

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Використання агентів великих мовних моделей
в бізнесі»

на здобуття освітнього ступеня бакалавра
зі спеціальності 122 Комп'ютерні науки

(код, найменування спеціальності)

освітньо-професійної програми Комп'ютерні науки
(назва)

*Кваліфікаційна робота містить результати власних досліджень.
Використання ідей, результатів і текстів інших авторів мають
посилання
на відповідне джерело*

(підпис)

Гліб ХАРИТОНОВ
(Ім'я, ПРІЗВИЩЕ здобувача)

Виконав:
здобувач вищої освіти
група КНД-42

Гліб ХАРИТОНОВ

Керівник: Щербина Ірина Сергіївна
науковий ступінь,
вчене звання

К. Т. Н

Рецензент:
науковий ступінь,
вчене звання

(Ім'я, ПРІЗВИЩЕ)

Київ 2024

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**
Навчально-науковий інститут інформаційних технологій

Кафедра Комп'ютерних наук

Ступінь вищої освіти Бакалавр

Спеціальність Комп'ютерні науки

Освітньо-професійна програма Комп'ютерні науки

ЗАТВЕРДЖУЮ

Завідувач кафедру Комп'ютерних наук

_____ Віктор ВИШНІВСЬКИЙ
« _____ » _____ 2024 р.

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ**

_____ Харитонов Гліб Євгенович

(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: Використання агентів великих мовних моделей в бізнесі.

керівник кваліфікаційної роботи Ірина ЩЕРБИНА, к. т. н., доцент.

(Ім'я, ПРІЗВИЩЕ науковий ступінь, вчене звання)

затверджені наказом Державного університету
інформаційно-комунікаційних технологій від «27» 02.2024р. №36

2. Строк подання кваліфікаційної роботи «31» травня 2024р.

3. Вихідні дані до кваліфікаційної роботи: науково-технічна література.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які
потрібно розробити)

Вступ до агентів великих мовних моделей (LLM)

Розробка та конфігурація агентів LLM

Впровадження агентів LLM в бізнес та виробництво

5. Перелік графічного матеріалу: *презентація*
1. Сфери використання LLM агентів
 2. Алгоритм розробки LLM агента
 3. Компоненти архітектури агента
 4. API для впровадження агентів LLM в бізнес-середовище
 5. Можливості застосування LLM-агентів в аналізі даних
6. Дата видачі завдання «23» лютого 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз наявної науково-технічної літератури	23.02-27.02.24	Виконано
2	Дослідження великих мовних моделей	28.02-07.03.24	Виконано
3	Дослідження LLM-агентів	08.03-12.03.24	Виконано
4	Дослідження ключових ресурсів та технології розробки агентів LLM	13.03-28.03.24	Виконано
5	Дослідження процесу впровадження LLM-агентів у бізнесі	28.03-02.04.24	Виконано
6	Визначення бізнес-потреб і можливостей для розробки агентів LLM	02.04-13.04.24	Виконано
7	Оформлення роботи: вступ, висновки, реферат	14.04-29.04.24	Виконано
8	Розробка демонстраційних матеріалів	29.04-14.05.24	Виконано

Здобувач вищої освіти

(підпис)

Гліб ХАРИТОНОВ

(Ім'я, ПРІЗВИЩЕ)

Керівник
кваліфікаційної роботи

(підпис)

Ірина ЩЕРБИНА

(Ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня бакалавра: 81 стор., 15 рис., 30 джерел.

Актуальність роботи. Агенти на основі великих мовних моделей (LLM), що знаходяться в авангарді цієї технологічної революції, пропонують безпрецедентні можливості в обробці та створенні людського тексту, що робить їх безцінними в різних секторах.

Мета роботи - дослідження застосування LLM-агентів у різних бізнес-середовищах, оцінка їхнього впливу та окреслення стратегічних підходів до їхньої інтеграції.

Основним завданням цієї роботи є детальний аналіз LLM-агентів, від їхніх базових технологій до практичного застосування в конкретних галузевих сценаріях. Це передбачає ретельне вивчення того, як ці агенти використовуються в даний час, які переваги вони пропонують і які виклики вони створюють.

Об'єктом цього дослідження є широкий спектр застосувань агентів LLM у бізнесі, а предметом - механізми, за допомогою яких ці агенти впливають на операційну ефективність, обслуговування клієнтів та загальну бізнес-аналітику.

Методи дослідження та досягнення розуміння будуть використані різноманітні методи дослідження, включаючи якісний аналіз за допомогою тематичних досліджень та експертних інтерв'ю, а також кількісні методи для вимірювання ефективності та впливу LLM-агентів у реальних умовах.

КЛЮЧОВІ СЛОВА: LLM, АГЕНТ, БІЗНЕС, ВПРОВАДЖЕННЯ, ВИКОРИСТАННЯ

ABSTRACT

The text part of the qualification work for obtaining a bachelor's degree: 77 pages, 15 pictures, 30 sources.

Relevance of work. Agents based on large language models (LLM), at the forefront of this technological revolution, offer unprecedented capabilities in the processing and creation of human text, making them invaluable in various sectors.

The purpose of the work is to study the application of LLM-agents in various business environments, assess their impact and outline strategic approaches to their integration.

The main objective of this work is a detailed analysis of LLM agents, from their basic technologies to practical applications in specific industry scenarios. This involves a careful study of how these agents are currently used, what advantages they offer, and what challenges they pose.

The focus of this research is the wide range of applications of LLM agents in business, and the mechanisms by which these agents affect operational efficiency, customer service and general business intelligence.

Research Methods and Gaining Understanding A variety of research methods will be used, including qualitative analysis through case studies and expert interviews, as well as quantitative methods to measure the effectiveness and impact of LLM agents in real-world settings.

KEY WORDS: LLM, AGENT, BUSINESS, IMPLEMENTATION, USE

ЗМІСТ

ВСТУП.....	9
1 ВСТУП ДО АГЕНТІВ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ (LLM).....	11
1.1 Дослідження великих мовних моделей.....	11
1.2 Дослідження LLM-агентів.....	13
1.3 Важливість LLM-агентів у сучасних технологіях.....	15
1.4 Застосування агентів LLM у різних галузях.....	18
1.5. Завдання.....	21
Висновки.....	25
2 РОЗРОБКА ТА КОНФІГУРАЦІЯ АГЕНТІВ LLM.....	26
2.1 Дослідження ключових ресурсів та технології розробки агентів LLM.....	26
2.2 Розробка базового агента LLM.....	30
2.3 Навчання агентів LLM.....	33
2.4 Тестування та оцінка агентів LLM.....	40
2.5 Розширені конфігурації та налаштування.....	46
Висновки.....	52
3 ВПРОВАДЖЕННЯ АГЕНТІВ LLM В БІЗНЕС ТА ВИРОБНИЦТВО..	53
3.1 Визначення бізнес-потреб і можливостей для розробки агентів LLM.	53
3.2 Дослідження процесу впровадження LLM-агентів у бізнесі.....	56
3.3 Використання агентів LLM для аналізу даних та прийняття рішень...	59
3.4 Порівняльний аналіз впроваджених результатів.....	63
3.5 Майбутні тенденції та потенційні розробки.....	66
ВИСНОВКИ.....	71
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	73

ВСТУП

Штучний інтелект (AI) має численні переваги, включаючи здатність швидко й точно обробляти великі обсяги даних. Він може автоматизувати рутинні завдання, знижуючи трудові витрати та підвищуючи ефективність роботи. AI також виявляє приховані закономірності в даних, допомагаючи приймати більш обґрунтовані рішення в різних галузях, таких як бізнес і медицина. Великі мовні моделі (LLM), на кшталт GPT-4, спеціалізуються на обробці природної мови. Вони можуть розуміти та генерувати текст, що робить їх корисними для автоматичного перекладу, створення контенту, підтримки клієнтів та навчання. LLM здатні адаптуватися до різних контекстів, що дозволяє їм ефективно відповідати на широкий спектр запитань і завдань.

LLM агенти поєднують можливості AI та LLM для виконання складніших завдань. Вони можуть діяти як інтелектуальні помічники, виконуючи багатокрокові завдання, що потребують контекстуального розуміння. Це дозволяє автоматизувати планування зустрічей, управління проектами чи проведення досліджень.

Перевага LLM агентів полягає в їх здатності інтегруватися з іншими системами та сервісами, що розширює їх функціональність. Вони можуть збирати дані з різних джерел, аналізувати їх та надавати користувачам узагальнену інформацію або рекомендації, що сприяє більш ефективному управлінню інформацією та прийняттю рішень.

LLM агенти можуть працювати в режимі реального часу, надаючи користувачам миттєві відповіді та підтримку. Це особливо корисно у сферах, де швидкість реакції є критичною, наприклад, у службах підтримки або медичних додатках. Крім того, AI, LLM та LLM агенти можуть бути налаштовані під специфічні потреби користувача або організації, що робить їх дуже гнучкими інструментами. Вони знаходять застосування у різних галузях, від фінансів і маркетингу до освіти і розваг. Завдяки постійному розвитку технологій, можливості AI та LLM агентів постійно розширюються, відкриваючи нові горизонти для їх використання. Це робить їх важливим інструментом у сучасному

світі, який допомагає вирішувати складні проблеми та підвищувати ефективність різних процесів.

1 ВСТУП ДО АГЕНТІВ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ (LLM)

1.1 Дослідження великих мовних моделей

Великі мовні моделі (Large Language Model - LLM) - це клас моделей штучного інтелекту, які були навчені на величезних обсягах текстових даних генерувати людський текст, перекладати з однієї мови на іншу, відповідати на запитання та виконувати різні інші завдання з обробки природної мови (NLP). Ці моделі базуються на архітектурах глибокого навчання, таких як трансформатори, які дозволяють їм фіксувати та використовувати складні закономірності та взаємозв'язки в навчальних даних. Розвиток LLM був зумовлений наявністю величезних текстових корпусів, розвитком обчислювальних ресурсів та інноваціями в архітектурі нейронних мереж. Архітектура трансформатора мала особливий вплив на створення LLM. Трансформатор спирається на механізм самоуваги, який дозволяє моделі зважувати важливість різних частин вхідної послідовності при генерації кожного вихідного токена. Це досягається шляхом обчислення балів уваги $a_{\{ij\}}$ між кожною парою tokenів i та j у вхідній послідовності:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^n \exp(e_{ik})}, \quad (1.1)$$

де e_{ij} оцінка сумісності між токенами i та j та n довжина послідовності. Ці оцінки уваги потім використовуються для обчислення зваженої суми представлень вхідних лексем, яка формує контекст для генерації наступної лексеми. Однією з ключових характеристик LLM є їхня здатність генерувати зв'язний і контекстуально релевантний текст. Навчаючись на основі різноманітних текстових

джерел, LLM можуть створювати граматично правильний, семантично значущий і стилістично відповідний до певного контексту текст. Ця здатність призвела до розвитку таких додатків, як мовний переклад, реферування текстів і діалогові системи. Ще однією важливою особливістю магістрів є їхня здатність до навчання з кількох спроб або навіть з нуля. Це означає, що LLM можуть адаптуватися до нових завдань з мінімальними додатковими навчальними даними або взагалі без них, використовуючи знання та навички, набуті під час попередньої підготовки. Наприклад, LLM, навчений на великому корпусі англійських текстів, може бути налаштований на невеликий набір даних для виконання аналізу настроїв або розпізнавання іменованих об'єктів без необхідності тривалого навчання для виконання конкретних завдань[6]. Один з варіантів архітектури LLM моделі представлений на рисунку 1.1.

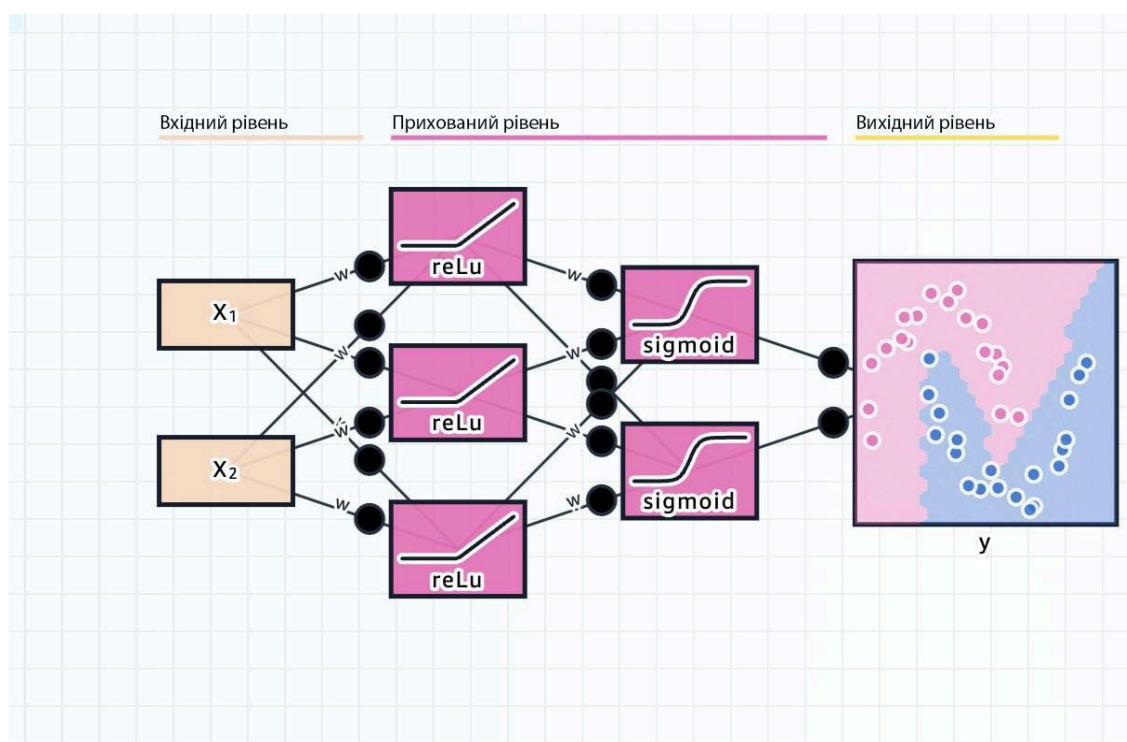


Рис 1.1 - Архітектура LLM моделі

Однак, LLM також мають кілька обмежень, які слід враховувати. Однією з головних проблем є потенційна можливість отримання упереджених результатів, оскільки LLM можуть успадковувати та посилювати упередження, наявні в їхніх

навчальних даних. Це може призвести до створення тексту, який увічніює стереотипи або дискримінаційну мову. Дослідники активно вивчають методи пом'якшення цих упереджень, такі як фільтрація даних, регуляризація моделі та методи постобробки. Ще одним обмеженням LLM є відсутність у них можливостей явного представлення знань і міркувань. Хоча LLM можуть генерувати текст, який виглядає обізнаним і зв'язним, вони не володіють справжнім розумінням концепцій і взаємозв'язків у тексті. Це може призвести до непослідовності, фактичних помилок і труднощів у виконанні складних міркувань. Крім того, обчислювальні ресурси, необхідні для навчання і впровадження LLM, можуть бути значними, що може обмежити їхню доступність і застосовність у певних контекстах. Споживання енергії та вплив на навколишнє середовище, пов'язані з навчанням великих моделей, також викликають занепокоєння, що спонукає до розробки більш ефективних і стійких підходів [4].

Отже, великі мовні моделі - це значний прогрес в обробці природної мови, що дозволяє генерувати текст, схожий на людський, і виконувати різні завдання NLP. Їх здатність навчатися на величезних обсягах даних і адаптуватися до нових завдань відкрила широкий спектр застосувань. Однак обмеження ШНМ, такі як потенційні упередження і відсутність явного представлення знань, повинні бути ретельно розглянуті і вирішені, щоб забезпечити їх відповідальне і ефективне використання в реальних сценаріях.

1.2 Дослідження LLM-агентів

Агенти LLM - це системи AI, які використовують можливості LLM для виконання складних завдань і взаємодії з користувачами природною мовою. Ці агенти поєднують потужність LLM з додатковими компонентами і методами, щоб забезпечити цілеспрямовану поведінку, розуміння контексту і прийняття рішень. Архітектура агентів LLM, як правило, складається з трьох основних компонентів. LLM слугує основним компонентом агента, забезпечуючи можливість генерувати текст, подібний до людського, на основі підказки введення та контексту. Вибір

LLM залежить від конкретних вимог агента, таких як бажаний рівень вільного володіння мовою, знання предметної області та багатомовна підтримка. Популярні LLM, що використовуються в агентних системах, включають. Для адаптації LLM до конкретної задачі або предметної області використовується модуль, орієнтований на конкретну задачу. Цей модуль може бути реалізований з використанням різних підходів, таких як точне налаштування, швидке проектування або поєднання обох. Точне налаштування передбачає навчання МНМ на меншому наборі даних, що відповідає цільовому завданню, що дозволяє моделі набути специфічних для предметної області знань і навичок. З іншого боку, інженерія підказок фокусується на розробці ефективних підказок, які спрямовують МНМ на отримання бажаного результату без додаткового навчання. Вибір між точним налаштуванням і швидким інжинірингом залежить від таких факторів, як наявність специфічних для завдання даних, складність завдання і необхідні обчислювальні ресурси. Система управління діалогом відповідає за взаємодію між користувачем і LLM-агентом. Вона обробляє вхідні дані користувача, підтримує контекст розмови і генерує відповідні відповіді на основі поточного стану діалогу. Система управління діалогом може бути змодельована як частково спостережуваний марковський процес прийняття рішень, де дії агента визначаються політикою π яка відображає поточний стан переконань b_t на дію a_t :

$$\pi(b_t) = \arg \arg Q(b_t, a_t), \quad (1.2)$$

де $Q(b_t, a_t)$ очікувана кумулятивна винагорода за виконання дії a_t у стані переконання b_t . Стан переконання представляє розуміння агентом контексту розмови та намірів користувача, яке оновлюється на основі введених користувачем даних та попередніх дій агента. Інтеграція LLM, спеціального модуля та системи управління діалогами дозволяє агентам LLM брати участь у цілеспрямованих діалогах і допомагати користувачам у різних сферах. Наприклад, LLM-агент,

призначений для підтримки клієнтів, може розуміти запити користувачів, отримувати доступ до відповідної інформації з бази знань і надавати точні та корисні відповіді. Аналогічно, LLM-агент для персональної допомоги може керувати розкладом, давати рекомендації та виконувати завдання на основі вподобань користувача та контексту.

