КРИТЕРІЇ ОЦІНЮВАННЯ ЯКОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Юрій Хома, д.т.н., доцент
Національний Університет "Львівська Політехніка", Україна,
e-mail: yurii.v.khoma@lpnu.ua
Іван Щудло, аспірант
Національний Університет "Львівська Політехніка", Україна,
e-mail: ivan.o.shchudlo@lpnu.ua

Анотація

Розвиток великих мовних моделей (LLM - large language models) з кожною новою ітерацією демонструє суттєве вдосконалення їх здатності до розуміння та генерування тексту, що відкриває все ширші можливості для їх інтеграції в системи обробки інформації та цифрові бізнес-процеси підприємств та установ. В умовах постійного зростання складності та функціональних можливостей LLM, розробка надійних методик їх оцінювання стає фундаментальним викликом для дослідницької спільноти, оскільки традиційні метрики оцінювання текстової інформації часто не здатні повністю охопити всю глибину та багатогранність їхніх потенційних спроможностей та характеристик. Створення комплексних систем оцінювання якості LLM покликане забезпечити не лише об'єктивне порівняння різних моделей, але й сформувати критичний зворотний зв'язок для цілеспрямованого вдосконалення технологій та запобігання потенційним ризикам, пов'язаним із їх широкомасштабним впровадженням. Стаття присвячена систематизації критеріїв оцінювання якості великих мовних моделей, що набули широкого поширення в задачах обробки природної мови. Метою статті є створення комплексного підходу до оцінювання LLM, що охоплює основні аспекти їх функціонування. У роботі визначено критерії оцінювання LLM, такі як прецизійність та змістовність відповідей, природність мовлення, послідовність, токсичність, заангажованість, безпекові вразливості, тощо. Детально проаналізовано три основні підходи до оцінки якості LLM: експертні оцінки, порівняння з еталонними даними та автоматизовані методи без еталонних даних. Для кожного критерію якості визначено найефективніші методи оцінювання, зазначено їхні переваги та недоліки в різних контекстах застосування. У підсумку підкреслено, що, попри високу надійність експертних оцінок, автоматизовані методи стають усе важливішими для масштабної оцінки LLM, особливо для суб'єктивних критеріїв типу токсичності чи заангажованості.

Ключові слова

Велика Мовна Модель, Оцінювання Якості Великих Мовних Моделей, Критерії Оцінювання Якості.

1. Вступ

LLM (LLM, англ. "Large Language Models") це штучні нейронні мережі які здійснюють обробку природної мови, приймають на вхід текстову інформацію, або аудіо сигнал з усним мовлення у випадку найновіших мультимодальних LLM. На виході LLM генерує текст, прогнозуючи наступні слова чи фрази, які найбільш імовірно відповідають заданому контексту. Це може включати відповіді на запитання, створення описів, або формулювання рекомендацій.

Промпти (від англ. "prompts") в контексті великих мовних моделей — це текстові запити або інструкції, які вводяться в модель для отримання бажаних відповідей або інформації. Вони можуть варіюватися від простих запитів до складних структурованих інструкцій, що вказують моделі, як саме слід взаємодіяти з вхідними даними. Системні промпти — це специфічні типи промптів, які задають загальні правила або налаштування для взаємодії з моделлю. Вони можуть включати в себе інструкції або параметри, які регулюють, як модель повинна реагувати на запити.

RAG (англ. Retrieval-Augmented Generation) - це підхід до побудови програмних рішень, який поєднує пошук релевантної інформації з зовнішніх джерел даних з генеративними можливостями LLM [1]. Архітектура RAG складається з трьох основних компонентів:

- векторного сховища для зберігання та індексації документів,
- системи семантичного пошуку для знаходження релевантної інформації
- механізму інтеграції знайденої інформації з генеративною моделлю через доповнений промпт.

Основними перевагами RAG є підвищена точність відповідей завдяки використанню актуальних зовнішніх джерел, зменшення галюцинацій моделі, та можливість роботи з приватними корпоративними даними. До технічних обмежень RAG відносять залежність від якості індексованих даних, обмеження розміру контекстного (вхідного) текстового вікна та додаткові обчислювальні витрати на пошук і обробку зовнішньої інформації. Практичне застосування RAG систем охоплює широкий спектр задач, включаючи корпоративні системи підтримки користувачів, аналіз документації, автоматизацію бізнес-процесів та наукові дослідження, де критично важлива точність та актуальність інформації.

