Кафедра *інформаційно-вимірювальних технологій*.

ПОЯСНЮВАЛЬНА ЗАПИСКА

до бакалаврської кваліфікаційної роботи на тему

Оцінка якості генеративних систем штучного інтелекту

Студент *152, МТ-41 Шовак Р. М.*
(група, шифр, прізвище та ініціали)

Керівник проекту _____ */ д.т.н., доцент Хома Ю.В. /*

Консультанти _____ */ к.е.н., доцент Рачинська Г.В. /*

_____ */ д.т.н., проф. Кочан О.В. /*

Завідувач кафедри _____ */ д.т.н., проф. Бубела Т.З. /*

«___» _____ 2025 р

Національний університет «Львівська політехніка»

(назва вищого навчального закладу)

Інститут *ІКТА* Кафедра *ІВТ*

Спеціальність 152 *Метрологія та інформаційно-вимірювальна техніка*

“ЗАТВЕРДЖУЮ”

Завідувач кафедри ІВТ

Т.З. Бубела

«__» 2025 р.

ЗАВДАННЯ

на бакалаврську кваліфікаційну роботу студентові

Шовак Ростислав Михайлович

(прізвище, ім’я по батькові)

1. Тема БКР Оцінка якості генеративних систем штучного інтелекту.

затверджена наказом по університету від *8 квітня 2025 р. № 1282-4-08*

2. Термін подання студентом закінченого проекту *16 червня 2025 р.*

3. Вихідні дані до проекту: *Здійснити порівняльний аналіз генеративних моделей за допомогою розробленої системи автоматичної оцінки згенерованого тексту та провести якісне оцінювання результатів відповідно до розроблених критеріїв.*

4. Зміст розрахунково – пояснювальної записки (перелік питань, що їх треба розробити):

1. Дослідження архітектур генеративних моделей

2. Аналіз особливостей та недоліків автоматичних методів оцінки Великих Мовних моделей

3. Формування критеріїв та розробка системи оцінювання якості

4. Експериментальні дослідження та оцінка розробленої системи

5. Економічне обгрунтування.

6. Перелік графічного матеріалу

Презентація виконана у редакторі Google Presentation

7. Консультанти, із зазначенням розділів БКР, що стосуються їх

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв
Економічна частина	*к.е.н. Рачинська Галина Василівна*		

8. Дата отримання завдання 14.03.2025

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

Пор. №	Назва етапів бакалаврської кваліфікаційної роботи	Термін виконання етапів БКР	Примітки
1	*Отримання завдання*	*14.03.25 р.*	
2	*Здійснити огляд і аналіз науково-технічної літератури*	*15.03.25 р.-22.03.25 р.*	
3	*Робота над першим розділом*	*23.03.25 р.-14.04.25 р.*	
4	*Формування та обробка набору даних*	*15.04.25 р.-24.05.25 р.*	
5	*Розробка системи оцінювання ВВМ*	*25.04.25 р.-05.05.25 р.*	
6	*Розрахунок розділу економічного обґрунтування*	*06.05.25 р.-10.05.25 р.*	
7	*Оформлення пояснювальної записки та графічної частини. Опрацювання висновків та рекомендацій*	*11.05.25 р.-30.05.25 р.*	

Студент – дипломник _____
(підпис)

Керівник проекту _____
(підпис)

ABSTRACT

Due to the rapid development of generative artificial intelligence systems, particularly large language models (LLMs), there is a growing need for objective and comprehensive evaluation of their performance. In this thesis we explore current approaches to assessing generative text, with a focus on accuracy, relevance, logical coherence, and factual consistency. A prototype of an automated evaluation system was developed based on the RAGAS framework, utilizing Retrieval-Augmented Generation. Experimental studies were conducted to evaluate the quality of two of the most well-known models, using a custom-built dataset. Each model was analyzed using key quality metrics, including human evaluation. The results confirm the effectiveness of the chosen approach and highlight key differences between the models. Possible reasons for such outcomes are discussed, and suggestions for future work are provided.

АНОТАЦІЯ

У зв’язку із стрімким розвитком генеративних систем штучного інтелекту, особливо великих мовних моделей (ВММ), зростає потреба в об’єктивній та комплексній оцінці якості їхньої роботи. У цій дипломній роботі досліджено сучасні підходи до оцінювання генеративного тексту, зокрема з точки зору точності, релевантності, логічної послідовності та фактичної достовірності. Було розроблено прототип автоматизованої системи оцінки якості генерації на основі фреймворку RAGAS із використанням Retrieval-Augmented Generation. Проведено експериментальні дослідження з оцінювання якості двох найвідоміших моделей, побудовані на основі власного датасету. Для кожної моделі було проведено аналіз за основними метриками якості, включаючи людське оцінювання. Результати підтверджують ефективність використаного підходу та демонструють ключові відмінності між моделями. Можливі причини таких результатів розглянуто. Та надано пропозиції щодо подальшої роботи.

ЗМІСТ

ВСТУП

Сучасна епоха характеризується стрімким розвитком технологій штучного інтелекту, зокрема генеративних систем, серед яких особливе місце посідають великі мовні моделі (ВММ). Ці системи, побудовані на основі архітектури Transformer, демонструють надзвичайні можливості у сферах обробки природної мови, генерації тексту, аналізу документів та взаємодії з користувачами. Моделі GPT-4 від OpenAI, Gemini від Google DeepMind та інші подібні системи вже знайшли широке застосування у медицині, освіті, фінансах, чат-ботах, дослідницькій діяльності та багатьох інших галузях.

Однак швидке впровадження генеративних систем штучного інтелекту у критично важливі сфери людської діяльності актуалізує питання об'єктивної та комплексної оцінки їхньої продуктивності. Традиційні підходи до оцінювання, що базуються на простих метриках подібності тексту або статистичних показниках, виявляються недостатніми для всебічного аналізу якості сучасних генеративних моделей. Особливо постає проблема оцінки систем, що використовують підхід Retrieval-Augmented Generation (RAG), де якість відповіді залежить не лише від генеративних здібностей моделі, але й від ефективності пошуку та використання зовнішньої інформації.

Ключовими викликами в оцінюванні генеративних систем є забезпечення фактичної точності відповідей та запобігання галюцинаціям. Існуючі автоматизовані метрики, такі як BLEU, ROUGE чи BERTScore, часто не здатні адекватно відобразити характеристики тексту, що сприймаються людиною.

В даній роботі буде досліджено можливості застосування методології комплексної оцінки якості генеративних систем штучного інтелекту на прикладі великих мовних моделей. Описано основні критерії оцінки, проведення порівняльного аналізу якості відповідей моделей GPT-4 та Gemini в умовах RAG-архітектури, їхні переваги та недоліки.

РОЗДІЛ 1. ОГЛЯД СУЧАСНИХ ГЕНЕРАТИВНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

1.1 Аналіз архітектури та принципів роботи Великих Мовних Моделей (ВММ)

Великі мовні моделі (ВММ) побудовані на основі передових архітектур нейронних мереж, переважно на архітектурі Transformer, яку представили Vaswani та інші у 2017 році. [1]

Модель Transformer обробляє текст, спочатку перетворюючи слова у числові токени, а потім відображаючи ці токени у неперервні векторні представлення через шар ембедінгу.

Основною інновацією є механізм self-attention, який дозволяє моделі контекстуалізувати кожен токен відносно всіх інших у послідовності. На кожному шарі Transformer токени взаємодіють через multi-head self-attention, яка посилює значення важливих токенів та зменшує значущість менш релевантних, на основі вивчених взаємозв’язків. Завдяки використанню механізму multi-head attention функція уваги отримує інформацію з різних частин представлення, що неможливо при використанні self-attention.

На відміну від попередніх рекурентних архітектур (RNNs), які обробляли слова послідовно, Transformers працюють паралельно з послідовностями (без рекурентності), що дає значно швидше навчання і кращу обробку довготривалих залежностей. Ця паралельність та здатність захоплювати довгий контекст зробили трансформери основою для сучасних ВВМ, які навчаються на великих текстових наборах даних.

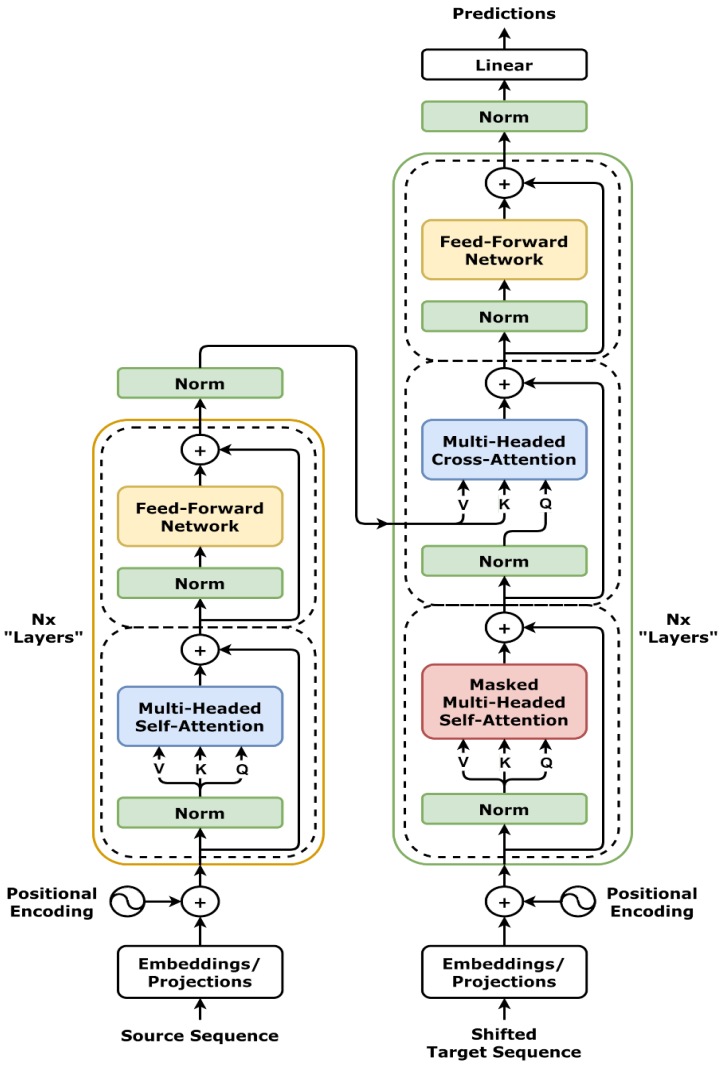

Рисунок 1.1 – Ілюстрація Transformer та механізму уваги.

Ключові компоненти архітектури ВВМ на основі Transformers включають:

- **Токенізація** - процес поділ вхідного тексту на послідовність токенів (слів, підслів або окремих символів). Це визначає словник моделі й деталізацію, з якою текст обробляється.
- **Шар ембедінгів** - перетворення токенів багатовимірні векторні представлення. Кожному токену відповідає навчене векторне представлення, що містить семантичні властивості слова. Для врахування порядку слів додаються позиційні ембедінги, оскільки механізм self-attention сам по собі не фіксує порядок.
- **Multi-head self-attention** – дозволяє моделі аналізувати взаємний вплив токенів один на одного. Вектори ембедінгів перетворюються у вектори

запиту (Query), ключа (Key) та значення (Value). Пізніше модель обчислює скалярні добутки між Query та Key для отримання коефіцієнтів уваги.

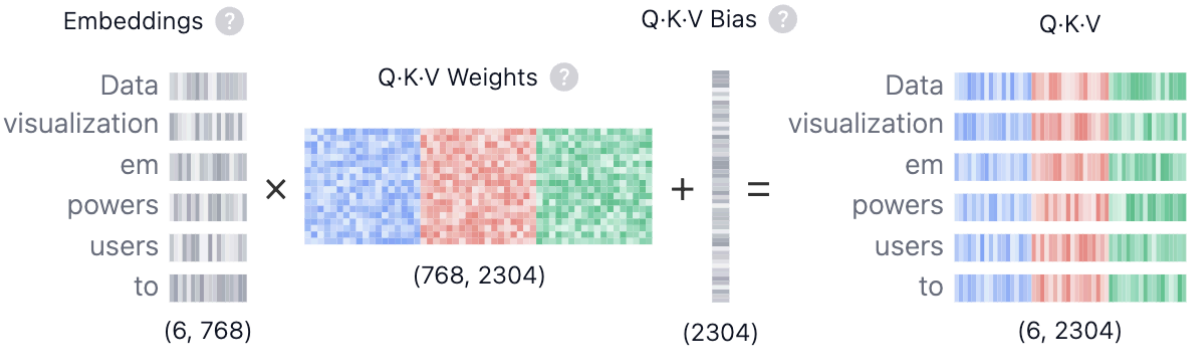

Рисунок 1.2 – *Обчислення матриць Query, Key та Value з оригінальних ембедінгів.*

Цей процес відбувається паралельно в кількох «головах», кожна з яких аналізує різні аспекти інформації, дозволяючи захоплювати складні взаємозв'язки та довготривалі залежності.

- **Feed-Forward Network (мережа прямого поширення)** – двошарова повнозв'язна нейронна мережа з нелінійною активацією (наприклад, ReLU або GELU), яка застосовується до векторів після етапу уваги. Вона спочатку розширює, а потім зменшує розмірність векторів, удосконалюючи індивідуальні властивості токенів.
- **Residual Connections та нормалізація** – кожен підшар уваги та FFN обгорнутий резидуальними пропускними зв'язками та нормалізацією шару, які допомагають стабілізувати навчання глибоких моделей та зберегти потік інформації. Це означає, що вхідні дані підшару додаються до його виходу перед нормалізацією, що запобігає надмірному відхиленню моделі від початкових представлень.
- **Вихідний шар** – виконує лінійну трансформацію для перетворення фінальних векторів моделі у логіти для кожного токена словника. За допомогою функції Softmax ці логіти конвертуються в розподіл

ймовірностей, що використовується для генерації тексту шляхом вибору наступного токена.