Однією з ключових проблем при розробці LLM-агентів є забезпечення узгодженості та послідовності згенерованих відповідей. LLM, хоч і здатні створювати вільний текст, іноді можуть генерувати непослідовну або нерелевантну інформацію. Щоб пом'якшити цю проблему, можна використовувати такі методи, як фільтрація відповідей, оцінювання послідовності та навчання в умовах змагальності. Ці методи допомагають виявити і відфільтрувати неякісні або непослідовні відповіді, підвищуючи загальну надійність агента. Важливим моментом у розробці агента LLM є врахування етичних принципів і заходів безпеки. Оскільки LLM навчаються на великих і різноманітних наборах даних, вони можуть ненавмисно засвоїти і відтворити упередження або генерувати неприйнятний контент. Щоб вирішити ці проблеми, LLM-агенти повинні бути оснащені механізмами фільтрації контенту, етичними настановами та заходами безпеки для запобігання шкідливим або образливим результатам[9].

На закінчення, агенти LLM представляють собою потужне застосування великих мовних моделей у створенні інтелектуальних, інтерактивних систем. Поєднуючи LLM з модулями для конкретних завдань і системами управління діалогами, ці агенти можуть брати участь у взаємодії природною мовою, допомагати користувачам у різних сферах і виконувати складні завдання. Однак розробка LLM-агентів також пов'язана з проблемами, пов'язаними з узгодженістю, послідовністю та етичними міркуваннями, які необхідно вирішувати шляхом ретельного проектування та впровадження.

1.3 Важливість LLM-агентів у сучасних технологіях

Поява агентів LLM стала важливою віхою в галузі штучного інтелекту, що має далекосяжні наслідки для різних технологій і галузей промисловості. Ці

агенти, що працюють на основі найсучасніших мовних моделей і передових архітектур, мають потенціал революціонізувати способи взаємодії людини з машинами та доступу до інформації. Важливість LLM-агентів у сучасних технологіях можна пояснити їхньою здатністю розуміти і генерувати людський текст, виконувати складні завдання та адаптуватися до різних сфер. Однією з головних причин зростаючого значення LLM-агентів є їхній потенціал для покращення взаємодії між людиною і машиною. Традиційні інтерфейси, такі як графічні інтерфейси користувача (GUI) та інтерфейси командного рядка (CLI), часто вимагають від користувачів вивчення конкретних команд або навігації по складних меню. На відміну від них, агенти LLM дозволяють користувачам спілкуватися з машинами за допомогою природної мови, що робить взаємодію більш інтуїтивно зрозумілою і доступною. Розуміючи наміри користувача і надаючи контекстно-релевантні відповіді, LLM-агенти можуть подолати розрив між людським і машинним спілкуванням, що призводить до більш ефективного і зручного для користувача досвіду. Важливим аспектом LLM-агентів є їхня здатність виконувати широкий спектр завдань у різних сферах. На відміну від вузькоспеціалізованих систем AI, призначених для виконання конкретних завдань, LLM-агенти можуть бути адаптовані до різних застосувань шляхом точного налаштування або швидкого проектування. Така універсальність робить їх цінними інструментами в таких галузях, як охорона здоров'я, фінанси, освіта та обслуговування клієнтів. Наприклад, LLM-агент може допомагати лікарям діагностувати захворювання, аналізуючи дані про пацієнта і надаючи рекомендації, що ґрунтуються на фактах. У фінансовому секторі LLM-агенти можуть допомагати інвесторам приймати обґрунтовані рішення, аналізуючи ринкові тенденції, новини та фінансові звіти[9].

Адаптивність агентів LLM ще більше підвищується завдяки їхнім можливостям навчання з кількох пострілів. Навчання з кількох спроб - це здатність моделі навчатися новим завданням на обмеженій кількості прикладів. LLM-агенти, особливо ті, що базуються на трансформаторних архітектурах, таких як GPT-3, демонструють чудові показники навчання з кількох спроб. Це означає,

що вони можуть бути швидко налаштовані для конкретних завдань або доменів з мінімальними навчальними даними, що скорочує час і ресурси, необхідні для розробки. Здатність LLM-агентів до навчання з кількох спроб можна формалізувати наступним чином:

$$p(F_b x, F_c D) = \frac{p(F_c D)}{p(F_c D)}, \quad (1.3)$$

де x вхідні дані, y вихідні дані, а D множина прикладів з кількома пострілами. Модель навчається генерувати вихідні дані y в залежності від вхідних даних x та прикладами з кількома пострілами D . Важливість LLM-агентів виходить за рамки їхніх технічних можливостей, оскільки вони також мають потенціал для стимулювання інновацій та створення нових можливостей. Автоматизуючи завдання, які раніше вимагали людського досвіду, LLM-агенти можуть вивільнити людські ресурси, щоб зосередитися на більш високому рівні, творчій та стратегічній діяльності. Це може призвести до підвищення продуктивності, економії коштів та розробки нових продуктів і послуг. Крім того, LLM-агенти можуть допомогти демократизувати доступ до знань і досвіду, надаючи інтелектуальну допомогу користувачам незалежно від їхнього досвіду та рівня кваліфікації[4]. Однак розгортання LLM-агентів у реальних додатках також пов'язане з певними проблемами, які необхідно вирішити. Однією з головних проблем є потенційна можливість отримання упереджених або оманливих результатів. Навчаючись на великих наборах даних, ШНМ можуть успадковувати і посилювати упередження, присутні в навчальних даних. Це може призвести до дискримінаційних або образливих відповідей, які можуть мати серйозні наслідки в таких чутливих сферах, як охорона здоров'я або юридичні послуги. Щоб зменшити ці ризики, важливо розробити методи виявлення упередженості, пом'якшення її наслідків і забезпечення справедливості та прозорості результатів роботи агентів LLM.

Через складну природу моделей глибокого навчання може бути важко зрозуміти, як LLM-агент приходить до певної відповіді або рішення. Така

непрозорість може перешкоджати довірі та підзвітності, особливо в додатках з високими ставками. Проводяться дослідження з метою розробки методів інтерпретації та пояснення поведінки LLM-агентів, таких як візуалізація уваги, зондування та генерування зрозумілих для людини пояснень. Агенти LLM є значним досягненням у сучасних технологіях, що має потенціал для трансформації різних галузей і покращення взаємодії між людиною і машиною. Їх здатність розуміти і генерувати людський текст, виконувати складні завдання і адаптуватися до різних областей робить їх цінними інструментами для автоматизації, підтримки прийняття рішень і поширення знань. Однак розробка і розгортання LLM-агентів також створює проблеми, пов'язані з упередженістю, справедливістю та інтерпретованістю, які необхідно вирішувати за допомогою постійних досліджень і відповідальних практик. Оскільки агенти LLM продовжують розвиватися і вдосконалюватися, вони готові відігравати все більш важливу роль у формуванні майбутнього технологій і суспільства.

1.4 Застосування агентів LLM у різних галузях

Універсальність та адаптивність агентів великих мовних моделей (LLM) призвела до їх широкого впровадження в різних галузях, революціонізуючи спосіб роботи бізнесу та взаємодії з клієнтами. Від охорони здоров'я та фінансів до освіти та обслуговування клієнтів, агенти LLM розгортаються для автоматизації завдань, надання інтелектуальної допомоги та покращення процесів прийняття рішень. У цьому розділі досліджується поточний ландшафт використання LLM-агентів у різних сферах бізнесу, аналізуються специфічні для кожної галузі випадки використання, переваги та виклики.

У галузі охорони здоров'я LLM-агенти використовуються для покращення догляду за пацієнтами, оптимізації клінічних робочих процесів та підтримки медичних досліджень. Одним з відомих застосувань є розробка віртуальних медичних асистентів, які можуть вести розмови природною мовою з пацієнтами, надаючи їм персоналізовану інформацію про стан здоров'я, оцінку симптомів та

рекомендації щодо сортування. Ці асистенти також можуть допомогти пацієнтам орієнтуватися в складних системах охорони здоров'я, планувати зустрічі та контролювати хід лікування. Агенти LLM можуть допомагати медичним працівникам, створюючи клінічні нотатки, узагальнюючи записи пацієнтів і витягуючи відповідну інформацію з медичної літератури. У медичних дослідженнях агенти LLM можуть прискорити відкриття ліків, аналізуючи величезні обсяги біомедичних даних, генеруючи гіпотези та визначаючи потенційних кандидатів у ліки[6].

Фінансова індустрія також використовує LLM-агентів для покращення клієнтського досвіду, автоматизації фінансового аналізу та виявлення шахрайських дій. Чат-боти, керовані LLM-агентами, можуть надавати персоналізовані фінансові поради, відповідати на запити клієнтів і направляти користувачів у складних фінансових процесах, таких як подача заявок на кредити або прийняття інвестиційних рішень. В алгоритмічній торгівлі агенти LLM можуть аналізувати ринкові дані в режимі реального часу, настрої в новинах і стрічки соціальних мереж, щоб приймати торгові рішення на основі даних. Здатність LLM-агентів обробляти та розуміти неструктуровані дані, такі як фінансові звіти та новини, дозволяє їм отримувати цінну інформацію та формувати інвестиційні стратегії. Крім того, агенти LLM можуть допомогти виявити та запобігти фінансовому шахрайству, виявляючи закономірності та аномалії в даних про транзакції.

В освітньому секторі агенти LLM трансформують способи навчання студентів та викладання викладачів. Інтелектуальні системи навчання, що працюють на базі агентів LLM, можуть забезпечити персоналізований досвід навчання, адаптуючись до індивідуальних потреб і стилів навчання студентів. Ці системи можуть залучати студентів до інтерактивних діалогів, відповідати на запитання, надавати пояснення та пропонувати цілеспрямований зворотній зв'язок для покращення результатів навчання. Агенти LLM також можуть допомагати викладачам у створенні навчального контенту, формуванні запитань для вікторин, автоматизованому оцінюванні та наданні зворотного зв'язку щодо студентських

завдань. У сфері досліджень та академічного письма агенти LLM можуть допомогти науковцям знайти відповідну літературу, узагальнити статті і навіть створити початкові проекти наукових робіт. Обслуговування клієнтів - ще одна сфера, де LLM-агенти роблять значний вплив. Розгортаючи розмовних агентів AI, компанії можуть надавати підтримку клієнтам у режимі 24/7, обробляти великі обсяги запитів і покращувати час реагування на них. LLM-агенти можуть розуміти наміри клієнта, надавати точні та контекстно-релевантні відповіді, а також переадресовувати складні питання людям, коли це необхідно. Використання LLM-агентів в обслуговуванні клієнтів може призвести до підвищення рівня задоволеності клієнтів, зниження операційних витрат і підвищення ефективності. Крім того, LLM-агенти можуть збирати цінну інформацію про клієнтів та їхні відгуки, допомагаючи компаніям визначати сфери для вдосконалення та приймати рішення на основі даних[5].

Застосування LLM-агентів у різних галузях не позбавлене викликів. Однією з головних проблем є можливість отримання упереджених або оманливих результатів, що може мати серйозні наслідки в таких чутливих сферах, як охорона здоров'я або фінанси. Щоб зменшити ці ризики, важливо розробити методи виявлення та пом'якшення упередженості, такі, наприклад, як змагальний дебайсинг (Zhang та ін., 2018) або доповнення контрфактичних даних (Kaushik та ін., 2020). Ці методи спрямовані на зменшення впливу упереджених патернів у навчальних даних і забезпечення справедливості та надійності результатів роботи LLM-агентів. Іншим викликом є інтерпретованість та пояснюваність роботи агентів LLM. У таких галузях, як охорона здоров'я та фінанси, де рішення можуть мати значні наслідки, дуже важливо розуміти, як LLM-агент приходить до певної рекомендації або рішення. Такі методи, як візуалізація уваги (Vaswani та ін., 2017) і зондування (Conneau та ін., 2018), можуть дати уявлення про внутрішню роботу LLM-агентів, допомагаючи визначити найбільш релевантні вхідні характеристики і підвищити прозорість. Крім того, генерування пояснень, зрозумілих для людини, поряд з результатами моделі може підвищити довіру та підзвітність[8]. Конфіденційність і безпека даних також є важливими факторами при розгортанні

агентів LLM у різних галузях. Оскільки агенти LLM обробляють і генерують великі обсяги конфіденційних даних, важливо впроваджувати надійні заходи захисту даних, такі як шифрування, контроль доступу та безпечне зберігання. Дотримання галузевих норм, таких як HIPAA в охороні здоров'я або GDPR в Європейському Союзі, має вирішальне значення для забезпечення відповідального та етичного використання агентів LLM. На рисунку 1.2 зображено сфери використання LLM агентів



Рисунок 1.2. - Сфери використання LLM агентів

На закінчення, застосування агентів LLM у різних галузях має потенціал для стимулювання інновацій, підвищення ефективності та покращення досвіду клієнтів. Від охорони здоров'я та фінансів до освіти та обслуговування клієнтів, LLM-агенти розгортаються для автоматизації завдань, надання інтелектуальної допомоги та підтримки процесів прийняття рішень. Однак успішне впровадження агентів LLM вимагає вирішення проблем, пов'язаних з упередженістю, інтерпретованістю і конфіденційністю даних. Очікується, що з розвитком досліджень і передових практик, впровадження агентів LLM в різних галузях буде зростати, трансформуючи спосіб роботи бізнесу та взаємодії з клієнтами.

1.5. Завдання

Основне завдання бакалаврської роботи - дослідити застосування агентів великих мовних моделей (LLM) у різних сферах бізнесу, зосередившись на їхньому потенціалі для трансформації галузей, вдосконалення процесів прийняття

рішень та покращення взаємодії з клієнтами. Дослідження має на меті надати всебічний огляд поточного стану використання агентів LLM, проаналізувавши специфічні для кожної галузі випадки використання, переваги та виклики. Вивчаючи найсучасніші методи та найкращі практики розробки та розгортання агентів LLM, цей диплом має на меті зробити внесок у зростаючий обсяг знань у галузі штучного інтелекту та його практичного застосування в бізнесі. Для досягнення цих цілей диплом буде зосереджений на декількох ключових областях дослідження. По-перше, він забезпечить ґрунтовний вступ до магістерських програм, обговорюючи їхню архітектуру, навчальні процеси та можливості. Цей фундамент буде необхідним для розуміння ролі LLM у створенні інтелектуальних агентів та їхнього потенціалу для вирішення складних бізнес-проблем. Потім учасники заглиблюються в концепцію агентів LLM, вивчаючи їх компоненти, функціональні можливості та механізми взаємодії. У цьому розділі також розглядаються різні підходи до розробки LLM-агентів, такі як точне налаштування, швидке проектування та навчання з кількох пострілів[7].

Значна частина диплому буде присвячена дослідженню поточного ландшафту використання LLM-агентів у різних галузях. Це включатиме аналіз конкретних галузевих випадків використання, висвітлення переваг та викликів, пов'язаних із впровадженням LLM-агентів у кожній галузі. Охоплені галузі включатимуть, але не обмежуватимуться ними, охорону здоров'я, фінанси, освіту та обслуговування клієнтів. Надаючи всебічний огляд застосувань агентів LLM, цей диплом має на меті запропонувати цінну інформацію для компаній, які розглядають можливість впровадження цих технологій, а також для дослідників, зацікавлених у вивченні нових шляхів для інновацій. На додаток до галузевого аналізу, диплом також досліджує ширший вплив агентів LLM на бізнес-процеси та прийняття рішень. Це передбачає вивчення того, як агенти LLM можуть бути інтегровані в існуючі робочі процеси, такі як системи управління взаємовідносинами з клієнтами (CRM) або програмне забезпечення для планування ресурсів підприємства (ERP). Диплом також досліджує потенціал агентів LLM для підтримки прийняття рішень на основі даних шляхом обробки та

аналізу великих обсягів неструктурованих даних, таких як відгуки клієнтів, пости в соціальних мережах або фінансові звіти. Також буде обговорено інтеграцію LLM-агентів з іншими методами AI, такими як алгоритми машинного навчання для предиктивної аналітики[7].

Щоб забезпечити відповідальну та етичну розробку та розгортання LLM-агентів, у дипломі будуть розглянуті виклики та міркування, пов'язані з цими технологіями. Сюди входить вивчення методів виявлення та пом'якшення упередженості, таких як змагальний дебайсинг (Zhang et al., 2018) або доповнення контрфактичних даних (Kaushik et al., 2020). Диплом також вивчатиме методи підвищення інтерпретованості та пояснюваності агентів LLM, такі як візуалізація уваги (Vaswani та ін., 2017) або зондування (Conneau та ін., 2018). Ці методи мають на меті забезпечити розуміння процесів прийняття рішень агентами LLM, покращуючи прозорість та підзвітність[2]. Також будуть розглянуті питання конфіденційності та безпеки даних, підкреслюючи важливість впровадження надійних заходів захисту даних та дотримання галузевих нормативно-правових актів. У дипломі будуть розглянуті кращі практики безпечного зберігання даних, контролю доступу та шифрування при розгортанні агентів LLM у чутливих сферах, таких як охорона здоров'я та фінанси. Хоча диплом має на меті надати всебічний огляд LLM-агентів у бізнесі, деякі аспекти будуть виходити за його межі. Диплом не зосереджуватиметься на технічних деталях архітектури LLM або алгоритмах навчання, оскільки ці теми широко висвітлені в літературі. Натомість, у ньому буде зроблено акцент на практичному застосуванні та значенні LLM-агентів у бізнес-контексті. Крім того, диплом не буде вичерпно охоплювати всі галузі або випадки використання, а скоріше зосередиться на репрезентативних прикладах, які демонструють потенціал та виклики агентів LLM у різних сферах.