Великі мовні моделі (LLM) мають два основні типи застосування на основі характеру задач:

- структуровані відповіді, які включають чітко форматований вивід (наприклад, заповнення форм, класифікація тексту, або витяг конкретних даних з документів) [2]
- неструктуровані відповіді, що передбачають генерацію вільного тексту (як-от написання статей, створення описів, або ведення діалогу).

В контексті роботи з запитами LLM моделі можуть обробляти:

- конкретні (closed-ended) запити, які вимагають точних, фактичних відповідей (наприклад, "Хто написав 'Кобзар'?" або "Яка столиця Франції?"), та
- відкриті (open-ended) питання, що потребують розгорнутих, аналітичних відповідей з елементами міркування та синтезу інформації (наприклад, "Як штучний інтелект може вплинути на майбутнє освіти?").

Структуровані відповіді часто використовуються в бізнес-процесах, де потрібна стандартизація та автоматизація обробки даних, тоді як неструктуровані відповіді більш поширені в творчих та аналітичних задачах. При роботі з конкретними запитами LLM демонструють високу точність та надійність, оскільки такі запити мають чітко визначені правильні відповіді, які можна верифікувати. Відповіді на відкриті питання, навпаки, вимагають від моделі більшої гнучкості, здатності до аналізу та синтезу інформації, а також вміння генерувати обґрунтовані судження та висновки. Важливим аспектом застосування LLM є їхня здатність адаптуватися до різних контекстів та форматів взаємодії, що дозволяє використовувати їх як універсальний інструмент для широкого спектру завдань - від простої класифікації до складного аналізу та генерації контенту.

Оцінка якості великих мовних моделей являє собою комплексний процес, що охоплює аналіз їхньої здатності до розуміння контексту, генерації релевантного та корисного контенту, розмірковування (англ. reasoning), вирішення проблем, а також етичної надійності та безпеки [3]. Систематизація критеріїв оцінювання якості великих мовних моделей є критично важливою для забезпечення об'єктивного порівняння моделей та визначення їхніх переваг і недоліків у різних прикладних аспектах таких як безпека, етичність, змістовність. Сучасний стан проблеми характеризується фрагментарністю підходів до оцінювання, де використовуються різноманітні еталонні метрики “бенчмарки” (від англ. benchmark), але відсутня єдина комплексна методологія, що охоплювала б усі значущі аспекти функціонування LLM. Дослідницька спільнота активно працює над створенням багатовимірних систем оцінювання, таких як HELM (від англ. Holistic Evaluation of Language Models) [4], які намагаються інтегрувати показники точності, обґрунтованості, безпеки, калібрування впевненості та соціального впливу моделей. Труднощі стандартизації оцінювання посилюються швидкою еволюцією можливостей мовних моделей, що вимагає постійного оновлення “бенчмарок” для запобігання їх застаріванню та втраті дискримінаційної здатності при оцінювання моделей різної якості. Наприклад в попередніх поколіннях генеративних мовних моделей, така метрика як кількість помилок орфографії та пунктуації на 1000 символів могла б бути інформативним способом оцінювання якості моделі, але з появою LLM вона втрачає дискримінаційну здатність оскільки всі LLM генерують тексти без помилок і на всіх текстах очікувані результати якості по цьому критерію стабільно сягнули 100%. Створення комплексної, стандартизованої та динамічної системи оцінювання LLM залишається відкритим науковим викликом, вирішення якого матиме суттєвий вплив на розвиток галузі та практичне застосування цих технологій.

2. Критерії оцінки якості великих мовних моделей (LLM)

Як зазначалося вище, першим етапом для на шляху уніфікації оцінювання LLM є формування відповідних критеріїв оцінювання, які б змогли покрити багатовимірну природу функціональних можливостей і сфер застосування цих моделей. У цій роботі розглянуто та систематизовано відповідні критерії.

1. **Прецизійність згенерованої відповіді (Accuracy).** Критерій прецизійності обчислює відсоток правильних відповідей у таких завданнях, як класифікація чи відповідь на питання [5]. Сюди також можна включити такі метрики як:
    - **Повнота (recall)**: це відношення кількості правильно виявлених позитивних відповідей до загальної кількості фактичних позитивних відповідей у вибірці.