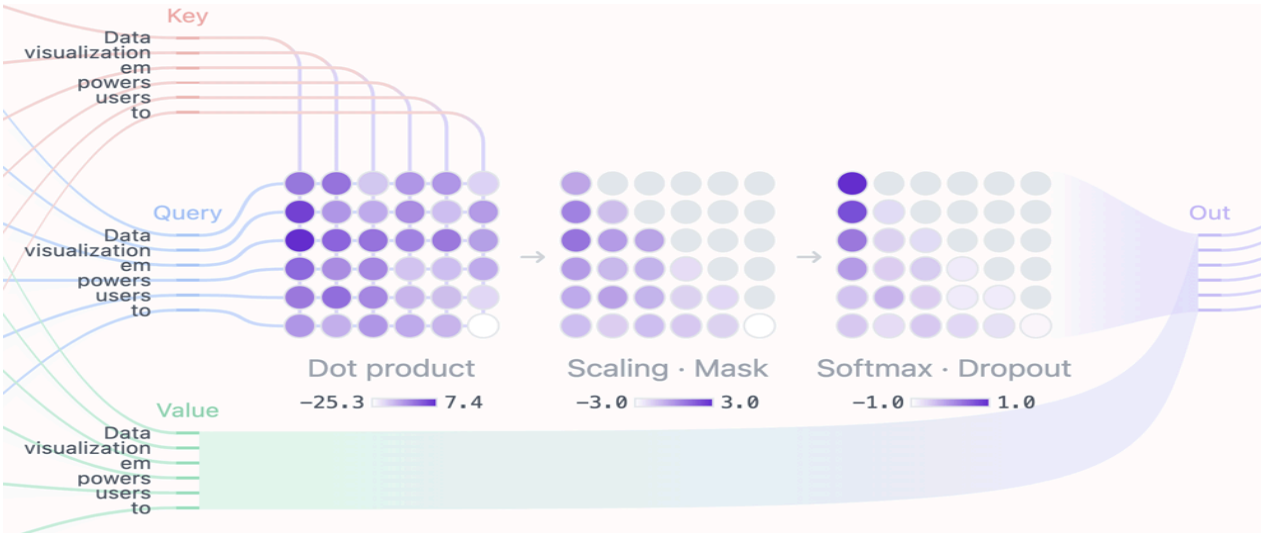

Рисунок 1.3 – *Ілюстрація self-attention в моделі Transformer (одна голова для наочності)*.

Кожен вхідний токен (наприклад, слова у фразі "Data visualization empowers users to") проектується у три вектори: запиту (синій), ключа (червоний) та значення (зелений).

Модель обчислює скалярні добутки між усіма парами векторів Query–Key, створюючи матрицю уваги (показану по центру), яка відображає, наскільки кожен токен (слово) повинен звертати увагу на кожен інший токен. Під час генерації тексту до цієї матриці застосовується маска, яка забороняє моделі враховувати майбутні токени. Потім оцінена матриця нормалізується через Softmax, формуючи ваги уваги (фіолетовий відтінок), які визначають, наскільки сильно токени впливають один на одного.

Ці ваги використовуються для обчислення зваженої суми векторів значень, що дає нове представлення (Out) для кожного токена, яке включає контекстну інформацію від токенів, на які звертається увага. Transformer використовує multi-head attention, виконуючи кілька паралельних обчислень уваги на різних навчальних проекціях (зазвичай 8 або 12 голів); результати

всіх голів конкатенуються і піддаються лінійній трансформації, для охоплення різних зв’язків у тексті. Після цього мережа прямого проходу (MLP) додатково трансформує вектор кожного токена. Цей процес повторюється протягом кількох шарів. [2]

Великі мовні моделі навчаються у два основних етапи: попереднє навчання на великих текстових корпусах та донавчання на конкретних задачах або із залученням зворотного зв’язку від людини. [3]

Під час попереднього навчання модель вивчає загальні мовні закономірності виконуючи самоконтрольовані завдання. Наприклад, моделі типу GPT навчаються передбачати наступний токен, враховуючи попередні токени, тоді як інші, як BERT, використовують завдання з маскуванням слів.

Останнім нововведенням у навчанні ВВМ є навчання з Reinforcement Learning from Human Feedback. Анотатори оцінюють відповіді моделі, що дозволяє навчити модель винагороди, яка визначає бажані результати. Потім модель оптимізується через алгоритми навчання з підкріпленням (часто Proximal Policy Optimization, PPO), що значно підвищує якість відповідей та знижує ризик неприйнятного контенту. [4]

1.2 Сфери застосування та типові проблеми систем на основі Великих Мовних Моделей

Великі мовні моделі стали універсальним інструментом для автоматизації обробки мови у багатьох сферах. Вони здатні генерувати, аналізувати, перекладати та узагальнювати тексти, що дозволяє скоротити потребу в ручній праці.

Попри значний прогрес, генеративні моделі мають низку характерних проблем, серед яких: галюцинації, упередженість, етичні дилеми, високі обчислювальні витрати та питання безпеки.

ВВМ можуть виконувати численні завдання в межах однієї моделі – від програмування до написання текстів і відповідей на запити. Це робить їх привабливими для впровадження у медицину, де вони допомагають підсумовувати лікарські записи, аналізувати медичну документацію чи взаємодіяти з пацієнтами через чат боти. У фінансовій сфері, вони використовуються для обробки великого обсягу текстової інформації чи взаємодії з клієнтами. [5]

1.2.1 Сфери застосування Великих Мовних Моделей

Створення контенту та копірайтинг

Такі моделі ефективно генерують тексти різного типу, беручи до уваги потреби бізнесу. Наприклад, статті, рекламні оголошення, публікації у соціальних мережах чи художні твори. Інструменти як ChatGPT, Gemini або Llama 2 дозволяють бізнесам масштабувати контентні стратегії, підвищувати персоналізацію та оптимізувати витрати часу й ресурсів.

Дослідження та аналіз даних

ВММ допомагають швидко обробляти великі обсяги інформації, виділяючи ключові висновки і знаходити потрібні дані навіть у складних документах, що підвищує ефективність дослідницької діяльності.

Вони можуть аналізувати звіти, наукові статті та інші джерела, деколи перевершуючи за точністю та швидкістю людину. Це сприяє виявленню трендів, закономірностей та створенню докладних звітів у сферах фінансів, медицини та науки.

Таким чином, генеративні системи автоматизують рутинні процеси, що

дозволяє фахівцям зосередитись на стратегії та аналітиці, що в свою чергу сприяє ефективному ухваленню нових рішень.

Освіта та персоналізоване навчання

Системи генеративного штучного інтелекту можуть адаптувати навчальні процеси до потреб окремого студента – формувати плани занять, надавати персональні рекомендації та відстежувати прогрес у реальному часі.

Також моделі автоматизують рутинні завдання, такі як перевірка робіт чи створення інтерактивних навчальних матеріалів, що робить процес навчання цікавішим та ефективнішим

Охорона здоров'я та медичні застосування

У сфері медицини такі моделі сприяють додатковій перевірці діагнозів, аналізуючи великі обсяги медичних даних – електронні медичні записи пацієнта чи результати обстежень. Вони допомагають підбирати варіанти лікування з урахуванням індивідуальних особливостей пацієнтів і сучасних рекомендацій.

Також сучасні генеративні системи прискорюють розробку нових лікарських засобів, прогнозуючи їхню ефективність і безпечність для людини.

Інші помітні застосування

- Удосконалення пошукових систем – ВММ покращують якість пошукових результатів, краще розуміючи запити користувачів.
- Віртуальні асистенти – в основі роботи Alexa, Google Assistant, Siri та інших помічників лежать великі мовні моделі, що забезпечують розуміння мовлення та виконання завдань.
- Класифікація текстових даних – моделі групують тексти за змістовою чи емоційною подібністю, що є корисним для обробки відгуків або

тематичного пошуку інформації.

- Рекомендаційні системи – моделі аналізують вподобання користувачів і формують персоналізовані пропозиції товарів, послуг або контенту.

1.2.2 Типові проблеми систем на основі Великих Мовних Моделей

Попри значні переваги, системи на основі великих мовних моделей стикаються з низкою притаманних проблем, що створюють ризики для достовірності результатів, безпеки та етичності застосування генеративних систем, які необхідно враховувати під час їхнього проектування та впровадження.

Упередження у навчальних даних та етичні проблеми

Великі мовні моделі схильні успадковувати та відтворювати упередження, закладені в їхніх навчальних даних, що містять як корисну, так і шкідливу інформацію. Як результат, це може спричинити викривлені уявлення або несправедливе ставлення до представників різних соціальних груп за ознаками статі, раси, мови чи культури.

Наприклад, такі моделі можуть несвідомо закріплювати гендерні стереотипи щодо певних професій або демонструвати расову упередженість. Подібні упередження здатні призводити до дискримінаційних наслідків, зокрема у сферах працевлаштування чи кредитування.

Також без належної фільтрації ВММ здатні генерувати образливі чи неприйнятні висловлювання, особливо при провокаційних запитах. Хоча сучасні моделі проходять додаткове навчання із використанням RLHF та фільтрацію токсичних патернів, повністю позбутися упереджень поки не вдається.

Подолання цієї проблеми вимагає детального аналізу навчальних

корпусів та постійного моніторингу для забезпечення справедливості в роботі моделей.

Галюцинації та фактичні неточності

Модель може генерувати переконливі, але хибні твердження – наприклад, вигадувати джерела або помилятися у фактах. Оскільки моделі прогнозують наступні слова на основі ймовірностей, без розуміння фактичної істини, вони можуть генерувати граматично правильні, але змістовно помилкові відповіді. Це пов’язано з тим, що вона не має прямого зв’язку із зовнішніми джерелами – лише закономірності з навчальних даних.

Стратегії зменшення цієї проблеми вимагають підключення зовнішніх баз знань (retrieval augmentation) та автоматичну перевірку фактів.

Обчислювальні та ресурсні обмеження

Тренування таких моделей вимагає величезних обчислювальних ресурсів, великих наборів даних та часу. Наприклад, навчання GPT-3 (175 мільярдів параметрів) може коштувати мільйони доларів тільки за хмарні обчислення.

Етичні аспекти застосування ВММ

Окрім упередженості й ризику дезінформації, генеративні системи породжують низку інших етичних викликів, серед яких – захист конфіденційності користувачів, безпечне використання даних, запобігання створенню шкідливого контенту та відповідальність за результати роботи моделей. Відсутність чітких механізмів визначення відповідальності ускладнює регулювання використання таких моделей та підвищує ризик зловживань. Це зумовлює необхідність розробки етичних стандартів та нормативних актів, що забезпечуватимуть безпечне та відповідальне впровадження таких технологій.

Безпека та надійність систем на основі ВММ

Із поширенням генеративних систем зростає кількість потенційних кіберзагроз, серед яких особливе занепокоєння викликають:

- Prompt injection – маніпулювання поведінкою моделі через шкідливі вхідні тексти.
- Витік конфіденційної інформації – ненавмисне відтворення приватних даних користувачів.

Крім того, моделі схильні до нестабільності – на дуже схожі запити вони можуть давати суттєво різні відповіді, що ускладнює гарантування надійності в критичних сферах застосування.

1.3 Огляд наборів даних для оцінювання генеративних систем штучного інтелекту

Оцінка генеративних систем штучного інтелекту, зокрема ВВМ, потребує різноманітних наборів даних та бенчмарків для аналізу різних аспекти якості та здібностей.

За останні роки розроблено кілька основних бенчмарків :

- **GLUE (General Language Understanding Evaluation)** – це набір із дев’яти завдань, спрямованих на перевірку розуміння природної мови в різних доменах. Він включає класифікацію речень, визначення схожості, парафрази, логічні висновки (MNLI, QNLI, RTE, WNLI). Кожне завдання оцінюється за окремими метриками (наприклад, точність або F1), а загальний бал підсумовує продуктивність моделі по всіх завданнях. [6]
- **TruthfulQA** – спеціальний бенчмарк для оцінки правдивості відповідей мовних моделей, створений OpenAI у 2021 році. Він містить 817 питань у 38 категоріях (медицина, право, фінанси,

поширені міфи тощо), сформульованих так, щоб перевірити, чи схильна модель повторювати поширені помилкові уявлення. TruthfulQA вимірює точність та інформативність відповідей, показуючи схильність моделей до галюцинацій. Важливо зазначити, що більші моделі не завжди краще відповідають на такі питання, якщо їх спеціально не налаштовано уникати типових помилок. [7]

- **Perplexity** – класична метрика, що оцінює здатність моделі передбачати послідовність тексту. Чим нижча заплутаність, тим краще модель передбачає текст, що свідчить про її базову мовну продуктивність. Однак заплутаність не обов'язково корелює з правдивістю, релевантністю або загальною якістю відповідей, а лише загальну ймовірність плавності.

Таким чином, для комплексної оцінки генеративних моделей необхідно використовувати кілька метрик та наборів даних (GLUE, TruthfulQA, MT-Bench), які враховують заплутаність, послідовність, релевантність, фактичну точність і відсутність упередженості. Жодна метрика не є достатньою сама по собі, а повна оцінка передбачає і кількісні порівняння, і якісний аналіз, включаючи оцінки людей. Саме такий комплексний підхід дозволяє не лише визначати продуктивність, а й спрямовувати покращення генеративних моделей.

