DOI: https://doi.org/10.23939/csn2024.02.267
УДК 004.8

МОЖЛИВОСТІ ТА ОБМЕЖЕННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

І.Ю. Юрчак, О.О. Кичук, В.М.Оксентюк, А.О.Хіч

Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування
E-mail: Iryna.Y.Yurchak@lpnu.ua, Olha.Kychuk.knm.2020@lpnu.ua, Vira.M.Oksentyuk@lpnu.ua, Andrii.O.Khich@lpnu.ua

Робота присвячена дослідженню великих мовних моделей (ВММ) та підходів для підвищення ефективності їх використання у новому сервісі. Стрімкий розвиток ВММ, заснованих на архітектурі трансформерів, відкрив нові можливості в галузі обробки природної мови та автоматизації різноманітних завдань. Однак, використання повного потенціалу цих моделей вимагає ретельного підходу та врахування численних факторів.

Здійснено огляд еволюції великих мовних моделей, виділено провідні компанії, які займаються дослідженнями та розробкою ефективних систем. Розглянуто будову цих моделей та способи представлення внутрішніх знань. Описано ключові підходи до навчання, включаючи збирання та попередню обробку даних, а також вибір відповідної архітектури нейронних мереж, що застосовуються у великих мовних моделях. Зазначено, що найбільшого прориву досягнуто за допомогою нейромережі Трансформер, заснованої на механізмі уваги.

Проведено дослідження та наведено порівняння популярних моделей на базі архітектури трансформера, а саме: GPT, Claude та Gemini. Визначено метрики останніх версій з відкритими API, унікальні особливості, сильні та слабкі сторони, можливості та обмеження.

Актуальність теми полягає в стрімкому розвитку технологій обробки природної мови та зростанні попиту на великі мовні моделі в різних галузях. Ефективне використання цих моделей має величезний потенціал для підвищення продуктивності та якості роботи з текстовими даними. Однак, через складність архітектури та великі обсяги даних, необхідних для навчання, вибір та налаштування оптимальної моделі для конкретної задачі є непростим завданням.

Як результат дослідження наведено рекомендації для розробників щодо використання популярних моделей з відкритим кодом у новому сервісі або інтеграції зі сторонніми програмами. Зазначено особливості моделей, їх сильні сторони, обмеження та певні застереження щодо довіри до отриманих результатів.

Ключові слова: великі мовні моделі, GPT, Claude, Gemini, архітектура трансформер, нейронні мережі, чат-бот, генерування контенту.

1. Вступ

За останні роки сфера штучного інтелекту значно еволюціонувала, а великі мовні моделі (ВММ) стали рушійною силою цієї трансформації. Новітні потужні інструменти відкривають широкі горизонти в галузях, що варіюються від чат-ботів до пошукових систем і інструментів для творчого письма. Однак, щоб розкрити потенціал ВММ для створення корисних продуктів, потрібні ґрунтовні знання та навички.

2. Огляд літературних джерел

Використання великих мовних моделей для вирішення інтелектуальних завдань набуває стрімкої популярності надають багатообіцяючі результати. Однак, мало уваги приділено розробці стратегій для оцінки та порівняльного аналізу корисності конкретних ВММ. У статті [1] проведено аналітичний огляд популярних ВММ і розглядаються три різні питання: який мінімальний набір функціоналу необхідний для ВММ; які стратегії порівняння допомагають вибрати оптимальні ВММ; як можна оцінити результат ВММ на конкретних завданнях? Відповіді на ці запитання є основоположною для розробки комплексних тестів інтелектуального аналізу процесів на ВММ, що охоплюють різні завдання та парадигми впровадження.

Роботи зі створення перших ВММ почалися з використання векторного представлення слів у просторі. Це допомогло застосувати машинне навчання для мовних моделей на великому корпусі тексту. Під час навчання модель вчиться передбачати оточуючі слова, що задані цільовим словом [2]

Наступні дослідження вже стосувалися вибору архітектури нейронних мереж. Самою оптимальною на той момент виявилися рекурентні нейронні мережі (Recurrent Neural Networks, RNN) та їх модифікації. Рекурентні мережі здатні вловлювати залежності у всій послідовності слів і навчаються в режимі самоконтролю на великих текстових корпусах для передбачення наступного слова заданої послідовності. [3]

Архітектура Transformer перевершує моделі на основі RNN за обчислювальною ефективністю. Моделі GPT і BERT продемонстрували ефективність трансформенних моделей у різних завданнях, використовуючи мовні моделі, що попередньо навчені на великих корпусах. У статті [4] досліджено ефективні архітектури Transformer для мовної моделі, включаючи додавання додаткових рівнів LSTM (Long Short-Term Memory) для кращого захоплення послідовного контексту, зберігаючи при цьому ефективність обчислень.

Сьогодні існує багато платформ, які намагаються запропонувати свої рішення для генерації контенту, автоматизації процесів і багато іншого. Основними гравцями на цьому ринку є OpenAI, Anthropic, Google та інші. Ці компанії мають свої унікальні пропозиції, і варто розглянути їх докладніше, щоб зрозуміти, який інструмент може бути корисним у різноманітних проектах [5-7].

Використання великих мовних моделей має значний потенціал для трансформації суспільства та життя людей. Однак, бурхливий розвиток різноманітних застосувань з використанням мовних моделей створюють ризики, які можуть негативно вплинути на окремих осіб, групи, організації, спільноти, суспільство, навколишнє середовище та планету. Ці ризики можуть виникати в різний спосіб та можуть бути охарактеризовані як довготермінові чи короткострокові, з високою чи низькою ймовірністю, системні чи локалізовані, а також із сильним чи низьким впливом. [9] У статті [10] проведено структурування ризиків, пов’язаних із великими мовними моделями. Щоб сприяти розвитку відповідальних інновацій, потрібне глибоке розуміння потенційних ризиків, пов’язаних із цими моделями.

Покоління великих мовних моделей

Перше покоління розвивалось у 1950-ті – 1980-ті роки [1]. Ранні мовні моделі використовували статичні методи, такі як n-грами, для передбачення ймовірності наступного слова в послідовності. Одним із ключових внесків статистичних мовних моделей є те, що вони показали, як слова можна представляти векторами чисел. Ці вектори, які називаються розподілами слів, відображають ймовірність того, що слово з'являтиметься в різних контекстах. Наприклад, слово «кішка» може мати високий розподіл у контексті слів «собака», «тварини» або «пухнастий», але низький розподіл у контексті слів «машина», «число» або «твердий».