Насамкінець, цей диплом має на меті дослідити застосування LLM-агентів у різних сферах бізнесу, надаючи всебічний огляд галузевих випадків використання, переваг та викликів. Вивчаючи поточний ландшафт використання агентів LLM, досліджуючи інтеграцію цих технологій у бізнес-процеси та розглядаючи етичні та практичні міркування, пов'язані з їх розробкою та розгортанням, це

дослідження має на меті внести свій внесок у зростаючий обсяг знань у галузі штучного інтелекту та його трансформаційний потенціал у світі бізнесу.

Висновки

В розділі досліджено сутність та сфери застосування агентів великих мовних моделей (LLM) та їх зростаючої важливості в сучасних технологіях та бізнес-додатках. Огляд LLM висвітлив їх ключові характеристики, можливості та обмеження, заклавши основу для розуміння ролі цих моделей у створенні інтелектуальних агентів. У вступі до LLM-агентів було розглянуто їхню архітектуру, компоненти та функціональні можливості, а також різні підходи до їхньої розробки, такі як точне налаштування, швидке проектування та навчання з кількох пострілів.

2 РОЗРОБКА ТА КОНФІГУРАЦІЯ АГЕНТІВ LLM

2.1 Дослідження ключових ресурсів та технології розробки агентів LLM

Розробка агентів LLM потребує надійної системи ресурсів, що включає програмне забезпечення, апаратне забезпечення, набори даних та інші необхідні інструменти. Основою розробки агентів LLM є передові середовища та інструменти програмування. Ключовим серед них є використання фреймворків машинного навчання, таких як TensorFlow, PyTorch та бібліотека Transformers від Hugging Face. Ці фреймворки пропонують комплексну підтримку для побудови та навчання нейронних мереж, включаючи готові моделі, які можна адаптувати під конкретні завдання. Крім того, програмне забезпечення для обробки природної мови, таке як spaCy або NLTK, має вирішальне значення для попередньої обробки текстових даних і реалізації лінгвістичних функцій, необхідних для агентів LLM[4].

Обчислювальні вимоги до навчання LLM є значними, що вимагає потужного апаратного забезпечення. Графічні процесори (GPU) або тензорні процесори (TPU) мають вирішальне значення завдяки їхній здатності виконувати паралельні обчислення, що прискорює процес навчання. Наприклад, для навчання LLM на кшталт GPT-3, який має мільярди параметрів, потрібні кластери високопродуктивних GPU або TPU для ефективної обробки великих обчислень. Апаратне налаштування часто поширюється на середовища високопродуктивних обчислень (HPC), особливо для навчання в масштабі. Ефективність агента LLM значною мірою залежить від якості та обсягу набору даних, що використовується для навчання. Набори даних повинні бути масштабними, різноманітними та репрезентативними для мови та завдань, які виконуватиме агент LLM. Поширеними наборами даних є WebText, Common Crawl і дампи Вікіпедії, які надають широкий спектр загальних знань і використання мови. Для специфічних застосунків необхідні спеціальні набори даних, щоб точно налаштувати

LLM-агентів на спеціалізовані знання, що вимагає ретельного збору та курації даних.

Окрім основних компонентів програмного забезпечення, обладнання та наборів даних, розробка агентів LLM також вимагає систем контролю версій, таких як Git для управління кодом, а також Docker або Kubernetes для контейнеризації та оркестрування навчальних процесів. Ці інструменти підтримують масштабованість і відтворюваність, які є важливими для спільної розробки і розгортання в різних середовищах. Інтеграція та оптимізація цих ресурсів інкапсульована рівнянням для обчислювальних витрат на підготовку LLM, яке має вигляд

$$[C = n \times d \times s \times e], \quad (2.1)$$

де C представляє обчислювальні витрати, n - кількість циклів навчання, d - розмірність моделі (наприклад, кількість шарів, розмір кожного шару), s - розмір набору даних, і e - ефективність використовуваного апаратного забезпечення.

Після того, як агент LLM розроблений, інтеграція в існуючі системи має вирішальне значення. Це передбачає використання API (інтерфейсів прикладного програмування) і SDK (комплектів розробки програмного забезпечення), які сприяють безперешкодній інтеграції агентів LLM з бізнес-додатками, системами управління взаємовідносинами з клієнтами та іншими цифровими платформами. Ці інструменти повинні підтримувати надійний, безпечний зв'язок між LLM-агентом і хост-додатками, щоб забезпечити ефективне функціонування агента в його операційному середовищі.

Інструменти моніторингу та обслуговування: Після розгортання безперервний моніторинг і обслуговування є важливими для забезпечення максимальної ефективності і точності роботи агента LLM. Такі інструменти, як Prometheus для моніторингу і Grafana для візуалізації, є життєво важливими для відстеження показників ефективності агентів LLM. Ці інструменти допомагають виявити будь-яке погіршення продуктивності або потенційні проблеми, які

можуть вплинути на користувацький досвід або точність відповідей агента. Враховуючи широку взаємодію з потенційно чутливими даними, впровадження суворих заходів безпеки не підлягає обговоренню. Протоколи шифрування, стандарти безпечної передачі даних і дотримання правил захисту даних (таких як GDPR або HIPAA, залежно від географічних і галузевих вимог) є обов'язковими для захисту даних, що обробляються агентами LLM. Крім того, для захисту від кіберзагроз, що постійно розвиваються, необхідні регулярні аудити та оновлення системи безпеки.

Інструменти проектування користувацького досвіду: Ефективність LLM-агента визначається не лише його технічними можливостями, але й тим, наскільки добре він може взаємодіяти з користувачами. Інструменти для розробки інтерактивних і зручних інтерфейсів, такі як Adobe XD або Sketch, відіграють вирішальну роль у створенні привабливого досвіду. Ці інструменти допомагають у створенні прототипів і тестуванні різних моделей взаємодії користувача з агентом LLM, гарантуючи, що кінцевий продукт буде не тільки функціональним, але й інтуїтивно зрозумілим для кінцевих користувачів[4].

Середовища навчання та розробки: Окрім початкового налаштування, для постійного вдосконалення LLM-агентів вирішальне значення має забезпечення постійного навчання і середовища розробки. Це включає доступ до оновлених наборів даних, перепідготовку фреймворків для включення нових даних, а також тестові середовища, де можна безпечно тестувати оновлення, не впливаючи на реальні системи. Хмарні рішення, такі як AWS, Google Cloud або Microsoft Azure, пропонують масштабовані середовища, які можна використовувати для постійного навчання та розвитку. Цілісний підхід до розподілу ресурсів та управління ними гарантує, що агенти LLM не лише створюються з використанням найсучасніших технологій, але й є стійкими та ефективними у довгостроковому розгортанні. Ця комплексна ресурсна стратегія підтримує життєвий цикл LLM-агентів від концепції до розробки, інтеграції та постійного вдосконалення, гарантуючи, що ці агенти залишаються передовими інструментами, здатними сприяти значному прогресу в різних галузях[2].

Отже, розробка агентів LLM є ресурсномістким процесом, який вимагає ретельно організованого поєднання програмного забезпечення, апаратних засобів і ресурсів даних. Кожен компонент відіграє ключову роль у навчанні та функціонуванні LLM-агентів, впливаючи на їхню здатність розуміти та ефективно генерувати текст, подібний до людського. Системний підхід до визначення цих вимог гарантує, що процес розробки є не лише здійсненим, але й оптимізованим для створення високопродуктивних LLM-агентів, здатних трансформувати різноманітні додатки в бізнес- та технологічному середовищі. На рисунку 2.1 зображена схема ресурсів розробки агентів LLM



Рисунок 2.1 - схема ресурсів розробки агентів LLM

2.2 Розробка базового агента LLM

Створення базового агента великої мовної моделі (LLM) передбачає системний підхід, який інтегрує різні компоненти штучного інтелекту, зокрема, в області обробки природної мови. Цей розділ містить детальний, покроковий опис процесу, гарантуючи, що кожна частина сприяє всебічному розумінню побудови функціонального агента LLM.

- Крок 1: Визначення мети та сфери застосування Початковим кроком у створенні LLM-агента є чітке визначення його мети та сфери застосування. Це передбачає визначення конкретних завдань, які виконуватиме агент, таких як відповіді на запити клієнтів, надання рекомендацій або автоматизація певних типів електронних листів. Визначення сфери застосування включає визначення мови та домену, в якому працюватиме агент, що безпосередньо впливає на дані, необхідні для навчання. Модель послідовної взаємодії з LLM агентом представлено на рисунку 2.2

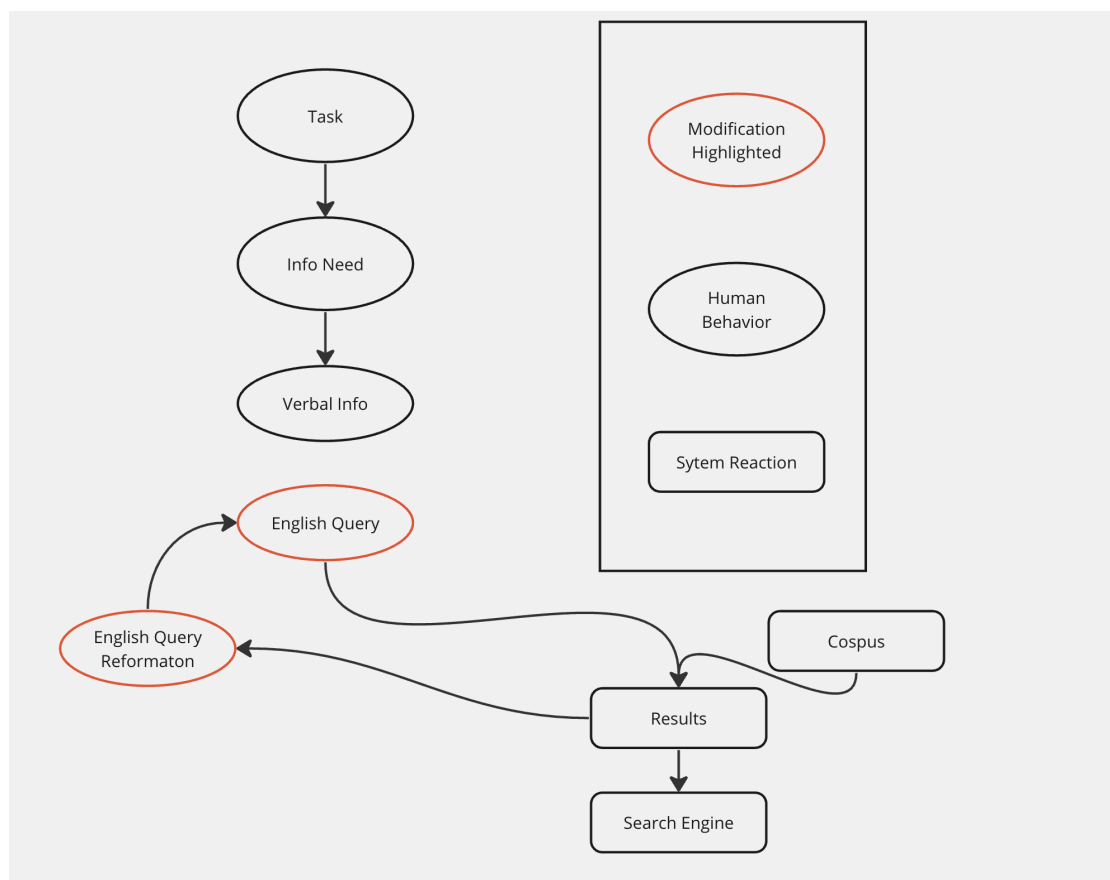


Рис 2.2 Послідовна взаємодія з LLM агентом

- Крок 2: Вибір правильної моделі та фреймворку Вибір правильної моделі має вирішальне значення. Для базового LLM-агента зазвичай використовують такі моделі, як GPT (Generative Pre-trained Transformer) або BERT (Bidirectional Encoder Representations from Transformers), завдяки їхній надійності та універсальності у вирішенні різноманітних мовних завдань. Вибір фреймворку, такого як TensorFlow або PyTorch, залежить від конкретних вимог проекту, включаючи сумісність з іншими інструментами, простоту використання та підтримку спільноти[3].

- Крок 3: Збір та попередня обробка даних Після вибору моделі та фреймворку наступним кроком є збір даних. Дані мають бути релевантними до предметної області та завдань, визначених на Кроці 1. Це передбачає збір великого масиву текстових даних, на яких модель буде навчатися. Щоб зробити ці дані придатними для навчання, необхідно попередньо обробити їх; це включає такі завдання, як токенизація, видалення стоп-слів і нормалізація тексту.

- Крок 4: Навчання моделі Навчання моделі - це обчислювально інтенсивний процес, який вимагає налаштування відповідного апаратного середовища, зазвичай з використанням графічних або паралельних процесорів для швидшої обробки. Процес навчання передбачає введення попередньо оброблених даних у модель та налаштування таких параметрів, як швидкість навчання, розмір партії та кількість епох. Навчання керується функцією втрат, яка у випадку LLM часто є перехресною ентропійною втратою, що визначається як

$$\left[L = - \sum_{i=1}^N y_i \log \log \left(\hat{y}_i \right) \right], \quad (2.2)$$

де (y_i) - істинна мітка, $(\hat{y}_i)$ - передбачена мітка, а (N) - кількість вибірок.

- Крок 5: Оцінювання та доопрацювання Після початкового навчання необхідно оцінити роботу моделі, використовуючи окремий валідаційний набір даних, який не використовувався на етапі навчання. Для вимірювання

продуктивності використовуються такі метрики, як точність, точність, пригадування та оцінка F1. На основі оцінки модель може потребувати доопрацювання, що передбачає коригування параметрів моделі або її подальше навчання на більш цілеспрямованому наборі даних для підвищення її точності та ефективності[3].

- Крок 6: Інтеграція та розгортання Навчену модель інтегрують у прикладне середовище, де працюватиме агент LLM. Це передбачає налаштування API для взаємодії моделі з іншими додатками та користувацькими інтерфейсами. Розгортання також включає впровадження інструментів моніторингу для відстеження продуктивності LLM-агента в режимі реального часу і забезпечення його коректної роботи.

- Крок 7: Безперервне навчання та оновлення Для підтримки актуальності та ефективності агента LLM важливо впровадити механізми безперервного навчання. Цього можна досягти, періодично перенавчаючи модель на нових даних або використовуючи такі методи, як онлайн-навчання, коли модель поступово навчається на нових даних, коли вони стають доступними.

- Крок 8: Відгуки користувачів та ітеративне вдосконалення Після розгортання збір відгуків користувачів має вирішальне значення для ітеративного вдосконалення LLM-агента. Зворотній зв'язок може дати уявлення про те, наскільки добре агент відповідає очікуванням користувачів, і висвітлити області, де він може бути недопрацьованим. Цей зворотний зв'язок слід систематично збирати та аналізувати, щоб інформувати про подальший розвиток. Коригування можуть включати уточнення відповідей моделі, розширення її бази знань або покращення можливостей взаємодії.

- Крок 9: Етичні міркування та пом'якшення упередженості Оскільки агенти ЗНМ навчаються на великих наборах даних, які можуть містити притаманні їм упередження, важливо включити кроки для виявлення та пом'якшення цих упереджень. Такі методи, як аудит упередженості, використання збалансованих наборів даних та впровадження алгоритмів справедливості, є критично важливими. Крім того, етичні міркування повинні керувати

розгортанням агентів LLM, гарантуючи, що вони поважають конфіденційність користувачів, їхню згоду та забезпечують прозорість у своїй діяльності[3].

- Крок 10: Масштабованість і обслуговування Нарешті, забезпечення масштабованості та постійного обслуговування агента LLM є життєво важливим для його довгострокового успіху. Масштабованість передбачає планування збільшення навантаження та розширення можливостей агента без погіршення продуктивності. Обслуговування включає регулярне оновлення базових моделей, бібліотек залежностей і протоколів безпеки для захисту від вразливостей.

Отже, створення базового агента LLM - це складний, але структурований процес, який включає визначення мети агента, вибір відповідних моделей та інструментів, підготовку даних, навчання моделі та розгортання агента у відповідному середовищі. Кожен крок має вирішальне значення і повинен бути виконаний з точністю, щоб забезпечити функціональність і ефективність агента LLM. Такий системний підхід не тільки полегшує розробку базового LLM-агента, але й закладає основу для більш просунутих адаптацій і функціональних можливостей. Такий комплексний підхід до створення базового агента LLM, від початкової концепції до постійного вдосконалення та етичного управління, гарантує, що агент не тільки ефективно виконує свої завдання, але й адаптується до вимог, що змінюються, та підтримує відповідність етичним стандартам. Кожен крок у цьому процесі взаємопов'язаний, утворюючи цілісний робочий процес, який підкреслює важливість точності, орієнтованого на користувача дизайну та етичних міркувань при розробці рішень, керованих AI. Дотримуючись цієї структурованої методології, розробники можуть створювати LLM-агентів, які є не лише функціональними та ефективними, але й відповідальними та адаптивними до мінливої динаміки потреб користувачів і технологічного прогресу.