1.4 Формулювання мети, завдань та гіпотез дослідження

Метою цієї бакалаврської роботи є дослідження та розробка методології оцінки якості сучасних генеративних систем штучного інтелекту на прикладі великих мовних моделей. Методологія повинна поєднувати автоматизовані метрики із експертною людською оцінкою для більш точного аналізу результатів генерації. Цілі включають:

1) Проведення огляду сучасних генеративних моделей, таких як Transformer-архітектури, моделі з навчанням з підкріпленням (RLHF) та підходи Retrieval-Augmented Generation (RAG). Дослідити сфери застосування та ключових викликів у оцінці якості результатів.
2) Пошук та аналіз існуючих бенчмарків і наборів даних, які використовуються при оцінці ВММ.
3) Формулювання гіпотез щодо ефективності поєднання автоматичних метрик та людської оцінки при аналізі згенерованих текстів.
4) Вибір та обґрунтування метрик, які використовуються для оцінки відповіді моделі, включаючи критерії точності, достовірності, релевантності та повноти.
5) Проведення серії експериментів із застосуванням великих мовних моделей у різних конфігураціях генерації, аналіз отриманих результатів та перевірка висунутих гіпотез.
6) Формулювання висновків та ідей щодо удосконалення підходів до оцінювання генеративних систем штучного інтелекту, а також можливих напрямів подальших досліджень – зокрема розширення тестових корпусів і впровадження мультимодальних підходів до аналізу.

РОЗДІЛ 2. РОЗРОБКА СИСТЕМИ ОЦІНКИ ЯКОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

2.1 Визначення критеріїв оцінки системи

Оцінювання якості вихідних даних генеративних моделей штучного інтелекту вимагає застосування комплексного набору критеріїв, які можна поділити на об'єктивні (автоматизовані) та суб'єктивні (людські оцінки). Об’єктивні метрики забезпечують кількісне та відтворюване вимірювання певних характеристик тексту, тоді як суб’єктивне оцінювання людьми враховує нюанси розуміння, контексту і якості, що важко формалізувати.

Ключові критерії оцінки включають:

Таблиця 2.1.

Перелік критеріїв оцінки якості генеративних систем штучного інтелекту.

Група критеріїв	**Метрика / показник**	**Сутність**
Коректність	*Answer Correctness*	Чи збігаються твердження з еталоном (фактична точність)
Релевантність	*Answer Relevancy*	Визначення зв’язку відповіді із початковим запитом
Повнота	*Context Recall*	Охоплення всіх важливих аспектів запиту
Узгодженість	*Coherence*	Логічність побудови відповіді та відсутність суперечностей
Безпека	*Toxicity, Bias*	Перевірка на упередженість або шкідливість змісту

Ефективність	*Latency, Token Efficiency*	Швидкість відповіді та вартість обчислювальних ресурсів
Робастність	*Adversarial Robustness*	Стійкість до некоректних або навмисно викривлених запитів
Людська оцінка	Ручна 5-бальна шкала	Суб’єктивна оцінка якості за шкалою

Підсумовуючи, ефективна оцінка генеративних моделей потребує поєднання об’єктивних тестів (стандартизовані метрики, бенчмарки) та суб’єктивних людських оцінок. Сучасні підходи включають використання семантичних метрик (наприклад, BERTScore, BLEURT) та оцінювання за допомогою мета-експертів (іншими потужними моделями). Лише такий комплексний підхід, що включає різні критерії, забезпечує всебічне розуміння сильних і слабких сторін цих систем, спрямовуючи подальший розвиток і вдосконалення. [8] [9]

2.2 Розробка підходу для оцінки ефективності та коректності генерації тексту

Для цілісної оцінки результатів роботи великих мовних моделей пропонується комбінований підхід, який об’єднує автоматизовані метрики з людським оцінюванням. Автоматичні методи гарантують об’єктивність і швидкість аналізу, тоді як експертна оцінка дозволяє глибше зрозуміти аспекти, які важко кількісно визначити. Такий гібридний підхід дає змогу перевірити як якість (наскільки вдало модель генерує змістовно доречний результат), так і достовірність створеного контенту.

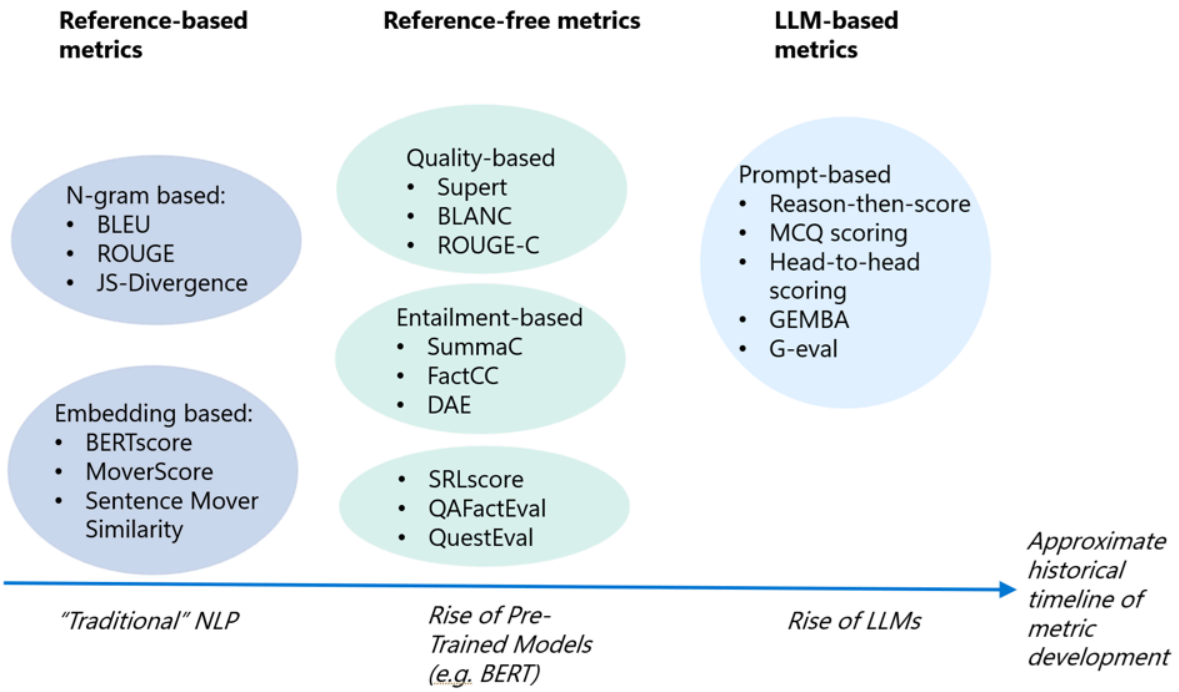

Рисунок 2.1 – *Ілюстрація категорій метрик оцінки для згенерованого тексту*

Автоматичні метрики оцінки

Автоматичні метрики дають можливість швидко та стандартизовано оцінювати характеристики згенерованих текстів. Більшість таких метрик спочатку були створені для конкретних NLP-задач (переклад чи узагальнення), але сьогодні активно застосовуються для ВВМ.

Метрики на основі еталону (Reference-based)

Цей підхід порівнює результат моделі з попередньо визначеними еталонними (правильними) відповідями. Такі метрики визначають, наскільки згенерований текст схожий до очікуваної відповідь. Найбільш відомі метрики цього типу:

- **N-gram overlap metrics** – BLEU та ROUGE є класичними прикладами. BLEU (Bilingual Evaluation Understudy) оцінює збіг коротких фрагментів тексту, між згенерованою відповіддю та еталонною відповіддю. Він широко використовується для оцінювання якості генерації, зокрема в задачах машинного перекладу та фактологічних

QA. Для запитань, що мають одну правильну відповідь, вищий показник BLEU часто корелює з точністю. Однак важливо враховувати, що BLEU здебільшого фокусується на лексичну подібність, тобто на поверхневу схожість формулювань, а не на смислову відповідність. Він є чутливим до варіацій у формулюванні, синонімів або перефразувань – і внаслідок цього модель може дати абсолютно правильну відповідь, але через використання інших слів BLEU буде низьким. Таким чином, BLEU може занижувати оцінку, особливо якщо відповідь правильна за змістом, але відрізняється стилістично.

ROUGE, зокрема ROUGE-N і ROUGE-L, краще підходить для завдань із відкритою відповіддю, де важливо, наскільки добре передано зміст оригінального тексту. На відміну від BLEU, ROUGE більше орієнтований на виявлення збігу за змістом, а не лише за формою, що робить його більш адаптованим для завдань зі складною або розгорнутою відповіддю.

- **Semantic similarity metrics** – щоб подолати обмеження поверхневих методів оцінки, дедалі частіше застосовуються метрики, орієнтовані на семантичну схожість. Вони також зазвичай потребують наявності еталонного тексту, однак замість прямого зіставлення рядків аналізують змістовні представлення тексту. Наприклад, метрика BERTScore визначає семантичну подібність між згенерованим текстом і референсом через порівняння ембедінгів, отриманих із попередньо натренованої моделі трансформера BERT. Таким чином, семантичні метрики дають змогу точніше оцінювати відповідність відповідей, акцентуючи увагу на переданому змісті, а не тільки на збігові слів.

Метрики без еталонних значень (Reference-free metrics)

Ці метрики дозволяють оцінити тексти у випадках, коли єдино правильної відповіді не існує або генерація є дуже варіативною (наприклад, у випадках діалогових систем).

- **Фактична коректність** – використовуються зовнішні перевірені джерела знань (відповідь може бути перевірена шляхом пошуку відповідних тверджень у Wikipedia або інших джерелах) або NLI-моделі, що дозволяє визначити, чи є твердження логічно узгодженими з контекстом. Такий метод допомагає мінімізувати ризик генерації неправдивої інформації, оцінюючи підтвердження фактів наявними даними.
- **Instruction-following and relevance** – автоматичні перевірки визначають, наскільки відповідь дотримується заданого формату, стилю, уникає заборонених тем і повністю охоплює поставлені питання. Зокрема, аналізується відповідність тону ролі, кількість необхідних пунктів, дотримання умовних обмежень, а також семантична близькість між запитом і відповіддю за допомогою косинусної схожості ембедінгів.

У деяких випадках самі ВВМ використовуються як інструменти оцінювання, граючи роль мета-експертів. У цьому випадку потужніша модель (або та ж модель зі спеціалізованим промптом) аналізує відповідь, створену меншою моделлю, та виставляє їй оцінку за визначеними критеріями. Наприклад, можна залучити GPT-4 для оцінки відповіді GPT-3 за фактичною точністю та логічною узгодженістю, сформулювавши завдання як: `"Оціни наведену відповідь за критеріями фактичної точності та логічної послідовності, вистав оцінку від 1 до 5 і обґрунтуй свій вибір."`

Такий метод (іноді називається GPT-score або оцінювання на основі ВВМ) демонструє перспективні результати, оскільки великі моделі здатні

виявляти помилки подібно до людини. Однак, повна довіра до моделі-оцінювача є недоцільною через можливість її власних помилок або упереджень, тому цей підхід часто поєднується з вибірковою перевіркою людиною.

Оцінювання систем на основі Retrieval-Augmented Generation (RAG):

Окремо слід розглянути випадок, коли Великі Мовні Моделі поєднуються з механізмом пошуку – системи з retrieval-augmented generation (RAG). У таких системах якість відповіді залежить не лише від генерувальної здатності моделі, але й від роботи модуля пошуку інформації. Відповідно, і критерії оцінки включають додаткові компоненти: оцінку релевантності знайденого контексту та того, наскільки модель правильно використала цей контекст при формуванні відповіді. [10]

Для оцінювання якості RAG-процесу розроблено спеціалізований фреймворк RAGAS (Retrieval-Augmented Generation Assessment). Він пропонує набір метрик для окремого вимірювання якості пошукового модуля та генеруючого модуля, а також інтегрованої оцінки системи без необхідності ручних розміток. Зокрема, RAGAS дозволяє автоматично оцінити:

- релевантність витягнутого контексту (чи дійсно знайдені документи стосуються запиту)
- фактичну достовірність відповіді на основі цього контексту (чи не містить відповідь інформації, якої немає в знайдених джерелах)
- загальну якість генерованого тексту (наскільки відповідь відповідає запиту і сформульована правильно)

Рисунок 2.2 – *Ключові метрики оцінювання в системі RAGAS, розділені на модулі генерації та пошуку*

Оцінки надаються в межах від 0 до 1, де вищі значення вказують на кращу якість. Наприклад, значення Faithfulness = 1.0 означає, що відповідь не містить жодного твердження, яке б не було підтверджено витягнутим контекстом.

В наступних розділах ми обрали саме RAGAS як основний інструмент оцінювання, оскільки він дозволяє фіксувати критичні аспекти якості, які часто ігноруються стандартними метриками, зокрема наявність "галюцинацій" у відповідях. Важливо, що цей підхід дає змогу розмежувати джерело помилки – чи вона виникла на етапі витягування інформації, чи вже під час генерації відповіді.