Ранні методи векторного представлення слів мали перевагу в простоті реалізації та низьких вимогах до обчислювальних ресурсів. Проте, їх ключовим недоліком була нездатність враховувати контекст слів, що обмежувало точність у складних завданнях, де значення слова може трактуватися залежно від його оточення, що призводило до низької точності в складних завданнях.

Друге покоління, нейронні мовні моделі (1990-ті – 2010-ті роки), було скероване на дослідження контекстних векторних представлень слів та подальше вдосконалення моделей з першого покоління.

Воно стало революційним етапом у розвитку обробки природної мови, відзначившись значним прогресом у розумінні та генеруванні мови [2].

На відміну від попередніх моделей, що ґрунтувалися на статистичних методах, моделі другого покоління використовували нейронні мережі, які здатні навчатися на великих обсягах даних та виявляти складні закономірності в мові. Виникли такі архітектури, як рекурентні нейронні мережі та довга короткострокова пам'ять, які дозволили моделям ефективніше навчатися на даних з довгостроковими залежностями [3]. Цьому також посприяли потужні комп'ютери, що призвело до кращої їх продуктивності.

До цього напрямку відносяться такі моделі, як CoVe, ELMo, а також відомі прориви від OpenAI, такі як GPT-2 та BERT. Вони відображають високий рівень контекстуального розуміння та дозволяють краще узгоджувати слова з їхнім контекстом, що покращує якість результатів у завданнях обробки природної мови. Ці моделі стали основою для багатьох сучасних застосунків, таких як чат-боти, голосові помічники, системи машинного перекладу та інструменти для творчого письма. [5-7]

Третє покоління сучасних ВММ (2010-ті – 2020-ті роки) характеризується значними покращеннями у розумінні контексту та виявленні складних семантичних відносин. Саме в цьому поколінні почали застосовувати архітектуру Transformer, яка значно відрізняється від своїх попередників. Модель використовує механізм уваги, який дозволяє їй враховувати зв'язки між будь-якими двома словами в реченні, незалежно від їх відстані. Це робить Transformer значно потужнішою моделлю для розуміння природної мови [4].

Це покоління відзначається початком інтеграції інформації з різних джерел та модальностей, таких як текст, зображення та аудіо, що дозволяє моделям забезпечити більш глибоке розуміння вхідних даних. Спостерігається підвищена продуктивність завдяки методам стиснення даних та досягнуто значні покращення у розумінні лінгвістичних та структурних властивостей текстів, тобто їх граматики та синтаксису мови. Найпоширенішими моделями третього покоління є GPT-3 та LaMDA [6].

Четверте покоління великих мовних моделей є найновішим, містить основні досягнення третього покоління, а також додаткові інновації. Моделі четвертого покоління охоплюють більший обсяг даних для навчання та більшу кількість параметрів, що дозволяють розуміти ширший спектр мовних конструкцій та нюансів.

Покращені можливості розуміння та генерації текстів призводять до створення більш зв'язних та контекстуально точних відповідей. Спостерігається вдосконалене точне налаштування та трансферне навчання, під час якого використовуються знання, що набуті на одному наборі даних для виконання нового завдання на іншому наборі. Такий підхід дозволяє збільшити продуктивність моделей у конкретних завданнях, таких як переклад, реферування чи відповіді на запитання. Збільшення масштабованості та ефективності також є характеристиками цього покоління.

На даний момент, однією з найвідоміших мовних моделей, яка належить до четвертого покоління великих мовних моделей, є GPT-4 (Generative Pre-trained Transformer 4) від компанії OpenAI. Ця модель побудована на основі трансформерної архітектури та має вражаючі можливості у генерації текстів, розумінні мови та вирішенні різних завдань у галузі обробки природної мови [5].

Детальний аналіз різних поколінь мовних моделей демонструє поступовий розвиток від простих статистичних методів до складних нейронних архітектур, здатних виявляти тонкі семантичні зв'язки та враховувати широкий контекст. Реалізовано ключові проривні моменти, такі як застосування новітніх архітектур нейронних мереж, механізмів уваги та інтеграції різномодальних даних [1].

3. Постановка задачі

Основна ідея мовних моделей полягає в тому, що слова в мові не є випадковими. Існує певна закономірність у тому, як слова з'являються разом утворюючи фрази, фрази відповідно утворюють речення, а моделі навчаються виявляти цю закономірність. Це робиться шляхом аналізу великих обсягів текстових даних, таких як книги, статті та веб-сайти. Зазвичай, вони базуються на нейронних мережах, які є складними математичними моделями, що навчаються на великих обсягах текстових даних і здатні виявляти закономірності в тому, як слова з'являються разом.

Ці закономірності використовуються для передбачення ймовірності появи наступного слова в послідовності. Саме тому, їх можна використовувати для перетворення розмовної мови в текст, переклад, генерацію статей чи творів, надання відповідей на запитання, визначення емоційного тону тексту, тощо.

Метою статті є проведення аналізу сучасних великих мовних моделей, визначити їхні ключові можливості, обмеження, сфери застосування та застереження для максимально ефективного використання цих потужних інструментів в нових сервісах чи сторонніх програмах. Для досягнення цієї мети необхідно вирішити такі основні завдання:

- Провести огляд існуючих великих мовних моделей, їх архітектури, принципів роботи, переваг та недоліків.
- Визначити критерії оцінювання ефективності великих мовних моделей у певних задачах та провести порівняльний аналіз різних моделей за цими критеріями.
- Визначити можливості та обмеження сучасних великих мовних моделей, а також виявити проблеми та шляхи їх подолання.
- Надати рекомендації для розробників щодо використання чи втілення ВММ з відкритим кодом у новий продукт або інтеграції зі сторонніми програмами.

4. Архітектура ВММ

Великі мовні моделі ґрунтуються на складних нейронних мережах, які здатні навчатися на величезних обсягах даних і виявляти складні закономірності між словами та фразами. Нижче наведено огляд найпоширеніших архітектур ВММ.

Трансформер (Transformer) – тип нейронної мережі, який використовується в обробці природної мови. Він з'явився у 2017 році і швидко став однією з найпопулярніших архітектур для ВММ [4]. Архітектура трансформера складається з двох основних компонентів: кодувальника (encoder) та декодувальника (decoder) (рис.1). Кодувальник приймає послідовність вхідних токенів і створює послідовність прихованих представлень. Потім декодувальник отримує ці приховані представлення та генерує послідовність вихідних токенів. Обидва компоненти містять стек шарів самоуваги.