2.3 Навчання агентів LLM

Ефективне навчання агентів великих мовних моделей (LLM) на даних конкретної предметної області включає в себе ряд структурованих кроків і

методологій, які гарантують, що агент не тільки добре розуміє загальну мову, але й вміє виконувати конкретні завдання, що стосуються певної області. У цьому розділі описано процес і найкращі практики підготовки LLM-агентів, з акцентом на таких методах, як точне налаштування, оперативне проектування та навчання з підкріпленням. Діаграму процесу обробки даних агентами представлено на рисунку 2.3

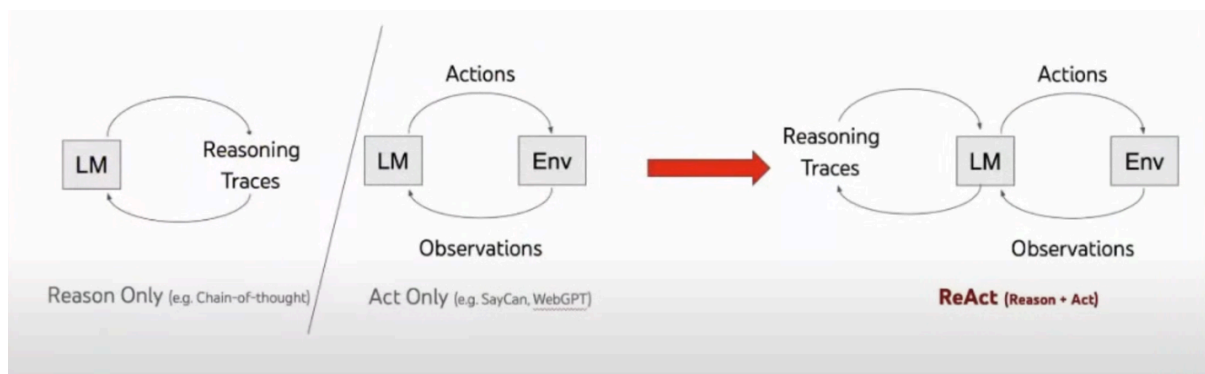


Рис 2.3 Діаграма процесу обробки даних агентами

Точне налаштування на даних, специфічних для домену Точне налаштування є критично важливим методом адаптації попередньо навчених LLM до конкретних доменів або завдань. Цей процес передбачає продовження навчання попередньо навченої моделі на новому наборі даних, який є меншим за обсягом, але дуже релевантним для потрібної задачі або предметної області. Ключ до ефективного тонкого налаштування полягає в ретельному відборі та підготовці цього набору даних для конкретної предметної області. Дані повинні бути репрезентативними для фактичних сценаріїв використання, з якими агент LLM зіткнеться під час розгортання. Процес точного налаштування коригує ваги попередньо навченої моделі так, щоб вона стала більш пристосованою до нюансів нових даних. Зазвичай це робиться шляхом ініціалізації моделі з попередньо навченими вагами, а потім її подальшого навчання з меншою швидкістю. Мета полягає в тому, щоб уточнити параметри моделі без перезапису цінних загальних можливостей, які вона вже вивчила. Математичне представлення функції втрат, що використовується при точному налаштуванні, можна виразити так:

$$\left[L(\theta) = - \sum_{(x,y) \in D} \log \log p(x; \theta) \right], \quad (2.3)$$

де L - функція втрат, (θ) - параметри моделі, D - набір даних для конкретної предметної області, x - вхідні дані, y - вихідні дані, і p - ймовірність y для x при параметрах моделі (θ) . На рисунку 2.4 зображено адаптацію LLM агента до конкретних завдань.

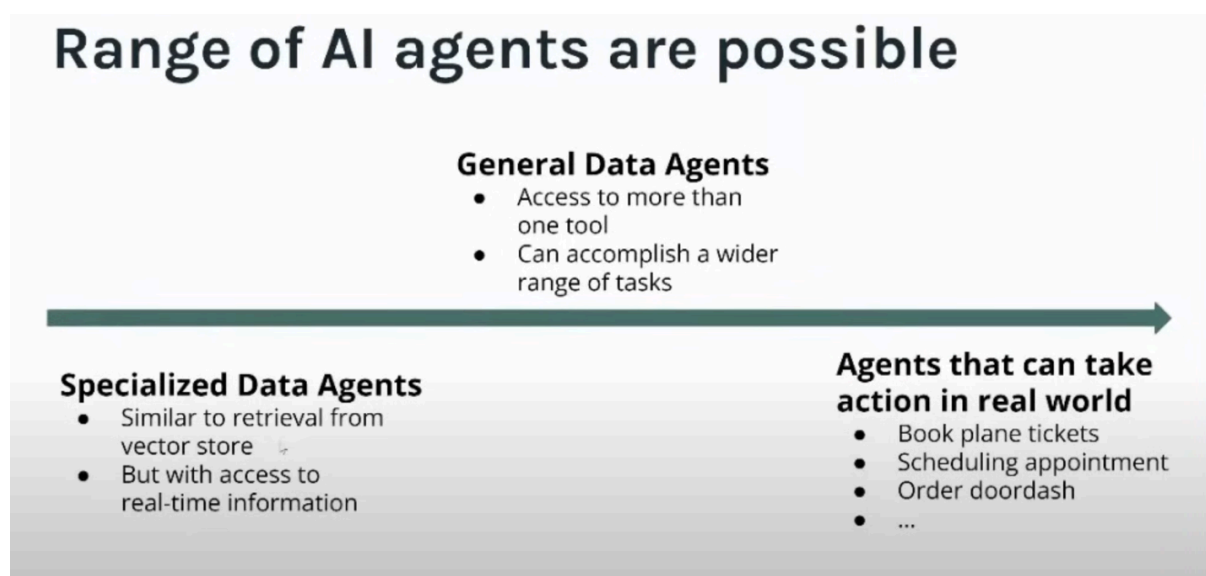


Рис 2.4 Адаптація LLM агента до конкретних завдань

Експрес-інжиніринг - це ще один метод, який використовується для адаптації LLM до конкретних завдань без необхідності значної перепідготовки. Цей підхід передбачає створення підказок, які спрямовують модель на отримання бажаного результату. Ефективний інжиніринг підказок вимагає глибокого розуміння можливостей і обмежень моделі, а також творчого підходу до розробки підказок, які можуть викликати правильні відповіді[8]. Підказки можуть бути простими, наприклад, пряме запитання, або більш складними, що включають сценарій або контекст, який спонукає модель генерувати певний тип відповідей. Успіх інженерії підказок значною мірою залежить від здатності імітувати контекст,

в якому будуть використовуватися результати роботи моделі, таким чином узгоджуючи реакції моделі з реальними вимогами. Навчання з підкріпленням (RL) вводить ще один рівень складності в навчання агентів LLM, дозволяючи їм вчитися на основі взаємодії з навколишнім середовищем. У контексті LLM середовище, як правило, включає симуляцію, де модель отримує зворотний зв'язок на основі її результатів. Модель вчиться виконувати завдання, максимізуючи сигнал винагороди, а не просто мінімізуючи функцію втрат[2].

Процес навчання RL можна описати наступним правилом оновлення, яке коригує параметри моделі для максимізації очікуваної винагороди:

$$\left[\theta_{t+1} = \theta_t + \alpha \nabla_{\theta E} [\theta_t]\right], \quad (2.4)$$

де (θ_t) - параметри в момент часу (t) , (α) - швидкість навчання, R - винагорода, і $(\nabla_{\theta E} [\theta_t])$ - градієнт очікуваної винагороди щодо параметрів. Щоб забезпечити ефективність цих методів навчання, слід дотримуватися кількох найкращих практик:

- **Якість та релевантність даних:** Переконайтеся, що навчальні дані мають високу якість, не містять помилок і тісно пов'язані з конкретними завданнями, які виконуватиме LLM-агент.
- **Регулярне оцінювання:** Постійно оцінюйте модель під час навчання, щоб відстежувати її продуктивність як на загальних, так і на специфічних для домену завданнях. Це допомагає виявити будь-які проблеми на ранніх етапах навчання.
- **Збалансоване навчання:** Під час точного налаштування слід дотримуватися балансу між засвоєнням нових знань про предметну область і збереженням загальних мовних навичок.
- **Етичні міркування:** Пам'ятайте про етичні міркування, особливо щодо упередженості даних і результатів моделювання, щоб переконатися, що LLM-агент працює чесно і відповідально.

Насамкінець, навчання агентів LLM на даних конкретної предметної області передбачає поєднання точного налаштування, швидкого проектування та навчання з підкріпленням, що підтримується найкращими практиками, які гарантують, що моделі будуть ефективними та етично узгодженими з потребами користувачів. Ці методи, при розумному застосуванні, дозволяють розробляти LLM-агентів, які є не лише технічно досконалими, але й практично цінними в конкретних галузях[5]. Ітеративне вдосконалення та цикли зворотного зв'язку Важливим аспектом підготовки LLM-агентів є впровадження ітеративного вдосконалення та циклів зворотного зв'язку. Цей процес передбачає безперервне вдосконалення моделі на основі зворотного зв'язку від її роботи в реальних сценаріях або змодельованих середовищах. Ітеративне вдосконалення допомагає точно налаштувати реакції моделі, щоб вони більше відповідали бажаним результатам, підвищуючи точність і релевантність. Зворотний зв'язок може бути отриманий від взаємодії з користувачами, коли відповіді агента LLM оцінюються кінцевими користувачами або експертами предметної області. Цей зворотний зв'язок має вирішальне значення для виявлення областей, в яких модель може працювати не так, як очікувалося, що дозволяє здійснювати цілеспрямовані коригування. Крім того, автоматизовані системи можуть бути використані для моделювання взаємодій і генерування зворотного зв'язку на основі заздалегідь визначених критеріїв, що сприяє швидким ітераціям[9]. На рисунку 2.5 зображено структуру запиту RAG.

From RAG to Agents



Рис 2.5 Структура запису RAG

Оптимізація гіперпараметрів Оптимізація гіперпараметрів відіграє важливу роль у процесі навчання агентів LLM. Гіперпараметри, такі як швидкість навчання, розмір партії та кількість епох навчання, суттєво впливають на динаміку навчання моделі та загальну продуктивність. Оптимізація цих параметрів може призвести до підвищення ефективності та результативності моделі. Для пошуку оптимального набору гіперпараметрів зазвичай використовують такі методи, як сітковий пошук, випадковий пошук або більш просунуті методи, такі як байєсівська оптимізація. Мета полягає в тому, щоб максимізувати продуктивність моделі на валідаційному наборі, який діє як проксі для того, наскільки добре модель буде працювати в реальних задачах.

$$[E(x, y) \in D_{val}[\log \log p(x; \theta(\lambda))]], \quad (2.5)$$

де (λ) представляє гіперпараметри, (D_{val}) є набором даних для перевірки, а $(\theta(\lambda))$ позначає параметри моделі, які залежать від гіперпараметрів. Хоча навчання агентів ШНМ на даних, специфічних для певної галузі, має вирішальне значення, не менш важливо підтримувати адаптивність у різних галузях, особливо для агентів, які, як очікується, виконуватимуть кілька типів

завдань. Такі методи, як багатозадачне навчання та адаптація до домену, можуть бути використані для підвищення гнучкості моделі та розширення її застосовності. Багатозадачне навчання передбачає одночасне навчання моделі декільком пов'язаним завданням. Цей підхід допомагає моделі засвоїти більш загальне представлення, яке може бути ефективним для різних завдань. З іншого боку, методи адаптації до предметної області зосереджені на налаштуванні моделі, навченої в одній області, щоб вона добре працювала в іншій, часто з обмеженими додатковими специфічними для цієї області даними або взагалі без них. Надійність та узагальненість моделі. Забезпечення надійності та узагальненості агентів LLM є життєво важливим для їх успішного розгортання. Надійність - це здатність моделі обробляти варіації вхідних даних і приклади, що суперечать один одному, без значного погіршення продуктивності. Такі методи, як навчання в умовах конкуренції, коли модель навчається як на нормальних, так і на несприятливих даних, можуть підвищити робастність. Узагальнення, тим часом, вимірює, наскільки добре модель працює на нових, небачених даних. Методи регуляризації, такі як відсіювання та L2 регуляризація, є ефективними методами запобігання надмірному пристосуванню та сприяють узагальненню[5]. На рисунку 2.6 зображено діапазон дій мультимодального агента.

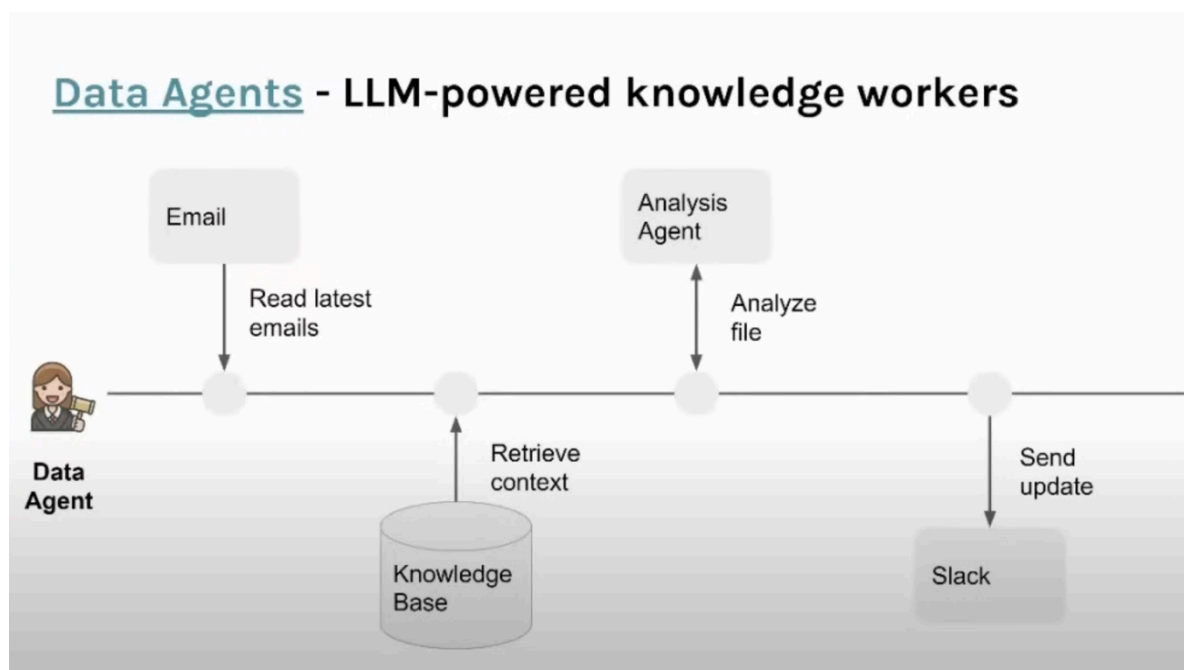


Рис 2.6 Діапазон дій мультимодального агента

Навчання LLM-агентів на даних конкретної предметної області передбачає комплексний підхід, який включає в себе точне налаштування, швидке проектування, навчання з підкріпленням і кілька найкращих практик. Кожен компонент - від ітеративного уточнення до оптимізації гіперпараметрів і міждоменної адаптивності - відіграє вирішальну роль у розробці всебічно розвиненого, ефективного LLM-агента. Дотримуючись цих структурованих методів і постійно оцінюючи та вдосконалюючи модель, розробники можуть створювати LLM-агентів, які є не лише технічно досконалими, але й універсальними та надійними, здатними ефективно працювати в різних доменах і завданнях. Такий цілісний підхід гарантує, що LLM-агенти можуть задовольнити складні вимоги реальних додатків, надаючи цінну допомогу і знання в різних галузях.

2.4 Тестування та оцінка агентів LLM

Тестування та оцінка агентів великих мовних моделей (LLM) є критично важливими етапами процесу розробки, які гарантують, що ці системи відповідають необхідним стандартам продуктивності та надійності перед розгортанням. У цьому розділі описано систематичні методи, що використовуються для тестування та оцінювання агентів LLM, з детальним описом метрик і еталонів, необхідних для оцінювання їхньої якості[4].

Модульне тестування: Передбачає тестування окремих компонентів LLM-агента, щоб переконатися, що кожна частина функціонує коректно в ізоляції. Модульні тести, як правило, автоматизовані і призначені для охоплення широкого діапазону вхідних умов, щоб перевірити наявність помилок у коді LLM-агента.

Інтеграційне тестування: Після модульного тестування, інтеграційне тестування оцінює, як різні компоненти агента LLM працюють разом. Цей крок має вирішальне значення, оскільки допомагає виявити проблеми, що виникають

при об'єднанні окремих частин, які не є очевидними, коли ці частини тестуються окремо.

Системне тестування: Це комплексний етап тестування, на якому агент LLM оцінюється як єдине ціле. Системне тестування включає тестування агента в змодельованих реальних умовах, щоб переконатися, що він відповідає заданим вимогам.

Тестування продуктивності: Для оцінки ефективності та масштабованості агента LLM проводяться тести продуктивності. Ці тести вимірюють час відгуку, використання ресурсів і пропускну здатність за різних умов навантаження.

Тестування безпеки: Враховуючи чутливість даних, які обробляються агентами LLM, тестування безпеки є необхідним. Воно включає в себе оцінку стійкості агента до загроз і вразливостей безпеки.

Тестування прийнятності для користувача (User Acceptance Testing, UAT): Нарешті, UAT передбачає тестування агента LLM реальними користувачами в контрольованому середовищі. UAT допомагає переконатися, що агент відповідає очікуванням користувачів і готовий до розгортання. Для кількісної оцінки продуктивності LLM-агентів використовується кілька метрик. Це найпростіший показник, який вимірює частку правильних відповідей, наданих LLM-агентом. Вона розраховується наступним чином:

$$\left[\text{Точність} = \frac{\text{Кількість правильних прогнозів}}{\text{Загальна кількість прогнозів}} \right], \quad (2.6)$$

Ці метрики особливо корисні у сценаріях, де баланс між хибнопозитивними та хибнонегативними результатами має вирішальне значення. Точність вимірює точність позитивних прогнозів, тоді як пригадування вимірює здатність знаходити всі відповідні екземпляри.

$$\left[\text{Точність} = \frac{\text{Правильні спрацювання}}{\text{Правильні спрацювання} + \text{хибні спрацювання}} \right], \quad (2.7)$$

$$\left[Recall = \frac{\text{Істинно позитивні}}{\text{Істинно позитивні} + \text{хибно негативні}} \right], \quad (2.8)$$

Оцінка F1 - це середнє гармонійне значення точності та пригадування, що забезпечує єдину метрику, яка збалансовує обидва показники.

$$\left[F1 = 2 \times \frac{\text{Точність} \times Recall}{Precision + Recall} \right], \quad (2.9)$$

Вимірює час, необхідний агенту LLM для повернення відповіді, що є критично важливим для додатків, які вимагають взаємодії в реальному часі. Ця якісна метрика оцінює, наскільки користувачі задоволені роботою LLM-агентів, зазвичай збирається за допомогою опитувань та форм зворотного зв'язку.