Такий підхід ілюструє зростаючу потребу в адаптованих метриках для різних сценаріїв застосування ВВМ – від інформаційних агентів до діалогових систем із доступом до зовнішніх знань. [11]

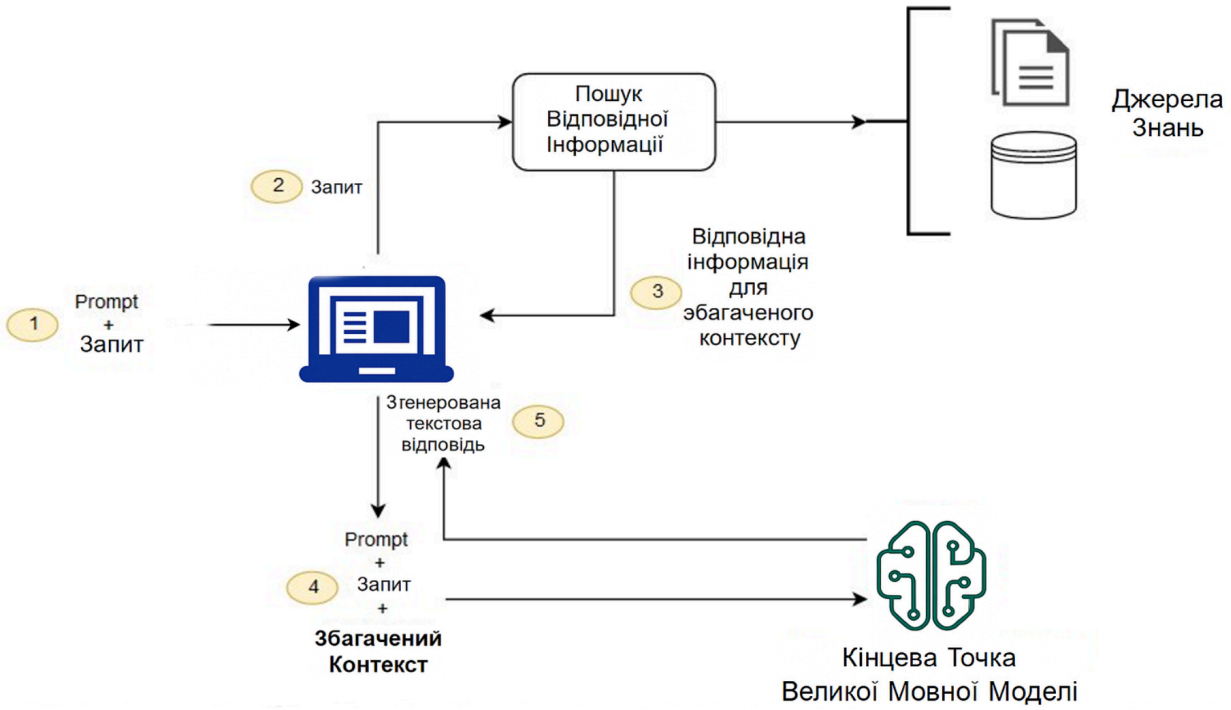

Рисунок 2.3 – *Схематичне зображення процесу генерації з використанням пошукового підсилення (RAG). Спочатку користувацький запит (2) надходить до пошукового модуля, який здійснює витяг релевантних даних із бази знань. Отриманий контекст (3) об'єднується із початковим запитом (1), формуючи розширений запит(4) для великої мовної моделі. На основі цього підсиленого запиту модель генерує фінальну відповідь (5), що враховує додаткову інформацію та забезпечує вищу фактичну точність.*

Людська оцінка

Незважаючи на прогрес автоматичних методів, оцінювання людиною залишається основним еталоном якості. Експерти або кінцеві користувачі можуть цілісно судити про такі параметри, як релевантність відповіді запиту, фактична точність, зрозумілість і логічна послідовність викладу, стилістична відповідність, а також корисність та доречність відповіді в контексті поставленого завдання. Люди здатні врахувати тонкі аспекти, наприклад,

наскільки відповідь задовольняє інформаційну потребу або чи не є вона образливою для певної аудиторії. Тому у випадках спірних або складних відповідей остаточне рішення про якість покладається на людину. Людське втручання є особливо важливим, якщо автоматичні метрики дають суперечливі результати чи мають високу невизначеність – наприклад, якщо модель згенерувала текст із потенційно токсичним забарвленням або сумнівною фактологією, де потрібна експертна перевірка. Зібрані від людей оцінки також слугують основою для подальшого навчання та налаштування моделей під очікування реальних користувачів.

Методології залучення людей:

- **Парне порівняння** (Pairwise ranking) – оцінювачі порівнюють кілька відповідей на одне запитання й обирають найкращу за заданими критеріями.
- **Likert-scale оцінювання** – відповіді оцінюються за шкалою (наприклад, від 1 до 5) за кількома параметрами: змістовність, релевантність, коректність тощо.
- **Annotation with explanation** – експерти надають обґрунтування, що дозволяє уточнити межі неоднозначності в оцінці.

Отже, поєднуючи автоматичні метрики та людські судження, підхід до оцінювання може охоплювати всі необхідні аспекти. Автоматизовані інструменти забезпечують масштабованість і швидкість, коли людська оцінка надає глибину аналізу, гарантуючи, що бажані аспекти правильно адаптовані.

2.3 Імплементація алгоритмів для аналізу якості згенерованого тексту

Після визначення критеріїв оцінювання, перейдемо до їх імплементації. Метою є створення алгоритмів, які можуть автоматично аналізувати

згенерований текст та надавати оцінки якості відповідно до обраних критеріїв.

Оцінювання за допомогою ВВМ

Один із ефективних способів автоматичного аналізу тексту – залучення потужніших мовних моделей в ролі мета-критика. Ідея полягає у використанні вже наявних можливостей великих моделей, таких як GPT-4, для оцінювання. Модель отримує на вхід згенеровану відповідь і спеціально сформульований промпт, що інструктує її проаналізувати цю відповідь за потрібними критеріями. Правильно створений запит може змусити модель оцінити текст за параметрами точності, логічності, стилю тощо та видати оцінки. Наприклад, промпт: «Оціни наведений текст за критеріями точності та логічності. Постав оцінку за шкалою від 1 до 5 та поясни цей вибір» – спонукає модель-експерта надати обґрунтовану оцінку. Такий алгоритм можна запрограмувати як функцію, що передає відповідь моделі-аналізатору і розбирає отриманий відгук (оцінки) на структуровані метрики. Важливо забезпечити узгодженість оцінювання: щоб зменшити випадковість, використовують фіксовані інструкції і, за можливості, середнє по кількох прогонах моделі-оцінювача. Якщо різні моделі-експерти дають суперечливі оцінки, або модель не впевнена у відповіді, передбачається перехід до ручної перевірки.

Класичні класифікатори

Окремі характеристики згенерованого тексту – токсичність, наявність упереджень, фактичні помилки. Аналіз відбувається за допомогою спеціалізованих моделей. Наприклад, для токсичності використовуються API типу Perspective або відкриті моделі такі як Detoxify. Для логічної

послідовності та перевірки фактів застосовуються NLI-моделі, наприклад RoBERTa або FactCC, які порівнюють твердження з контекстом.

Агрегація результатів

Оцінки (числові, бінарні, ймовірнісні) зводяться до одного підсумкового результату. Агрегація відбувається з використанням вагових коефіцієнтів, які визначають важливість кожного критерію для підсумкової оцінки. Ваги встановлюються відповідно до пріоритетів: наприклад, для систем, де критична правдивість, метриці фактичної точності (достовірності) надається найбільша вага. Також бали можуть бути агреговані за допомогою логічних правил (наприклад, якщо одна з ключових метрик дорівнює нулю – інші бали можуть бути знецінені). Такий підхід гарантує, що грубі помилки (пряма дезінформація або образливі вислови) не будуть компенсовані іншими хорошими якостями відповіді.

Інтеграція оцінювання через RAGAS

Для систем з використання підходу RAG оцінювання відповідей здійснюється за допомогою фреймворку RAGAS, який автоматизує багато описаних вище перевірок, орієнтованих на релевантність і достовірність. Ці метрики дозволяють системі об'єктивно порівнювати відповіді з еталонними даними, не залучаючи людських експертів.

Спочатку варто сформувати набір, де кожен приклад містить запит, витягнутий контекст та відповідь, згенеровану ВВМ. Після цього визначаються метрики, які будуть використанні.

Людський контроль

Для повної системи оцінки важливо інтегрувати перевірку людиною – особливо там, де автоматичні метрики можуть помилятися. Наприклад, якщо певна метрика падає нижче встановленого порогу, відповідь автоматично

маркується як така, що потребує перевірки. Ці дані згодом використовується для подальшого налаштування системи. Таким чином, людський контроль слугує додатковою перевіркою, для зменшення майбутніх помилок автоматичних алгоритмів і одночасно джерелом даних для навчання метрик, більш узгоджених із людськими оцінками.

2.4 Висновки

Комбінування автоматичних і ручних оцінок суттєво підвищує надійність результатів, оскільки метрики для ВВМ не завжди корелюють із людською якістю. Було визначено певну множину критеріїв якості (точність, релевантність, повнота, узгодженість, безпека, ефективність та ін.) та імплементовано алгоритми для їх кількісного вимірювання. Застосування метрик на основі еталону, семантичних та reference-free підходів дало змогу об’єктивно аналізувати генерацію моделей за різними показниками, тоді як інтеграція фреймворку RAGAS продемонструвала ефективність спеціалізованих оцінок для систем з доступом до зовнішніх знань. Такий комплексний підхід забезпечує більш об’єктивне та всебічне оцінювання роботи ВММ та створює основу для подальшого покращення моделей на основі зворотного зв’язку.

РОЗДІЛ 3. ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ ТА ОЦІНКА РОЗРОБЛЕНОЇ СИСТЕМИ

3.1 Реалізація прототипу системи оцінки якості ВВМ

У цьому розділі представлено структуру та реалізацію експериментального прототипу, розробленого для аналізу якості відповідей ВВМ. Система побудована на основі підходу Retrieval-Augmented Generation (RAG), що поєднує генерацію тексту з попереднім пошуком релевантної інформації. Такий підхід допомагає моделі отримати доступ до додаткових знань перед формуванням відповіді, що дозволяє значно підвищити її точність і зменшити ймовірність вигаданих фактів.

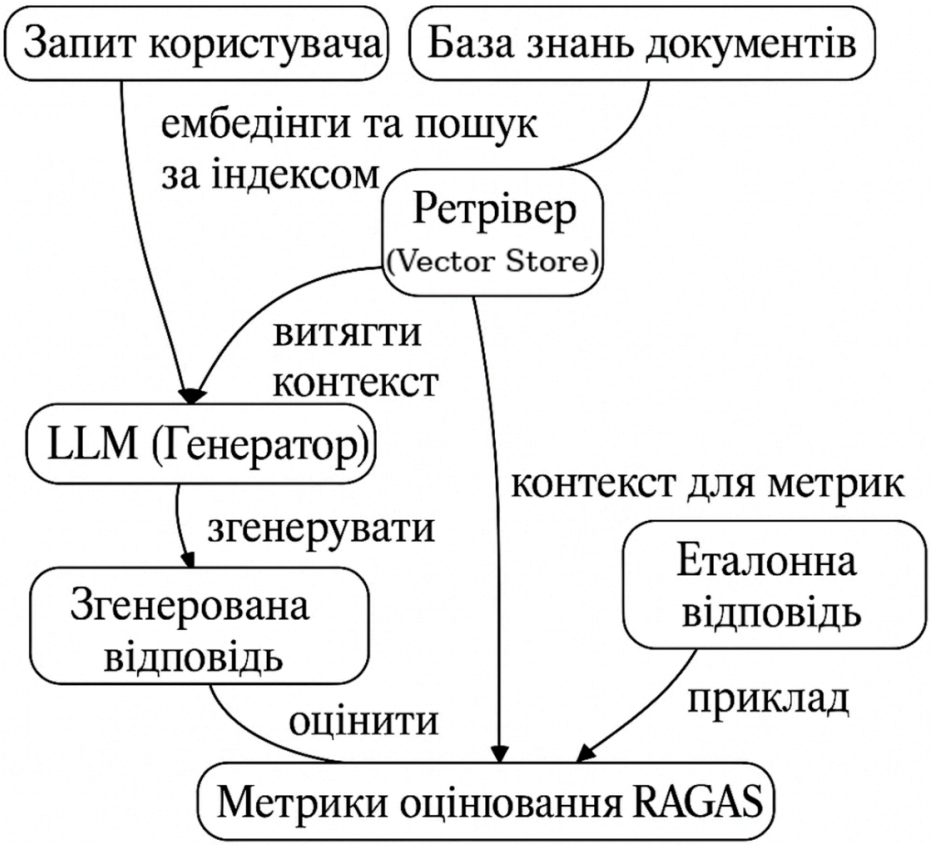

Рисунок 3.1 – *Ілюстрація прототипу архітектури автоматичної оцінки RAGAS*

Архітектура системи включає два основні модулі – генерації, який формує відповідь на основі запиту та знайденого контексту і оцінювання, що

автоматично визначає якість відповіді за заданими метриками. Набір технологій було обрано з метою максимальної сумісності із сучасними великими мовними моделями та для використання вже наявних інструментів. Для швидкого прототипу було використано мову програмування Python, а зв'язок між компонентами був реалізований з використанням фреймворку LangChain.

Базу знань реалізовано як векторний індекс Qdrant (для швидкого пошуку за схожістю), з використанням однієї й тієї ж моделі векторизації як для GPT-4, так і для Gemini, аби забезпечити узгодженість у пошуку.

На першому етапі запит користувача обробляється за допомогою пошуку по векторизованій базі знань, яка містить попередньо оброблені документи, перетворені на вектори за допомогою ембедингової моделі. Найбільш релевантні фрагменти додаються до запиту й передаються до LLM, яка генерує відповідь.

Далі відповідь разом із запитом та контекстом оцінюється за допомогою фреймворку RAGAS, який використовує низку метрик для аналізу якості без участі людини. Було обрано п'ять основних метрик, які відповідають цілям оцінки – Answer Relevancy, Answer Correctness, Faithfulness, Factual Correctness та Context Recall.

Для розрахунку цих метрик необхідно мати еталонну відповідь. Оскільки RAGAS працює за принципом порівняння – аналізує текстові збіги та схожість векторних представлень між відповіддю моделі та еталонними даними.

Структура прототипу забезпечує гнучкість – кожен модуль можна легко замінити або удосконалити. Система дозволяє додавати нові метрики, що робить її відкритою до розвитку.

У підсумку, прототип забезпечує повноцінний цикл – від обробки запиту й генерації відповіді до її автоматичної оцінки. Такий підхід дозволяє систематизовано й ефективно аналізувати відповіді ВВМ. Застосування

RAGAS дає змогу об’єктивно вимірювати різні аспекти без потреби в ручній перевірці кожної з них, що суттєво пришвидшує цикл оцінки RAG-систем. [12][13]

3.2 Формування власного набору даних для експериментальної оцінки генеративних моделей

Для проведення якісної оцінки відповідей різних мовних моделей, було сформовано власний датасет, який складається із 50 пар "питання–відповідь" (Q&A). Цей датасет включає широкий спектр питань і рівнів їх складності, щоб забезпечити повноцінне тестування можливостей генеративних моделей. До складу входять як фактичні питання – з чіткими та перевіреними відповідями, так і відкриті – що потребують аналітики, пояснень або більш розгорнутих міркувань. Завдяки різноманітності запитів ми можемо спостерігати, як моделі справляються з простим відтворенням фактів у порівнянні зі складнішими завданнями які вимагають роздумів або описовими завданнями.