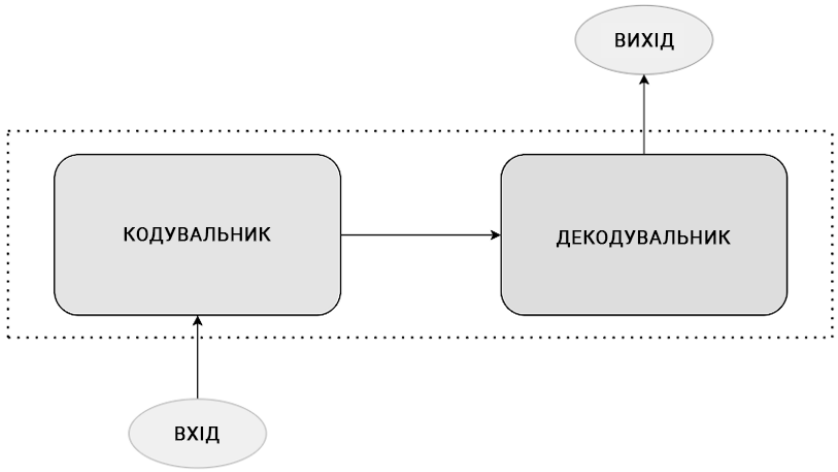

Рис. 1. Архітектура трансформера

Загалом, у галузі ВММ термін «токен» використовується для опису фрагмента тексту, який модель читає або створює. Токени не завжди є словом, вони можуть бути як і меншою одиницею: символом або частиною слова, так і більшою, як ціла фраза. На етапі розробки архітектури мовної моделі кількість токенів встановлюється і не може бути змінена.

Механізм уваги дозволяє зосередитися на найважливіших частинах вхідної послідовності. Це робить Трансформер більш ефективним, ніж інші архітектури ВММ, на таких завданнях, як узагальнення тексту, відповіді на запитання та машинний переклад. Трансформер використовує паралельну обробку та двоскеровану архітектуру, які дозволяють йому враховувати контекст слів у реченні. Це покращує його здатність розуміти складні речення.

Прикладами мовних моделей на архітектури Трансформер є:

Можливості та обмеження великих мовних моделей

- GPT-4 (OpenAI) – це велика мультимодальна модель, яка здатна обробляти запити у вигляді картинок та тексту і надавати текстові відповіді. GPT-4 працює на «рівні людини» у різних професійних та академічних тестах, і у середньому вона набирає у цих тестах 88% і більше. [5]
- Gemini (GoogleAI) – це набір генеративних моделей, що використовують обробку природної мови для динамічної інтерпретації та реагування на дані, які користувач вводить. Моделі можуть взаємодіяти з різними типами контенту – текстом, відео, аудіо та кодом на Python, Java, C++ та Go.
- Claude (AnthropicAI) – це мовна модель, що спроможна витягнути коротку анотацію з довгих статей, новин чи документів. Claude є відмінним помічником у програмуванні на всіх основних мовах програмування.

Рекурентна нейрона мережа (Recurrent Neural Networks, RNN) - це тип нейронної мережі, який використовується для обробки послідовних даних, таких як текст або аудіо [3]. RNN можуть навчатися на довгих послідовностях даних і враховувати контекст попередніх елементів у послідовності (рис.2).

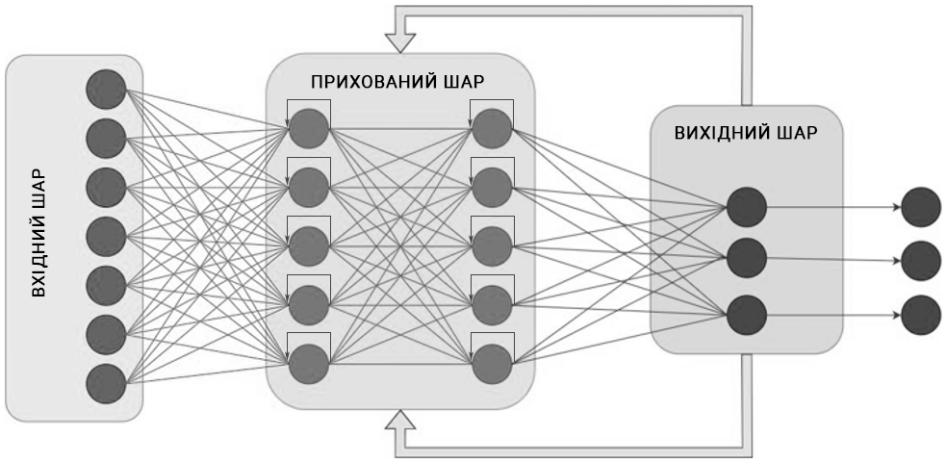

Рис. 2. Архітектура рекурентної нейронної мережі

Особливостями RNN є рекурентні з'єднання, які дозволяють їм передавати інформацію з одного часового кроку на наступний. Це робить їх придатними для обробки послідовних даних. Мережа отримує вхідні дані у вигляді послідовності слів, наприклад, «Я люблю свою родину». На першому часовому кроці мережа отримує слово «Я». Вона обробляє це слово та генерує прогноз наступного слова. На другому часовому кроці мережа отримує прогнозоване слово («люблю») та слово «Я». Вона обробляє ці два слова разом і генерує прогноз наступного слова. Цей процес повторюється для кожного слова в реченні, а мережа використовує інформацію з попередніх слів, щоб покращити свої прогнози.

Деякі RNN мають корисний механізм, який називається «Довга Короткочасна пам'ять» (Long Short-Term Memory, LSTM). RNN можуть страждати від проблеми, яка називається «зникнення градієнта». Ця проблема виникає, коли інформація з попередніх слів з часом стає все менш важливою. Тому, для подолання цієї проблеми добре підходить LSTM, що спроможна зберігати інформацію протягом більш тривалого періоду часу.

Прикладами ВММ на основі рекурентних нейронних мереж є:

- ELMo (Allen Institute for Artificial Intelligence) – модель, що навчена на великому наборі даних тексту. Вона може генерувати векторні представлення слів, які враховують контекст повідомлення. Ці вектори можуть бути використані для покращення продуктивності різних завдань обробки природної мови, таких як машинний переклад та аналіз емоцій.
- ULMFit (University of Washington) – RNN-модель, що також навчена на великому наборі даних тексту. Використовує метод часткового навчання, де лише невелика частина даних має мітки. Це робить її більш економною з точки зору обчислювальних ресурсів, ніж інші RNN-моделі.