Бенчмаркінг передбачає порівняння ефективності роботи агента LLM зі встановленими стандартами або аналогічними системами. Це може включати галузеві стандарти або аналіз конкурентів, що дає уявлення про відносну ефективність роботи агента. Контрольні показники слід вибирати, виходячи з їх відповідності конкретній галузі та завданням, які має виконувати LLM-агент. Після початкового тестування і розгортання безперервний моніторинг стає необхідним для підтримки продуктивності і надійності агентів LLM з плином часу. Це передбачає налаштування систем для безперервного відстеження продуктивності агента і виявлення будь-яких відхилень від очікуваної поведінки або деградації з плином часу. Компоненти архітектури агента представлені на рисунку 2.7

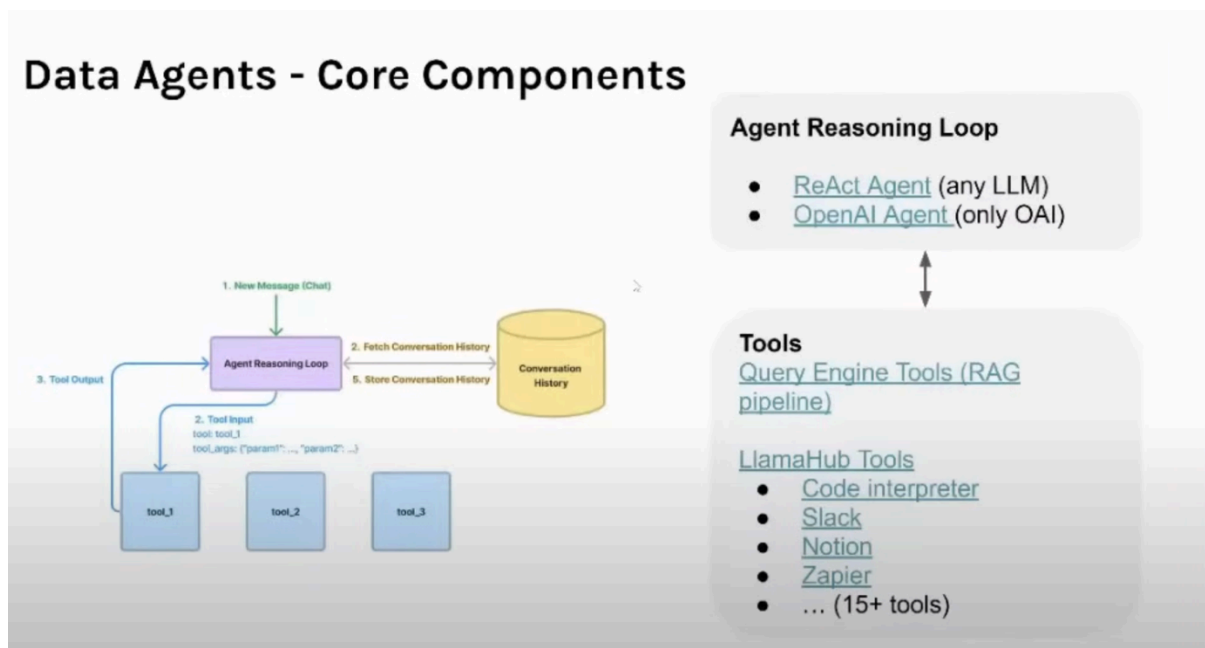


Рис 2.7 Компоненти архітектури агента

- **Моніторинг у режимі реального часу:** Впровадження інструментів моніторингу в реальному часі, які можуть забезпечити миттєвий зворотній зв'язок щодо показників ефективності агента. Це дозволяє негайно виявляти такі проблеми, як збільшення часу відгуку, зниження точності або збої в роботі системи.

- **Аналіз логів:** Регулярний аналіз журналів, що генеруються агентом LLM, може дати уявлення про його роботу і допомогти виявити закономірності, які можуть вказувати на основні проблеми. Автоматизовані інструменти можна використовувати для просіювання великих обсягів даних журналів, щоб виявити аномалії або тенденції, які потребують уваги.

- **А/В-тестування:** Періодично можуть розроблятися нові версії агента LLM для підвищення продуктивності або додавання нових функцій. А/В-тестування передбачає одночасний запуск старої та нової версій в однакових умовах для безпосереднього порівняння їхньої продуктивності. Цей метод допомагає переконатися, що оновлення агента призводять до реальних покращень перед повномасштабним розгортанням.

- Цикли зворотного зв'язку: Включення механізмів зворотного зв'язку з користувачами дозволяє безперервно вдосконалювати LLM-агента. Користувачі можуть надати цінну інформацію про продуктивність та зручність використання агента, яка може бути використана для вдосконалення та розширення його можливостей. На рисунку 2.8 представлено обробку та аналіз даних датасету.

How to retrieve & analyze data from knowledge base?

Use our [query engines](#) as "data tools" over your agent:

- Semantic search
- Summarization
- Text-to-SQL
- Document comparisons
- Combining Structured Data w/ Unstructured

"Simple" Interface - all agent has to infer is a query string!

```

query_engine_tools = [
  QueryEngineTool(
    query_engine=lyft_engine,
    metadata=ToolMetadata(
      name="lyft_10k",
      description="Provides information about Lyft financials for year 2021. "
      "Use a detailed plain text question as input to the tool."
    )
  ),
  QueryEngineTool(
    query_engine=uber_engine,
    metadata=ToolMetadata(
      name="uber_10k",
      description="Provides information about Uber financials for year 2021. "
      "Use a detailed plain text question as input to the tool."
    )
  )
]
  
```

Example Notebook:

- [OpenAI Agent + query engines \(as tools\)](#)
- [Analyzing structured + unstructured data](#)

Рис 2.8 Обробка та аналіз даних датасету

Для подальшого вдосконалення процесу оцінювання можна використовувати передові аналітичні методи, щоб отримати більш глибоке розуміння продуктивності та поведінки LLM-агентів: Застосування алгоритмів машинного навчання до даних про ефективність роботи агентів LLM може допомогти передбачити потенційні збої або виявити можливості для оптимізації. Наприклад, предиктивна аналітика може передбачити, коли продуктивність може погіршитися на основі історичних тенденцій, що дозволяє вжити превентивних заходів. Візуалізація складних даних за допомогою теплових карт або інших графічних зображень може допомогти зацікавленим сторонам швидко зрозуміти закономірності та аномалії продуктивності. Це особливо корисно для виявлення проблем юзабіліті або областей, де агент LLM не відповідає очікуванням користувачів. Для LLM-агентів, які беруть участь у взаємодії з клієнтами, аналіз

настроїв може бути цінним інструментом. Аналіз настроїв користувачів під час взаємодії може дати уявлення про задоволеність користувачів і виділити аспекти роботи агента, які можуть потребувати вдосконалення.

З огляду на потенційний вплив агентів LLM на конфіденційність, безпеку та справедливість, проведення регулярних етичних та комплаєнс-аудитів має вирішальне значення. Ці аудити повинні оцінювати Забезпечення дотримання агентами LLM відповідних законів і правил щодо захисту даних, таких як GDPR або HIPAA, залежно від географічного розташування та сфери застосування. Регулярне тестування агента LLM на упередженість його відповідей, особливо якщо він взаємодіє з різними групами користувачів. Повинні бути розроблені стратегії пом'якшення наслідків для усунення будь-яких виявлених упереджень. Оцінка наявних заходів безпеки, спрямованих на захист агента ПЗМ від зовнішніх загроз і забезпечення цілісності його операцій.

Отже, систематичне тестування та оцінка LLM-агентів передбачає поєднання автоматизованих і ручних методів тестування, а також набору кількісних і якісних показників. Ці методи гарантують, що агенти LLM є надійними, ефективними, безпечними і здатними задовольнити потреби кінцевих користувачів. Дотримуючись цих суворих протоколів тестування та оцінки, розробники можуть гарантувати, що їхні агенти LLM готові до успішного розгортання в реальних додатках. На завершення, тестування та оцінка агентів ПЗМ - це комплексний процес, який виходить за рамки перевірок перед розгортанням. Він охоплює постійний моніторинг, передову аналітику та регулярні аудити, щоб гарантувати, що агент залишається ефективним, дієвим та етичним протягом усього свого життєвого циклу. Підтримуючи суворі стандарти та адаптуючись до відгуків і нових викликів, розробники можуть гарантувати, що їхні LLM-агенти продовжуватимуть ефективно та відповідально служити своїм цілям.

2.5 Розширені конфігурації та налаштування

У сфері агентів великих мовних моделей (LLM) розширені конфігурації та налаштування відіграють ключову роль у підвищенні їхньої продуктивності та розширенні функціональності. У цьому розділі розглядаються складні методи та архітектури, які можна використовувати для оптимізації LLM-агентів, приділяючи особливу увагу генерації з розширеним пошуком та мультиагентним системам. Кожен з цих підходів проаналізовано, щоб зрозуміти їхній внесок у розвиток галузі та практичне значення. Приклад обробки фінансового звіту за допомогою RAG представлено на рисунку 2.9

Example: Financial Analysis with Agents + RAG

Question: "Compare and contrast Uber and Lyft's revenue growth"

- **Agent:** breaks down question into sub-questions over tools
- **Per-Document RAG:** Answer question over a given document via top-k retrieval.

Let's Try It Out!

```

agent.chat_repl()

===== Entering Chat REPL =====
Type "exit" to exit.

> == Calling Function ==
Calling Function: lyft_10k with args: {
  "input": "What was Lyft's revenue growth in 2021?"
}
Got output:
Lyft's revenue grew by 36% in 2021 compared to the prior year.
=====
> == Calling Function ==
Calling Function: uber_10k with args: {
  "input": "What was Uber's revenue growth in 2021?"
}
Got output:
Uber's revenue growth in 2021 was 57%.
=====
Assistant: In 2021, Lyft's revenue grew by 36% compared to the previous year, while Uber's revenue growth was higher at 57%. This indicates that Uber experienced a faster rate of revenue growth than Lyft in 2021.

```

Рис 2.9 Приклад обробки фінансового звіту за допомогою RAG

Генерація, доповнена пошуком (RAG) Генерація, доповнена пошуком, являє собою значний прогрес в архітектурі агентів LLM. Цей підхід поєднує генеративні можливості LLM з компонентами на основі пошуку, що дозволяє моделі динамічно отримувати доступ до зовнішніх баз знань під час процесу генерації. Інтеграція механізмів пошуку дозволяє LLM-агенту генерувати відповіді, які є не лише контекстуально релевантними, але й фактично точними і заснованими на

актуальних зовнішніх даних. Процес включає два основні компоненти: ретривер і генератор. Ретривер відповідає за запит до зовнішньої бази даних або бази знань для отримання релевантної інформації на основі вхідного запиту. Потім ця інформація передається генератору, який синтезує отримані дані зі своїми попередньо навченими знаннями, щоб сформулювати послідовну і контекстуально відповідну відповідь. Математичне формулювання процесу пошуку може бути представлене наступним чином:

$$\left[p(x) = \frac{\exp(\text{sim}(x,r))}{\sum_{r \in R} \exp(\text{sim}(x,r))} \right], \quad (2.10)$$

де x - вхідний запит, r - знайдений документ, R - множина всіх можливих знайдених документів, а $(\text{sim}(x,r))$ позначає функцію подібності, яка вимірює релевантність r до x .

Мультиагентні системи Мультиагентні системи (МАС) передбачають розгортання декількох агентів LLM, які взаємодіють і співпрацюють для виконання складних завдань. Така конфігурація особливо вигідна в сценаріях, коли завдання, що виконується, вимагає різноманітних знань або коли різні аспекти проблеми потрібно вирішувати одночасно. У МАС кожен агент, як правило, спеціалізується на певній підзадачі або предметній області, і вони координують свої зусилля для отримання єдиного результату.

Архітектура мультиагентної системи полегшує розподіл праці між агентами, підвищуючи ефективність і масштабованість. Крім того, вона забезпечує більш надійне реагування, оскільки кожен агент може обробляти винятки або аномалії у своїй галузі знань. Взаємодія між агентами в МАС може бути формалізована за допомогою принципів теорії ігор, де стратегія кожного агента полягає в

максимізації його функції корисності, що часто призводить до загальної оптимальної продуктивності системи. Функція корисності для агента (i) може бути виражена як

$$\left[U_i(s_i, s_{-i}) = \sum_{j \neq i} \alpha_{ij} \cdot f_j(s_i, s_{-i}) \right], \quad (1.11)$$

де (s_i) - стратегія, обрана агентом (i), (s_{-i}) - стратегії всіх інших агентів, (α_{ij}) - коефіцієнти, які відображають вплив стратегії агента (j) на корисність агента (i), а (f_j) - функція, яка кількісно оцінює цей вплив.

Для подальшого покращення продуктивності агентів LLM можна застосовувати різні методи оптимізації. До них належить налаштування гіперпараметрів, яке передбачає регулювання таких параметрів, як швидкість навчання або розміри шарів, для покращення ефективності навчання та якості вихідних даних моделі. Крім того, такі методи, як обрізання і квантування, можуть бути використані для зменшення розміру моделі і обчислювальних вимог, що робить LLM-агентів більш придатними для розгортання в середовищах з обмеженими ресурсами.

Окрім початкових конфігурацій та оптимізацій, LLM-агенти можуть бути вдосконалені за допомогою методів адаптивного навчання, які забезпечують постійне вдосконалення на основі реальних взаємодій та зворотного зв'язку. Адаптивне навчання передбачає динамічне оновлення параметрів моделі або навчальних даних у відповідь на нову інформацію або зміни в навколишньому середовищі. Такий підхід гарантує, що LLM-агенти залишатимуться актуальними та ефективними протягом тривалого часу, адаптуючись до нових тенденцій, вподобань користувачів та нових моделей даних, що з'являються. Одним з ефективних методів реалізації адаптивного навчання є онлайн-навчання, де модель поступово оновлюється без необхідності перенавчання з нуля. Це може

бути особливо корисно в середовищах, де дані швидко змінюються або де недоцільно часто проводити повні цикли перепідготовки. Правило оновлення в сценарії онлайн-навчання можна представити так:

$$\left[\theta_{t+1} = \theta_t + \eta \nabla_{\theta} \log \log p(x_t; \theta_t) \right], \quad (2.12)$$

де (θ_t) - параметри моделі у момент часу (t) , (x_t) і (y_t) - вхідні дані і цільовий вихід у момент часу (t) відповідно, (η) - швидкість навчання, а $(\nabla_{\theta} \log \log p(x_t; \theta_t))$ - градієнт логарифму ймовірності цільового виходу при заданих вхідних даних, з урахуванням параметрів моделі.

Налаштування є ще одним важливим аспектом розширених конфігурацій для агентів LLM. Інтегруючи відгуки користувачів безпосередньо в цикл навчання, LLM-агенти можуть бути налаштовані так, щоб краще відповідати індивідуальним або груповим уподобанням, підвищуючи задоволеність і залученість користувачів. Кастомізація на основі зворотного зв'язку не тільки допомагає точно налаштувати відповіді LLM-агентів, але й допомагає визначити сфери, в яких агент може не мати достатньої підготовки або здібностей. На рисунку 2.10 зображено обробку запиту з великим обсягом даних

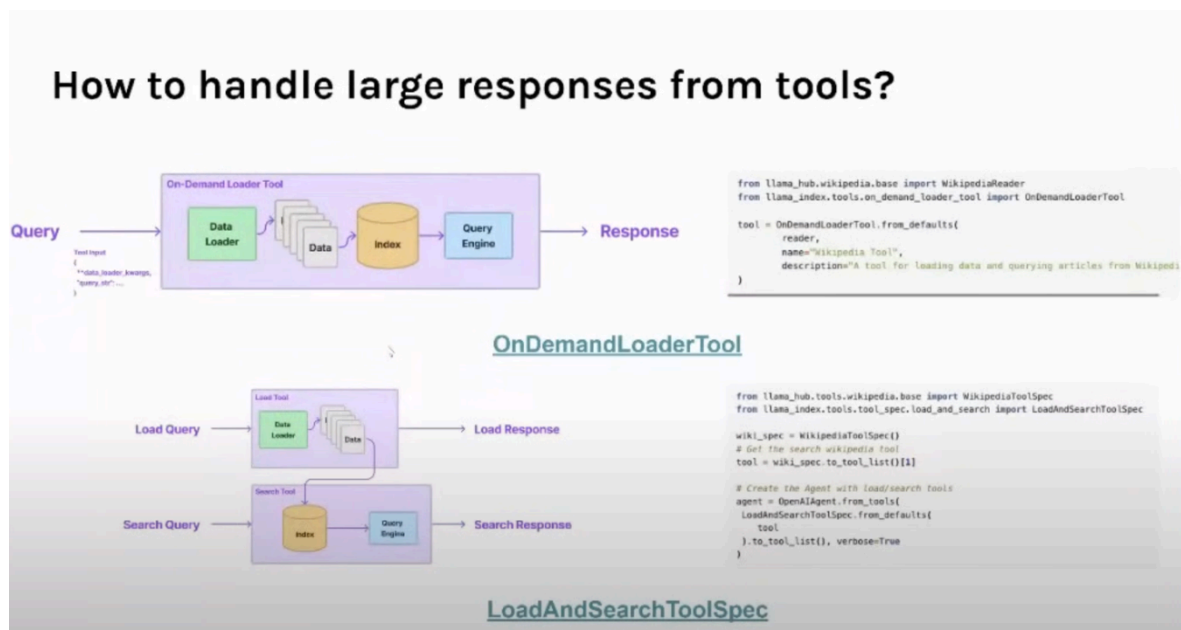


Рис 2.10 Обработка запроса с великим объемом данных

Щоб систематично враховувати відгуки користувачів, можна використовувати такі методи, як навчання з підкріпленням, де взаємодія з користувачем слугує винагородою або покаранням, що спрямовує процес навчання. Цей метод узгоджує роботу агента з очікуваннями користувачів, роблячи його більш чутливим до конкретних потреб і контекстів своїх користувачів. Модель навчання з підкріпленням можна описати за допомогою функції винагороди

$$[R(s, a) = \text{Задоволеність користувача}(s, a)], \quad (2.13)$$

де s - стан, що представляє контекст взаємодії, a - дія, виконана агентом LLM, і $(User\ Satisfaction(s, a))$ кількісно визначає задоволеність користувача діями агента в даному стані.

