IZDEVNIECĪBA
BALTIJA
PUBLISHING

TRANSFORMATION OF THE SCIENTIFIC AREA IN THE CONTEXT OF CONTEMPORARY CHALLENGES

Scientific monograph

Riga, Latvia
2025

UDK 001(082)
MO045

Title: Transformation of the scientific area in the context of contemporary challenges
Subtitle: Scientific monograph
Scientific editor and project director: Anita Jankovska
Authors: Yuliia Vakal, Olha Molodchenkova, Yaroslav Fanin, Maksym Hanchuk, Viktoriia Skyba, Olena Venhrina, Nataliia Dankevych, Oksana Dobrovolska, Ganna Solodovnyk, Olena Shapovalova, Yakiv Vorobiov, Olesia Pavlenko, Yuri Bandazheuski, Nataliia Dubovaya, Olena Serhieieva, Volodymyr Pundiev, Hanna Ivaniuk, Svitlana Tsybulska, Nataliia Pavliuk, Tetiana Spychak, Oleksandr Korchev, Nataliia Bilous, Liudmyla Diachuk, Inna Dovzhenko, Roksolana Verbova, Oleg Baklan, Tetiana Fedorenko, Larysa Yerofieienko, Iryna Ihnatchenko, Yaroslava Ryabchenko, Yakym Yakymovych, Piotr Zięba, Paulina Kolisnichenko, Inna Boychuk, Yevhen Vorobets, Taisiia Martynova, Andrii Tatarchenko
Publisher: Publishing House "Baltija Publishing", Riga, Latvia
Available from: http://www.baltijapublishing.lv/omp/index.php/bp/catalog/book/676
Year of issue: 2025

Transformation of the scientific area in the context of contemporary challenges : Scientific monograph. Riga, Latvia : Baltija Publishing, 2025. 692 p.

ISBN: 978-9934-26-631-7

DOI: https://doi.org/10.30525/978-9934-26-631-7-1

This scientific monograph presents research on theoretical and practical areas of science in the context of contemporary challenges. The publication encompasses a wide range of subjects within the natural and agricultural sciences, legal and economic sciences, as well as philological and pedagogical sciences. The publication is intended for a wide range of readers, including scientists, educators, postgraduate students, and students.

Izdevniecība “Baltija Publishing”
Valdeķu iela 62 – 156, Rīga, LV-1058
E-mail: office@baltijapublishing.lv

Iespiests tipogrāfijā SIA “Izdevniecība “Baltija Publishing”
Parakstīts iespiešanai: 2025. gada 19. decembrī
Tirāža 300 eks.

Table of Contents

CHAPTER «ECONOMIC SCIENCES»

CHAPTER «ENGINEERING SCIENCES»

DO RAG SYSTEMS PROVIDE AN ADVANTAGE? AN EMPIRICAL STUDY OF CHATGPT AND NOTEBOOKLM EFFECTIVENESS IN GENERATING ACADEMIC TESTS

ЧИ НАДАЮТЬ RAG-СИСТЕМИ ПЕРЕВАГУ? ЕМПІРИЧНЕ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ CHATGPT ТА NOTEBOOKLM ПРИ СТВОРЕННІ ТЕСТІВ З АКАДЕМІЧНИХ ДИСЦИПЛІН

Olena Venhrina[1]

DOI: https://doi.org/10.30525/978-9934-26-631-7-4

Abstract. In the context of the rapid digitalization of education, automating the creation of assessment materials has become a critical task for educators. The emergence of Large Language Models (LLMs) has opened new perspectives for generating test items; however, the question of selecting the optimal toolkit remains unresolved within the educational community. Specifically, there is a need for empirical verification of the hypothesis regarding whether specialized tools with Retrieval-Augmented Generation (RAG) architecture, such as NotebookLM, provide a significant advantage in quality and reliability over universal chatbots (e.g., ChatGPT) when used by subject teachers lacking complex prompt engineering skills. *Objective.* The study aims to conduct a comparative analysis of the quality, structural correctness, and cognitive depth of multiple-choice test questions generated using three different AI usage scenarios. *Methods.* The experimental basis was the educational material of the "Database Organization and Storage" discipline (topics: SQL DDL, DML, Aggregation). A pool of 90 test questions was generated across