Гібридні архітектури ВММ поєднують різні типи архітектур [1]. Це може покращити продуктивність ВММ на певних завданнях. Тому, їх особливістю є можливість поєднання сильних сторін різних моделей. Наприклад, Трансформер може бути використаний для обробки довгих послідовностей даних, а RNN може бути використаний для обробки коротких послідовностей даних. Гібридні моделі можна налаштувати для конкретних завдань, оскільки вони є більш гнучкими, ніж інші типи архітектур.

Прикладом гібридної архітектури є наступна модель:

- WuDao 2.0 (Academy of Sciences of China) – це гібридна модель, яка поєднує Transformer та RNN архітектури. Вона має 1.75 трлн параметрів і навчена на наборі даних тексту та коду з китайської мови. WuDao 2.0 може виконувати багато завдань, подібних до англомовних ВММ, а також має кращу продуктивність у деяких областях, таких як генерування китайської поезії та коду.

Незважаючи на відмінності в деталях архітектури та підходах до навчання, більшість сучасних передових великих мовних моделей сьогодні використовують архітектуру Трансформера як базову основу. Трансформери продемонстрували високі здібності в таких завданнях обробки природної мови, як генерація художнього чи технічного тексту, машинний переклад, проведення змістовних діалогів, створення програмного коду, розпізнавання чи генерація зображень.

Одним з найбільш перспективних напрямків застосування мовних моделей на базі трансформера є розробка інтерактивних систем діалогу, зокрема чат-ботів. Здатність трансформерів генерувати зв'язні та контекстно-релевантні відповіді робить їх ідеальними кандидатами для створення природних конверсаційних інтерфейсів [3].

5. Алгоритми роботи ВММ

Великі мовні моделі працюють, використовуючи методи глибокого навчання та великі обсяги текстових даних. Вони складаються з багатошарової структури, де кожен шар володіє параметрами, які піддаються налаштуванню під час навчання. Додатково, для кращого фокусування на ключових аспектах даних, використовується «механізм уваги», що акцентує увагу на певних фрагментах наборів даних.

В процесі навчання ВММ прагнуть передбачити наступне слово в реченні, ґрунтуючись на контексті, наданому попередніми словами. Це досягається шляхом присвоєння ймовірнісних оцінок токенам – розбитим на менші фрагменти символам. Далі ці токени перетворюються на вставки, які чисельно відображають контекст. Для забезпечення точності ВММ навчаються на величезних масивах тексту (в мільярдах сторінок).

Для навчання великих мовних моделей застосовують три поширені моделі навчання:

- Навчання без прикладів або з малою кількістю прикладі (Zero-shot and Few-shot). Базові ВММ можуть розуміти широкий спектр запитів без спеціального навчання, часто за допомогою підказок, хоча точність відповідей в різних сесіях може різнитися.
- Навчання у кілька прийомів. Якщо навести кілька відповідних прикладів навчання, можна значно підвищити продуктивність базової моделі у конкретній галузі.
- Точне налаштування. Це розширення навчання у кілька прийомів, під час якого фахівці з аналізу даних навчають базову модель коригувати параметри за допомогою додаткових даних, які стосуються конкретного застосування.

Наслідком навчання є здатність ВММ генерувати текст та самостійно передбачати майбутнє слово, аналізуючи вхідний текст і використовуючи шаблони та знання. В результаті можна отримати зв'язне та контекстуально доречне генерування тексту, яке буде використовуватися для широкого спектру завдань з обробки природної мови та створення контенту.

Для навчання моделі застосовуються надвеликі масиви до десятків терабайтів тексту, що у певному сенсі надає універсальні знання практично про все. Знання закодовано у вагових коефіцієнтах нейронної мережі, які формуються в процесі навчання.[8] Сама модель має величезну кількість параметрів. Завдяки цьому можна «запам'ятати» всі стандартні конструкції великої кількості мов, включаючи мови програмування, сенс слів та термінів, стилі тексту та правила логічних висновків.

- Модель навчається передбачати наступне слово в тексті на основі попередніх слів.
- Модель аналізує статистичні закономірності та взаємозв'язки між словами у текстах.

- Ці взаємозв'язки запам'ятовуються у вагах нейронної мережі як розподілені числові представлення слів та контексту.

Так формується "узагальнена пам'ять", що дозволяє моделі робити логічні висновки та генерувати нові формулювання на основі внутрішніх представлень мови. Знання у мовних моделях є зазвичай статистичними, а не динамічними, що виведені з наявних даних у процесі самонавчання.

Оцінка та налаштування моделі

Після навчання модель оцінюється на тестовому наборі даних, який не використовувався під час навчання. Тестовий набір даних дозволяє зрозуміти, наскільки добре модель узагальнюється на невідомих даних. Залежно від результатів оцінки модель може бути налаштована різними способами.

Одним з них є тонке налаштування (Fine-Tuning) - донавчання на новому наборі даних, який стосується конкретного завдання. Наприклад, модель, що навчена на загальному текстовому корпусі, може бути тонко налаштована на наборі даних новинних статей для покращення її здатності генерувати реалістичні новини.

Іноді корисним може бути перенесення навчання (Transfer Learning). Тоді, модель, що навчена на одному завданні, використовується як початкова точка для навчання на іншому завданні. Це може бути корисним, якщо обсяг даних для нового завдання є обмеженим. Використовується для різних цілей: для створення ігор і веб-додатків, розробки внутрішніх інструментів для різноманітних проектів і написання чат-ботів. Моделі також широко застосовуються в науковій області для досліджень і розв’язування прикладних завдань [9].

6. Дослідження характеристик мовних моделей на базі архітектури трансформера

Нижче проведено порівняння ВММ на базі архітектури Трансформера, а саме: GPT, Gemini AI та Claude AI [5-7]. Для широкого кола користувачів, модель представлена у вигляді зручного інтерфейсу – чат-бота, який адаптований до ведення діалогів, спілкування та інтерактивності. Чат-боти на базі конкретної моделі здатні генерувати текст людського рівня, перекладати мови, створювати різноманітний творчий контент та надавати інформативні відповіді на запитання. Огляд моделей, їх функціоналу, переваг та обмежень допоможе глибше зрозуміти потенціал і виклики, пов'язані з використанням великих мовних моделей для побудови діалогових систем.

Вибір метрик для здійснення порівняння є важливим завданням, яке потребує ретельного та всебічного підходу. Вони повинні бути репрезентативними, інформативними та корисними для прийняття обґрунтованого рішення про те, який чат-бот найкраще підходить для конкретних потреб. Для цього аналізу обрано такі критерії як: розмір моделі, архітектура, дані навчання, можливості, основні обмеження, етика та безпека і сфери застосування.