Оскільки агенти LLM стають складнішими і розгортаються в масштабі, підтримка ефективності та ефективне управління обчислювальними ресурсами набувають першорядного значення. Такі методи, як дистиляція моделей, коли менша, ефективніша модель навчається відтворювати поведінку більшої, складнішої моделі, можуть мати вирішальне значення. Цей підхід допомагає

розгортати агентів LLM на різних платформах, в тому числі з обмеженими обчислювальними можливостями, без значних втрат у продуктивності.

Розширені конфігурації та налаштування агентів LLM, такі як генерація з розширенням пошуку та мультиагентні системи, пропонують значні можливості для підвищення їхньої функціональності та ефективності. Використовуючи ці складні архітектури, LLM-агенти можуть бути адаптовані до конкретних потреб і виконувати складні завдання більш ефективно. Постійне вивчення та впровадження цих передових методів є вирішальним для еволюції агентів LLM, розширюючи межі можливого в додатках штучного інтелекту. Крім того, використання передових стратегій кешування і пошуку даних в таких архітектурах, як генерація з доповненим пошуком, може оптимізувати час відгуку і зменшити обчислювальні накладні витрати. Ефективне управління зовнішніми базами знань і стратегічне кешування інформації, до якої часто звертаються, гарантує високу продуктивність агентів LLM навіть в умовах високого попиту.

Висновки

Розширені конфігурації та налаштування агентів LLM передбачають поєднання складних архітектурних рішень, механізмів безперервного навчання та стратегій налаштування, орієнтованих на користувача. Використовуючи ці передові методи, розробники можуть створювати LLM-агентів, які є не лише потужними та ефективними, але й адаптивними та персоналізованими для задоволення мінливих потреб користувачів. Постійний розвиток і вдосконалення цих систем мають важливе значення для використання повного потенціалу технологій LLM у різних сферах застосування, стимулювання інновацій і вдосконалення в галузі штучного інтелекту.

3 ВПРОВАДЖЕННЯ АГЕНТІВ LLM В БІЗНЕС ТА ВИРОБНИЦТВО

3.1 Визначення бізнес-потреб і можливостей для розробки агентів LLM

Інтеграція агентів великих мовних моделей (LLM) у бізнес-операції вимагає системного підходу до визначення та пріоритизації високоцінних варіантів використання. У цьому розділі описано комплексну систему оцінки потреб і можливостей бізнесу, яка гарантує, що розгортання агентів LLM відповідає стратегічним цілям і приносить максимальну користь.

Першим кроком у визначенні бізнес-потреб є проведення ретельного аналізу існуючих бізнес-процесів. Це передбачає складання карти робочих процесів, визначення ключових видів діяльності та виявлення сфер, де існує неефективність або вузькі місця. Для візуалізації та аналізу цих процесів можна використовувати такі методи, як картування процесів і картування потоків створення цінності. Взаємодія із зацікавленими сторонами на різних рівнях організації має вирішальне значення для отримання уявлення про больові точки та сфери для вдосконалення. До зацікавлених сторін належать працівники, менеджери та клієнти, які взаємодіють з бізнес-процесами. Опитування, інтерв'ю та фокус-групи можна використовувати для збору якісних даних про їхній досвід та очікування[7].

Аналіз історичних даних може надати цінну інформацію про ефективність бізнесу та висвітлити сфери, де агенти LLM можуть додати цінності. Це передбачає вивчення ключових показників ефективності (KPI), відгуків клієнтів та операційних показників для виявлення тенденцій і закономірностей. Інструменти і методи аналізу даних, такі як статистичний аналіз і машинне навчання, можуть бути використані для виявлення прихованих можливостей.

Порівняння ефективності бізнесу з галузевими показниками і показниками конкурентів може допомогти визначити сфери, в яких агенти LLM можуть забезпечити конкурентну перевагу. Це передбачає аналіз використання конкурентами технологій штучного інтелекту та LLM, розуміння найкращих галузевих практик та виявлення прогалин у поточних можливостях бізнесу.

Визначення пріоритетів для високоцінних кейсів використання зображено на рисунку 3.1

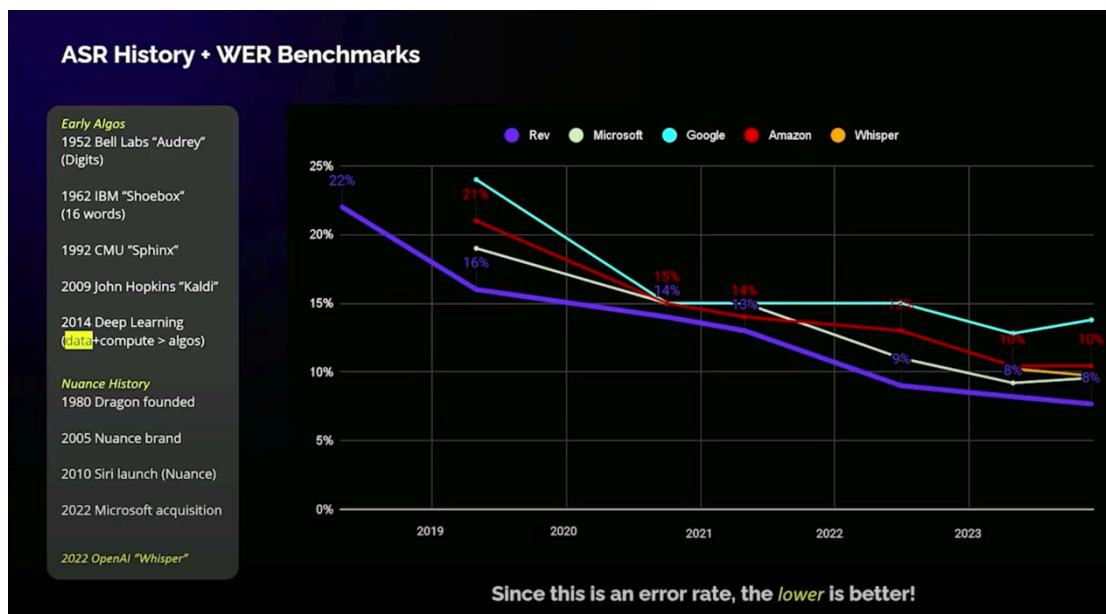


Рис 3.1. Визначення пріоритетів для високоцінних кейсів використання

Після визначення потреб бізнесу наступним кроком є визначення пріоритетності варіантів використання на основі їхньої потенційної цінності та реалістичності. Для оцінки та ранжування варіантів використання можна використовувати наступні критерії:

Оцінка потенційний вплив кожного варіанту використання на стратегічні цілі бізнесу. Це включає оцінку того, як варіант використання узгоджується з такими цілями, як збільшення доходу, зменшення витрат, підвищення задоволеності клієнтів або підвищення операційної ефективності. Вплив можна кількісно оцінити за допомогою таких показників, як рентабельність інвестицій (ROI) або чиста теперішня вартість (NPV).

$$\left[ROI = \frac{\text{Чистий прибуток}}{\text{Інвестиційні витрати}} \right], \quad (3.1)$$

Оцінка можливостей реалізації кожного варіанту використання, враховуючи такі фактори, як технічна складність, наявність ресурсів та організаційна

готовність. Це передбачає оцінку необхідної інфраструктури, доступності даних і навичок, необхідних для розробки і розгортання агента LLM[9].

Час, необхідний для створення цінності: Враховуйте час, необхідний для реалізації варіанту використання та реалізації його переваг. Пріоритетними мають бути ті варіанти використання, які можна швидко впровадити і які приносять негайну користь. Це можна оцінити за допомогою таких показників, як час виходу на ринок (TTM) або період окупності. Проведення оцінки ризиків, щоб визначити потенційні проблеми і невизначеності, пов'язані з кожним варіантом використання. Це включає оцінку технічних ризиків, таких як якість даних і продуктивність моделі, а також бізнес-ризиків, таких як дотримання нормативних вимог і сприйняття користувачами. Матриця ризиків може бути використана для класифікації та визначення пріоритетності ризиків на основі їхньої ймовірності та впливу. Щоб проілюструвати застосування цієї системи, розглянемо наступні приклади використання для агентів LLM у різних сферах бізнесу.

Обслуговування клієнтів: LLM-агентів можна використовувати для обробки запитів клієнтів, надання персоналізованої підтримки та ефективного вирішення проблем. Автоматизуючи рутинні завдання та надаючи миттєві відповіді, LLM-агенти можуть підвищити рівень задоволеності клієнтів та зменшити операційні витрати.

Продажі та маркетинг: LLM-агенти можуть допомогти у генерації лідів, сегментації клієнтів та проведенні персоналізованих маркетингових кампаній. Аналізуючи дані про клієнтів і створюючи цільовий контент, агенти LLM можуть підвищити ефективність маркетингу та стимулювати зростання продажів[12].

Управління персоналом: LLM-агенти можуть оптимізувати HR-процеси, такі як підбір персоналу, адаптація та залучення співробітників. Автоматизуючи адміністративні завдання та надаючи персоналізовану підтримку, LLM-агенти можуть підвищити ефективність роботи з персоналом та покращити досвід співробітників.

Фінанси та бухгалтерський облік: LLM-агенти можуть автоматизувати фінансову звітність, виявлення шахрайства та контроль за дотриманням

законодавства. Аналізуючи фінансові дані та виявляючи аномалії, LLM-агенти можуть підвищити точність і знизити ризик помилок і шахрайства.

Визначення та визначення пріоритетів високоцінних варіантів використання LLM-агентів у бізнесі вимагає системного підходу, який включає аналіз бізнес-процесів, консультації із зацікавленими сторонами, аналіз даних та конкурентний бенчмаркінг. Оцінюючи варіанти використання на основі їхнього впливу на бізнес-цілі, доцільності, часу, необхідного для створення цінності, та ризиків, компанії можуть гарантувати, що їхні інвестиції в LLM-агентів принесуть максимальну віддачу. Застосування цієї системи дозволяє компаніям стратегічно розгортати агентів LLM, стимулюючи інновації та підвищуючи ефективність у різних сферах діяльності.

3.2 Дослідження процесу впровадження LLM-агентів у бізнесі

Розгортання агентів LLM у бізнес-середовищі вимагає ретельного підходу для забезпечення їхньої ефективної інтеграції та роботи. У цьому розділі обговорюються ключові міркування та найкращі практики для впровадження агентів LLM, включаючи такі аспекти, як розробка API, масштабування, моніторинг та обслуговування. Кожен компонент проаналізовано, щоб забезпечити всебічне розуміння процесу розгортання, гарантуючи, що агенти LLM забезпечують оптимальну продуктивність і цінність. На рисунку 3.2 зображено дизайн API для впровадження агентів LLM в бізнес-середовище.

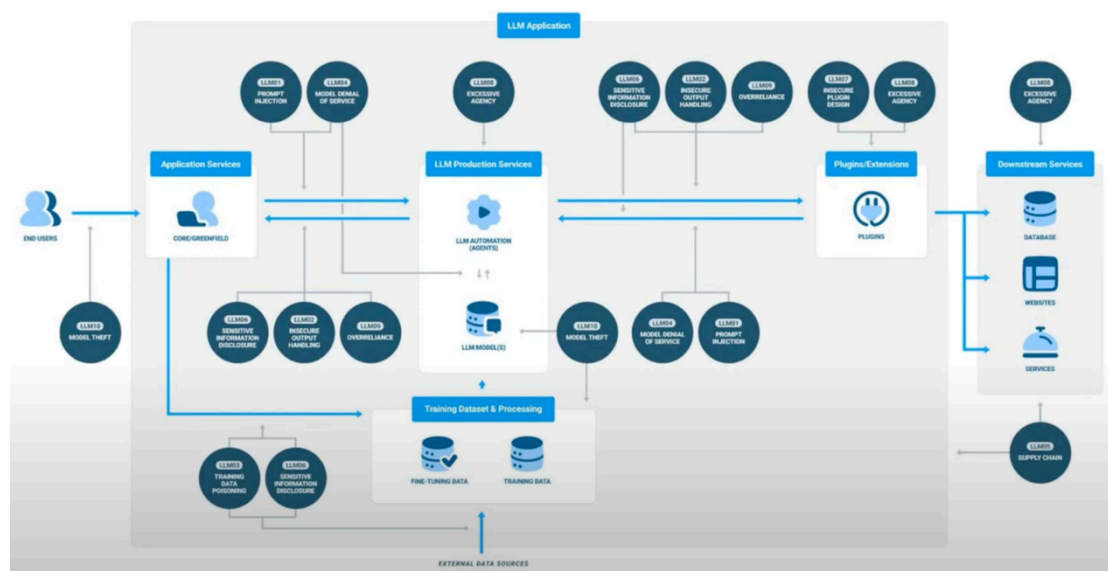


Рис 3.2 Дизайн API для впровадження агентів LLM в бізнес-середовище

Розробка інтерфейсів прикладного програмування (API) є критично важливим аспектом розгортання агентів LLM. API слугують комунікаційним мостом між агентом LLM та іншими системами або додатками в бізнес-середовищі. Ефективний дизайн API забезпечує безперебійну взаємодію, обмін даними та функціональність. API з передачею представницького стану (REST) широко використовуються завдяки своїй простоті та масштабованості. RESTful API використовують стандартні методи HTTP (GET, POST, PUT, DELETE) і не мають стану, що підвищує продуктивність і надійність. Дизайн повинен включати чіткі та послідовні кінцеві точки, правильне використання кодів стану HTTP та вичерпну документацію. Безпека має першорядне значення при розробці API. Впровадження надійних механізмів автентифікації, таких як OAuth 2.0, гарантує, що тільки авторизовані користувачі та програми можуть отримати доступ до агента LLM. Крім того, контроль доступу на основі ролей (RBAC) може бути використаний для обмеження доступу до певних функцій на основі ролей користувачів. Щоб запобігти зловживанням і забезпечити чесне використання, слід впровадити механізми обмеження швидкості та дроселювання. Ці механізми контролюють кількість запитів до API, які можуть бути зроблені протягом певного періоду часу, захищаючи агента LLM від перевантаження надмірною кількістю запитів[5].

Масштабування агентів LLM для обробки різних робочих навантажень і запитів користувачів має важливе значення для підтримки продуктивності та доступності. Ефективні стратегії масштабування гарантують, що LLM-агент може витримувати зростання та пікові періоди використання без погіршення якості обслуговування. Горизонтальне масштабування передбачає додавання більшої кількості екземплярів агента LLM для розподілу навантаження. Цей підхід підвищує відмовостійкість і дозволяє краще обробляти паралельні запити. Для рівномірного розподілу вхідних запитів між декількома екземплярами можна використовувати балансувальники навантаження. Вертикальне масштабування передбачає збільшення обчислювальних ресурсів (процесор, пам'ять) існуючих екземплярів. Хоча це може забезпечити негайне підвищення продуктивності, воно має обмеження і може бути не таким економічно ефективним, як горизонтальне масштабування. Впровадження механізмів автоматичного масштабування дозволяє системі динамічно регулювати кількість екземплярів на основі попиту в реальному часі. Політики автоматичного масштабування можна налаштувати так, щоб запускати дії масштабування на основі таких показників, як завантаження процесора, використання пам'яті або затримка запиту.

Безперервний моніторинг агентів LLM має вирішальне значення для забезпечення їх надійності, продуктивності та безпеки. Моніторинг дає уявлення про поведінку агента, допомагає виявити аномалії та полегшує проактивне обслуговування. Ключові показники ефективності, такі як час відгуку, пропускну здатність і рівень помилок, повинні відстежуватися для оцінки роботи агента. Такі інструменти, як Prometheus і Grafana, можна використовувати для збору та візуалізації цих показників у режимі реального часу. Повноцінне ведення журналів запитів API, відповідей та системних подій є важливим для усунення несправностей та аудиту. Централізовані рішення для керування журналами, такі як ELK Stack (Elasticsearch, Logstash, Kibana), забезпечують ефективну агрегацію, пошук та аналіз журналів. Налаштування механізмів оповіщення гарантує, що аномалії та проблеми з продуктивністю будуть швидко виявлені та вирішені. Сповіщення можна налаштувати на основі попередньо визначених порогових

значень ключових показників. План реагування на інциденти повинен бути в наявності, щоб керувати вирішенням проблем, мінімізуючи час простою і вплив на користувачів. Регулярне технічне обслуговування необхідне для забезпечення постійної ефективності та безпеки агентів LLM. Заходи з обслуговування включають оновлення моделей, виправлення вразливостей та оптимізацію продуктивності.