[1] Candidate of Technical Sciences, Associate Professor,
Department of Cybersecurity and Information Technologies,
Simon Kuznets Kharkiv National University of Economics, Ukraine

three scenarios: (A) generation in ChatGPT based solely on the topic title; (B) generation in ChatGPT based on uploaded lecture notes; (C) generation in NotebookLM based on an uploaded source. The quality of the obtained content was evaluated by three independent experts on a 5-point scale according to the following criteria: factual correctness, relevance to the topic, and quality of distractors (incorrect answer options). Additionally, questions were classified according to a simplified Bloom's taxonomy. To verify statistical hypotheses, the non-parametric Kruskal-Wallis H-test and Pearson's χ^2 test of independence were used. *Results.* Statistical analysis revealed no significant differences ($p > 0.05$) between the three scenarios for any of the quality criteria. A pronounced "ceiling effect" was recorded for the "relevance" criterion (mean scores 4.8–4.99), indicating the high competence of base models in standard academic topics even without providing context. At the same time, it was found that the NotebookLM tool demonstrated technical instability when generating content in Ukrainian, specifically omitting individual words in question formulations, which led to significantly higher variability in "correctness" scores ($SD = 1,01$) compared to the stable results of ChatGPT. Analysis using Bloom's taxonomy confirmed that switching to a RAG system does not automatically increase the cognitive complexity of tasks: the majority of questions in all groups remained at the levels of remembering and understanding. Furthermore, generating plausible distractors remains a weak point for all examined AI tools. *Conclusions.* The findings refute the assumption of the unconditional advantage of RAG systems for generating tests in standardized disciplines. For an educator, the use of universal chatbots with simple prompts is the most effective method in terms of the time-to-quality ratio. The use of specialized tools (NotebookLM) is advisable primarily for working with unique authorial materials; however, it requires increased attention to the verification of each generated question.

1. Вступ

Однією з актуальних задач, яку потрібно вирішувати викладачу протягом навчального року, є розробка та оновлення контрольно-вимірювальних матеріалів, зокрема тестів для поточного та підсумкового контролю. Створення високоякісних запитань вимагає від викладача значних ресурсів, професійної підготовки та досвіду [13, с. 2559].

Особливо це стосується розробки питань з декількома варіантами відповідей, де необхідно не лише сформулювати коректну відповідь, але й створити валідні дистрактори – правдоподібні, але хибні варіанти [6, с. 2]. Як зазначають [3, с. 243], створення доречних дистракторів є виснажливим завданням, що вимагає специфічної когнітивної навички: здатності розуміти та передбачати поширені помилки або хибні уявлення, характерні для студентів [5, с. 3068].

Стрімкий розвиток технологій штучного інтелекту, а саме великих мовних моделей (LLM), відкрив нові перспективи для автоматизації цього процесу. Інструменти на кшталт ChatGPT та NotebookLM, демонструють здатність виконувати широкий спектр завдань з генерації текстів, що актуалізує питання про доцільність їхнього застосування в освіті [10, с. 209].

Наукова спільнота вже активно досліджує проблему автоматичної генерації питань. Перший напрям досліджень фокусується на здатності універсальних LLM (наприклад, ChatGPT) створювати тестові завдання. Дослідники підтверджують, що такі моделі можуть швидко генерувати синтаксично правильні питання [12, с. 8]. Однак, [2, с. 6; 8, с. 5-7] наголошують на високих ризиках, пов’язаних із "галюцинаціями", неточностями та фокусуванням ШІ на завданнях нижчого когнітивного рівня, що робить людську верифікацію абсолютно необхідною.

Другий напрям досліджень присвячений оцінці якості та валідності згенерованого контенту. [6, с. 9-10] зазначають, що навіть фактологічно коректні питання не завжди відповідають педагогічним цілям. Водночас, [9, с. 374] виявили, що LLM мають потенціал у створенні якісних дистракторів, оскільки схильні обирати ті самі помилкові варіанти, що й студенти.

Третій, найбільш перспективний напрям, стосується використання архітектури Retrieval-Augmented Generation (RAG). Системи на кшталт NotebookLM доповнюють модель інформацією з конкретної бази знань, що дозволяє "заземлити" відповіді на фактичному матеріалі. За даними [11, с. 2; 4, с. 5-6], це теоретично має підвищувати точність, зменшувати ймовірність "галюцинацій" та забезпечувати верифікованість результату порівняно з генерацією "наосліп".

Попри значну кількість досліджень [1, с. 6-8], поява загальнодоступних RAG-систем кардинально змінює ландшафт, вперше відкри-

ваючи можливості для їх широкого застосування викладачами без технічних навичок [7, с. 4-5]. Однак залишається відкритим питання: чи дійсно використання спеціалізованих "заземлених" інструментів надає статистично значущу перевагу у якості та когнітивній глибині питань порівняно з універсальними чат-ботами при вирішенні стандартних викладацьких задач?

Мета дослідження – порівняльний аналіз ефективності застосування різних сценаріїв використання ШІ-помічників (від простих промтів до RAG-технологій) для створення банків тестових завдань.

Об'єктом дослідження є процес генерації тестових питань з використанням універсальних мовних моделей та спеціалізованих інструментів, орієнтованих на роботу з джерелами.

Предметом дослідження виступають якість згенерованого контенту (коректність, релевантність, дистрактори), його когнітивна складність та надійність інструментів при їх використанні викладачем-неспеціалістом в ІТ.