Формування датасету

Набір питань і відповідей сформовані на основі річного фінансового звіту компанії Amazon за 2022 рік (форма 10-K). Цей документ є обов’язковим звітом компанії, який містить детальний аналіз фінансового стану, ризиків та операційної діяльності компанії. З цього звіту було сформовано 50 запитань, із прямими відповідями в тексті. Звіт містить приблизно 100 сторінок, які було розбито на фрагменти по 1024 токенів та індексовано у векторному сховищі Qdrant. Таким чином, кожен запит проходив через пошук релевантного контексту, який потім використовувався

для генерації відповіді. Еталонні відповіді було отримано безпосередньо із тексту звіту.

Для формування первинного набору питань було використано напівавтоматизований підхід, що поєднує можливості ВВМ (GPT-4o) з людською перевіркою. Модель аналізувала частини документу та генерувала запитання. На наступному етапі згенеровані запитання проходили ручну перевірку та оцінювалась за зрозумілістю, релевантністю та відповідністю змісту документа. Частину запитань було відредаговано з метою усунення невизначеності або надмірної загальності. Таким чином, було сформовано збалансований набір, який охоплює як конкретні фактичні питання, так і більш відкриті запити з чіткою прив'язкою до тексту джерела.

Для фактичних запитів було обрано інформацію загальних знань, таких як рік створення компанії, ризики для акціонерів (наприклад, "Коли було створено компанію Amazon?" або "Хто є засновником компанії?"), де відповідь зазвичай однозначна. Також додані питання, пов'язані з поточними подіями на момент року документу, щоб перевірити здатність моделей орієнтуватися в актуальній інформації. У випадку відкритих запитань – таких як "Які основні регуляторні виклики були згадані у декларації Amazon за 2019 рік?" – вони не мають єдиної правильної відповіді, але охоплюють ключові тези. Кожне питання супроводжується еталонною відповіддю (ground truth). Для фактичних питань це була коротка, перевірена відповідь. Для пояснювальних або відкритих питань були підготовлені розгорнуті відповіді, які охоплювали ключові пункти. Наприклад:

- Питання: «Згідно з річним звітом компанії Amazon за 2022 рік, які основні фактори були вказані як причини зростання операційних витрат?»
- Відповідь: «У звіті зазначено, що основними факторами зростання операційних витрат були вищі витрати на

персонал через збільшення кількості працівників та зростання витрат на маркетинг.»

- Еталон: «У звіті Amazon за 2022 рік зазначено, що основними факторами зростання операційних витрат було збільшення витрат на персонал через збільшення кількості працівників та зростання витрат на маркетинг.»
- Контекст: Компанія Amazon, 10-K за 2022 рік, стор. 45, абзац 2…

Такі відповіді мають дві основні функції: виступають як еталон для метрик типу “Точність Відповіді” або Семантична Подібність, та забезпечують чіткий критерій оцінки якості відповіді моделі.

Метою цього експерименту є перевірка здатності мовних моделей ефективно використовувати зовнішню інформацію, отриману через механізм витягування, та порівняльний аналіз моделей GPT-4 від OpenAI та Gemini Flash від Google. Моделі не мають попереднього знання вмісту звіту – всі відповіді повинні базуватися виключно на витягнутих фрагментах. Хоча ці моделі теоретично могли бачити інформацію про Amazon під час навчання, точні цифри та специфічні формулювання звіту доступні лише через контекст, отриманий через Qdrant. Це надало змогу не лише оцінити їхню здатність адаптуватися до стилю документації, але й перевірити чи слідують моделі наданим їм інструкціям.

Розроблений прототип, підготовлений набір даних та оцінка якості моделей знаходиться у вільному доступі за посиланням [14].

3.3. Оцінювання якості мовної моделі OpenAI GPT-4

У даному розділі розглянуто оцінку якості роботи мовної моделі GPT-4, розробленої компанією OpenAI, у задачах питання-відповідь (retrieval-based Question Answering) використовуючи підхід Retrieval Augmented Generation (RAG). Метою аналізу є оцінка здатності ВВМ використовувати зовнішній контекст при формуванні відповідей. У межах дослідження було використано попередньо описану експериментальну систему, а також підготовлений набір даних, що дозволяє зрозуміти, наскільки ця модель є придатною для практичного використання в таких середовищах, та сформувати рекомендації для її подальшого налаштування або інтеграції.

Також враховано вплив параметрів генерації на якість згенерованої відповіді. Було проведено серію експериментів зі зміною ентропії генерації тексту (температура) (0, 0.5, 1.0, 1.5) та обсягу наданого контексту (від одного до п’яти документів). Це дозволило встановити, як саме дані параметри впливають на три основні критерії оцінки відповіді – достовірність, релевантність та повноту.

У межах експерименту GPT-4 отримувала шість релевантних фрагментів тексту разом із запитанням від користувача та попередньо сформованим системним промптом для моделі. Формат промпту був стандартизованим та містив наступні інструкції: `"Використай надану інформацію для відповіді. Якщо недостатньо - повідом про це."`. Ентропію встановлено на рівні 0.2 для кращої стабільності відповідей. Контекст обмежувався до 8 тисяч токенів, що забезпечило відповідність технічним можливостям моделі. Згенеровані відповіді мали довжину 1–5 речення та оцінювалися за визначеними метриками.

Аналіз результатів включає як кількісні показники за кожною метрикою, так і якісний розбір типових помилок GPT-4 – зокрема, випадки впевнених, але фактично не правильних відповідей (галюцинацій) попри

наявність правильної інформації в контексті. Зокрема, розглядається, наскільки модель схильна пропускати релевантну інформацію або перефразовувати питання.

Таблиця 3.1.
Результати GPT-4 за ключовими метриками.

Метрика	**Оцінка (RAGAS)**	**GPT-4o оцінка**	**Оцінка людиною**
Answer Relevancy	0.97	1	1
Answer Correctness	0.95	0.97	0.97
Factual Correctness	0.80	0.97	0.96
Faithfulness	0.94	0.91	0.98
Context Recall	0.98	0.95	1

Узагальнення результатів оцінки моделі GPT-4 за ключовими метриками. Нижче наведено стислий опис отриманих результатів:

Релевантність та правильність відповіді (Answer Relevancy, Answer Correctness) – модель продемонструвала високий рівень релевантності, із середнім значенням показника близько 0.97. Це означає, що у переважній більшості випадків відповіді прямо відповідали на поставлені запитання.

Фактична точність (Factual Correctness) – модель часто формує відповіді у дещо розширеній формі, додаючи додаткову уточнювальну інформацію, яка відсутня в еталонній відповіді. Наприклад, якщо еталонна відповідь виглядала як “$5.4 мільярдів”, а GPT-4 могла відповісти `"Це було $5.4 billion у 2022 році."`. Хоча такі варіанти є змістовно ідентичними, вони не вважаються точним збігом.

Достовірність (Faithfulness) – середній рівень достовірності відповіді становив близько 0.94. Хоча в окремих випадках спостерігались незначні відхилення, коли модель додавала загальні твердження, відсутні у вихідному

контексті – загальна оцінка свідчить про здатність GPT-4 зберігати відповідність інформації з контексту.

Повнота використаного контексту (Context Recall) – середня повнота вказує на те, що в окремих випадках модель не задіює всі релевантні елементи з наданого контексту. Наприклад, якщо у вихідних даних містилося три аргументи на підтримку певного рішення, GPT-4 могла згадати лише два з них, ігноруючи третій. Така ситуація не є критичною з погляду точності відповіді, однак свідчить про часткову втрату повноти відтворення інформації. У випадку застосування моделі в критично важливих галузях, подібна неповнота може мати суттєве значення, що варто враховувати при оцінюванні її ефективності.

Дослідження впливу ентропії генерації тексту (“температура”) на фактичну точність

Також було проведено додаткове дослідження залежності між температурою генерації тексту та фактичною точністю відповідей.

За стандартного значення температури генерації 0.2, модель GPT-4 демонструвала високу стабільність і забезпечувала правильні відповіді у близько 90% випадків. З метою подальшого аналізу, ті самі запитання були повторно надані моделі із використанням підвищених ентропії: $t = 0,5$ (помірний рівень креативності), $t = 1$ (високий рівень креативності) та $t = 1,5$ (дуже високий, з ознаками галюцинацій). Отримані результати свідчать про певну закономірність, а саме – зі зростанням ентропії генерації спостерігається зниження фактичної точності відповідей.

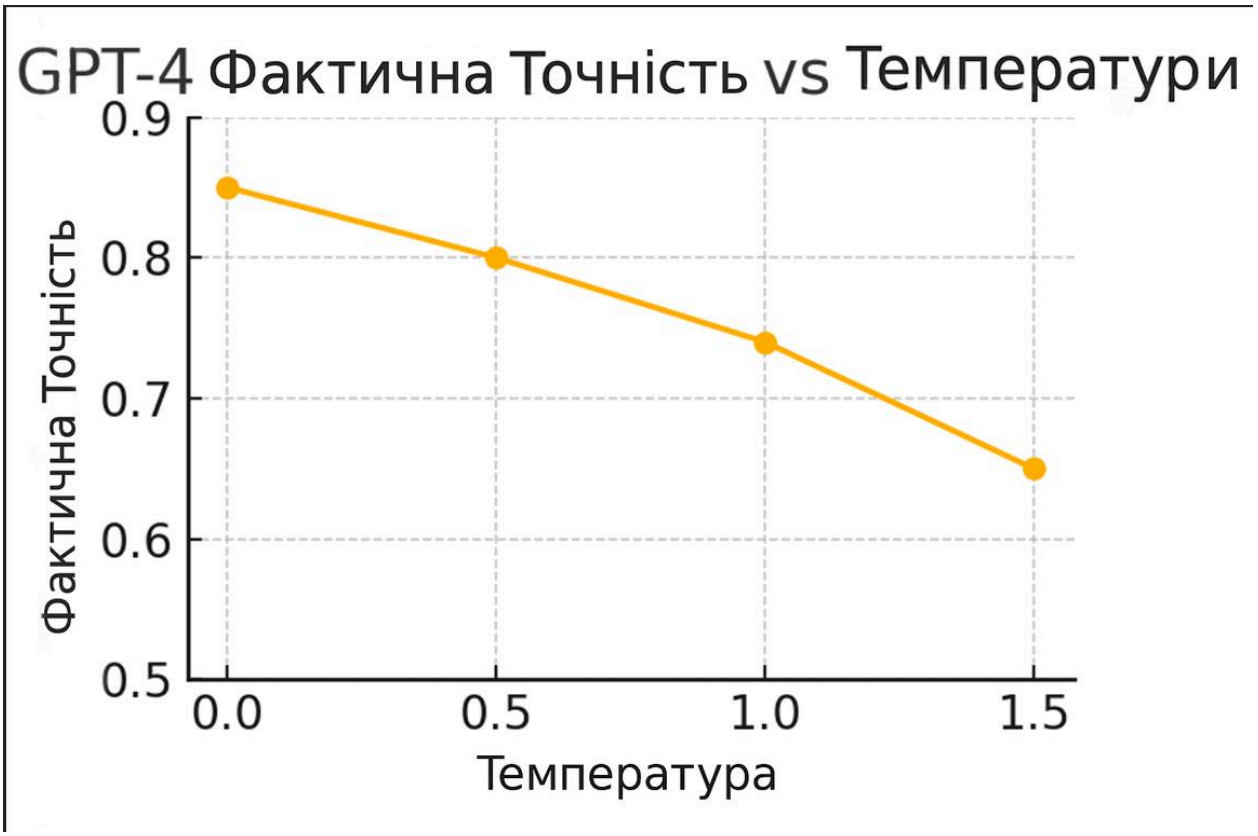

Рисунок 3.2 – *Вплив ентропії генерації тексту (“температури”) генерації на фактичну точність GPT-4*

Графік ілюструє відсоток відповідей, які були фактично правильними (вісь Y), залежно від ентропії генерації (вісь X). Найвищої точності GPT-4 досягає при низьких значеннях ентропії, тоді як підвищення впливає на можливість зростання кількості помилок та галюцинацій.

Зокрема, при значенні ентропії 1.5 спостерігаються випадки, коли модель більше використовувала гіпотетичні формулювання. Наприклад «Можна припустити, що...», що іноді супроводжувалося хибними числовими даними або некоректним представленням інформації із різних фрагментів контексту. За ентропії 0.5 більшість відповідей залишалася правильними, проте траплялися незначні фактичні неточності чи узагальнення.

В результаті, для завдань, що вимагають високої фактичної точності, доцільно застосовувати низькі значення ентропії генерації тексту.

Дослідження впливу обсягу контексту

Було проаналізовано вплив обсягу контекстної інформації на результати відповіді моделі. За замовчуванням використовувалася вибірка з трьох найбільш релевантних фрагментів тексту. Для оцінки впливу цього параметра було також протестовано вибірка із одним та шістьма документами. При використанні лише одного джерела спостерігалося зменшення повноти відповідей. Це пов’язано із тим, що частина інформації, яка містилася у другому або третьому фрагменті, не враховувалась. Наприклад, на питання, які передбачали наведення кількох причин певного значення, модель бачила лише ті, що були в першому документі, ігноруючи інші.

Збільшення обсягу до шести фрагментів призводило до незначного зниження релевантності через можливі менш значущі або суперечливі деталі. Проте, різниця між вибіркою з трьома та шістьма фрагментами була незначною – повнота відповідей зросла з 0.85 до приблизно 0.89, натомість точність знизилась з 0.92 до 0.90. Отже, обсяг у три документи є більш оптимальним значенням між повнотою та точністю.