Періодичні оновлення моделі агента LLM необхідні для включення нових даних, підвищення точності та усунення будь-яких виявлених упереджень. Процес оновлення слід ретельно контролювати, щоб уникнути збоїв, а перед розгортанням слід провести ретельне тестування. Для захисту від вразливостей критично важливим є постійне оновлення програмного забезпечення та інфраструктури за допомогою патчів безпеки. Регулярний аудит безпеки та тестування на проникнення може допомогти виявити та пом'якшити потенційні загрози. Постійна оптимізація продуктивності включає в себе точне налаштування моделі, коригування гіперпараметрів та оптимізацію використання ресурсів. Такі методи, як обрізання моделі та квантування, можуть бути використані для зменшення обчислювальних витрат і підвищення ефективності.

Розгортання агентів LLM в бізнес-середовищі вимагає комплексного підходу, який охоплює розробку API, масштабування, моніторинг та обслуговування. Дотримуючись найкращих практик у кожній з цих сфер, компанії можуть гарантувати, що їхні агенти LLM працюють ефективно, приносять користь, залишаються безпечними та надійними. Систематичне розгортання та управління агентами LLM дозволяє компаніям використовувати весь потенціал цих передових технологій, стимулюючи інновації та підвищуючи операційну ефективність.

3.3 Використання агентів LLM для аналізу даних та прийняття рішень

Застосування агентів великих мовних моделей (LLM) в процесах аналізу даних і прийняття рішень є значним досягненням в області штучного інтелекту. Ці

агенти здатні обробляти величезні обсяги неструктурованих даних, видобувати значущі ідеї та підтримувати прийняття обґрунтованих рішень у різних сферах бізнесу. У цьому розділі розглядається систематичне застосування LLM-агентів для аналізу даних і прийняття рішень, що підтверджується наскрізним тематичним дослідженням у секторі фінансових послуг. Можливості LLM-агентів в аналізі даних представлено на рисунку 3.3

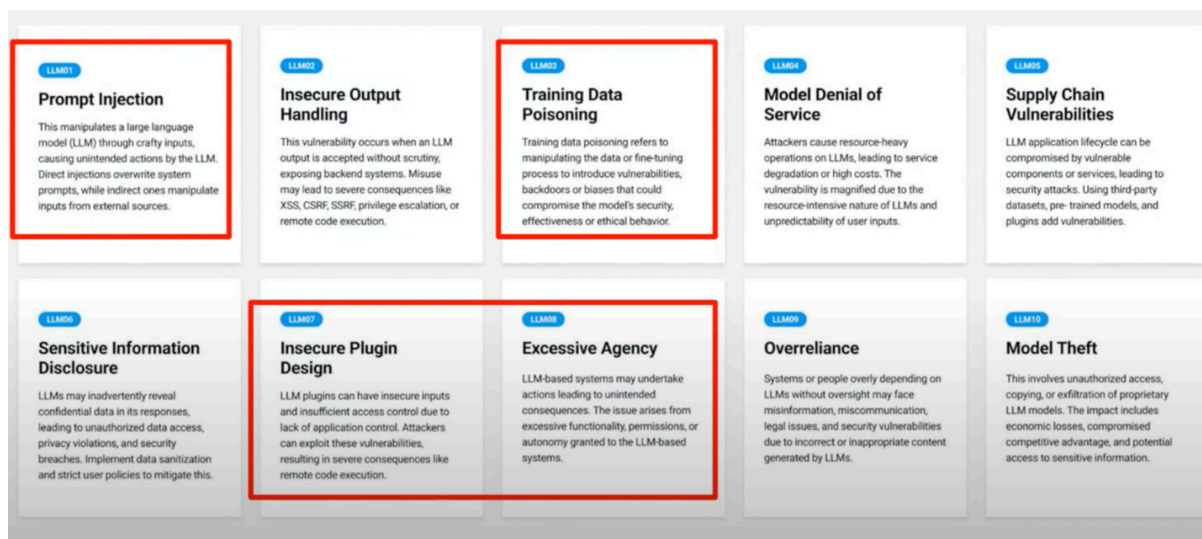


Рис 3.3. Можливості LLM-агентів в аналізі даних

LLM-агенти оснащені розширеними можливостями обробки природної мови (NLP), які дозволяють їм аналізувати та інтерпретувати великі обсяги текстових даних. Основні функціональні можливості включають:

1. **Класифікація тексту:** LLM-агенти можуть класифікувати текст за попередньо визначеними класами на основі його змісту. Це особливо корисно для організації та визначення пріоритетів інформації.
2. **Аналіз настроїв:** Аналізуючи настрої, виражені в тексті, агенти LLM можуть оцінити громадську думку, задоволеність клієнтів і ринкові настрої.
3. **Розпізнавання іменованих об'єктів (NER):** LLM-агенти можуть ідентифікувати і класифікувати такі об'єкти, як імена, дати і місця в тексті, полегшуючи вилучення структурованої інформації з неструктурованих даних.

4. Підсумовування: LLM-агенти можуть створювати стислі резюме об'ємних документів, що дозволяє швидко зрозуміти ключові моменти.

5. Відповіді на запитання: LLM-агенти можуть надавати точні відповіді на конкретні запити на основі інформації, що міститься в наборі даних.

Щоб проілюструвати застосування LLM-агентів в аналізі даних і прийнятті рішень, розглянемо приклад із сектору фінансових послуг. Мета - проаналізувати ринкові настрої та прийняти інвестиційні рішення на основі новинних статей і даних соціальних мереж.

Перший крок передбачає збір відповідних даних з різних джерел, включаючи веб-сайти фінансових новин, платформи соціальних мереж та ринкові звіти. Дані агрегуються в централізованому сховищі для подальшої обробки. Попередня обробка необхідна для забезпечення якості та узгодженості даних. Вона включає такі завдання, як видалення дублікатів, обробка відсутніх значень і нормалізація тексту. Для підготовки тексту до аналізу застосовуються токенізація, стеммінг і лематизація.

LLM-агент використовується для виконання аналізу настроїв на зібраних даних. Агент класифікує кожен сегмент тексту як позитивний, негативний або нейтральний, виходячи з висловлених почуттів. Оцінка настрою S для документа d може бути обчислена наступним чином:

$$\left[S(d) = \frac{\sum_{i=1}^n s_i}{n} \right], \quad (3.2)$$

де (s_i) - оцінка настрою i -го речення, а n - загальна кількість речень у документі[13].

LLM-агент виконує NER для виявлення в тексті ключових об'єктів, таких як назви компаній, біржові символи та фінансові терміни. Ця структурована

інформація витягується і зберігається в базі даних для подальшого аналізу. Використовуючи оцінки настроїв та ідентифіковані об'єкти, агент LLM проводить тренд-аналіз для виявлення закономірностей і кореляцій. Наприклад, агент може проаналізувати тенденції настроїв для конкретних компаній з плином часу і співвіднести їх з рухом цін на акції. LLM-агент генерує резюме ключових висновків, виділяючи значущі тенденції та ідеї. Ці резюме складаються у комплексний звіт, який містить огляд ринкових настроїв та потенційних інвестиційних можливостей.

Висновки, отримані в результаті аналізу, використовуються для прийняття інвестиційних рішень. Агент LLM може надати рекомендації на основі виявлених тенденцій та кореляцій. Наприклад, якщо аналіз настроїв вказує на позитивну тенденцію для певної компанії, агент може порекомендувати розглянути можливість інвестування в акції цієї компанії[13].

Процес прийняття рішень можна формалізувати за допомогою моделі підтримки прийняття рішень. Нехай (x) представляє набір ознак, отриманих в результаті аналізу даних (наприклад, оцінки настроїв, кількість підприємств), а (y) - інвестиційне рішення (наприклад, купувати, тримати, продавати). Модель підтримки прийняття рішень може бути виражена наступним чином:

$$[y = f(x) = \arg \arg E [x]], \quad (3.3)$$

де A - множина можливих дій, а $R(a)$ - очікуваний прибуток від дії a . Функція $f(x)$ зіставляє ознаки з оптимальною дією на основі очікуваного прибутку.

Застосування агентів LLM в процесах аналізу даних і прийняття рішень пропонує значні переваги з точки зору ефективності, точності та масштабованості. Використовуючи передові можливості NLP, LLM-агенти можуть обробляти і аналізувати величезні обсяги неструктурованих даних, витягувати значущі інсайти і підтримувати прийняття обґрунтованих рішень. Наскрізне тематичне дослідження у секторі фінансових послуг демонструє практичне застосування

LLM-агентів, висвітлюючи їхній потенціал для трансформації процесу прийняття рішень на основі даних у різних сферах бізнесу. Завдяки систематичному збору, попередній обробці, аналізу та звітності, агенти LLM дозволяють компаніям використовувати силу даних і приймати стратегічні рішення з упевненістю.

3.4 Порівняльний аналіз впроваджених результатів

Порівняльний аналіз реалізованих результатів роботи агентів великих мовних моделей (LLM) є важливим кроком в оцінці їхнього впливу на бізнес і рентабельності інвестицій (ROI). У цьому розділі представлено структурований підхід до оцінювання ефективності роботи агентів LLM, який порівнює результати з галузевими показниками та попередніми підходами. Аналіз ґрунтується на кількісних показниках та якісному аналізі, що забезпечує комплексну оцінку ефективності агентів LLM. Вплив LLM-агентів на бізнес можна оцінити за допомогою різноманітних показників, які відображають їхній внесок у ключові показники ефективності (KPI). Ці показники включають

Однією з основних переваг LLM-агентів є їхня здатність автоматизувати рутинні завдання, тим самим підвищуючи операційну ефективність. Для кількісної оцінки цього впливу можна використовувати такі показники, як час виконання завдань, рівень помилок і кількість автоматизованих завдань. Наприклад, скорочення часу виконання завдання з (T_1) до (T_2) можна виразити так:

$$[\Delta T = T_1 - T_2], \quad (3.4)$$

де (ΔT) представляє покращення ефективності. LLM-агенти часто взаємодіють безпосередньо з клієнтами, що робить задоволеність клієнтів вирішальною метрикою. Це можна виміряти за допомогою балів задоволеності клієнтів (CSAT),

балів чистого промоутера (NPS) та відгуків клієнтів. Збільшення CSAT з (S_1) до (S_2) може бути представлено наступним чином:

$$[\Delta S = S_2 - S_1], \quad (3.5)$$

де (ΔS) вказує на підвищення рівня задоволеності клієнтів. Здатність агентів LLM збільшувати дохід завдяки покращенню взаємодії з клієнтами, персоналізованим рекомендаціям та вдосконаленню процесів продажу є ще одним ключовим показником. Вплив на дохід можна виміряти, порівнюючи показники продажів до і після впровадження агентів LLM. Зміна доходу (ΔR) може бути розрахована наступним чином:

$$[\Delta R = R_{post} - R_{pre}], \quad (3.6)$$

де (R_{post}) та (R_{pre}) представляють дохід після та до впровадження відповідно. LLM-агенти можуть призвести до значної економії коштів за рахунок зменшення потреби в людській праці та мінімізації помилок. Економію витрат можна оцінити кількісно, порівнюючи операційні витрати до і після впровадження. Економія витрат (ΔC) може бути виражена наступним чином:

$$[\Delta C = C_{pre} - C_{post}], \quad (3.7)$$

де (C_{pre}) та (C_{post}) представляють витрати до та після впровадження відповідно. Рентабельність інвестицій є критично важливим показником для оцінки фінансових результатів діяльності агентів програми «Життєздатне місто». Він вимірює прибутковість інвестицій відносно їхньої вартості. ROI можна розрахувати за формулою:

$$\left[ROI = \frac{\text{Чистий прибуток}}{\text{Інвестиційні витрати}} \times 100 \right], \quad (3.8)$$

де чистий прибуток - це різниця між загальними вигодами (збільшенням доходу та економією витрат) і загальними інвестиційними витратами. Позитивна рентабельність інвестицій вказує на те, що інвестиції в агентів з управління персоналом були вигідними. Для контекстуалізації ефективності роботи LLM-агентів важливо порівняти результати з галузевими показниками. Галузеві бенчмарки забезпечують стандарт, за яким можна виміряти ефективність роботи агентів LLM. Це порівняння включає в себе наступне.

Визначення релевантних показників: Вибір контрольних показників, які стосуються конкретної галузі та сфери використання LLM-агентів. Ці показники можуть бути отримані з галузевих звітів, академічних досліджень та аналізу конкурентів.

Порівняння ефективності: Порівняння показників впроваджених агентів LLM з визначеними еталонами. Це порівняння допомагає зрозуміти, як агенти LLM працюють порівняно з галузевими стандартами. Наприклад, якщо галузевий стандарт задоволеності клієнтів становить 80 балів за тестом CSAT, а впроваджені агенти LLM досягають 85 балів, це свідчить про вищу продуктивність.

Аналіз прогалин: Проведення аналізу прогалин для виявлення сфер, в яких агенти LLM не досягають галузевих стандартів. Цей аналіз допомагає визначити конкретні аспекти, які потребують вдосконалення, та інформує про майбутні зусилля з оптимізації.

На додаток до галузевих стандартів, важливо порівняти ефективність агентів LLM з попередніми підходами, що використовувалися в організації. Це порівняння включає в себе наступне.

Базові показники: Встановлення базових показників на основі попередніх підходів, які слугуватимуть точкою відліку. Ці показники мають бути порівнянними з тими, що використовуються для оцінки агентів LLM.

Покращення ефективності: Оцінка покращення показників ефективності після впровадження агентів LLM. Наприклад, якщо попередній підхід мав

середній час виконання завдання 10 хвилин, а агенти LLM скоротили цей час до 5 хвилин, то покращення можна кількісно оцінити так:

$$\left[\text{Покращення} = \frac{10-5}{10} \times 100 = 50\% \right], \quad (3.9)$$

На додаток до кількісних показників, якісне розуміння від зацікавлених сторін та користувачів може надати цінний контекст для порівняння. Відгуки про простоту використання, надійність і загальну задоволеність агентами LLM порівняно з попередніми підходами можуть дати цілісне уявлення про їхній вплив.

Порівняльний аналіз результатів впровадження агентів LLM передбачає систематичне оцінювання їхнього впливу на бізнес, рентабельності інвестицій та продуктивності порівняно з галузевими показниками та попередніми підходами. Використовуючи комплексний набір показників і проводячи ретельні порівняння, організації можуть отримати чітке розуміння ефективності своїх LLM-агентів. Таке оцінювання не лише висвітлює успіхи та сфери для вдосконалення, а й дає інформацію для прийняття стратегічних рішень щодо майбутніх інвестицій та оптимізації. Ретельний процес бенчмаркінгу гарантує, що агенти LLM приносять відчутну користь і сприяють досягненню загальних цілей організації.

3.5 Майбутні тенденції та потенційні розробки

Стрімка еволюція агентів великих мовних моделей (LLM) відкрила нові можливості для їх застосування в бізнесі, стимулюючи інновації та трансформуючи операції в різних секторах. У цьому розділі досліджуються нові тенденції та майбутні напрямки розвитку агентів LLM, а також визначаються ключові сфери для подальших досліджень і розробок. Аналіз ґрунтується на поточних досягненнях і прогнозованих траєкторіях, забезпечуючи всебічний огляд потенційного розвитку в цій галузі.

Однією з важливих тенденцій у розвитку LLM є покращення розуміння контексту. Очікується, що майбутні агенти LLM матимуть глибше розуміння

контексту, що дасть змогу надавати більш точні та релевантні відповіді. Це передбачає вдосконалення розуміння природної мови (NLU) та інтеграцію мультимодальних даних, таких як текст, зображення та аудіо, для забезпечення цілісного розуміння введених користувачем даних. Персоналізація стає все більш важливою при розгортанні агентів LLM. Майбутні розробки будуть зосереджені на створенні агентів, які можуть адаптуватися до індивідуальних уподобань і поведінки користувачів, забезпечуючи індивідуальну взаємодію і рекомендації. Такі методи, як навчання з підкріпленням і цикли зворотного зв'язку з користувачем, відіграватимуть вирішальну роль у досягненні такого рівня персоналізації[7].

У міру того, як агенти ШНМ стають все більш поширеними, вирішення етичних проблем і пом'якшення упереджень набуває першорядного значення. Майбутні дослідження будуть зосереджені на розробці методів виявлення та усунення упереджень у навчальних даних і результатах моделювання. Це включає в себе впровадження алгоритмів справедливості і створення етичних принципів для розробки і розгортання AI. Інтеграція агентів LLM з інструментами бізнес-аналітики є новою тенденцією, яка сприятиме підвищенню ефективності прийняття рішень на основі даних. Використовуючи аналітичні можливості інструментів BI та можливості обробки природної мови (NLP) агентів LLM, компанії можуть отримати глибше розуміння та приймати більш обґрунтовані рішення. Така інтеграція сприятиме безперешкодному вилученню, аналізу та візуалізації даних.

Майбутні розробки будуть зосереджені на покращенні масштабованості та ефективності агентів LLM. Це включає вдосконалення методів стиснення моделей, таких як обрізання і квантування, щоб зменшити обчислювальні вимоги і уможливити розгортання на пристроях з обмеженими ресурсами. Крім того, розробка більш ефективних алгоритмів навчання дозволить скоротити час і витрати, пов'язані з навчанням великих моделей.

Однією з найважливіших сфер для майбутніх досліджень є пояснюваність та інтерпретованість агентів LLM. Оскільки ці моделі стають все більш складними,

розуміння процесів прийняття ними рішень стає все більш складним завданням. Дослідження будуть зосереджені на розробці методів, які забезпечать прозоре і зрозуміле пояснення результатів, що генеруються LLM-агентами. Це включає такі методи, як візуалізація уваги і методи діагностики інтерпретованості моделей.