GPT (Generative Pretrained Transformer), OpenAI

GPT-4 - це основна версія великої мовної моделі, розробленої OpenAI, яка є продовженням GPT-3.5. Вона здатна генерувати більш контекстуально точні відповіді на широкий спектр питань, обробляти складніші запити та працювати з більшим обсягом інформації порівняно з попередніми версіями. GPT-4 має покращені можливості щодо оброблення кількох завдань одночасно, що робить її корисною для вирішення різноманітних проблем. [5]

GPT-4-turbo (4o) - це оптимізована версія GPT-4, що має вищу продуктивність та швидкість обробки запитів. GPT-4o є більш досконалим, ніж його попередники, у трьох ключових сферах: креативність, візуальне введення та контекстний діапазон.

Модель спроможна у співпраці з користувачами працювати над творчими проектами: написання музики, сценаріїв, художнього чи технічного тексту. Можна обробляти до 25 000 слів тексту від користувача. Користувачі можуть надіслати до чату посилання на веб-сторінку та попросити взаємодіяти з текстом на цій сторінці, не вводячи його самостійно. Ця нова функція оптимізує можливості алгоритму для створення вмісту, а також його здатність вести довгі розмови. Покращено здатність моделі точно обробляти дані, пов'язані з датами, числами та таблицями. Ще однією важливою новою функцією є обробка зображень.

GPT-o1 – модель, що використовує навчання з підкріпленням, ефективні алгоритми оптимізації та надвеликий набір навчальних даних. За рахунок цього версія o1 ретельніше обмірковує рішення та витрачає більше часу на аналіз, перш ніж дати відповідь на запит. Вона реагує не так швидко, як

попередні моделі, оскільки створює ланцюжок роздумів для вирішення складного завдання. З виходом GPT-o1 вирішено проблему попередніх версій ChatGPT, які могли думати покроково лише завдяки додатковим запитам (промптам).

Всі версії GPT базуються на трансформенній архітектурі, але кожна з них оптимізована для різних сценаріїв використання, що робить їх гнучкими та придатними для різних завдань та програм.

Gemini, Google DeepMind

Gemini 1.5 є значним кроком у розвитку мовних моделей від Google. Точний розмір моделі Gemini не розголошується, але, за деякими оцінками, він може містити близько 540 мільярдів параметрів. Однією з ключових особливостей є те, що Gemini може отримувати доступ до актуальної інформації з інтернету через інтеграцію з Google Search, що робить його відповіді більш релевантними та фактологічними.

Збільшено швидкість обробки Gemini 1.5, що дозволяє отримувати відповіді на запити практично миттєво. Це особливо помітно під час роботи з великими обсягами даних. Модель стала економічнішою щодо споживання обчислювальних ресурсів, що дозволяє запускати їх у ширшому спектрі пристроїв. Gemini 1.5 здатна краще розуміти нюанси людської мови, включаючи сарказм, іронію та інші тонкощі. Модель спроможна краще впоратися зі складними логічними завданнями та вибудовувати причинно-наслідкові зв'язки. Gemini 1.5 може генерувати тексти в різних стилях від формального до розмовного, що робить її більш універсальною.

Модель навчилася працювати не лише з текстом, а й із зображеннями, відео та аудіо. Це відкриває нові можливості для створення креативних та інтерактивних програм.

Модель може бути адаптована під конкретні завдання, що дозволяє створювати спеціалізовані рішення для різних сфер діяльності. Gemini 1.5 є більш потужним, гнучким і універсальним інструментом, ніж його попередник Gemini 1. Він відкриває нові горизонти для застосування штучного інтелекту в різних областях, від створення контенту до розробки інтелектуальних помічників.

Claude AI, Anthropic

В основу моделі Claude покладено підхід Constitutional AI, що закладає у модель певні принципи, правила та обмеження на етапі навчання. Технічно це реалізовано через механізм навчання із підкріпленням на основі зворотного зв'язку від людини. Замість класичного навчання з вчителем на розмічених даних тут модель взаємодіє з реальними інженерами, отримує від них оцінки своїх дій і коригує свою поведінку, щоб максимізувати нагороду.

Інженер запитує, модель генерує кілька варіантів відповіді, які оцінюються за різними критеріями: релевантність, безпека, дотримання інструкцій тощо. На основі оцінок модель оновлює свої параметри, вчиться видавати відповідні відповіді. Дана модель не просто механічно генерує текст, а враховує персональні уподобання, дотримується заданих принципів, демонструє навички здорового глузду та етики. Саме тому Claude не генеруватиме небезпечний чи образливий контент, навіть якщо його про це просять - модель навчена цього не робити.

Claude має модульну архітектуру, де замість однієї великої моделі використовується сімейство моделей - Opus, Sonnet, Haiku, які оптимізовані під певний клас завдань. Технічно вони відрізняються кількістю параметрів, набором навчальних даних, але збережено основні принципи – трансформерна архітектура, навчання з підкріпленням і зворотній зв’язок з людиною.

Claude 3.5 Sonnet - це потужна та інноваційна модель, яка пропонує унікальні можливості для автоматизації та покращення багатьох процесів. Дана модель навчається на величезному ретельно підібраному наборі тексту та коду, щоб представляти широкий спектр тем і стилів. Спроможна працювати з довгим контекстом – до 200 000 токенів, що надає можливість аналізувати та враховувати інформацію з об'ємних документів, статей, діалогів. Модель виконує широкий спектр завдань обробки природної мови, включаючи художні або технічні тексти високого рівня, аналітика, дослідження та питання-відповіді, навіть якщо вони відкриті, складні або дивні.

До інтерфейсу моделі додано функцію Artifacts, яка дозволяє користувачам переглядати та редагувати вміст, згенерований Claude, такий як фрагменти коду, текстові файли та дизайн веб-сайтів, поряд із вікном діалогу. Цей динамічний робочий простір дозволяє інтегрувати згенерований вміст у проекти та робочі процеси користувачів.

Можливості та обмеження великих мовних моделей

Модель має три основні особливості, що перевершує GPT-4o у більшості завдань. Claude 3.5 показує найвищу продуктивністю при виконанні завдань машинного зору, вдвічі вищу швидкість генерації відповідей, зручний інтерфейс користувача для таких завдань, як генерація коду та анімація.