- **F1 score**: це гармонійне середнє показників precision (точності) та recall (повноти), яке надає збалансовану оцінку якості класифікації або передбачення моделі машинного навчання.
  Цей критерій застосовується для оцінки відповідей LLM на конкретні питання, а також для оцінки відповідей LLM в структурованому форматі.

Приклад питання: "У якому році народився Тарас Шевченко?"
Правильна відповідь LLM: "1814"

2. **Змістовність згенерованої відповіді.**
   Оцінка якості згенерованого тексту в LLM включає аналіз когерентності, семантичної зв'язності та відповідності контексту, який здійснюється через порівняння генерованого тексту з еталонними зразками або експертними оцінками. Додатково використовуються кількісні метрики, такі як BLEU, ROUGE [6] та перплексія, які дозволяють кількісно виміряти точність, інформативність та природність створеного машиною тексту.
   Цей критерій застосовується для оцінки відповідей LLM на відкриті питання, де відповідь може включати один або декілька абзаців.
   Приклад питання: "Чим відомий Тарас Шевченко?"
   Якісна відповідь LLM: "Тарас Шевченко відомий як великий український поет, який писав вірші рідною мовою і боровся за права простого народу. Він також був талановитим художником і створив багато картин та малюнків, що зображували життя українського народу."
3. **Природність мовлення (Natural Language).**
   Природність мовлення визначає здатність моделі генерувати граматично правильний та стилістично доречний текст, який важко відрізнити від написаного людиною. Це включає правильне використання сленгу, ідіом та контекстуально-залежних виразів, коли це доречно. Модель має адаптувати свій стиль спілкування відповідно до контексту - від формального до розмовного. Важливим аспектом є уникнення механічних, штучних конструкцій та здатність передавати емоційні нюанси через відповідний тон та лексику [7].
4. **Послідовність відповідей (Consistency)**
   Здатність моделі надавати однакові або логічно узгоджені відповіді на подібні запитання, навіть якщо вони сформульовані по-різному. Важливим аспектом є збереження стабільності тверджень моделі протягом усього діалогу, без раптових змін позиції чи суперечливих заяв [8]. Послідовність також включає здатність дотримуватися встановлених обмежень та інструкцій впродовж всієї взаємодії. Цей критерій особливо важливий для довготривалих діалогів та завдань, що вимагають послідовного застосування певних правил чи методології.
5. **Токсичність (Toxicity).**
   Токсичність вказує на наявність образливих, ненависницьких, расистських, сексистських або інших неприйнятних висловлювань у відповіді моделі. Оцінка токсичності є важливою для забезпечення етичного використання LLM. Необхідно мати механізми для виявлення і мінімізації таких висловлювань[9].
6. **Заангажованість (Bias).**
   Заангажованість є потенційною проблемою, яка може виникати в результаті навчання на нерепрезентативних або упереджених даних. Оцінка наявності заангажованості у відповідях моделі важлива для забезпечення справедливості і рівності в комунікації. Наприклад, модель повинна уникати стереотипів стосовно різних соціальних, культурних чи етнічних груп[10].
7. **Безпекові вразливості**
   Вразливості стосуються недоліків системи, які можуть бути використані для маніпуляцій або обману. Це може включати небажане розкриття моделлю чутливих даних, а також наявні вразливості що дозволяють обходити обмеження які накладають на модель системні промпти. Оцінка наявності таких вразливостей є ключовою для розробки безпечних LLM [11].

Релевантність деяких із зазначених вище критеріїв залежить від самої задачі, наприклад при оцінюванні системи пошуку документів для банківських чи юридичних установ, основними критеріями виступатимуть безпекові вразливості та заангажованість, тоді як для голосового бота в ресторані швидкої їжі найважливішими будуть природність мовлення та послідовність, а для агентської системи допомоги програмістам – безпекові вразливості та послідовність.