Це вказує на важливість правильного налаштування параметрів контексту в задачах генерації, оскільки як недостатній, так і надмірний обсяг інформації може негативно впливати на якість результатів, особливо в умовах недосконалого механізму пошуку. Під час аналізу помилок було виокремлено декілька типових категорій:

- **Неповні відповіді** – випадки коли модель генерувала релевантну, але не повну відповідь. Наприклад, у запиті, де еталонна відповідь містила шість пунктів, GPT-4 наводила лише два. Така відповідь оцінювалась як частково коректна через знижену повноту.

- **Неправильне тлумачення деталей** – інколи модель помилково обчислювала числові дані. Наприклад, у відповідь на питання про дохід за 2019 рік було надано показник за 2018 рік. Ймовірно через неправильний вибір рядка з таблиці. Подібні ситуації свідчать про ризик зниження достовірності навіть при наявності відповідного контексту.
- **Галюцинації при наявному контексту** – іноді траплялися випадки, коли модель додавала неіснуючі твердження. Така помилка, ймовірно, пов’язана із впливом попереднього навчання моделі, що включало загальні знання про подібні звіти. Цей тип помилки був виявлений за допомогою метрики достовірності (оцінка 0).

3.4 Оцінювання якості мовної моделі Google Gemini

У даному розділі розглянуто оцінку якості роботи мовної моделі Gemini від Google, та провести порівняльний аналіз двох великих мовних моделей. Gemini – новіша модель, розроблена Google DeepMind. Метою аналізу є оцінка здатності ВВМ використовувати зовнішній контекст при формуванні відповідей. У межах дослідження було використано попередньо описану експериментальну систему та набір даних. Що дозволяє зрозуміти, наскільки ця модель є придатною для практичного використання в таких середовищах, та сформувати рекомендації для її подальшого налаштування.

Результати оцінювання будуть подані не лише у вигляді кількісних показників, але й у форматі порівняльного аналізу, що висвітлює сильні та слабкі сторони Gemini у порівнянні з GPT-4.

У межах того ж підходу оцінювання замінено компонент ВВМ на Google Gemini. Зокрема, використовувалася модель Gemini Flash 2.0. Механізм пошуку залишився незмінним – здійснювався вибір трьох

найрелевантніших документів, параметр ентропії було встановлено на рівні 0.2, а також використовується той же промпт.

Під час тестування було виявлено, що відповіді, згенеровані моделлю Gemini, зазвичай є довшими, у порівнянні з тим, як відповідає GPT-4, навіть за відсутності такого налаштування. Модель демонструвала схильність до детальніших пояснень.

Узагальнені результати оцінки моделі Gemini Flash 2.0 за ключовими метриками представлені в наступній таблиці.

Таблиця 3.2.
Результати Gemini Flash 2.0 за ключовими метриками.

Метрика	**Оцінка (RAGAS)**	**GPT-4o оцінка**	**Оцінка людиною**
Answer Relevancy	0.93	0.91	1
Answer Correctness	0.90	0.92	0.94
Factual Correctness	0.78	0.81	0.87
Faithfulness	0.90	0.88	0.93
Context Recall	0.84	0.88	0.9

Порівняння результатів Gemini та GPT-4 демонструє майже однакові показники за більшістю метрик, з різницею у кількох пунктах. Зокрема, рівень актуальності відповідей моделей практично ідентичний: 0.93 у Gemini та 0.97 у GPT-4. Це свідчить про те, що обидві моделі здебільшого коректно відстворюють поставлене запитання, рідко відхиляючись від теми.

За показником достовірності (Faithfulness) Gemini дещо поступається – 0.90 проти 0.94 у GPT-4. В окремих випадках Gemini додавав припущення, які не підтверджувалися контекстом. У ряді випадків спостерігалося додавання моделлю припущень, які не випливали безпосередньо з наданого контексту. Наприклад, на запитання щодо планів компанії GPT-4 зазначив, що

звіт не містить відповідної інформації, тоді як Gemini додатково інтерпретував цю відсутність як ознаку зосередженості компанії на поточній діяльності. Подібні висновки, що не підтверджені контекстом, хоч і поодинокі, знижують загальний рівень достовірності відповіді.

У частині аналізу точності та повноти відповідей Gemini показав дещо нижчі показники. Це свідчить про тенденцію Gemini надавати ширші, детальніші відповіді, однак іноді з включенням другорядної або недостатньо релевантної інформації, що частково знижує точність. Таким чином, спостерігається прагнення моделі до більш повного охоплення контексту, навіть за рахунок певної надлишковості.

За результатами оцінювання на основі Фактичної точності обидві моделі продемонстрували майже однаковий рівень в межах 0.80. Варто зазначити, що допущені помилки стосувалися різних запитань, що свідчить про подібний загальний рівень правдивості.

Загальний профіль оцінювання свідчить про незначну перевагу GPT-4. При усередненні основних метрик (RAGAS) результат оцінки можна представити у наступному вигляді: GPT-4 середній бал – 0.92, Gemini — 0.87, що підтверджує майже рівнозначний загальний рівень продуктивності.

Для візуального представлення основних відмінностей між моделями GPT-4 та Gemini було побудовано наступний графік, де здійснено порівняння за ключовими метриками.

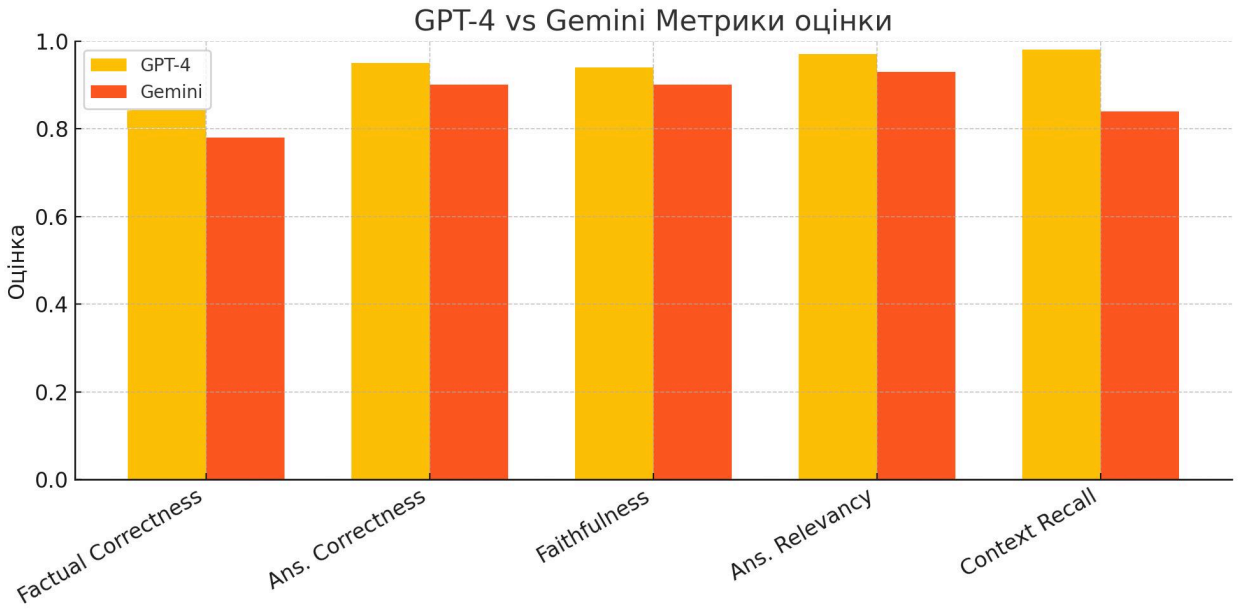

Рисунок 3.3 – *Порівняльна оцінка GPT-4 та Gemini за ключовими метриками. Кожна пара стовпців демонструє результат, отриманий GPT-4 (жовтий) і Gemini (помаранчевий) за відповідною метрикою.*

Проведене оцінювання моделі Google Gemini засвідчило здатність запропонованої системи ефективно здійснювати порівняльну оцінку, а також продемонструвало, що створений датасет має достатній рівень складності для виявлення відмінностей між сучасними ВВМ. Отримані результати свідчать про те, що Gemini демонструє ефективність, співставну з GPT-4, що на практиці надає можливість вибору між цими моделями з урахуванням економічної доцільності або вимог до інтеграції. Незначні відмінності в стилі відповідей (зокрема, схильність Gemini до розгорнутих формулювань) та проведений аналіз згенерованих відповідей надає глибше розуміння особливостей моделей.

3.5 Висновки

У цьому розділі представлено результати експериментального дослідження розробленої системи, яка була застосована для оцінювання

якості двох сучасних великих мовних моделей – GPT-4 та Gemini Flash 2.0. За визначеними критеріями оцінки, обидві моделі показали високі результати. Вони правильно відповідали на більшість питань, демонструючи рівень, близький до людського.

GPT-4 краще дотримувалася заданого контексту та надавала дещо точніші відповіді, тоді як Gemini забезпечувала ширше охоплення контекстної інформації та надавала більш повні відповіді.

У підсумку, експериментальні дослідження підтвердили ефективність і гнучкість розробленої системи оцінювання. Продемонстровано, що запропонований прототип здатен інтегруватися з сучасними великими мовними моделями, забезпечуючи детальний аналіз якості відповідей для порівняння моделей і вдосконалення систем.

4 ЕКОНОМІЧНА ЧАСТИНА

Дана науково-дослідна робота присвячена дослідженню методів оцінювання якості сучасних генеративних моделей, зокрема великих мовних моделей (ВММ), що активно застосовуються в системах штучного інтелекту. Оцінювання результатів генерації є ключовим етапом при впровадженні нових моделей у практичну експлуатацію. Отримані результати можуть бути використані для покращення контролю якості штучно створеного контенту та підвищення надійності генерації у пошукових агентах та чат-ботах. Це забезпечує економічну доцільність використання таких моделей у прикладних задачах та підвищує ефективність у роботі фахівців, які працюють з генеративними технологіями.

У процесі дослідження було реалізовано прототип системи оцінювання якості згенерованих відповідей та проведено експериментальні дослідження з аналізом ефективності популярних ВММ від компаній OpenAI та Google DeepMind.

4.1 Розрахунок витрат на виконання НДР

Розрахунки витрат на виконання бакалаврської кваліфікаційної роботи здійснюється шляхом складання калькуляції кошторисної вартості НДР та аналізу витрат за наступними статтями витрат:

1. витрати на оплату праці
2. відрахування на соціальні заходи
3. витрати на матеріали
4. витрати на використання комп’ютерної техніки
5. витрати на використання спецобладнання для наукових (експериментальних) робіт;
6. інші витрати
7. накладні витрати

Розрахунок усіх витрат виконується з точністю до сотих (до копійок).

4.1.1 Розрахунок витрат на оплату праці

Розрахунок витрат на оплату праці включає визначення заробітної плати керівника, консультанта та студента, який виконує дипломну роботу.

Для кожної категорії виконавців на першому етапі визначається середньоденна ставка. Середньоденна ставка визначається як частка місячного окладу та обсягу виконаних робіт за наступними формулами:

$$\text{Сд}_\text{і} = \text{ЗП}_\text{і}/Fр \ , \qquad (4.1)$$

де $ЗП_i$ – місячний посадовий оклад, Fp – фонд робочого часу (23 дні).

Витрати на оплату праці виконавців даної роботи визначаються за формулою:

$$\text{В}_\text{оп} = \sum_{i}^{N} n_\text{і} \cdot t_\text{і} \cdot \text{Сд}_\text{і} \ , \qquad (4.2)$$

де n_i – чисельність виконавців i-ої спеціальності, які приймають участь в проектуванні, осід; t_i – час, безпосередньо затрачений на розробку проекту кожним виконавцем i-ої спеціальності, днів; $\text{Сд}_\text{і}$ – денна заробітна плата i-ої спеціальності, грн.

У таблиці 4.1 наведено вихідні дані, які було використано для подальших розрахунків. Сам розрахунок витрат представлено в табл. 4.2.

Таблиця 4.1

Вихідні дані для розрахунку витрат на оплату праці

№ п/п	Найменування та посади виконавців	Місячний оклад, грн.	Середньоденна ставка, грн.
1	Керівник БКР, доцент	23000	1000,00
2	Консультант з економіки, доцент	23000	1000,00
4	Студент	2000	86,96

Таблиця 4.2

Розрахунок витрат на оплату праці

№ п/п	Посада виконавців	Час розробки, год	Погодинна заробітна плата, грн.	Витрати на розробку, грн.
1	Керівник БКР, доцент	18,5	125	2312.5
2	Консультант з економіки, доцент	0,5	125	62,5
4	Студент	180	10,87	1956,60
Разом				4331,60

4.1.2 Відрахування на соціальні заходи

Наступний етап після визначення витрат на заробітну плату — це обчислення єдиного соціального внеску (ЄСВ). Він нараховується на повну суму заробітної плати, зокрема на оплату праці всіх учасників, задіяних у реалізації проєкту. Відповідно до чинного законодавства, відрахування до державних соціальних фондів здійснюються за ставкою 22% від загального фонду оплати праці.

Розрахунок ЄСВ проводиться за формулою:

$$В_{ЄСВ} = \frac{22}{100} \cdot В_{ОП} \quad (4.3)$$

За умови, що загальні витрати на оплату праці становлять 4331,60 грн, сума внеску на соціальне страхування дорівнює:

$$В_{ЄСВ} = 0,22 \cdot 4331,60 = 952,95 \text{ грн}$$

4.1.3. Розрахунок витрат на матеріали

У цій частині розглядаються витрати, пов’язані з використанням матеріалів, необхідних для завершення виконання БКР. Зокрема, маються на увазі витрати на матеріали, які використовуються для оформлення та підготовки остаточного варіанту проєкту до подання на захист. До цієї категорії належать витрати на друк повного тексту роботи, виготовлення

титульних аркушів, підшивку, а також інші супутні витратні матеріали, що забезпечують відповідність документації встановленим вимогам.

