського варіанту завдання, відмінностей не має бути.

Адаптований текст треба переписати таким чином, що б він відповідав граматичним та орфографічним нормам української мови, був чітким та лаконічним.

Таблиця 7 – Результати ChatGPT v4 Omni після застосування більш детальної підказки

Параметр	Результат
% граматичних зауважень для адаптації	28%
% помилок	10%

Якість адаптації в кращу сторону майже не змінилась, але, вочевидь, підріс процент коректних розв'язків. Аналіз помилкових розв'язків показав, що всі вони лежать в аналітичній площині.

Перевіримо, як саме LLM розв'язує та детально пояснює шлях своїх «роздумів» за адаптованим варіант (якщо він розв'язав задачу коректно):

«Покажи детальний розв'язок задачі:

Щодня Людмила годує кожну зі своїх курей трьома склянками змішаного курячого корму, що містить насіння, хробаків та овочі, щоб зберегти їх здоров'я. Вона дає курям корм тричі на день. Вранці вона дає своєму стаду курей 15 склянок корму. Вдень вона дає своїм курям ще 25 склянок корму. Скільки склянок корму вона повинна дати своїм курям у остан-

ній прийом їжі на день, якщо розмір стада Людмили становить 20 курей?

Розв'яжемо задачу крок за кроком.

Спочатку визначимо, скільки склянок корму Людмила повинна дати одній курці за день. За умовою, кожна курка отримує 3 склянки корму на день. Оскільки у Людмили 20 курей, то весь стадо потребує:

$3 \text{ склянки на курку} \times 20 \text{ курей} = 60 \text{ склянок корму за день}$

Тепер розглянемо, скільки склянок корму Людмила вже дала своїм курям протягом дня. Вранці вона дає 15 склянок, а вдень - 25 склянок. Разом за перші дві частини дня:

$15 \text{ склянок} + 25 \text{ склянок} = 40 \text{ склянок}$

Таким чином, щоб забезпечити курей необхідною кількістю корму на день, вона повинна дати їм ще:

$60 \text{ склянок} - 40 \text{ склянок} = 20 \text{ склянок}$

Отже, Людмила повинна дати своїм курям 20 склянок корму у останній прийом їжі на день.

Відповідь: 20 склянок»[13]

Ми бачимо, що у цьому випадку, LLM успішно розв'язала завдання та логічно пояснила, як вона дійшла до відповіді. Український текст міг би бути більш літературним, та є певні граматичні неточності, але збережено сенс матеріалу, подача розв'язку послідовна та не залишає «білих плям».

Висновки

Чи здатна LLM адаптувати англомовне завдання? Так, з обмеженнями. Треба врахувати, що близько 30% таких завдань будуть містити граматичні зауваження. Залучення людини-редактора та зовнішньої системи для контролю граматики на фінальному етапі є обов'язковим.

Чи здатна система на базі LLM коректно розв'язати оригінальне завдання? Не рекомендовано використовувати LLM без незалежного контролю якості. LLM надзвичайно залежні від формату запиту та кількості (точності) даних, які вона має згенерувати. Наприклад, якщо вимагати детального розв'язку, а не тільки числового розв'язку, то точність виросте від ~50% до 90+%.

Чи здатна система коректно розв'язати адаптоване завдання? Так. Обмеження такі ж самі, як і до англійської версії, за умови коректності самої адаптації.

Чи впливає мова завдання на коректність розв'язку і наскільки великий цей вплив? У випадку ChatGPT вплив коливається в межах 5%.

Чи здатна система детально пояснити розв'язок конкретної задачі? Здатна, але у LLM відсутній самоконтроль відповіді, тому є можливість отримати некоректний розв'язок. Не можна рекомендувати без наявності незалежного контролю від викладача або зовнішньої системи.

Результати доволі красномовні: ChatGPT 4 Omni справляється з поставленими завданнями (як адаптацією, так і розв'язком) краще, ніж Gemini, останню важко рекомендувати для практичного застосування.