3. Підходи до оцінки якості відповідей LLM

Наступним кроком після визначення і категоризації конкретних критеріїв оцінювання, є систематизація підходів на основі яких цим критеріям можна присвоїти відповідні числові значення. Загалом способи оцінки якості LLM можна поділити на три категорії:

1. Використання експертних оцінок - "ручний" спосіб

Оцінювання якості LLM за допомогою експертних оцінок передбачає аналіз відповідей моделей фахівцями для визначення їхньої ефективності. В свою чергу, способи експертних оцінок можна поділити на індивідуальні та групові, а також на оцінювання з вказівками і без них.

Індивідуальне експертне оцінювання залучає одного спеціаліста, який оцінює відповіді за визначеними критеріями, забезпечуючи методологічну послідовність, швидкість проведення оцінки та економічну ефективність процесу. Даний підхід є оптимальним для оцінки об'єктивних критеріїв, таких як фактична точність, відповідність запиту та дотримання конкретних інструкцій, де існує чітка межа між правильними та неправильними відповідями.

Групове експертне оцінювання залучає декількох фахівців, що дозволяє досягти вищої надійності результатів завдяки усередненню індивідуальних розбіжностей та статистичному визначенню узгодженості між експертами. Для оцінки суб'єктивних критеріїв, таких як етичність контенту, стилістична відповідність та загальна корисність, групове оцінювання є більш методологічно обґрунтованим.

Оцінювання з вказівками передбачає надання експертам детальних інструкцій із чіткими дескрипторами для кожного рівня оцінки, що підвищує узгодженість між ними та зменшує вплив суб'єктивних інтерпретацій критеріїв.

Натомість, **оцінювання без вказівок** дозволяє експертам застосовувати власні критерії та підходи, сприяючи виявленню непередбачених аспектів якості роботи моделі, хоча й ускладнює кількісне порівняння різних систем. Сучасні методології на різних етапах дослідження часто комбінують індивідуальне та групове оцінювання, з та без вказівок, з метою забезпечення комплексності аналізу.

Адаптивні методи оцінювання передбачають поступове уточнення критеріїв на основі зворотного зв'язку від експертів та аналізу проміжних результатів. Для підвищення надійності експертних оцінок рекомендується проводити калібрувальні сесії перед початком основного етапу оцінювання. Дедалі популярнішим стає залучення експертів із різних галузей знань для формування міждисциплінарних панелей, здатних оцінити широкий спектр різних аспектів функціонування LLM. Застосування статистичних методів, таких як коефіцієнт Кендалла або коефіцієнт Спірмена[12], дозволяє квантифікувати узгодженість між експертами та перевірити надійність оцінок.

2. Порівняння з еталонними даними - напів-автоматизований спосіб

Оцінювання якості LLM через порівняння з еталонними відповідями є напівавтоматичним методом, оскільки вимагає початкового створення бази еталонних відповідей експертами, що є трудомістким ручним процесом, який забезпечує еталон для подальшого автоматизованого порівняння. Ефективність цього методу залежить від якості та репрезентативності зібраних еталонних даних, оскільки обмежений набір зразків може призвести до хибної оцінки здатності моделі до узагальнення та роботи з новими типами запитів. Попри певні обмеження, порівняння з еталонами залишається важливим компонентом комплексного оцінювання LLM, особливо для завдань із чітко визначеними критеріями правильності відповіді.

Оцінювання якості великих мовних моделей (LLM) через порівняння з еталонними відповідями можна поділити на два підходи: бінарне та небінарне оцінювання.

Бінарне оцінювання застосовується переважно для структурованих відповідей, де модель повинна надати конкретну інформацію в чітко визначеному форматі, наприклад, при витяганні фактів, класифікації, або розв'язанні математичних задач з єдиною правильною відповіддю. При бінарному оцінюванні відповідь моделі вважається правильною лише у випадку повного співпадіння з еталоном, що дозволяє об'єктивно визначити точність моделі та розрахувати метрики на кшталт accuracy, precision та recall.

Проте бінарне оцінювання має суттєві обмеження при роботі з генеративними моделями, які можуть формулювати коректні відповіді різними способами, що призводить до необхідності застосування небінарних методів оцінювання.