Для досягнення поставленої мети та перевірки висунутих гіпотез було сформульовано наступне дослідницьке питання:

ДП1. Якими є рівень якості (за критеріями фактологічної коректності, релевантності та правдоподібності дистракторів) та когнітивна складність (за таксономією Блума) тестових питань, згенерованих за допомогою трьох різних ШІ-сценаріїв з використанням простих розмовних промтів: (А) ChatGPT на основі виключно назви теми лекції, (Б) ChatGPT на основі завантаженого конспекту лекції та (В) NotebookLM на основі того ж конспекту?

2. Методи

2.1 Матеріали

Експериментальною основою для генерації тестових питань слугував набір з трьох конспектів лекцій з академічної дисципліни "Організація та збереження баз даних", яка викладається українською мовою. Обрані матеріали охоплювали наступні теми: "Мова визначення даних (DDL) та типи даних", "Основи вибірки даних (SELECT)" та "Агрегація і групування даних". Дисципліна викладається для здобувачів вищої освіти бакалаврського рівня третього року навчання і містить в собі базові аспекти мови SQL (на прикладі MySQL).

Для імітації типового робочого процесу викладача і для забезпечення універсальності експерименту, всі конспекти лекцій було підготовлено у поширеному форматі .docx. Загальний обсяг навчальних матеріалів склав 36 сторінок / 9500 слів. Лекції містять текст, таблиці та приклади коду.

Дослідження було проведено 05.11.2025. В якості інструментів генерації було обрано загальнодоступні вебзастосунки: NotebookLM від Google та ChatGPT від OpenAI. Дані вебзастосунки використовувались на умовах безкоштовного тарифного плану. На момент проведення дослідження інформація про конкретні моделі LLM, які використовуються для безкоштовних тарифних планів в цих застосунках, відсутня.

Оскільки NotebookLM не підтримує формат .docx, конспекти лекцій були конвертовані в формат .pdf за допомогою вбудованого у MS Word конвертера.

2.2 Процедура експерименту

Для відповіді на дослідницьке питання було проведено порівняльний аналіз трьох окремих сценаріїв генерації, що імітують різні підходи викладача до використання ШІ-інструментів (табл. 1). Кожен сценарій застосовувався до трьох визначених у підрозділі 2.1 конспектів лекцій.

Після застосування цих трьох сценаріїв для кожної з трьох тем лекцій було отримано загальний пул із 90 унікальних тестових питань. Отриманий "сирий" текстовий вивід було анонімізовано та підготовлено для подальшого експертного аналізу якості.

2.3 Аналіз якості та експертна оцінка

Для відповіді на дослідницьке питання весь отриманий пул із 90 унікальних питань та використані конспекти лекцій були передані на експертну оцінку.

Оцінку проводили три незалежні експерти з досвідом викладання дисциплін, пов’язаних з базами даних.

Для усунення упередженості експеримент було осліплено: питання були анонімізовані та перемішані, тож експерти не знали, яким зі сценаріїв було згенероване кожне конкретне питання.

Таблиця 1

Опис сценаріїв генерації питань

Сценарій	Інструмент	Вхідні дані (джерело)	Точне формулювання запиту (промту)
А	Вебзастосунок ChatGPT (безкоштовна версія)	Лише назва теми (конспект лекції не надавався)	Згенеруй 10 питань з чотирма варіантами відповідей для навчальної дисципліни "Організація та збереження баз даних" на тему [тема лекції]. Лише один варіант може бути правильним. Правильний варіант познач.
Б	Вебзастосунок ChatGPT (безкоштовна версія)	Конспект лекції у форматі .docx, завантажений до чату	Згенеруй 10 питань з чотирма варіантами відповідей для цієї лекції. Лише один варіант може бути правильним. Правильний варіант познач.
В	Вебзастосунок NotebookLM (безкоштовна версія)	Конспект лекції, завантажений в форматі .pdf як джерело	Згенеруй 10 питань з чотирма варіантами відповідей для цієї лекції. Лише один варіант може бути правильним. Правильний варіант познач.

Кожен експерт отримав два завдання для кожного з 90 питань:

– Використовуючи п’ятибальну рейтингову шкалу (табл. 2), якісно оцінити питання за трьома ключовими критеріями.

– Використовуючи представлену в табл. 3 шкалу, створену на основі таксономії Блума, класифікувати кожне питання відповідно до його когнітивного рівня.

– Використана шкала (табл. 3) є адаптованою (спрощеною) версією таксономії освітній цілей Блума. Враховуючи специфіку тестових завдань з множинним вибором для технічної дисципліни класифікація здійснювалась за трьома узагальненими категоріями.

До рівня "Запам'ятовування" було віднесено питання, що вимагають від здобувача вищої освіти актуалізації знань з довготривалої пам'яті. У контексті SQL це завдання на впізнавання ключових слів (наприклад, "Яка команда використовується для видалення даних?"), ідентифікацію коректного синтаксису або визначення типів даних. Відповідь на такі питання базується на механічному відтворенні вивченого матеріалу без необхідності його глибокої обробки.