Claude пропонує значно безпечніший підхід до використання штучного інтелекту. На його мовні моделі накладається більше обмежень щодо жорстких етичних норм. Однак, є й певні обмеження - модель не може безпосередньо взаємодіяти з інтернетом, відкривати посилання чи мультимедіа.

На підставі проведеного дослідження визначено основні характеристики останніх версій мовних моделей та проведено їх порівняння у таблиці 1.

Таблиця 1.

Порівняльна таблиця характеристик мовних моделей на базі архітектури трансформер

Метрика	GPT-4o	GPT-o1	Gemini 1.5	Claude 3.5
Розмір моделі	~ 200 млрд. параметрів	~ 2 трлн. параметрів	~ 540 млрд. параметрів	~ 2 трлн. параметрів
Архітектура	Трансформерна з InstructGPT	Модульна трансформерна, навчання з підкріпленням	Трансформерна	Модульна трансформерна, навчання з підкріпленням
Дані для навчання	Веб-дані, що були актуальні рік до поточної дати	Веб-дані, що були актуальні рік до поточної дати	Онлайн-джерела, з доступом до Google Search	Веб-дані, що були актуальні півроку до поточної дати
Контекстне вікно	128 000 токенів	128 000 токенів	~ 1 млн. токенов	200 000 токенів
Можливості	Широкий спектр завдань NLP	Завдання NLP, робота з документами, обробка зображень, інтеграція з зовнішніми інструментами	Завдання NLP, комп'ютерний зір, розуміння аудіоповідомлень	Завдання NLP, комп'ютерний зір, робота з документами, генерація програмного коду
Обмеження	Статичні знання, розуміння лише тексту, можлива упередженість	Може вигадувати факти та робити помилки у роздумах	Потенційна упередженість, недостатня контекстуалізація	Можлива упередженість
Наявність API	Так	Так	Так	Так
Етика та безпека	Заходи контролю від OpenAI	Заходи контролю від OpenAI	Заходи контролю від Google	Особливий акцент на етичність від Anthropic
Сфери застосування	Художні та технічні тексти, змістовні діалоги, програмування	Розуміння запитів, перетворення ідеї користувача на докладний запит, розпізнавання та генерація зображень.	Науково-технічна документація, аналітика, генерація програмного коду	Академічні роботи, дослідження, тексти високого рівня, програмний код, код інтерфейсів

7. Можливості та обмеження ВММ

Незважаючи на революційний потенціал, ВММ мають певні обмеження та верхню планку можливостей. Розуміння цього є ключовим для ефективного використання всієї потужності мовних моделей. Серед широкого спектру потенційних можливостей ВММ можна виділити кілька найважливіших.

По-перше, це розуміння природної мови. ВММ відрізняються в здатності інтерпретувати людську мову в різних областях та демонструють високі результати у генерації зв'язного тексту, що спирається на контекст розмови. Вони можуть витягати інформацію, робити логічні висновки та адекватно реагувати на різноманітні запити.

По-друге, використовуючи великі бази знань, ВММ вміють генерувати різні творчі текстові формати, програмний код, сценарії, електронні листи та загалом аналізувати складну інформацію і

виявляти закономірності та тенденції, які можуть бути невидимі для людей. Моделі виконують завдання, такі як побудова графів знань, семантичний пошук, і виявлення та оцінка ризиків. Це робить їх цінними інструментами для досліджень, аналітики та прийняття рішень.

По-третє, мовні моделі можуть використовуватися для створення персоналізованого контенту, який відповідає інтересам та потребам конкретного користувача. Ця адаптованість надає змогу отримувати індивідуальні рекомендації, персоналізованих помічників та доцільний контент. Google використовує ВММ для персоналізації результатів пошуку та рекомендацій продуктів для користувачів, а Netflix надає користувачам поради щодо фільмів та серіалів на основі їх історії переглядів та інших даних [1].

По-четверте, ВММ спроможні до автоматизації завдань, які раніше вимагали багато ресурсів та займали багато часу для людей. Використання мовних моделей призводить до значної економії часу та коштів, а також розвантажує користувачів для виконання більш творчої та стратегічної роботи. Провідні корпорації світу використовують ВММ для допомоги користувачам: відповідати на запитання користувачів, відтворювати музику, керувати розумними будинками тощо.

Незважаючи на значний прогрес в розробці ВММ, також існують певні обмеження, які необхідно враховувати і знати задля досягнення найбільш ефективної взаємодії [10].

Обмеження 1. Навчальні дані для великих мовних моделей

Мовні моделі навчаються на великому обсязі текстів, зібраних з різних джерел, таких як інтернет, цифрові книги, новини. Ці тексти містять багато інформації про різні аспекти світу, такі як факти, думки, правила, жарти тощо. Знання, що витягнуті з навчальних текстів використовують для генерації нових текстів.

Проте, є кілька проблем із цим підходом:

- Тексти не завжди є достовірними, актуальними чи узгодженими між собою. Моделі можуть переймати помилки, застарілі дані чи суперечливі твердження з джерел. Це може призводити до генерації неправдиву або оманливої інформації.
- Тексти не завжди є повними чи представницькими. ВММ можуть ігнорувати чи недооцінювати важливість деяких тем чи груп людей, які мало представлені у їхніх даних. Це може призводити до того, що моделі виявляють дискримінацію стосовно них.
- Тексти не завжди є єдиним чи найкращим способом передачі знань. ВММ можуть не враховувати чи втрачати нюанси, контекст чи емоції, що можуть бути виражені через інші модальності, такі як зображення, звуки чи жести. Це може призводити до того, що моделі неправильно інтерпретують чи виражають сенс.

Підходи до подолання

- Покращити якість та різноманітність даних для навчання мовних моделей. Наприклад, можна використовувати методи фільтрації, перевірки фактів або анотації даних для усунення помилок чи невідповідностей у текстах. Можна використовувати методи аугментації даних для розширення тематичного складу текстів.
- Покращити способи оцінки та контролю знань у мовних моделях. Наприклад, можна використовувати методи тестування, аналізу чи візуалізації знань для перевірки їх правильності, актуальності та узгодженості. Можна використовувати методи інтерпретації для розуміння логіки виведеного контенту.
- Покращити способи інтеграції та узагальнення знань у мовних моделях. Наприклад, можна використовувати методи мультимодального навчання, об’єднання чи адаптацію знань із різних джерел. Можна використовувати методи мета-навчання для швидкого чи ефективного вивчення нових знань.