Небінарне оцінювання включає широкий спектр методів для вимірювання часткової відповідності між відповіддю моделі та еталоном, починаючи від простих лексичних метрик і завершуючи складними семантичними порівняннями. Прості метрики лексичної подібності, такі як відстань Левенштейна, BLEU, ROUGE або METEOR [5], вимірюють подібність на рівні символів або n-грам, але не враховують семантичну

еквівалентність і контекстуальні відмінності. Більш складні методи, засновані на семантичному аналізі, використовують векторні представлення текстів (англ. vector embeddings) для порівняння змістовної подібності, навіть коли формулювання суттєво відрізняються. Сучасні підходи часто поєднують кілька метрик для комплексної оцінки якості відповідей, наприклад, BERTScore [13] використовує контекстуальні embeddings для порівняння семантичної подібності на рівні токенів.

Для оцінки відповідей, де можливі різні формулювання з однаковим змістом, застосовуються методи, засновані на нейромережевих моделях, які навчені визначати семантичну еквівалентність текстів, такі як Sentence-BERT або Universal Sentence Encoder. Критичним аспектом при використанні еталонних відповідей є формування репрезентативного набору еталонів, що охоплює різноманітні потенційно правильні формулювання. Гібридні підходи поєднують автоматичні метрики з експертним оцінюванням, де спочатку застосовуються алгоритмічні методи для фільтрації очевидно неправильних відповідей, а потім експерти аналізують складні випадки, що потребують глибшого розуміння контексту.

Важливим методологічним інструментом є контекстно-залежне оцінювання, яке враховує специфіку конкретного завдання та домену знань при порівнянні відповідей з еталонами. Сучасні дослідження демонструють ефективність методів, що використовують LLM як оцінювачі інших LLM (LLM-as-judge)[14, 15], де потужна модель порівнює відповідь оцінюваної моделі з еталоном, враховуючи семантику, контекст та прагматичні аспекти комунікації.

3. Автоматизовані методи оцінювання якості LLM без еталонних даних

Повністю автоматизоване оцінювання якості великих мовних моделей (LLM) без використання еталонних даних представляє інноваційний напрям досліджень, що дозволяє об'єктивно вимірювати продуктивність моделей у різноманітних контекстах на об'ємах даних значно більших ніж реально опрацювати в випадку експертних оцінок чи порівнянні з еталоними даними.

Основним методом, що набув поширення останнім часом, є використання інших **LLM як оцінювачів** (LLM-as-judges), коли одна мовна модель, зазвичай більш потужна, аналізує та оцінює відповіді іншої моделі за попередньо визначеними критеріями якості, такими як релевантність, фактична точність та логічна послідовність. Для забезпечення об'єктивності такого оцінювання розробляються спеціалізовані промпти-інструкції для моделі-оцінювача, які чітко визначають критерії оцінки, процедуру аналізу та шкалу оцінювання, мінімізуючи вплив потенційних упереджень самої моделі-оцінювача [14, 15].

Самооцінювання (self-evaluation) є методом, коли та сама модель, що генерує відповідь, аналізує власні результати, виявляючи потенційні помилки, неточності чи неповноту відповіді. Самооцінка великих мовних моделей здійснюється через їхню здатність аналізувати власні відповіді за допомогою спеціально розроблених промптів, які спонукають модель критично оцінити різні аспекти згенерованого контенту, включаючи фактичну точність, логічну послідовність та відповідність контексту. В рамках цього підходу моделі можуть використовувати техніки chain-of-thought reasoning для покрокового аналізу власних відповідей, а також застосовувати внутрішні метрики впевненості для оцінки надійності своїх висновків. Важливим компонентом самооцінки є також здатність моделі ідентифікувати потенційні обмеження та невизначеності у власних відповідях, що досягається через використання спеціальних промптів для аналізу так званих граничних випадків (англ. edge cases).

Спеціалізовані моделі для оцінки конкретних аспектів якості LLM представляють собою вузькоспрямовані нейронні мережі, навчені на специфічних наборах даних для виявлення таких характеристик як токсичність, упередженість, агресивність чи недоречний контент, що дозволяє проводити детальний автоматизований аналіз згенерованих текстів за конкретними критеріями якості. Такі моделі часто використовують комбінацію методів машинного навчання, включаючи класифікатори на основі нейромережевих трансформерів та спеціалізовані архітектури для виявлення тонких нюансів мови, що можуть вказувати на проблематичний контент. При цьому важливим аспектом є постійне оновлення та донавчання цих моделей на нових прикладах для підтримки їх актуальності та ефективності. Ключовою перевагою використання спеціалізованих моделей є їхня здатність працювати в режимі реального часу та масштабуватися для аналізу великих обсягів згенерованого контенту, забезпечуючи при цьому високу точність та специфічність оцінки за конкретними параметрами якості.