Таблиця 2

Шкала оцінки якості згенерованих питань

Критерій	1 бал (дуже погано)	3 бали (задовільно)	5 балів (відмінно)
Коректність	Питання або правильна відповідь містить грубі фактологічні помилки.	Питання коректне, але сформульоване нечітко, що може заплутати.	Питання та правильна відповідь є чіткими та абсолютно правильними згідно з лекцією.
Релевантність	Питання не стосується лекції або фокусується на неважливій дрібниці.	Питання стосується теми лекції, але перевіряє другорядну інформацію.	Питання перевіряє ключове поняття або навичку з лекційного матеріалу.
Якість дистракторів	Дистрактори очевидно безглузді, не з цієї теми або граматично не узгоджуються.	Дистрактори стосуються теми, але є очевидно неправильними для студента, який читав лекцію.	Дистрактори є правдоподібними та вимагають від здобувача вищої освіти реального розуміння, щоб обрати правильну відповідь.

Таблиця 3

Шкала для класифікації за когнітивним рівнем

Рівень	Назва	Опис та типові маркери питань
1	Запам’ятовування	Перевірка знання фактів, термінів, визначень (наприклад, "Що таке …?", "Перелічіть …", "Дайте визначення …").
2	Розуміння	Перевірка здатності пояснити, описати, інтерпретувати (наприклад, "Опишіть різницю …", "Чому …", "Поясніть …").
3	Застосування та вищі рівні	Перевірка здатності застосувати знання на практиці, проаналізувати (наприклад, "Який буде результат …?", "Оберіть правильний синтаксис …", "Вирішіть …").

Рівень "Розуміння" охоплює завдання, спрямовані на перевірку здатності інтерпретувати, порівнювати та пояснювати матеріал. Сюди увійшли питання, де необхідно пояснити призначення певного опе-

ратора, розрізнити схожі поняття (наприклад, різницю між CHAR та VARCHAR) або інтерпретувати результат виконання простого фрагмента коду. Ключовим критерієм тут виступає розуміння семантики коду, а не лише його синтаксису.

Враховуючи обмеження формату "питання з вибором однієї правильної відповіді", рівні застосування, аналізу та оцінювання було об'єднано в одну категорію "Застосування та вищі рівні". До неї увійшли ситуаційні задачі, що вимагають використання процедурних знань для вирішення конкретної проблеми (наприклад, "Який запит поверне список співробітників, прізвище яких починається на 'A'?"), прогнозування результату виконання складного запиту (з декількома умовами або об'єднанням таблиць) або виявлення логічної помилки у наведеному коді.

Для зменшення навантаження на експертів, а також для зручності подальшої обробки результатів, згенеровані питання і оціночні шкали були оформлені у вигляді індивідуального для кожного експерта файлу .xlsx.

2.4 Методи аналізу даних

Статистична обробка та аналіз емпіричних даних здійснювалися з використанням MS Excel. Процес аналізу складався з трьох етапів:

– Первинна обробка та агрегація оцінок.

– Описова статистика

– Перевірка статистичної значущості гіпотез.

Оскільки кожне з 90 згенерованих питань оцінювалося трьома незалежними експертами, для подальшого аналізу використовувались агреговані показники:

– Для критеріїв якості (коректність, релевантність, якість дистракторів) розраховувалося середнє арифметичне значення оцінок трьох експертів для кожного окремого питання.

– Для класифікації за таксономією Блума визначалося медіанне значення рівня, наданого трьома експертами, що дозволило отримати єдину категорію для кожного питання.

– Для загальної характеристики результатів за кожним сценарієм було розраховано описові статистики. Враховуючи порядковий характер даних, отриманих за 5-бальною рейтинговою шкалою, окрім

середнього арифметичного (M) та стандартного відхилення (SD), розраховувалися медіана (Me) та міжквартильний інтервал (IQR). Це дозволило коректно оцінити центральну тенденцію та узгодженість результатів навіть за умов асиметричного розподілу.

Для визначення надійності виявлених відмінностей застосовувалися методи інференційної статистики з рівнем значущості $\alpha < 0,05$:

– Н-критерій Краскела-Уолліса використовувався для порівняння трьох незалежних груп (сценаріїв) за кількісними показниками якості (коректність, релевантність, якість дистракторів). Цей непараметричний тест було обрано як найбільш адекватний для порядкових даних, розподіл яких відрізняється від нормального. Нульова гіпотеза (H_0) передбачала, що медіани оцінок у трьох групах рівні.

– Критерій узгодженості χ^2 Пірсона застосовувався для аналізу категоризованих даних з метою визначення статистичної значущості відмінностей у розподілі питань за когнітивними рівнями (таксономія Блума) між різними сценаріями генерації.