Хоча ШНМ загального призначення продемонстрували вражаючі можливості, зростає потреба в моделях для конкретних доменів, пристосованих до певних галузей або додатків. Майбутні дослідження будуть присвячені розробці LLM, які точно налаштовані на дані конкретної галузі, що підвищить їхню релевантність і точність у спеціалізованих контекстах. Це стосується таких галузей, як охорона здоров'я, фінанси, юриспруденція та освіта. Посилення співпраці між людиною і AI є перспективним напрямком для майбутнього розвитку. Дослідження будуть зосереджені на створенні агентів LLM, які можуть працювати разом з людьми, розширюючи їхні можливості та сприяючи більш ефективній співпраці. Це передбачає розробку інтерфейсів і моделей взаємодії, які забезпечать безперешкодне спілкування і співпрацю між людьми і AI-агентами.

Попит на обробку в режимі реального часу і взаємодію з низькою затримкою стимулює дослідження периферійних обчислень для агентів LLM. Майбутні розробки будуть зосереджені на розгортанні LLM на периферійних пристроях, що дозволить обробляти дані в реальному часі і приймати рішення в джерелі. Такий підхід зменшить залежність від централізованої хмарної інфраструктури і покращить час реагування. Вплив підготовки і розгортання великих LLM на навколишнє середовище викликає все більше занепокоєння. Майбутні дослідження надаватимуть пріоритет розробці стійких та енергоефективних моделей штучного інтелекту. Це включає вивчення альтернативних методів навчання, таких як федеративне навчання, а також оптимізацію апаратного та програмного забезпечення для зменшення споживання енергії.

Оптимізацію агентів LLM для ефективності та масштабованості можна формалізувати за допомогою математичних моделей. Наприклад, цільова функція для стиснення моделі може бути виражена як:

$$[(L(\theta) + \lambda \cdot Complexity(\theta))], \quad (3.10)$$

де (θ) - параметри моделі, $(L(\theta))$ - функція втрат, (λ) - параметр регуляризації, а $(Complexity(\theta))$ - кількісна оцінка складності моделі, наприклад, кількість параметрів або обчислювальна вартість.

Майбутні тенденції та потенційні розробки агентів LLM готові революціонізувати їх застосування в бізнесі, стимулюючи інновації та підвищуючи операційну ефективність. Нові тенденції, такі як покращене контекстне розуміння, персоналізація, етичний AI, інтеграція з інструментами BI та масштабованість, формуватимуть наступне покоління агентів LLM. Одночасно зусилля в галузі досліджень і розробок будуть зосереджені на пояснюваності, моделях для конкретних галузей, співпраці людини і AI, обробці даних в реальному часі і стійкості. Звертаючись до цих сфер, сфера агентів LLM продовжуватиме розвиватися, пропонуючи нові можливості та рішення для бізнесу в різних галузях. Систематичне вивчення цих тенденцій і розробок гарантує, що LLM-агенти залишатимуться на передовій технологічного прогресу, забезпечуючи значну цінність і вплив.

Висновки

У розділі 3 було проведено всебічне дослідження розгортання та використання агентів великих мовних моделей (LLM) у бізнес-середовищі, висвітлено їхній трансформаційний потенціал та стратегічні міркування, необхідні для їхнього ефективного впровадження. Нижче наведено основні ідеї та висновки, отримані з цього розділу. У цьому розділі було детально розглянуто розгортання, використання та майбутній потенціал LLM-агентів у бізнесі. Ідеї та найкращі практики, викладені в цьому розділі, слугують цінним посібником для компаній, які прагнуть використати можливості агентів LLM для стимулювання інновацій, підвищення операційної ефективності та прийняття рішень на основі даних. Системний підхід до визначення можливостей, розгортання агентів, аналізу даних, порівняльного аналізу результатів і вивчення майбутніх тенденцій гарантує, що агенти LLM залишатимуться на передовій технологічного прогресу, забезпечуючи значну цінність і вплив.

ВИСНОВКИ

Проведена дослідницька робота забезпечила комплексний аналіз розгортання та використання агентів великих мовних моделей (LLM) у бізнес-середовищі, висвітливши їх трансформаційний потенціал та стратегічні міркування, необхідні для ефективного впровадження. Основна мета полягала в тому, щоб дослідити визначення потреб і можливостей бізнесу, процес розгортання, застосування агентів LLM для аналізу даних і прийняття рішень, порівняльний аналіз впроваджених результатів, а також майбутні тенденції і потенційні розробки. Методи дослідження включали аналіз бізнес-процесів, консультації із зацікавленими сторонами, аналіз даних, конкурентний бенчмаркінг та тематичні дослідження. Отримані результати свідчать про те, що системний підхід до визначення високоцінних випадків використання агентів LLM має вирішальне значення для максимізації їхнього впливу.

Це передбачає проведення ретельного аналізу бізнес-процесів, взаємодію із зацікавленими сторонами, аналіз даних та порівняльний аналіз з галузевими стандартами, щоб визначити сфери, в яких агенти LLM можуть забезпечити значні покращення. Розгортання агентів LLM вимагає ретельного планування і виконання, включаючи розробку API, масштабування, моніторинг і обслуговування. Ефективний дизайн API сприяє безперешкодній інтеграції та безпечній комунікації, а стратегії масштабування гарантують, що агенти зможуть впоратися з різним робочим навантаженням. Безперервний моніторинг та проактивне обслуговування мають важливе значення для підтримки продуктивності та надійності. Застосування агентів LLM для аналізу даних і прийняття рішень продемонструвало їхні передові можливості в обробці та аналізі неструктурованих даних, що дозволяє компаніям отримувати значущі інсайти і приймати обґрунтовані рішення. Наскрізне тематичне дослідження у секторі фінансових послуг проілюструвало, як агенти LLM можуть аналізувати ринкові настрої, виявляти тенденції та підтримувати інвестиційні рішення. Системний підхід до збору, попередньої обробки, аналізу та звітності даних гарантує, що

агенти LLM надають дієві проведення суворого бенчмаркінгу для оцінки продуктивності та інформування про майбутні оптимізації. Доцільність впровадження цих рекомендацій підтверджується фактами і аналізом, представленими в дослідженні, які демонструють значну цінність і вплив агентів LLM в бізнес-середовищі. Подальше використання агентів LLM вимагатиме постійного моніторингу, адаптивного навчання та ітеративних вдосконалень для підтримки їхньої ефективності та актуальності. Дослідження підкреслює важливість врахування етичних міркувань і забезпечення прозорості та справедливості при розробці та розгортанні агентів LLM.

Науковий апарат, використаний у цьому дослідженні, включає передові методи обробки природної мови, алгоритми машинного навчання, статистичний аналіз і методологію тематичних досліджень. Результати проведеного аналізу дають чітке і конкретне розуміння потенціалу та викликів, пов'язаних з агентами LLM, пропонуючи цінну інформацію для бізнесу, який прагне використати їхні можливості для стимулювання інновацій, підвищення операційної ефективності та прийняття рішень на основі даних. Єдина термінологія та чіткість викладу гарантують, що висновки є обґрунтованими та практичними, забезпечуючи міцний фундамент для майбутніх досліджень і розробок у сфері LLM-агентів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Brockman G., Cheung V., Pettersson L., Schneider J., Schulman J., Tang J., Zaremba W. OpenAI's GPT-3 and Beyond: The Future of Large Language Models // AI Magazine. – 2020. – Vol. 41, No. 4. – P. 38–45.
2. Chollet F. Deep Learning with Python. – New York : Manning Publications Co., 2017. – 384 p.
3. Deng L., Yu D. Deep Learning: Methods and Applications. – Bristol : Morgan & Claypool Publishers, 2014. – 256 p.
4. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge : MIT Press, 2016. – 800 p.
5. Hinton G.E., Salakhutdinov R.R. Reducing the Dimensionality of Data with Neural Networks // Science. – 2006. – Vol. 313, No. 5786. – P. 504–507.
6. Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems 25 (NIPS 2012). – 2012. – P. 1097–1105.
7. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. – 2015. – Vol. 521, No. 7553. – P. 436–444.
8. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR. – 2013.
9. Rumelhart D.E., Hinton G.E., Williams R.J. Learning representations by back-propagating errors // Nature. – 1986. – Vol. 323, No. 6088. – P. 533–536.
10. Schmidhuber J. Deep Learning in Neural Networks: An Overview // Neural Networks. – 2015. – Vol. 61. – P. 85–117.
11. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. – 2014. – Vol. 15, No. 1. – P. 1929–1958.
12. Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. Going Deeper with Convolutions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 1–9.

13. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems 30 (NIPS 2017). – 2017. – P. 5998–6008.

14. Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Xie W., Liu R., Kaiser Ł., Gouws S., Kato K., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M., Dean J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation // arXiv preprint arXiv:1609.08144. – 2016.

15. Zhang X., Zhao J., LeCun Y. Character-level Convolutional Networks for Text Classification // Advances in Neural Information Processing Systems 28 (NIPS 2015). – 2015. – P. 649–657.

16. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ritter S., Stinson P., Gretton A., Gray J., Chesson P., Moffat J., Spiers H., Paine T.B., Glover N., Conerly C., Plumb J., Bridgland D., Selvam V., Snoswell S., Rolfe J.T., Baumgartner R., Papierniak M., Galactica Team., Ganguli D., Song F., Ahmad Z., Metz L., Nash C., Young P., Peng L., Webb M., Desmaison A., Kumar K., Xu J., Kelcey J., Steinhardt J., Patry J., Rao R., Li Y., Waghmare S., Noonan J.P., Venkatesh A., Bonde A.P., Chandrasekhar V., Ayers C., Mechery V., Kohli P., Henriques D., Fitzgerald J., Tabak D.D., Franklin L., Cheng L., Subramanian S., Rogers L., Dhuper S., Qureshi N., Lee J.H., Mokrov E., Zhu L., Goyal S., van Steenkiste S., Huang E., Huang E., Dalvi F., Tomašev N., Sanh V., Upadhyay S., Speilmeyer M., Menezes E., Munir R., Zhang T., Powers B., Potapenko A., Sharma S., Feinberg M., Tayem H., Ettinger A., Haque A., Hosseini M., Baier F., Sexton J., Gansky R., Zhou T., Karmarkar A., Rocki K., Pavlov M., Callison-Burch C., Jiang X., Guo J., Kim B., Lakhota K., Lin X.A., Ou J., Yang W., Amir H., Bhattacharya L., Luo C., Rashid A., Tan M., Tao B., Vaswani A., Wei J., Wu X., You G., Abramovitz L., Geva M., Goldberg Y., Niton J., Zheng S., Press D., Smith N.A., Lewis M., Zettlemoyer L., Strobel H., Gehrmann S., Dai A.M., Hashimoto T.B., Welleck S., Ippolito D. Language Models are

Few-Shot Learners // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 1877–1901.

17. Radford A., Narasimhan K., Salimans T., Sutskever I. Improving Language Understanding by Generative Pre-Training // arXiv preprint arXiv:1801.07704. – 2018.

18. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. – 2018.

19. Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Strobelt H. RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv preprint arXiv:1907.11692. – 2019.

20. Yang Z., Dai Z., Yang Y., Carbonell J.G., Salakhutdinov R.R., Le Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding // Advances in Neural Information Processing Systems 32 (NeurIPS 2019). – 2019. – P. 5754–5764.

21. Clark K., Luong M.-T., Le Q.V., Manning C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators // International Conference on Learning Representations (ICLR). – 2020.

22. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2016. – P. 770–778.

23. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition // arXiv preprint arXiv:1409.1556. – 2014.

24. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // Proceedings of the 32nd International Conference on Machine Learning (ICML). – 2015. – P. 448–456.

25. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization // International Conference on Learning Representations (ICLR). – 2015.

26. Sutskever I., Martens J., Dahl G., Hinton G. On the importance of initialization and momentum in deep learning // Proceedings of the 30th International Conference on Machine Learning (ICML). – 2013. – P. 1139–1147.

27. Zeiler M.D., Fergus R. Visualizing and Understanding Convolutional Networks // European Conference on Computer Vision (ECCV). – 2014. – P. 818–833.
28. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – P. 580–587.
29. Long J., Shelhamer E., Darrell T. Fully Convolutional Networks for Semantic Segmentation // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2015. – P. 3431–3440.
30. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks // Advances in Neural Information Processing Systems 28 (NIPS 2015). – 2015. – P. 91–99.



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ УНІВЕРСИТЕТ ТЕЛЕКОМУНІКАЦІЙ
ДИПЛОМНА РОБОТА
на ступінь вищої освіти магістр
із спеціальності 122 Комп'ютерні технології

Використання агентів великих мовних
моделей у бізнесі

Виконав: студент 4 курсу, групи КНД-42

Харитонов Гліб Євгенович

Керівник: доцент кафедри Комп'ютерних наук

к.т.н. Щербина Ірина Сергіївна

ЗАГАЛЬНА ХАРАКТЕРИСТИКА ДИПЛОМНОЇ РОБОТИ

Тема	Використання агентів великих мовних моделей у бізнесі
Мета дослідження	Застосування LLM-агентів у різних бізнес-середовищах, оцінка їхнього впливу та окреслення стратегічних підходів до їхньої інтеграції
Наукове завдання	Дослідити можливості вткористання агентів LLM у бізнесі
Об'єкт дослідження	Агенти великих мовних моделей

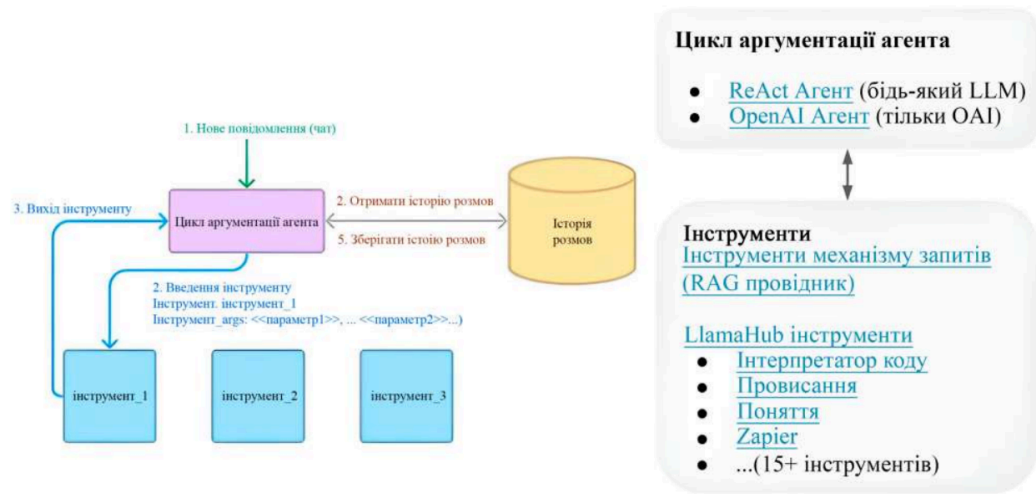
Сфери використання LLM агента



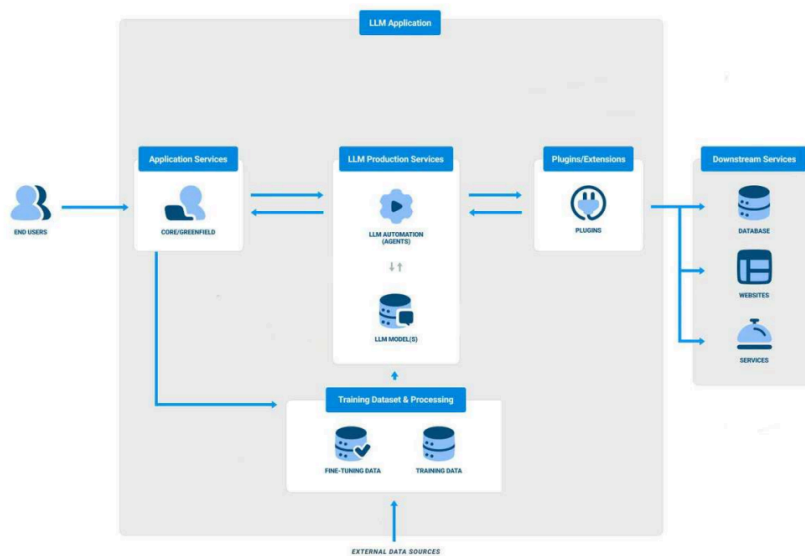
Алгоритм работы ллм агента



Компоненти архітектури агента



API для впровадження агентів LLM в бізнес-середовище



Можливості застосування LLM-агентів в аналізі даних



Висновки

- ✓ Проведена дослідницька робота забезпечила комплексний аналіз розгортання та використання агентів великих мовних моделей (LLM) у бізнес-середовищі, висвітливши їх трансформаційний потенціал та стратегічні міркування, необхідні для ефективного впровадження. Отримані результати свідчать про те, що системний підхід до визначення високоцінних випадків використання агентів LLM має вирішальне значення для максимізації їхнього впливу.
- ✓ Розгортання агентів LLM вимагає ретельного планування і виконання, включаючи розробку API, масштабування, моніторинг і обслуговування. Ефективний дизайн API сприяє безперешкодній інтеграції та безпечній комунікації, а стратегії масштабування гарантують, що агенти зможуть впоратися з різним робочим навантаженням. Безперервний моніторинг та проактивне обслуговування мають важливе значення для підтримки продуктивності та надійності. Застосування агентів LLM для аналізу даних і прийняття рішень продемонструвало їхні передові можливості в обробці та аналізі неструктурованих даних, що дозволяє компаніям отримувати значущі інсайти і приймати обґрунтовані рішення.
- ✓ Науковий апарат, використаний у цьому дослідженні, включає передові методи обробки природної мови, алгоритми машинного навчання, статистичний аналіз і методологію тематичних досліджень.

Апробація

Харитонов Гліб Євгенович,

“Використання LLM агентів в сучасному світі.”,

Конференція “Сучасні інтелектуальні інформаційні
технології в науці та освіті.”,

Конференція відбулась у дистанційному форматі у
Державному університеті інформаційно-комунікаційних
технологій 15 травня 2024 році.

Дякую за увагу!