ChatGPT можна використовувати для адаптації англійських колекцій питань в українські системи дистанційної освіти — він прискорить роботу редактора-перекладача в десятки разів. Наприклад, якщо редактор витратить 3-4 хвилини на лише одне питання (нехай навіть з допомогою автоматичного перекладача), то на 1300 питань доведеться зачекати 3900 хвилин (~65 годин). ChatGPT ж впорається за максимальні три години — тобто, у 20 разів швидше. Відразу необхідно зауважити, що у дослідженні для умови «якісної адаптації» було достатньо, аби LLM коректно зрозуміла та розв'язала завдання в адаптованому варіанті, хоча в реальності вимоги будуть жорсткіші. Процент граматичних зауважень від LanguageCheck є істотно більшим, ніж процент неправильних розв'язків (30% проти 10%), отже людині потрібно буде витратити час на перевірку якості адаптації шляхом звіряння результатів зворотного контролю з оригіналом. Спочатку — з неправильними відповідями, а вже потім — перевірити граматику.

Звичайно, навіть в умовно «правильних» завданнях можуть бути стилістичні помилки — але вони, вочевидь, не будуть настільки критичними, що заважатимуть учневі розв'язати задачу. Такі виправлення можуть бути вже внесені в робочому режимі. Результати адаптації слід вважати *чернеткою*, коли ChatGPT виконав монотонну та механічну роботу, але без допомоги редактора фінального результату не досягти. Комбінований підхід зменшить навантаження на викладача, підвищить якість матеріалу для учнів та «відкриє двері» для інтеграції іншомовних колекцій завдань в вітчизняні системи дистанційної освіти.

Щодо роботи помічником викладача, то тут необхідно підходити з великою обережністю. Результат такої допомоги перед демонстрацією учневі обов'язково має проходити перевірку або в автоматичному режимі (якщо це можливо), або з залученням людини-викладача.

Якість розв'язку підвищується, якщо явно вказати ChatGPT вказати на правильну відповідь.

У підсумку, можна сформулювати наступні практичні рекомендації:

1. Ставити перед LLM за один раз лише одне, конкретне завдання, а не поєднувати декілька (наприклад, не просити одночасно і перекласти задачу, і розв'язати її);

2. Давати максимально деталізовану підказку для системи: що саме їй треба зробити — чим точніше вона буде, тим якісніший буде результат;

3. Для неоднозначних моментів бажано вказати, на яку аудиторію буде розрахована відповідь (школярі, студенти) — відповідно, LLM буде використовувати лексику, яка більш підходить для цього випадку;

4. Пам'ятати, що результати обчислень мають бути обов'язково перевірені зовнішніми засобами.

5. Необхідна граматична перевірка отриманих результатів (наприклад, в автоматичному режимі). Це особливо стосується інтеграції LLM-моделей у вітчизняні системи дистанційного навчання.

Література

1. Леонтєва І. В. ChatGPT в освітньому процесі вищої школи: заборонити не можна використовувати. *Освіта та педагогічна наука*. 2023. № 1 (183). С. 13–23. URL: <http://eps.luguniv.edu.ua/index.php/eps/issue/view/17> (дата звернення: 30.04.2024)

2. Балик Н. Р., Шмигер Г. П. Впровадження штучного інтелекту в освіту шляхом використання ChatGPT. *Актуальні аспекти розвитку STEAM-освіти в умовах євроінтеграції*: зб. матеріалів Міжнар. наук.-практ. інтернет-конф., м. Кропивницький, 21 квіт. 2013 р. 2023. С. 147–149. URL: <http://dSPACE.tnpu.edu.ua/handle/123456789/28824> (дата звернення: 30.04.2024)

3. Івашкевич І. Застосування мовної моделі ChatGPT для перекладу: переваги та недоліки. *Дискурс професійної і творчої комунікації: лінгво-культурний, когнітивний, перекладацький та методичний аспекти*: зб. матеріалів VIII Міжнар. наук.-практ. конф., м. Київ, 18–19 трав. 2023 р. Київ, 2023. С. 110–113. URL: https://ktpnm.kpi.ua/sites/default/files/2023-11/ДПТК_блок_тези_18_05_23.pdf (дата звернення: 30.04.2024)

4. Тарановський А. О., Самойлов В. Д. ChatGPT і можливість його використання для безекспертного створення тестів. *Elektronnoe*

Modelirovanie. 2023. Т. 45, № 2. С. 44–60. DOI: <https://doi.org/10.15407/emodel.45.02.044> (дата звернення: 30.04.2024)

5. Шуляк І. М. Chat-GPT у теорії й практиці перекладу: виклики й перспективи. *Міжкультурна комунікація і перекладознавство: точки дотику та перспективи розвитку*: матеріали VI Міжнар. наук.-практ. конф.: зб. наук. пр., м. Переяслав, 12–13 трав. 2023 р. Переяслав, 2023. С. 101–104. URL: <http://dspace.tnpu.edu.ua/handle/123456789/31746> (дата звернення: 30.04.2024)

6. Одуха В. С. Спосіб визначення якості автоматизованого перекладу з використанням нейронної мережі : кваліфікаційна робота бакалавра: 122. Хмельницький, 2023. 83 с. URL: <https://elar.khmnpu.edu.ua/handle/123456789/13796>. (дата звернення: 30.04.2024)

7. Huang et al. J. Large Language Models Cannot Self-Correct Reasoning Yet / URL: <https://arxiv.org/abs/2310.01798> (date of access: 03.06.2024).