3. Результати

3.1 Аналіз якості згенерованих питань

Результати експертної оцінки якості питань за трьома критеріями представлені в табл. 4 – 6. Загалом, результати демонструють високу узгодженість оцінок ($Me = 5$ та $IQR = 0$ за більшістю критеріїв), що підтверджує надійність зведених показників.

За критерієм "Коректність" (табл. 4) усі сценарії продемонстрували високий середній рівень ($M > 4,5$). Однак, сценарій "В" показав найнижчий середній бал ($M = 4,59$) та, що є ключовим, найвище стандартне відхилення ($SD = 1,01$). Це вказує на низьку внутрішню узгодженість результатів. Аналіз виводу незалежними експертами підтверджує, що такий високий розкид викликаний нестабільністю генерації, коли NotebookLM у декількох випадках пропустив важливі слова у питанні. Це сигналізує про те, що копіювання "сирого" виводу NotebookLM без додаткової верифікації є ризикованим. Сценарії "А" і "Б", в яких використовувався ChatGPT були стабільними, пропущених слів виявлено не було.

Таблиця 4

Результати для критерію "Коректність"

Сценарій	$M \pm SD$	Me	IQR
А	4,92 ± 0,3	5	0
Б	4,94 ± 0,3	5	0
В	4,59 ± 1,01	5	0

Найкращій показник за критерієм "Релевантність" (табл. 5) продемонстрував сценарій "В" із середнім балом $M = 4,99$. Такий майже ідеальний результат є очікуваним, оскільки NotebookLM розроблений для тісної прив'язки до джерел, що практично усуває ризик генерації нерелевантної інформації. Сценарії "А" та "Б" також показали високі результати, що свідчить про те, що модель, яка використовується в безкоштовному тарифному плані ChatGPT на момент проведення експерименту, добре визначає ключові концепції, навіть якщо їй надано лише назву теми. Але слід зауважити, що з 30 питань, які були створені в межах реалізації сценарію "А" лише одне питання було таким, яке не відповідає матеріалу лекції. Як було зазначено вище, лекції містять загальновідомий базовий матеріал і призначені для здобувачів вищої освіти, які тільки починають своє знайомство з базами даних. Ймовірно, саме це і було причиною високих показників релевантності згенерованих "наосліп" питань.

Таблиця 5

Результати для критерію "Релевантність"

Сценарій	$M \pm SD$	Me	IQR
А	4,8 ± 0,6	5	0
Б	4,92 ± 0,26	5	0
В	4,99 ± 0,06	5	0

Створення якісних дистракторів (табл. 6) виявилось найскладнішим завданням для всіх ШІ-інструментів, про що свідчать найнижчі середні бали у порівнянні з іншими критеріями. Значення $IQR = 0,67$ демонструє значно більшу варіативність в оцінках експертів. Незважаючи на це, NotebookLM (сценарій "В") продемонстрував дещо кращій, хоча і не статистично значущий результат.

Таблиця 6

Результати для критерію "Якість дистракторів"

Сценарій	$M \pm SD$	Me	IQR
А	4,16 ± 0,62	4,33	0,67
Б	4,31 ± 0,47	4,33	0,67
В	4,32 ± 0,41	4,33	0,67

3.2 Статистична перевірка значущості відмінностей

Для визначення того, чи є відмінності у якості питань між трьома сценаріями статистично значущими, було проведено Н-тест Краскела-Уолліса для кожного з трьох критеріїв. Результати аналізу наведено в табл. 7.

Таблиця 7

Результати порівняння трьох сценаріїв генерації

Критерій якості	Н-статистика	Ступені вільності (df)	p-значення	Статистичний висновок
Коректність	1,44	2	0,487	Відмінності не є значущими ($p > 0,05$)
Релевантність	1,30	2	0,522	Відмінності не є значущими ($p > 0,05$)
Якість дистракторів	1,10	2	0,577	Відмінності не є значущими ($p > 0,05$)

Як видно з табл. 7 Н-тест Краскела-Уолліса показав, що для всіх трьох критеріїв $p > 0,05$. Це означає, що нульова гіпотеза про рівність груп не може бути відхилена, і статистично значущих відмінностей у якості між сценаріями не виявлено.

3.3 Аналіз когнітивного рівня питань

Для визначення того, чи впливає обраний інструмент на когнітивну складність згенерованих питань, було проаналізовано їх розподіл за спрощеною таксономією Блума (табл. 8).

Таблиця 8

Розподіл питань за когнітивними рівнями

Сценарій	Рівень 1	Рівень 2	Рівень 3+	Разом
А	22 (73,3%)	6	2	30
Б	22 (73,3%)	7	1	30
В	16 (53,3%)	11	3	30

Статистична перевірка за допомогою критерію χ^2 Пірсона показала результат $\chi^2(4) = 3.95, p = 0.413$. Оскільки $p > 0.05$, відмінності у розподілах не є статистично значущими.