4. Ефективність методів оцінювання якості для різних критеріїв

Наведені вище методи оцінювання будуть мати різну ефективність для різних критеріїв оцінювання. В Таб 1. наведено порівняльний аналіз методів оцінювання для критеріїв якості, вказано які саме види кожного типу методів будуть найбільш ефективними для кожного з критеріїв.

	Експертні оцінки	Еталонні дані	Автоматичні
Прецизійність відповіді	Індивідуальна експертна оцінка	Бінарне оцінювання, або небінарне з простими метриками лексичної подібності.	LLM-as-judge або самооцінювання (малоефективні у порівнянні з підходом з еталонними даними)
Змістовність відповіді	Індивідуальна експертна оцінка, з вказівками щодо шкали оцінювання змістовностік	Небінарне оцінювання з семантичним аналізом подібності.	LLM-as-judge або самооцінювання (малоефективні у порівнянні з підходом з еталонними даними)
Природність мовлення	Групова експертна оцінка, з вказівками щодо шкали оцінювання змістовності, або групова оцінка.	Неефективний	Спеціалізовані моделі для оцінки природності мовлення.
Послідовність відповідей	Індивідуальна експертна оцінка.	Бінарне або небінарне порівняння, в залежності від формату відповідей.	LLM-as-judge.
Токсичність	Групова експертна оцінка, з вказівками щодо шкали оцінювання змістовності.	Неефективний	Спеціалізовані моделі для оцінки токсичності.
Заангажованість	Групова експертна оцінка, з вказівками щодо шкали оцінювання змістовності.	Неефективний	Спеціалізовані моделі для оцінки заангажованості.
Безпекові вразливості	Індивідуальна експертна оцінка.	Неефективний	LLM-as-judge для генерування промптів з потенційними вразливостями і оцінки поведінки моделі.

Таб 1. Найбільш ефективні способи оцінювання якості LLM для кожного з критеріїв

Експертні оцінки є найбільш надійним методом оцінювання якості LLM для будь-якого із критеріїв. Саме експертні оцінки є базою для формування еталонних наборів даних та навчання моделей які здійснюють автоматичну оцінку. В той же час очевидним недоліком експертних оцінок є вартість та обмеженість людських ресурсів у порівнянні з машинними.

Метод з еталонними даними найкраще підходить для оцінок прецизійності та змісту відповідей LLM, коли існує набір даних з еталонними відповідями, з яким можна порівнювати реальні відповіді.

Автоматичні методи можна застосовувати для всіх критеріїв, але найбільш ефективними вони будуть для нечітких оціночних критеріїв, таких як природність, токсичність чи неззангажованість, для яких неможливо сформулювати репрезентативний еталонний набір даних.

5. Висновки

Стаття присвячена вирішенню нової проблеми, яка з'явилася в контексті стрімкого розвитку генеративних моделей штучного інтелекту та великих мовних моделей, — багатовимірному та всеохоплюючому оцінюванню якості згенерованих текстів. В роботі запропоновано набір з семи критеріїв оцінювання LLM, що відкриває можливість формалізувати та уніфікувати комплексну систему оцінки якості великих мовних моделей.

Запропонований підхід дозволяє системно охопити різноманітні аспекти функціонування LLM – від технічної точності до етичних та безпекових параметрів.

Проведений аналіз методів оцінювання якості LLM виділив три основні підходи: експертні оцінки (ручний спосіб), порівняння з еталонними даними (напівавтоматизований спосіб) та повністю автоматизовані методи без використання еталонів. В ході досліджень та аналізу встановлено, що експертні оцінки залишаються найбільш надійним, хоча й найменш масштабованим методом оцінювання для всіх критеріїв якості. Методи порівняння з еталонними даними показують високу ефективність для оцінки прецизійності та змістовності відповідей, проте мають обмеження при роботі з суб'єктивними критеріями, такими як токсичні чи заанглажованість. Автоматизовані методи, такі як LLM-as-judge та спеціалізовані оцінювальні моделі, демонструють найкращі результати для нечітких критеріїв, де створення репрезентативних еталонних наборів даних практично неможливе.