8. Cobbe et al. K. Training Verifier to Solve Math Word Problems / URL: <https://arxiv.org/abs/2110.14168> (date of access: 30.04.2024).

9. Федорович І. Процес оновлення словника української мови для морфологічного аналізатора Руморфy2. *VI Міжнародна студентська науково-технічна конференція "Природничі та гуманітарні науки. Актуальні питання"*: матеріали Міжнар. студент. наук.-техн. конф., м. Київ, 28–29 квіт. 2022 р. Київ, 2022. С. 187–188. URL: <https://elartu.tntu.edu.ua/bitstream/lib/41264/2/187-188.pdf> (дата звернення: 30.04.2024)

10. Мінгальова Ю. Огляд он-лайн сервісів, що варто використовувати для саморедкування наукового тексту. *Актуальні питання сучасної інформатики*, 2019. С. 122–126. URL: <http://eprints.zu.edu.ua/id/eprint/28104> (дата звернення: 30.04.2024)

11. Побудова словника предметної області на основі автоматизованого аналізу текстів українською мовою / О. Кунгурцев та ін. *Технічні науки та технології*. 2016. № 3 (5). С. 164–174. URL: http://nbuv.gov.ua/UJRN/tnt_2016_3_21 (дата звернення: 30.04.2024)

12. Automatic correction system of english-ukrainian computer translation for technical texts in the field of automation of technological processes / A. O. Stopakevych et al. *Informatics and mathematical methods in simulation*. 2023. Vol. 13, No 3-4. P. 322–334. DOI: <https://doi.org/10.15276/imms.v13.no3-4.322> (date of access: 03.06.2024).

13. ChatGPT v4o (12 червня 2024 р.). Покажи детальний розв'язок задачі: Щодня Людмила годує кожну зі своїх курей трьома склянками змішаного курячого корму, що містить насіння, хробраків та овочі, щоб зберегти їх здоров'я. Вона дає курям корм тричі на день. Вранці вона дає своєму стаду курей 15 склянок корму. Вдень вона дає своїм курям ще 25 склянок корму. Скільки склянок корму вона повинна дати своїм курям у останній прийом їжі на день, якщо розмір стада Людмили становить 20 курей? Джерело: <https://chatgpt.com/share/922ef0ca-5aba-41a0-9620-5fddbc79cf3d>

References

1. Leontieva I. V. ChatGPT v osvitnomu protsesi vyshchoi shkoly: zaboronyty ne mozhna vykorystovuvaty. *Osvita ta pedahohichna nauka*. 2023. No 1 (183). P. 13–23. URL: <http://eps.luguniv.edu.ua/index.php/eps/issue/view/17> (date of access: 30.04.2024) [in Ukrainian]

2. Balyk N. R., Shmyher H. P. Vprova-dzhennia shtuchnogo intelektu v osvitu shliakhom vykorystannia ChatGPT. *Aktualni aspekty rozvytku STEAM-osvity v umovakh yevrointehratsii* : zb. materialiv Mizhnar. nauk.-prakt. internet-konf., m. Kropyvnytskyi, 21 kvit. 2013 r. 2023. P. 147–149. URL: <http://dspace.tnpu.edu.ua/handle/123456789/28824> (date of access: 30.04.2024) [in Ukrainian]

3. Ivashkevych I. Zastosuvannia movnoi modeli ChatGPT dlia perekladu: perevahy ta nedoliki. *Dyskurs profesiinoi i tvorchoi komunikatsii: linhvokulturnyi, kohnityvnyi, perekladatskyi ta metodychnyi aspekty* : zb. materialiv VIII Mizhnar. nauk.prakt. konf., m. Kyiv, 18–19 trav. 2023 r. Kyiv, 2023. P. 110–113. URL: https://ktpnm.kpi.ua/sites/default/files/2023-11/DPTK_blok_tezy_18_05_23.pdf (date of access: 30.04.2024) [in Ukrainian]