3.4 Якісний аналіз типових кейсів генерації

Статистичний аналіз надав узагальнену картину якості, проте для глибшого розуміння природи виявлених тенденцій, зокрема, технічної нестабільності NotebookLM та "ефекту стелі" у ChatGPT, доцільно розглянути конкретні приклади згенерованих тестових завдань. Нижче наведено детальний розбір випадків, які ілюструють ключові знахідки дослідження.

Технічна нестабільність RAG-системи була виявлена під час реалізації сценарію "В", коли було виявлено наступне згенероване питання: "Яка клауза використовується для записів, які будуть включені в результуючий набір?". Коментар експерта: "У питанні пропущено слово "фільтрації" або "відбору". Без цього дієслова питання втрачає сенс, оскільки будь-яка клауза так чи інакше впливає на записи. Оцінка за коректність знижена до 2 балів".

У той же час, сценарії "А" та "Б" (ChatGPT), базуючись на тій самій темі, сформулювали аналогічне питання коректно: "Який оператор використовується для фільтрації записів у команді SELECT?". Цей випадок підтверджує висновок про те, що RAG-архітектура NotebookLM (у поточній версії) має проблеми з лінгвістичною стабільністю при генерації українською мовою, що вимагає від викладача обов'язкової перевірки кожного питання перед їх використанням.

Високі оцінки за критерієм "Релевантність" у сценарії "А" пояснюються тим, що базові концепції SQL є фундаментальною частиною навчальних даних будь-якої LLM.

Прикладом успішної генерації "наосліп" є питання: "Яка команда використовується для видалення таблиці з бази даних?" Варіанти:

A) DELETE TABLE; B) DROP TABLE (Правильна); C) REMOVE TABLE; D) ERASE TABLE. Коментар експерта: "Питання абсолютно коректне, дистрактори підібрані вдало, оскільки вони є синонімами слова 'видалити', але не є валідними SQL-командами. Оцінка 5/5".

Цей приклад демонструє, що для перевірки знань термінології та синтаксису (рівень "Запам'ятовування") моделі не потребують додаткового контексту у вигляді лекцій. Надання конспекту в сценаріях "Б" та "В" не змінило формулювання цього питання, що й призвело до статистичної відсутності різниці між групами.

Найнижчі бали у всіх групах отримав критерій "Якість дистракторів". Аналіз показує, що ШІ схильний генерувати варіанти відповідей, які є занадто очевидно неправильними для студента, що володіє мінімальними знаннями, або ж використовує неіснуючі терміни, які легко відсіюються методом виключення.

Прикладом використання "слабких" дистракторів є наступне згенероване питання: "Який тип даних використовується для зберігання рядків змінної довжини?" Варіанти: A) VARCHAR (Правильна); B) VAR; C) CHARVAR; D) VARIABL. Коментар експерта: "Варіанти B, C і D є вигаданими словами, які не схожі на реальні типи даних SQL. Студент може вгадати правильну відповідь, просто згадавши, що бачив слово CHAR, навіть не розуміючи суті. Краще було б використати реальні, але невідповідні типи, наприклад TEXT або CHAR, щоб перевірити розуміння різниці".

Цей приклад ілюструє спільну проблему для всіх досліджених інструментів: вони добре володіють фактажем (правильна відповідь), але погано моделюють типові помилки студентів (дистрактори). RAG-технологія не вирішила цю проблему, оскільки лекційний матеріал зазвичай містить правильні твердження, а не перелік можливих помилок, тому модель змушена вигадувати дистрактори так само як і звичайна LLM.

Переважна більшість згенерованих питань (понад 50-70%) у всіх сценаріях належала до рівня "Запам'ятовування". Типовим прикладом занадто простого питання є наступне: "Що робить оператор DISTINCT?" (Перевірка визначення).

Для переходу на рівень "Застосування" або "Аналіз" питання мало б виглядати інакше, наприклад: "Дано таблицю Orders з 5 запи-

сами (дублікати присутні). Скільки рядків поверне запит SELECT DISTINCT...?". Жоден із протестованих інструментів за умов використання простого промту ("Згенеруй 10 питань...") не створив задач такого типу. Вони обмежувалися питаннями на знання синтаксису. Це підтверджує висновок про те, що для підвищення когнітивного рівня тестів важливою є зміна структури промту (наприклад, додавання вимоги "створити ситуаційні задачі" або "надати фрагмент коду для аналізу"), а не просто зміна інструменту генерації.

4. Обговорення

Результати статистичного аналізу вказують на відсутність системної переваги будь-якого з досліджуваних методів у контексті загальної якості питань для дисципліни "Організація та збереження баз даних". Це дозволяє зробити кілька важливих спостережень.

Для критеріїв "Коректність" та "Релевантність" відсутність значущих відмінностей пояснюється високим базовим рівнем компетентності LLM, яка використовується у загальнодоступній безкоштовній версії вебзастосунку ChatGPT на момент проведення експерименту, у стандартних академічних темах. Модель здатна генерувати контент задовільної якості навіть без надання опорного конспекту. Це свідчить про те, що для загальновідомих дисциплін етап підготовки та завантаження файлів не є важливим для забезпечення коректності і релевантності згенерованих питань.