Також підкреслено відмінності в ефективності методів оцінювання для різних типів запитів та відповідей. Для структурованих відповідей та конкретних запитів найбільш ефективними є методи з використанням еталонних даних, тоді як для неструктурованих відповідей та відкритих питань більш доцільним є застосування комбінованих підходів з елементами експертного оцінювання.

Підсумовуючи, важливим є розвиток комбінованих методологій оцінювання, що поєднують переваги різних підходів для всебічної оцінки якості LLM з урахуванням конкретної специфіки їх застосування. Створення стандартизованих, комплексних і динамічних систем оцінювання LLM залишається відкритим науковим викликом, вирішення якого матиме вагомий вплив на розвиток галузі штучного інтелекту та впровадження цих технологій у критично важливі системи обробки інформації.

Список літератури

1. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A.. A Comprehensive Overview of Large Language Models. 2023. [Online]. Available: https://arxiv.org/abs/2307.06435
2. Li, Diya & Zhao, Yue & Wang, Zhifang & Jung, Calvin & Zhang, Zhe. (2024). Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. ISPRS International Journal of Geo-Information. 13. 405. 10.3390/ijgi13110405.
3. Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., & Xiong, D.. Evaluating Large Language Models: A Comprehensive Survey. 2023. [Online]. Available: https://arxiv.org/abs/2310.19736
4. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Hudson, D. A., Zelikman, E., . Koreeda, Y.. Holistic Evaluation of Language Models. 2022. [Online]. Available: https://arxiv.org/abs/2211.09110
5. Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S.. Benchmarking LLM powered Chatbots: Methods and Metrics. 2023. [Online]. Available: https://arxiv.org/abs/2308.04624
6. Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D.. Large Language Models are Diverse Role-Players for Summarization Evaluation. 2023. [Online]. Available: https://arxiv.org/abs/2303.15078
7. Ni, X., & Li, P.. A Systematic Evaluation of Large Language Models for Natural Language Generation Tasks. 2024. [Online]. Available: https://arxiv.org/abs/2405.10251
8. Cui, W., Zhang, J., Li, Z., Damien, L., Das, K., Malin, B., & Kumar, S.. DCR-Consistency: Divide-Conquer-Reasoning for Consistency Evaluation and Improvement of Large Language Models. 2024. [Online]. Available: https://arxiv.org/abs/2401.02132
9. Zhao, Y., Zhu, J., Xu, C., & Li, X.. Enhancing LLM-based Hatred and Toxicity Detection with Meta-Toxic Knowledge Graph. 2024. [Online]. Available: https://arxiv.org/abs/2412.15268
10. L. Baresi, C. Criscuolo and C. Ghezzi, "Understanding Fairness Requirements for ML-based Software," 2023 IEEE 31st International Requirements Engineering Conference (RE), Hannover, Germany, 2023, pp. 341-346, doi: 10.1109/RE57278.2023.00046.
11. Kolchenko, V., Khoma, V., Sabodashko, D., & Perepelytsia, P. (2024). Exploring large language models' security threats with automated tools. Social Development and Security, 14(6), 81-96. https://doi.org/10.33445/sds.2024.14.6.9
12. Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods* (5th ed.). Oxford University Press. https://archive.org/details/rankcorrelationm0000kend
13. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. 2019. [Online]. Available: https://arxiv.org/abs/1904.09675

14. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y.. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. 2024. [Online]. Available: https://arxiv.org/abs/2412.05579
15. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J.. A Survey on LLM-as-a-Judge. 2024. [Online]. Available: https://arxiv.org/abs/2411.15594