4. Taranovskyi A. O., Samoilo V. D. ChatGPT i mozhlyvist yoho vykorystannia dlia bezekspertnoho stvorennia testiv. *Elektronnoe Modelirovanie*. 2023. Т. 45, No 2. P. 44–60. DOI: <https://doi.org/10.15407/emodel.45.02.044> (date of access: 30.04.2024) [in Ukrainian]

5. Shuliak I. M. Chat-GPT u teorii y praktytsi perekladu: vyklyky y perspektyvy. *Mizhkulturna komunikatsiia i perekladoznavstvo: tochky dotyku ta perspektyvy rozvytku*: materialy VI Mizhnar. nauk.-prakt. konf.: zb. nauk. pr., m. Pereiaslav, 12–13 trav. 2023 r. Pereiaslav, 2023. P. 101–104. URL: <http://dspace.tnpu.edu.ua/handle/123456789/31746> (date of access: 30.04.2024) [in Ukrainian]

6. Odukha V. S. Sposib vyznachennia yakosti avtomatyzovanoho perekladu z vykorystanniam neironnoi merezhi : kvalifikatsiina robota bakalavra: 122. Khmelnytskyi, 2023. 83 p. URL: <https://elar.khmnu.edu.ua/handle/123456789/13796> (date of access: 30.04.2024) [in Ukrainian]
7. Huangetal J. Large Language Models Cannot Self-Correct Reasoning Yet / URL: <https://arxiv.org/abs/2310.01798> (date of access: 03.06.2024).
8. Cobbeetal K. Training Verifiersto Solve Math Word Problems / URL: <https://arxiv.org/abs/2110.14168> (date of access: 30.04.2024).
9. Fedorovych I. Protses onovlennia slovnyka ukrainskoi movy dlia morfolohichnoho analizatora Pymorphy2. VI Mizhnarodna studentska naukovo - tekhnichna konferentsiia "Pryrodnychi ta humanitarni nauky. Aktualni pytannia": materialy Mizhnar. student. naukovo-tekhn. konf., m. Kyiv, 28–29 kvit. 2022 r. Kyiv, 2022. P. 187–188. URL: <https://elartu.tntu.edu.ua/bitstream/lib/41264/2/187-188.pdf> (date of access: 03.06.2024) [in Ukrainian]
10. Minhalova Yu. Ohliad on-lain servisiv, shcho varto vykorystovuvaty dlia samoredahuvannia naukovoho tekstu. *Aktualni pytannia suchasnoi informatyky*. 2019. P. 122–126. URL: <http://eprints.zu.edu.ua/id/eprint/28104> (date of access: 03.06.2024) [in Ukrainian]
11. Pobudova slovnyka predmetnoi oblasti na osnovi avtomatyzovanoho analizu tekstiv ukrainskoiu movoiu / O. Kunhurtsev ta in. *Tekhnichni nauky ta tekhnolohii*. 2016. No 3 (5). P. 164–174. URL: http://nbuv.gov.ua/UJRN/tnt_2016_3_21 (date of access: 03.06.2024) [in Ukrainian]
12. Automatic correction system of english-ukrainian computer translation for technical texts in the field of automation of technological processes / A. O. Stopakevych ta in. *Informatics and mathematical methods in simulation*. 2023. Vol. 13, No 3-4. P. 322–334. DOI: <https://doi.org/10.15276/imms.v13.no3-4.322> (date of access: 03.06.2024).
13. ChatGPT v4o (12 chervnia 2024 r.). Pokazhy detalnyi rozviazok zadachi: Shchodnia Liudmyla hoduie kozhnu zi svoikh kurei trom skliankamy zmishanoho kuriachoho kormu, shcho mistyt nasinnia, khrobakiv ta ovochi, shchob zberehty yikh zdorovia. Vona daie kuriam korm trychi na den. Vrantsi vona daie svoiemu stadu kurei 15 sklianok kormu. Vden vona daie swoim kuriam shche 25 sklianok kormu. Skilky sklianok kormu vona povynna daty swoim kuriam u ostannii pryiom yizhi na den, yakshcho rozmir stada Liudmyly stanovyt 20 kurei? Dzherelo: <https://chatgpt.com/share/922ef0ca-5aba-41a0-9620-5fddbc79cf3d> [in Ukrainian]