Статистична схожість результатів за критерієм "Якість дистракторів" демонструє, що надання контексту (конспекту лекції) не покращило здатність ШІ генерувати правдоподібні хибні відповіді. Це завдання залишається слабким місцем усіх сценаріїв, оскільки вимагає креативності у створенні помилкових тверджень, яку RAG-технології (NotebookLM) не підсилюють.

Хоча тест не виявив різниці у середніх рангах, важливо розглядати цей результат у комплексі з описовою статистикою, де сценарій “B” продемонстрував вищу варіативність для критерія “Коректність”.

Виявлена вища варіативність в сценарії “B” для критерія “Коректність” вказує на важливий феномен: хоча NotebookLM здебільшого генерує точні питання, він схильний до спорадичних грубих помилок, наприклад, пропуск важливих слів у питанні, чого не було в сцена-

ріях “А” і “Б” (ChatGPT). З цього можна зробити висновок, що використання NotebookLM на момент проведення експерименту вимагає більш ретельної верифікації кожного окремого питання викладачем, ніж використання вебзастосунку ChatGPT, незважаючи на кращу контекстну обізнаність першого.

Основна маса питань, генерованих сучасними загальнодоступними LLM у "ледачому" режимі, залишається на рівні перевірки базових знань. Для отримання питань високого рівня, недостатньо просто змінити інструмент – необхідна зміна стратегії формування запитів, що потребує подальшого дослідження.

На основі проведеного аналізу можна сформулювати ряд практичних рекомендацій для викладачів закладів вищої освіти, які планують інтегрувати ШІ-інструменти у процес створення контрольно-вимірювальних матеріалів.

Стратегії вибору інструменту:

1. Для поточного контролю (експрес-тести) рекомендується використання універсальних чат-ботів (ChatGPT, Claude, Gemini) без завантаження файлів. Як показало дослідження, для базових тем вони забезпечують високу релевантність при мінімальних витратах часу на підготовку.

2. Для підсумкового контролю та авторських курсів використання RAG-систем (NotebookLM) є виправданим лише у випадках, коли тестові питання мають базуватися на специфічній термінології, унікальних класифікаціях або вузькоспеціалізованих даних, які відсутні у загальному доступі. При цьому викладач повинен закладати додатковий час на ретельну лінгвістичну верифікацію кожного питання через ризик технічних збоїв генерації.

Стратегії підвищення якості генерації:

1. Оскільки ШІ демонструє слабкість у створенні хибних відповідей, викладачам рекомендується використовувати "гібридний" підхід: генерувати основу питання та правильну відповідь автоматично, а дистрактори редагувати або замінювати вручну, спираючись на власний досвід типових помилок студентів.

2. Для отримання питань рівня "Застосування" та вище необхідно відмовитися від простих промтів. Рекомендується використовувати техніку "Few-Shot Prompting", надаючи моделі 1-2 приклади складних ситуаційних задач як зразок для наслідування.

Варто зазначити низку обмежень даної роботи, які відкривають перспективи для подальших розвідок.

По-перше, дослідження проводилося на матеріалі технічної дисципліни з чітко формалізованою структурою знань (SQL). Результати можуть відрізнятися для гуманітарних дисциплін (філософія, історія), де інтерпретація джерел має більше значення, і де RAG-системи можуть проявити себе краще.

По-друге, експеримент обмежувався використанням безкоштовних версій моделей. Платні версії, особливо ті, які здатні більш глибоко міркувати під час генерації відповіді, можуть демонструвати вищу якість "креативності" при створенні дистракторів.

По-третє, обсяг вибірки (90 питань) та кількість експертів (3 особи), хоча і є достатніми для пілотного дослідження, можуть бути розширені у майбутніх роботах для підвищення статистичної потужності висновків.

5. Висновки

У дослідженні було проведено порівняльний аналіз якості та когнітивної складності тестових питань, згенерованих за допомогою трьох різних сценаріїв використання загальнодоступних ШІ-інструментів (ChatGPT та NotebookLM) викладачем без глибоких навичок промт-інжинірінгу.

На основі отриманих результатів можна зробити наступні висновки:

1. Статистичний аналіз не виявив значущих відмінностей у якості питань (за критеріями коректності, релевантності та якості дистракторів) між генерацією на основі лише назви теми (ChatGPT) та генерацією з використанням завантажених конспектів лекцій (ChatGPT та NotebookLM). Це свідчить про те, що для стандартизованих навчальних дисциплін, таких як “Організація і збереження баз даних”, сучасні мовні моделі володіють достатнім обсягом знань, що робить етап підготовки та завантаження файлів необов’язковим для отримання задовільного результату.