Розрахунок загальної суми витрат на матеріали здійснюється за формулою:

$$В_М = \sum_{i=1}^{n} Н_i \cdot Ц_i \cdot (1 + К_{ТЗ}) , \quad (4.4)$$

де $Н_і$ – кількість одиниць і-го виду матеріалу; $Ц_і$ – ціна за одну одиницю відповідного матеріалу; $К_і$ – коефіцієнт транспортно-заготівельних витрат (приймається у межах 0,10-0,12).

Підсумкові значення витрат на матеріали наведено у таблиці 4.3.

Таблиця 4.3

Розрахунок витрат на основні та допоміжні матеріали

№ з/п	Найменування (вид) матеріалу	Одиниця виміру	$Н_і$, од.	$Ц_і$, грн/од.	$В_{Мі}$, грн	$К_{тз}$ (0,10)	Загальна сума
1.	Друк монохромний(формат А4)	шт.	90	3	270	27	297
2.	Папка для паперів	шт.	1	20	20	2	22
Разом							319

4.1.4 Витрати на використання комп’ютерної техніки

Цей розділ охоплює витрати, пов’язані з експлуатацією комп’ютерного обладнання протягом усього циклу підготовки бакалаврської кваліфікаційної роботи. Сюди включено знос та амортизацію комп’ютерної техніки, витрати на електроенергію, а також використання відповідного програмного забезпечення. Відповідно до розрахунків, прийнятих у розрахунковому центрі Національного університету “Львівська політехніка”, середня вартість однієї

години роботи персонального комп’ютера стандарту IBM PC/ATX становить 10,5 грн.

У межах виконання цієї роботи комп’ютерна техніка була задіяна протягом 180 години. Загальні витрати розраховуються за формулою:

$$В_{КТ} = Т \cdot С_{КТ}, \quad (4.5)$$

де Вкт — загальна сума витрат на використання комп’ютерної техніки, грн; Т — загальна кількість годин експлуатації, год; Скт — вартість однієї години роботи комп’ютера, грн/год.

Проведений розрахунок представлено у таблиці 4.4.

Таблиця 4.4

Розрахунки витрат на використання комп’ютерної техніки

№ з/п	Назва етапів робіт, при виконанні яких використовується комп’ютер	Час використання комп’ютера		Витрати на використання комп’ютера, грн
		днів	годин	
1.	Опрацювання наукових джерел та інформаційний пошук	12	36	378
2.	Реалізація прототипу та перевірка функціоналу моделей	12	36	378
3.	Аналіз отриманих результатів	4	8	84
4.	Оформлення БКР	12	52	546
Разом		40	132	1326

4.1.5. Розрахунок витрат за статтею “Спецобладнання” для наукових (експериментальних) робіт

У процесі реалізації експериментальної частини роботи було використано платні онлайн-сервіси з доступом до API провідних хмарних моделей штучного інтелекту, зокрема від OpenAI та Google. Надання доступу до розширених функцій генеративних моделей через API дозволило

автоматизувати низку обчислювальних завдань, пришвидшити процес оцінки та забезпечити вищу ефективність проведених досліджень. Це також дало можливість тестувати моделі в реальному часі та інтегрувати їх у процес оцінки та виводу результатів.

Загальні витрати на використання спеціалізованого програмно-апаратного забезпечення подано в таблиці 4.5 .

Таблиця 4.5

Розрахунок витрат на “Спецобладнання” для наукових (експериментальних) робіт

Назва сервісу/ інструмента	Кількість запитів	Загальна сума, грн.
OpenAI (GPT-4 + ada-002 embeddings)	4000	415,90
Google (Gemini Flash 2.0)	1000	26,80
Разом		442,70

4.1.6 Інші витрати

Інші витрати включають ті статті, які не охоплюються основними пунктами кошторису, але виникають у процесі реалізації проєкту. У рамках даної роботи не здійснювалось окреме придбання фізичного обладнання або ліцензійного програмного забезпечення. Враховуючи це, величина інших витрат умовно приймається на рівні 10% від суми загальних витрат на оплату праці зайнятих працівників. Оскільки останні становили 4331,60 грн, обсяг інших витрат складає:

$$В_{І} = 0,1 \cdot В_{ОП} = 0,1 \cdot 4331,60 = 431,60 \text{ грн}$$

4.1.6 Накладні витрати

Накладні витрати охоплюють усі непрямі витрати, що супроводжують організацію та координацію роботи над проєктом. До цих категорії належать:

1) витрати на управління;
2) загальногосподарські витрати;
3) невиробничі витрати.

У спрощеному вигляді накладні витрати можуть обчислюватися за середніми нормативами у відсотковому співвідношенні до загального фонду оплати праці всіх працівників, задіяних у виконанні науково-дослідної роботи. Зазвичай такий відсоток встановлюється у межах 150–200%. У даному випадку було прийнято мінімально допустиме значення — 150%, що дозволяє здійснити обережну, але обґрунтовану оцінку.

Розрахунок накладних витрат проводиться за формулою:

$$В_{Н} = В_{ОП} \cdot \frac{150}{100} = 4331,60 \cdot 1,5 = 6497,40 \text{ грн}$$

4.1.7 Калькуляція кошторисної вартості виконання БКР

У результаті здійснених розрахунків за всіма відповідними статтями витрат було сформовано підсумкову калькуляцію кошторисної вартості виконання бакалаврської кваліфікаційної роботи. Зведені дані подано в таблиці 4.6.

Таблиця 4.6

Обчислення кошторисної вартості БКР

№ з/п	Статті витрат	Сума витрат, грн
1.	Витрати на оплату праці	4331,60
2.	Відрахування на соціальні заходи	952,95
3.	Витрати на матеріали	319
4.	Витрати на використання комп’ютерної техніки	1326
5.	Інші витрати	442,70

6.	Розрахунок витрат за статтею “Спецобладнання”	431.6
7.	Накладні витрати	6497,40
Всього		14301,25

4.2 Розрахунок договірної ціни та прибутку БКР

Договірна ціна виконання науково-дослідної роботи визначається з урахуванням повної кошторисної вартості проєкту, встановленого рівня рентабельності та погоджених умов співпраці між виконавцем і потенційним замовником. Ця ціна має компенсувати всі витрати, пов’язані з реалізацією дослідження, і водночас гарантувати економічну доцільність для сторони-виконавця.

Для цього розрахунку було прийнято рівень рентабельності у розмірі 20%. Формула розрахунку договірної ціни має вигляд:

$$Ц = К \cdot (1 + 0,2) , \qquad (4.6)$$

де К — кошторисна вартість виконання БКР, грн.

Враховуючи попередньо розраховану суму кошторису (К = 14301,25 грн), остаточна договірна ціна становить:

$$Ц = 14301,25 \cdot 1,20 = 17161,50 \text{ грн}$$

Очікуваний прибуток обчислюється як різниця між договірною ціною та повною собівартістю:

$$П = Ц - К = 17161,50 - 14301,25 = 2860.25 \text{ грн}$$

4.3 Оцінка наукової та науково-технічної результативності БКР

Оцінка наукової та науково-технічної ефективності для НДР проводиться за допомогою коефіцієнтів, які обчислюються за формулами:

$$k_{\text{н.р.}} = \sum_{i=1}^{n} (k_{\text{зн.і}} \cdot k_{\text{д.і}}) , \quad (4.7)$$

$$k_{\text{н.т.р.}} = \sum_{j=1}^{m} (k_{\text{зн.}j} \cdot k_{\text{д.}j}) , \quad (4.8)$$

де $k_{\text{н.р.}}$, $k_{\text{н.т.р}}$ та – відповідно, коефіцієнти наукової та науково-технічної результативності; $k_{\text{зн}}$ – коефіцієнт значущості фактора; $k_{\text{д}}$ – коефіцієнт досягнутого рівня реалізації цього фактора. *n, m* – відповідно, кількість факторів наукової та науково-технічної результативності.

Усі результати оцінки результативності зводяться, відповідно, в табл. 4.7 і 4.8, на основі даних яких розраховуються відповідні коефіцієнти результативності.

На основі поданих значень розраховано узагальнені коефіцієнти згідно з формулами:

- коефіцієнт наукової результативності:

$$k_{\text{н.р.}} = (0,5 \cdot 0,7) + (0,35 \cdot 0,6) + (0,15 \cdot 0,6) = 0,65$$

- коефіцієнт науково-технічної результативності:

$$k_{\text{н.т.р.}} = (0,50 \cdot 0,8) + (0,30 \cdot 0,9) + (0,20 \cdot 0,5) = 0,77$$

Таблиця 4.7

Вибрані показники для розрахунку наукової результативності БКР

Фактор наукової результативності	Коефіцієнт значимості фактора	Якість фактора	Характеристика фактора	Коефіцієнт досягнутого рівня
Новизна отриманих чи	0,5	Середня	Проведено порівняння двох сучасних генеративних	0,7

прогнозованих результатів			моделей та запропоновано інтегровану систему оцінки для моделей типу RAG.	
Глибина наукового опрацювання	0,35	Середня	Детально розглянуто критерії оцінювання, реалізовано прототип системи, сформовано власний набір даних для оцінки, проведено серію порівняльних експериментів з аналізом типових помилок моделей.	0,6
Міра вірогідності успіху	0,15	Помірна	Експериментальна система успішно працює з декількома моделями, результати підтверджують доцільність обраного підходу для практичного використання.	0,6

Таблиця 4.8

Вибрані показники науково-технічної результативності проведеної БКР

Фактор науково-технічної результативності	Коефіцієнт значимості фактора	Якість фактора	Характеристика фактора	Коефіцієнт досягнутого рівня
Перспективність використання результатів	0,5	Важлива	Розроблений підхід можна застосовувати для систем оцінювання відповідей у пошукових агентах, чат-ботах, освітніх платформах та експертних системах.	0,8
Масштаб можливої реалізації результатів	0,3	Галузевий	Система оцінки адаптована до різних мовних моделей і може масштабуватись у межах проєктів із впровадження LLM у корпоративні чи дослідницькі середовища.	0,9

Завершеність отриманих результатів	0,2	Достатня	Повністю реалізовано прототип, проведено тестування, сформовано методологію оцінювання, запропоновано напрями подальшого розвитку.	0,5

ВИСНОВОК

У цій роботі проведено дослідження сучасних підходів до оцінювання якості великих мовних моделей, зокрема в умовах використання Retrieval-Augmented Generation (RAG). Розглянуто архітектуру трансформерних моделей, основні проблеми генеративних систем, а також актуальні метрики оцінки якості. Аналіз охоплює як класичні підходи до оцінювання (BLEU, ROUGE), так і сучасні інструменти, зокрема фреймворк RAGAS, який дозволяє більш точно оцінювати достовірність відповідей у контексті зовнішньої інформації.

Було розроблено прототип системи автоматичного оцінювання, що поєднує генерацію відповіді з попереднім пошуком інформації. Для дослідження використано мовні моделі GPT-4 та Gemini. Експерименти проводились на основі власного сформованого датасету, що містив питання з реального документа. Цей датасет охоплює як фактичні, так і відкриті запитання, що дало змогу перевірити здатність моделей працювати з різними типами запитів. Застосовано метрики, які дозволяють оцінити точність, повноту та відповідність відповідей наданому контексту.

Запропонований підхід показав свою ефективність для автоматизованого оцінювання відповідей мовних моделей. Було підтверджено, що сучасні системи можуть демонструвати високий рівень фактичної точності за умови належного налаштування процесу витягування контексту.

Результати мають практичну цінність і можуть бути використані в задачах аналізу відповідей генеративних систем, а також у побудові систем перевірки якості контенту. Зокрема, така система може стати корисною у сфері освіти, автоматизації документообігу або створення інформаційних агентів. Подальші дослідження можуть включати адаптацію запропонованого підходу до мультимодальних систем і розширення наборів даних для підвищення

надійності оцінювання. Також перспективним є впровадження більш складних моделей-експертів для оцінки відповідей для покращення результатів.

СПИСОК ДЖЕРЕЛ

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008.
2. Cho A., Kim G. C., Karpekov A., Helbling A., Wang J., Lee S., Hoover B., Chau P. Transformer Explainer: LLM Transformer Model Visually Explained. https://poloclub.github.io/transformer-explainer/
3. Юрчак І., Хіч А., Оксентюк В. (2024). Розуміння Великих мовних моделей: Майбутнє штучного інтелекту. Computer Design Systems: Theory and Practice, 51–55 ст.
4. Bergmann, D. (2023, 10 December). What is reinforcement learning from human feedback (RLHF)? https://www.ibm.com/think/topics/rlhf
5. Nucci, A. Large Language Models in Healthcare. Aisera. https://aisera.com/blog/large-language-models-healthcare/
6. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP https://paperswithcode.com/dataset/glue
7. Ip, J. (2025). TruthfulQA. In DeepEval – The Open-Source LLM Evaluation Framework. Confident AI. https://docs.confident-ai.com/docs/benchmarks-truthful-qa
8. Microsoft Learn. A list of metrics for evaluating LLM-generated content. https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-llms/evaluation/list-of-eval-metrics

9. RAGAS. Metrics documentation. https://docs.ragas.io/en/latest/concepts/metrics/

10. Balaji, M., & Parambath, B. K. Evaluate RAG responses with Amazon Bedrock, LlamaIndex and RAGAS. AWS Machine Learning Blog.