References

16. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A.. A Comprehensive Overview of Large Language Models. 2023. [Online]. Available: https://arxiv.org/abs/2307.06435
17. Li, Diya & Zhao, Yue & Wang, Zhifang & Jung, Calvin & Zhang, Zhe. (2024). Large Language Model-Driven Structured Output: A Comprehensive Benchmark and Spatial Data Generation Framework. ISPRS International Journal of Geo-Information. 13. 405. 10.3390/ijgi13110405.
18. Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., & Xiong, D.. Evaluating Large Language Models: A Comprehensive Survey. 2023. [Online]. Available: https://arxiv.org/abs/2310.19736
19. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Hudson, D. A., Zelikman, E., . Koreeda, Y.. Holistic Evaluation of Language Models. 2022. [Online]. Available: https://arxiv.org/abs/2211.09110
20. Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S.. Benchmarking LLM powered Chatbots: Methods and Metrics. 2023. [Online]. Available: https://arxiv.org/abs/2308.04624
21. Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D.. Large Language Models are Diverse Role-Players for Summarization Evaluation. 2023. [Online]. Available: https://arxiv.org/abs/2303.15078
22. Ni, X., & Li, P.. A Systematic Evaluation of Large Language Models for Natural Language Generation Tasks. 2024. [Online]. Available: https://arxiv.org/abs/2405.10251
23. Cui, W., Zhang, J., Li, Z., Damien, L., Das, K., Malin, B., & Kumar, S.. DCR-Consistency: Divide-Conquer-Reasoning for Consistency Evaluation and Improvement of Large Language Models. 2024. [Online]. Available: https://arxiv.org/abs/2401.02132
24. Zhao, Y., Zhu, J., Xu, C., & Li, X.. Enhancing LLM-based Hatred and Toxicity Detection with Meta-Toxic Knowledge Graph. 2024. [Online]. Available: https://arxiv.org/abs/2412.15268
25. L. Baresi, C. Criscuolo and C. Ghezzi, "Understanding Fairness Requirements for ML-based Software," 2023 IEEE 31st International Requirements Engineering Conference (RE), Hannover, Germany, 2023, pp. 341-346, doi: 10.1109/RE57278.2023.00046.
26. Kolchenko, V., Khoma, V., Sabodashko, D., & Perepelytsia, P. (2024). Exploring large language models' security threats with automated tools. Social Development and Security, 14(6), 81-96. https://doi.org/10.33445/sds.2024.14.6.9
27. Kendall, M.G. & Gibbons, J.D. (1990). *Rank Correlation Methods* (5th ed.). Oxford University Press. https://archive.org/details/rankcorrelationm0000kend
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. 2019. [Online]. Available: https://arxiv.org/abs/1904.09675
29. Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z., & Liu, Y.. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. 2024. [Online]. Available: https://arxiv.org/abs/2412.05579

    Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J.. A Survey on LLM-as-a-Judge. 2024. [Online]. Available: https://arxiv.org/abs/2411.15594

Конфлікт інтересів

Автори зазначають про відсутність фінансових, чи інших потенційних конфліктів щодо цієї роботи.

Інформація про авторів

Хома Юрій Володимирович
Доктор технічних наук, доцент
Кафедра «Інформаційно-вимірювальні технології»
Національний університет «Львівська політехніка»

вул. С. Бандери, 12, м. Львів, Україна, 79013
E-mail: yurii.v.khoma@lpnu.ua
Контактний тел.: +38 093 85 41 619
Кількість статей у загальнодержавних базах даних – 35
Кількість статей у міжнародних базах даних – 18
Номер ORCID: https://orcid.org/0000-0002-4677-5392

Yuriy Khoma
Doctor of Technical Sciences, Docent
Department of Information Measurement Technologies
Lviv Polytechnic National University
Bandera str., 12, Lviv, Ukraine, 79013
E-mail: yurii.v.khoma@lpnu.ua
Contact tel.: +38 093 85 41 619
The number of articles in national databases – 35
The number of articles in international databases – 18
Number ORCID: https://orcid.org/0000-0002-4677-5392

Щудло Іван Олексійович
аспірант
Кафедра «Інформаційно-вимірювальні технології»
Національний університет «Львівська політехніка»
вул. С. Бандери, 12, м. Львів, Україна, 79013
E-mail: ivan.o.shchudlo@lpnu.ua
Контактний тел.: +38 093 77 92 265
Кількість статей у загальнодержавних базах даних – 0
Кількість статей у міжнародних базах даних – 0
Номер ORCID: https://orcid.org/0009-0003-0412-7682

Ivan Shchudlo
PhD student
Department of Information Measurement Technologies
Lviv Polytechnic National University
Bandera str., 12, Lviv, Ukraine, 79013
E-mail: ivan.o.shchudlo@lpnu.ua
Contact tel.: +38 093 77 92 265
The number of articles in national databases – 0
The number of articles in international databases – 0
Number ORCID: https://orcid.org/0009-0003-0412-7682