Обмеження 2. Розуміння сенсу

ВММ навчаються на основі статистичних закономірностей у мові, які вони знаходять у навчальних даних. Вони навчаються передбачати наступне слово або фразу в тексті на основі попереднього контексту. Це дозволяє генерувати зв'язані та правдоподібні тексти.

Проте, є кілька проблем із цим підходом:

- Статистичне передбачення не означає семантичне розуміння. Моделі можуть не враховувати

значення або наслідки поєднання слів. Це може призвести до генерації безглуздих, нерелевантних або небезпечних текстів.

- Статистичне передбачення не означає логічний висновок. Моделі можуть не перевіряти правила, факти чи докази у своїх текстах. Це може призводити до генерації суперечливих чи оманливих текстів.
- Статистичне передбачення не означає творче вираження. Моделі можуть не розвивати власний стиль чи перспективу у згенерованих текстах. Це може призводити до генерації повторюваних або плагіатних текстів.

Підходи до подолання

- Покращити способи вимірювання та підвищення розуміння мови. Наприклад, можна використовувати методи оцінювання або змагань для перевірки та порівняння здібностей мовної моделі до розуміння мови на різних рівнях (слова, фрази, речення, тексти) та в різних завданнях (класифікація, аналіз тональності, відповіді на питання тощо). Можна використовувати методи навчання з підкріпленням, активного навчання або інтерактивного навчання для покращення розуміння сенсу через зворотний зв'язок від користувачів чи розробників.
- Покращити способи використання логіки та здорового глузду в мовних моделях. Наприклад, можна використовувати методи символьного обчислення та семантичних мереж для поєднання статистичного та логічного подання знань. Можна використовувати методи впровадження, збагачення чи перевірки знань для уточнення логічних чи загальнозначущих фактів.
- Покращити способи заохочення та оцінки творчості і оригінальності мовної моделі. Наприклад, можна використовувати методи генерації різноманітності та новизни для вимірювання чи покращення творчого потенціалу. Можна використовувати методи персоналізації та адаптації для розвитку стилю мовної моделі.

Обмеження 3. Взаємодія з людиною

ВММ можуть бути використані для створення або покращення сервісів, що пов'язані з мовою: пошукові системи, чат-боти, асистенти, редактори. Вони можуть спілкуватися з людьми через текстові або голосові інтерфейси, надаючи їм різноманітну інформацію.

Проте, є кілька проблем:

- ВММ не завжди задовольняють потреби чи очікування людей. Вони можуть не розуміти або не враховувати контекст, цілі та настрій співрозмовників. Це може призвести до генерації невідповідних, небажаних чи образливих текстів.
- ВММ не завжди дотримуються або сприяють етичним, соціальним, юридичним нормам та цінностям. Вони можуть порушувати права, свободи, інтереси користувачів і генерувати шкідливі, небезпечні чи незаконні тексти.
- ВММ не завжди відповідають за свої дії чи наслідки. Мовні моделі можуть не усвідомлювати чи не визнавати свою відповідальність за свої слова. Це може призводити до генерації помилкових, маніпулятивних або злочинних текстів.

Підходи до подолання

- Покращити способи вивчення та задоволення потреб і очікувань людей від використання ВММ. Наприклад, можна використовувати методи аналізу користувачів для визначення характеристик, цілей та переваг різних груп користувачів. Можна використовувати методи зворотного зв'язку від користувачів, покращення реакції та оцінки користувачів на згенеровані тексти.
- Покращити способи дотримання та підтримки етичних, соціальних, юридичних норм та цінностей. Наприклад, можна використовувати методи аналізу етики, соціальної справедливості або правового регулювання для визначення та врахування основних принципів, правил та стандартів, яким повинні дотримуватися ВММ. Можна використовувати методи аудиту, моніторингу для перевірки та усунення порушень, пов'язаних із текстами ВММ.
- Покращити способи прийняття та несення відповідальності за дії та наслідки, що спричинені ВММ. Наприклад, можна використовувати методи ідентифікації для визначення та підтвердження джерела, авторства чи справжності текстів. Можна використовувати методи пояснення для надання та обґрунтування мотивів чи цілей текстів ВММ.

8. Результати дослідження

Рішення щодо використання ВММ для компаній приймається індивідуальною, виходячи з поставлених цілей. Але, варто знати, як популярні типи моделей можуть змінити операційну спроможність і робочий процес.

Загалом, існує дві категорії великих мовних моделей – з відкритим та закритим вихідним кодом. Продукти з відкритим вихідним кодом розвиваються завдяки колективним зусиллям щодо їх покращення та модифікації, тоді як альтернативи найчастіше є комерційними за своєю природою. Варто також розрізняти типи моделей щодо їх призначення.

- Загального призначення - це стандартні рішення, які можуть виконувати багато завдань з різних сфер і навчаються на доступних текстах в інтернеті.
- Багатомовні – такі моделі підтримують багато мов і ефективні у виконанні багатомовних завдань.
- Специфічні завдання - моделі, що адаптовані для успішної роботи у певній області, наприклад онлайн переклад.
- Галузеві - ці моделі краще допомагають у конкретних сферах. Вони спеціально навчені на сайтах, присвячених охороні здоров'я, фінансам і надають більш поглиблені відповіді.
- Few-shot — такі моделі не потребують багато даних для надання відмінних результатів. Вони спеціально розроблені для навчання на невеликій кількості інформації.

У таблиці 2 наведено характеристики, які допоможуть обрати доцільну модель для втілення у новий сервіс.

Таблиця 2.

Характеристики ВММ щодо використання у новому сервісі

Модель	Унікальні особливості	Сильні сторони	Обмеження
GPT-4o	Висока ефективність навчання	Глибокий аналіз різних форматів даних, включаючи графіки та зображення	Етичні проблеми та упередженість даних*
GPT-o1	Перевершує попередню версію щодо креативного потенціалу	Високі здібності самонавчання під час спілкування з користувачами.	Етичні проблеми та упередженість даних*
Gemini 1.5	Контекстна логіка та легка модифікація запитів. Пошук інформації в Інтернеті	Експорт відповідей до сторонніх додатків і програм. Наявність нативних інструментів генерації зображень і мови.	Ризик упередженого самонавчання
Claude 3.5	Стеження за потоком діалогу та спроможність вести розлогі розмови. Відмінні результати в задачах генерації програмного коду.	Слідування правовим вимогам та етичним стандартам.	Невисока ефективність у створенні певного контенту

* Дані, які використовуються для навчання моделей або прийняття рішень, містять систематичні помилки або перекоси, що може призвести до викривлень у висновках або результатах моделі.