11. Shahul Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217.

12. ExplodingGradients. (2023). ragas: Automated evaluation of retrieval augmented generation https://github.com/explodinggradients/ragas

13. Couchbase Developer Portal. Evaluate RAG Responses using Ragas.

14. https://github.com/rostyslavshovak/RAG-Retrieval-Augmented-Generation/tree/main

ДОДАТКИ

Додаток 1. Промпт для моделей GPT-4 та Gemini Flash 2.0

```
PROMPT_TEMPLATE = """<instructions>
You are an expert assistant. Your answer must rely
exclusively on the information provided in the <context>
section. Do not incorporate any external knowledge or
details.
1. Use only the information provided in the <context>
section. If the necessary details are not present, respond
exactly: "I do not see that information in the context."
2. When referring to details in the context, indicate where
the information is found (for example, "as noted in the
context", page 26, etc) if it supports your answer.
3. For questions involving numerical values or any type of
calculations, provide only the final answer without
intermediate calculations. Ensure that all numbers are
accompanied by clear and correct units (e.g., "million" or
"billion") exactly as stated in the context, and
double-check these details for accuracy.
4. Before answering, check the headings from **Table of
contents** or **Table Data** for any additional unit
specifications provided in brackets (for example, "(in
millions)" or "(in billions)"). Use these units exactly as
indicated. For instance, if the heading specifies that
values are "(in millions)", your answer must reflect that
unit.
5. Ensure that your final answer is factually correct and
fully supported by the context. Do not include any extra
commentary, page references, assumptions, inferences, or
external data.
6. Be clear, concise, and accurate in your answer while
maintaining a friendly, professional tone.
7. If the context provides ambiguous or conflicting
information, explicitly state that the information is
ambiguous.
8. Do not speculate or add any information that is not
```

```
explicitly stated in the context. If a detail is missing,
respond with: "I do not see that information in the
context."
</instructions>

<context>
{context}
</context>

<conversation_history>
{chat_history}
</conversation_history>

<question>
{question}
</question>

<answer>
"""
```

Додаток 2. Перелік типових Q&A даних для оцінки моделей

```
- user_input: "What was Amazon's net sales for 2019?"
  expected_response: "Amazon's net sales for 2019 were
$280,522 million."

- user_input: "How many employees did Amazon have at the
end of 2019?"
  expected_response: "Amazon employed approximately 798,000
full-time and part-time employees as of December 31, 2019."

- user_input: "When was Amazon founded?"
  expected_response: "Amazon was founded in 1994."

- user_input: "What was Amazon's net income in 2019?"
  expected_response: "Amazon's net income in 2019 was
$11,588 million."
```

```
- user_input: "How much did Amazon spend on technology and
content in 2019?"
  expected_response: "Amazon spent $35,931 million on
technology and content in 2019."

- user_input: "What was Amazon's total long-term debt at
the end of 2019?"
  expected_response: "The face value of Amazon's total
long-term debt at the end of 2019 was $23,513 million."

- user_input: "What were Amazon's total operating expenses
in 2019?"
  expected_response: "Amazon's total operating expenses in
2019 were $265,981 million."

- user_input: "Which segments contributed most to Amazon's
net sales in 2019?"
  expected_response: "In 2019, the North America segment
contributed the most to Amazon's net sales, with $170,773
million."

- user_input: "What portion of net sales were from AWS in
2019?"
  expected_response: "In 2019, AWS accounted for 12% of the
consolidated net sales."

- user_input: "What are Amazon's primary revenue streams?"
  expected_response: "Amazon's primary revenue streams
include the sale of a wide range of products and services
to customers. These include merchandise and content
purchased for resale, products offered by third-party
sellers, the manufacture and sale of electronic devices,
and the production of media content. Additionally, Amazon
offers services such as compute, storage, and database
offerings, fulfillment, advertising, publishing, and
digital content subscriptions."
```

```
- user_input: "What was Amazon's net cash provided by
operating activities in 2019?"
  expected_response: "The net cash provided by operating
activities for Amazon in 2019 was $38.5 billion."

- user_input: "How much did Amazon spend on marketing in
2019?"
  expected_response: "Amazon spent $18,878 million on
marketing in 2019."

- user_input: "What kind of assets does Amazon report as
major?"
  expected_response: "Amazon reports major assets include
cash and cash equivalents, marketable securities,
inventories, accounts receivable, property and equipment,
operating leases, goodwill and other assets."

- user_input: "How much inventory did Amazon carry at the
end of 2019?"
  expected_response: "Amazon's inventories were $20,497
million at the end of 2019."

- user_input: "What are Amazon's main long-term
liabilities?"
  expected_response: "Amazon's main long-term liabilities
include long-term lease liabilities, long-term debt, and
other long-term liabilities."

- user_input: "How did Amazon's operating income (EBIT)
change from 2018 to 2019?"
  expected_response: "Amazon's operating income increased
from $12.4 billion in 2018 to $14.5 billion in 2019."

- user_input: "What was Amazon's net product sales in
2019?"
  expected_response: "Amazon's net product sales in 2019
```

```
were $160,408 million."

- user_input: "What was Amazon's total interest expense in
2019?"
  expected_response: "Amazon's total interest expense in
2019 was $1.6 billion."

- user_input: "Which geographic segments does Amazon
typically report in its statements?"
  expected_response: "AmazAmazon typically reports the
following geographic segments in its statements: North
America, International, and AWS (Amazon Web Services)."

- user_input: "What was Amazon's earnings per share (EPS)
in 2019?"
  expected_response: "The diluted earnings per share for
Amazon in 2019 was $23.01."


- user_input: "What was Amazon's free cash flow in 2019?"
  expected_response: "Amazon's free cash flow in 2019 was
$25,825 million."

- user_input: "What were Amazon's capital expenditures in
2019?"
  expected_response: "Amazon's capital expenditures in 2019
were $16,861 million."

- user_input: "What were Amazon's cash equivalents and
marketable fixed income securities at the end of 2019?"
  expected_response: "Amazon's cash equivalents and
marketable fixed income securities at the end of 2019 were
$45,359 million."

- user_input: "How did Amazon's subscription services
perform in 2019?"
  expected_response: "Amazon's subscription services net
```

```
sales in 2019 were $19,210 million."

- user_input: "What were the primary risks mentioned in
Amazon's 2019 10-K filing?"
  expected_response: "The primary risks mentioned in
Amazon's 2019 10-K filing include: Intense competition
across various industries and geographies, which could
impact Amazon's business operations and financial results.
Risks related to the company's fulfillment network and
inventory, particularly during periods of high demand,
which could affect cash flow and operating results.
Commercial agreements, strategic alliances, and other
business relationships that may expose Amazon to various
operational risks and affect the company's ability to
maintain and develop these relationships. Risks related to
adequately protecting Amazon's intellectual property rights
and potential accusations of infringing on the intellectual
property rights of third parties. System interruptions and
lack of redundancy, which could make Amazon's websites and
services unavailable or slow, potentially reducing sales
and harming the company's reputation. Significant inventory
risks that could adversely affect Amazon's operating
results due to factors like seasonality, rapid changes in
product cycles, and changes in consumer demand. Legal
proceedings and claims that could result in financial
liabilities and affect the company's reputation and
financial condition."

- user_input: "What percentage of Amazon's net sales were
generated from international operations in 2019?"
  expected_response: "The percentage of Amazon's net sales
generated from international operations in 2019 was 27%."

- user_input: "How did subscription services contribute to
Amazon's revenue in 2019?"
  expected_response: "In 2019, subscription services
contributed $19,210 million to Amazon's revenue."
```

```
- user_input: "What was the reported growth rate of
Amazon's AWS segment in 2019?"
  expected_response: "The reported growth rate of Amazon's
AWS segment in 2019 was 37%."

- user_input: "What was the percentage increase in Amazon's
capital expenditures from 2018 to 2019?"
  expected_response: "The percentage increase in Amazon's
capital expenditures from 2018 to 2019 was approximately
12.39%."

- user_input: "What percentage of Amazon's total assets was
comprised of cash and cash equivalents in 2019?"
  expected_response: "The percentage of total assets
comprised of cash and cash equivalents in 2019 is
approximately 16.02%."

- user_input: "What was the impact of currency fluctuations
on Amazon's international revenue in 2019?"
  expected_response: "The impact of currency fluctuations
on Amazon's international revenue in 2019 was a decrease of
$2.4 billion."

- user_input: "What were the key growth drivers for
Amazon's international segment as described in the 2019
report?"
  expected_response: "The key growth drivers for Amazon's
international segment in 2019 were increased unit sales,
driven largely by efforts to reduce prices for customers,
enhanced shipping offers, improved in-stock inventory
availability, and an expanded selection of products."

- user_input: "What were Amazon's total assets at the end
of 2019?"
  expected_response: "At the end of 2019, Amazon's total
assets were approximately $225.2 billion."
```

```
- user_input: "What were the main regulatory challenges
mentioned in Amazon's 2019 filing?"
  expected_response: "The main regulatory challenges
mentioned in Amazon's 2019 filing include compliance with
the U.S. Foreign Corrupt Practices Act and other applicable
U.S. and foreign laws prohibiting corrupt payments,
restrictions on foreign investment and operation in sectors
like the Internet and retail in China and India, and the
need to meet local ownership and cybersecurity requirements
in these countries. Additionally, there are uncertainties
regarding the interpretation of laws and regulations in
China and India, which could impact Amazon's business
operations and structures in these countries."

- user_input: "What supply chain risks were highlighted in
Amazon's 2019 filing?"
  expected_response: "The report identified several supply
chain risks, including potential disruptions in global
supply networks, rising logistics costs, and a reliance on
third-party carriers for fulfillment."

- user_input: "What method does Amazon use to value its
inventories in the 2019 10-K?"
  expected_response: "Amazon values its inventories using
the first-in, first-out (FIFO) method and the lower of cost
or net realizable value."

- user_input: "Into which segments does Amazon organize its
operations as disclosed in the 2019 10-K?"
  expected_response: "The 10-K divides Amazon's operations
into three segments: North America, International, and
Amazon Web Services (AWS)."

- user_input: "How does Amazon manage foreign exchange risk
as noted in the 2019 filing?"
  expected_response: "Amazon manages foreign exchange risk
```

```
by monitoring its foreign-denominated cash, cash
equivalents, and marketable securities, as well as its
intercompany balances denominated in various foreign
currencies. The company assesses potential adverse changes
in foreign exchange rates and estimates the impact these
changes could have on the fair value of its foreign funds
and the losses on its intercompany balances. Additionally,
Amazon includes fluctuations in foreign exchange rates in
its comprehensive income, and these are recorded in
Accumulated other comprehensive income (loss), a separate
component of stockholders' equity. Changes in foreign
exchange rates are also recognized in net income when they
involve equity securities with readily determinable fair
values."

- user_input: "How are operating expenses categorized in
Amazon's 2019 10-K?"
  expected_response: "Operating expenses in Amazon's 2019
10-K are categorized as follows: Cost of sales: $165,536,
Fulfillment: $40,232, Technology and content: $35,931,
Marketing: $18,878, General and administrative: $5,203,
Other operating expense (income), net: $201, Total
operating expenses amounted to $265,981 million."

- user_input: "How does the 2019 10-K describe Amazon's
approach to legal proceedings?"
  expected_response: "The 2019 10-K describes Amazon's
approach to legal proceedings by stating that the company
disputes allegations of wrongdoing and intends to defend
itself vigorously in these matters. This approach is
reflected in several specific legal cases mentioned, where
Amazon consistently disputes the claims against it and
expresses its intention to vigorously defend itself."

- user_input: "What are Amazon's core guiding principles as
described in the 2019 10-K?"
  expected_response: "Amazon's core guiding principles, as
```

```
described in the 2019 10-K, include customer obsession
rather than competitor focus, passion for invention,
commitment to operational excellence, and long-term
thinking."

- user_input: "Who is the CEO of Amazon Web Services as
mentioned in the 2019 10-K?"
  expected_response: "The CEO of Amazon Web Services as
mentioned in the 2019 10-K is Andrew R. Jassy."

- user_input: "How many physical Amazon stores were there
as of December 31, 2019?"
  expected_response: "There were 564 North America and 7
International physical Amazon stores as of December 31,
2019."

- user_input: "What was the total leased square footage of
Amazon's facilities as of December 31, 2019?"
  expected_response: "Amazon operated approximately 318,171
thousand square feet (about 318 million square feet) of
leased space as of December 31, 2019."

- user_input: "What was the total owned square footage of
Amazon's facilities as of December 31, 2019?"
  expected_response: "The total owned square footage of
Amazon's facilities as of December 31, 2019, was 15,615."

- user_input: "Which acquisition was included in Amazon's
consolidated financial statements starting in 2017?"
  expected_response: "The acquisition of Whole Foods
Market, Inc., completed on August 28, 2017, was included in
Amazon's consolidated financial statements."

- user_input: "What is the address of Amazon's principal
executive offices as provided in the 2019 10-K?"
  expected_response: "Amazon's principal executive offices
address is 410 Terry Avenue North, Seattle, Washington
```

```
98109-5210."

- user_input: "On which stock exchange is Amazon's common
stock traded, and what is its trading symbol?"
  expected_response: "Amazon's common stock is traded on
the Nasdaq Global Select Market under the symbol 'AMZN.'"

- user_input: "How many shareholders of record did Amazon
have as of January 22, 2020, according to the 2019 10-K?"
  expected_response: "As of January 22, 2020, Amazon had
3,169 shareholders of record of their common stock."

- user_input: "What was the aggregate market value of
voting stock held by non-affiliates as of June 30, 2019?"
  expected_response: "The aggregate market value of voting
stock held by non-affiliates was approximately $786.3
billion as of June 30, 2019."

- user_input: "What percentage of Amazon's annual revenue
was recognized in the fourth quarter of 2019?"
  expected_response: "Approximately 31% of Amazon's annual
revenue was recognized during the fourth quarter of 2019."

- user_input: "What are some of the electronic devices that
Amazon manufactures and sells, as mentioned in the 2019
10-K?"
  expected_response: "Amazon manufactures and sells devices
such as the Kindle, Fire tablet, Fire TV, Echo, and Ring."

- user_input: "What is Amazon's investor relations website
as noted in the 2019 10-K?"
  expected_response: "Amazon's investor relations website
is amazon.com/ir."
```