2. Попри високу релевантність, інструмент NotebookLM продемонстрував технічну нестабільність при генерації українською мовою, що виражалось у пропуску слів у декількох питаннях. Це призвело до вищої варіативності оцінок за критерієм “Коректність” і свідчить про

необхідність ретельної ручної верифікації результатів його роботи, що може нівелювати переваги від автоматизації.

3. Використання "простих" розмовних промтів, незалежно від обраного інструменту, призводить до генерації питань переважно низького когнітивного рівня (запам'ятовування та розуміння). Сама лише зміна інструменту (наприклад, перехід на RAG-систему NotebookLM) автоматично не підвищує когнітивну складність завдань.

Для швидкого створення банку тестових питань із загальновідомих дисциплін найбільш ефективним (з точки зору співвідношення зусиль та якості) є використання вебзастосунку ChatGPT із простими запитами на основі теми. Використання спеціалізованих інструментів на кшталт NotebookLM є доцільним лише за умови роботи з унікальними авторськими матеріалами, які відсутні у навчальних даних моделі, проте вимагає підвищеної уваги до контролю якості виводу.

Перспективи подальших досліджень полягають у вивченні впливу методів складного промтингу (наприклад, "Chain-of-Thought" або надання прикладів "few-shot") на підвищення когнітивного рівня питань та якість дистракторів.

References:

1. Artsi, Y., Sorin, V., Konen, E., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2024). Large language models for generating medical examinations: Systematic review. *BMC Medical Education*, 24(1). doi: 10.1186/s12909-024-05239-y

2. Choi, W. (2023). Assessment of the capacity of ChatGPT as a self-learning tool in medical pharmacology: A study using MCQs. *BMC Medical Education*, 23(1). doi: 10.1186/s12909-023-04832-x

3. Dutulescu, A., Ruseti, S., Iorga, D., Dascalu, M., & McNamara, D. S. (2024). Beyond the Obvious Multi-choice Options: Introducing a Toolkit for Distractor Generation Enhanced with NLI Filtering. *Lecture Notes in Computer Science*, 14830 LNAI, 242–250. doi: 10.1007/978-3-031-64299-9_18

4. Eskenazi, J., Krishnan, V., Konarzewski, M., Constantinescu, D., Lobaton, G., & Dodds, S. D. (2025). Evaluating retrieval augmented generation and ChatGPT's accuracy on orthopaedic examination assessment questions. *Annals of Joint*, 10. doi: 10.21037/aoj-24-49

5. Feng, W., Lee, J., McNichols, H., Scarlatos, A., Smith, D., Woodhead, S., … Lan, A. (2024). Exploring Automated Distractor Generation for Math Multiple-choice Questions via Large Language Models. 3067–3082. doi: 10.18653/v1/2024.findings-naacl.193

6. Grévisse, C., Pavlou, M. A. S., & Schneider, J. G. (2024). Docimological Quality Analysis of LLM-Generated Multiple Choice Questions in Computer Science and Medicine. *SN Computer Science*, 5(5). doi: 10.1007/s42979-024-02963-6

7. Hadzhikoleva, S., Rachovski, T., Ivanov, I., Hadzhikolev, E., & Dimitrov, G. (2024). Automated Test Creation Using Large Language Models: A Practical Application. *Applied Sciences (Switzerland),* 14(19). doi: 10.3390/app14199125

8. Law, A. K. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-Ching Hung, K., & Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: A cohort study in a high-stakes examination. *BMC Medical Education*, 25(1). doi: 10.1186/s12909-025-06796-6

9. Liu, N., Sonkar, S., & Baraniuk, R. (2025). Do LLMs Make Mistakes Like Students? Exploring Natural Alignments Between Language Models and Human Error Patterns. *Lecture Notes in Computer Science*, 15880 LNAI, 364–377. doi: 10.1007/978-3-031-98459-4_26

10. Malathi, S., Hemamalini, S., Muralidharan, M., & Benny, R. (2024). Knowledge Navigator: Revolutionizing Education through LLMs in Generative AI. *Fusion: Practice and Applications*, 16(1), 209–222. doi: 10.54216/FPA.160114

11. Randolph, C., Michaleas, A., & Ricke, D. O. (2025). Large language models for closed-library multi-document query, test generation, and evaluation. *Frontiers in Artificial Intelligence*, 8. doi: 10.3389/frai.2025.1592013

12. Rivera-Rosas, C. N., Calleja-López, J. R. T., Ruibal-Tavares, E., Villanueva-Neri, A., Flores-Felix, C. M., & Trujillo-López, S. (2024). Exploring the potential of ChatGPT to create multiple-choice question exams. *Educacion Medica*, 25(4). doi: 10.1016/j.edumed.2024.100930

13. Wang, L., Song, R., Guo, W., & Yang, H. (2025). Exploring prompt pattern for generative artificial intelligence in automatic question generation. *Interactive Learning Environments*, 33(3), 2559–2584. doi: 10.1080/10494820.2024.2412082