Для використання ВММ у побудові нового сервісу або втіленні у розроблений продукт, важливо враховувати, що ВММ може згенерувати все, що завгодно. Відповіді сучасних моделей не проходять перевірку на достовірність, тому можна отримати у відповідь недостовірну, небезпечну або токсичну інформацію. Ця проблема отримала назву AI-галюцинацій і привертає увагу провідних вчених.

- Використання інформації, що не пройшла перевірку, може призвести до великих репутаційних втрат.
- Недостовірна інформація, що створена за допомогою ВММ, може бути причиною судових позовів.
- Люди схильні довіряти відповідям ВММ і це створює загрози у сфері кібербезпеки.

Щоб прийняти рішення про використання мовної моделі у новому продукті, потрібно переконатися, що згенерована неправдоподібна, хибна або токсична інформація буде безпечною для користувача. Якщо існує ризик небезпечності, тоді краще утриматися або продумати і реалізувати необхідні правила фільтрації такого контенту, щоб мінімізувати ризики.

Можливості та обмеження великих мовних моделей

Для використання ВММ у продуктах та сервісах найбільш безпечними підходами буде додаткова обробка відповідей моделі, що використовується як асистент-помічник для користувача, що мають усвідомлювати обмеження цієї технології.

Отже, розуміння цих факторів і активні дослідження дозволять розкрити повний потенціал ВММ, зменшуючи їх вроджені обмеження і створюючи умови для більш відповідального користування.

9. Висновки

Досліджено теоретичні засади великих мовних моделей, різні типи архітектур ВММ, зокрема трансформери, рекурентні нейронні мережі та гібридні підходи. Подано узагальнений опис алгоритму функціонування мовних моделей, кроків навчання та оцінки результатів. Наведено способи додаткового налаштування моделі для отримання більш якісних результатів.

Окрему увагу приділено популярним чат-ботам на базі архітектури трансформерів, а саме ChatGPT, Gemini AI та Claude AI. Проведено порівняння їхніх характеристик, яке дозволило визначити сильні та слабкі сторони кожного чат-бота.

Дослідження виявило, що ВММ мають значний потенціал у розумінні природної мови, генеруванні контенту, аналітиці, персоналізації та автоматизації завдань. Однак їм також притаманні певні обмеження, такі як висока вартість створення, залежність від великих обсягів даних, упередженість та етичні проблеми. Тому, подальші дослідження та розробки у сфері ВММ мають бути спрямовані на подолання цих викликів, щоб забезпечити відповідальне та ефективне використання таких технологій.

Як результат дослідження подано рекомендації щодо втілення ВММ у новий сервіс або інтеграції зі сторонніми програмами. Наведено порівняння популярних моделей щодо їх особливостей для використання у створенні нового сервісу або інтеграції зі сторонніми програмами.

Загалом, великі мовні моделі є потужним інструментом в обробці природної мови та створенні діалогових систем. Однак їх розвиток та застосування вимагають ретельного підходу, врахування етичних аспектів та безперервного вдосконалення для розкриття їхнього повного потенціалу.

Список літератури

1.Alessandro Berti, Humam Kourani, Hannes Hafke, Chiao-Yun Li, Daniel Schuster (2024) Evaluating Large Language Models in Process Mining: Capabilities, Benchmarks, and Evaluation Strategies https://doi.org/10.48550/arXiv.2403.06749.

2.Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation.(2014) " Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. https://doi.org/10.3115/v1/D14-1162

3.Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, Shahrokh Valaee. Recent Advances in Recurrent Neural Networks (2017). https://doi.org/10.48550/arXiv.1801.01078

4.Wang, Chenguang, Mu Li, and Alexander J. Smola. "Language models with transformers." arXiv preprint arXiv:1904.09408 (2019). https://doi.org/10.48550/arXiv.1904.09408

5.OpenAI, URL: https://platform.openai.com/docs/introduction, (Accessed: 13 September 2024).

6.Google AI, URL: https://ai.google.dev/gemini-api/docs/model-tuning, (Accessed: 13 September 2024).

7.Anthropic, URL: https://docs.anthropic.com/claude/docs/intro-to-claude, (Accessed: 13 September 2024).

8.T. Brown, B. Mann, N. Ryder "Language models are few-shot learners." (2020) arXiv preprint arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

9.Artificial intelligence risk management framework (2023) https://doi.org/10.6028/NIST.AI.100-1.

10. Laura Weidinger, John Mellor, Maribeth Rauh. Ethical and social risks of harm from Language Models (2021) https://doi.org/10.48550/arXiv.2112.04359

І.Ю. Юрчак, О.О. Кичук, В.М.Оксентюк, А.О.Хіч

CAPABILITIES AND LIMITATIONS OF LARGE LANGUAGE MODELS

I.Yu.Yurchak, O.O. Kychuk, V.M. Oksentyuk, A.O. Khich

Lviv Polytechnic National University,
Department of "Computer Design Systems"
E-mail: Iryna.Y.Yurchak@lpnu.ua, Olha.Kychuk.knm.2020@lpnu.ua,
Vira.M.Oksentyuk@lpnu.ua, Andrii.O.Khich@lpnu.ua

The work is dedicated to the study of large language models (LLMs) and approaches to improving their efficiency in a new service. The rapid development of LLMs based on transformer architecture has opened up new possibilities in natural language processing and the automation of various tasks. However, fully utilizing the potential of these models requires a thorough approach and consideration of numerous factors.

A review of the evolution of large language models was conducted, highlighting leading companies engaged in the research and development of efficient systems. The structure of these models and ways of representing internal knowledge were examined. Key approaches to training were described, including data collection, preprocessing, and selecting appropriate neural network architectures used in large language models. It was noted that the greatest breakthrough was achieved with the Transformer neural network, which is based on the attention mechanism.

A comparison of popular transformer-based chatbots was presented, namely: ChatGPT, Claude AI, and Gemini AI. Their metrics, capabilities, and limitations were identified.

The relevance of the topic lies in the rapid development of natural language processing technologies and the growing demand for large language models across various industries. The effective use of these models has tremendous potential to improve productivity and the quality of work with textual data. However, due to the complexity of the architecture and the large amounts of data required for training, selecting and configuring the optimal model for a specific task is a challenging process.

As a result of the study, recommendations for developers were provided on the use of popular open-source models in the new service or integration with third-party programs. The characteristics of the models, their strengths, limitations, and certain caveats regarding trust in the generated results were indicated.

Keywords: large language models, transformer architecture, neural networks, chatbot, content generation.