DOI: https://doi.org/10.15276/ict.03.2026.10
УДК 004.451:502/504

Багатовимірна оцінка використання великих мовних моделей у задачах узагальнення текстів

Єгоров Олег Йосипович[1)]
ORCID: https://orcid.org/0000-0002-8260-9463; egoroffoleg@ukr.net
Борщенко Володимир Олександрович[1)]
ORCID: https://orcid.org/0009-0007-1089-5580; v.o.borshchenko@ust.edu.ua
[1)] Український державний університет науки і технологій, вул. Лазаряна, 2. Дніпро, 49010, Україна

АНОТАЦІЯ

Стрімке зростання обсягів цифрових текстових даних у різних галузях, зокрема науці, медицині, бізнесі та інформаційних технологіях, зумовлює необхідність розроблення ефективних методів автоматичного узагальнення інформації. Великі мовні моделі демонструють високий потенціал для розв'язання цієї задачі завдяки здатності генерувати зв'язні, змістовні та контекстно релевантні тексти. Водночас проблема комплексного та об'єктивного оцінювання їх якості залишається недостатньо дослідженою, оскільки більшість існуючих підходів базується на обмеженому наборі метрик і не враховує багатовимірний характер якості узагальнення. Метою роботи є розроблення й апробація методики комплексного порівняння моделей у задачах автоматичного узагальнення текстів на основі дворівневої системи оцінювання, яка поєднує різні аспекти якості та ефективності. У дослідженні здійснено експериментальне оцінювання чотирьох моделей (GPT, Claude, Gemini та LLaMA) на текстах різних типів. Для забезпечення об'єктивності результатів використовувалися однакові умови генерації узагальнень, зокрема фіксовані параметри та різні довжини вихідних текстів. Оцінювання якості узагальнень здійснювалося із застосуванням комплексу метрик: ROUGE для визначення лексичної подібності, BERTScore для оцінювання семантичної відповідності, SummaC для аналізу фактичної узгодженості, а також оцінювання за допомогою мовних моделей, що імітує людське сприйняття якості тексту. Додатково враховувалися показники ефективності, зокрема час генерації узагальнень та вартість їх отримання. Отримані результати свідчать про суттєві відмінності між моделями за всіма досліджуваними критеріями. Зокрема, встановлено, що різні моделі мають спеціалізацію: одні демонструють вищу фактичну точність, інші – кращу читабельність або ефективність. Також виявлено залежність якості узагальнення від довжини тексту: коротші узагальнення характеризуються вищою фактичною узгодженістю, тоді як довші забезпечують кращу повноту та зрозумілість викладу. Отримані результати дозволяють сформулювати практичні рекомендації щодо вибору великих мовних моделей залежно від вимог конкретних прикладних задач. Запропонований підхід до багатовимірного оцінювання може бути використаний як основа для подальших досліджень у галузі обробки природної мови, зокрема для вдосконалення методів оцінювання якості та адаптації моделей до спеціалізованих предметних областей..

Ключові слова: інформаційні технології; нейронні мережі; штучний інтелект; великі мовні моделі; обробка природної мови; автоматичне узагальнення тексту

Для цитування: **Єгоров О. Й., Борщенко В. О. "Багатовимірна оцінка використання великих мовних моделей у задачах узагальнення текстів". *Інформатика. Культура. Техніка*. 2026; Том 3 № 1 (3): 120–126. DOI: https://doi.org/10.15276/ict.03.2026.10**

Актуальність. Стрімке збільшення обсягів цифрової текстової інформації у науці, освіті, медицині, бізнесі та інформаційних технологіях зумовлює потребу в засобах її швидкого й змістовного опрацювання. Ручний аналіз великих масивів документів потребує значних часових і людських ресурсів, тому автоматичне узагальнення текстів стає важливим інструментом для скорочення інформації, виділення ключових положень і підтримки прийняття рішень.

Розвиток великих мовних моделей суттєво розширив можливості автоматичного узагальнення. Сучасні моделі здатні не лише виокремлювати окремі фрагменти вихідного документа, ай формувати зв'язні та контекстно релевантні узагальнення, адаптовані до заданого обсягу й типу тексту. Водночас результати їх застосування можуть відрізнятися за повнотою, семантичною відповідністю, читабельністю та фактичною достовірністю. Особливого значення набуває проблема появи у згенерованих узагальненнях тверджень, які не підтверджуються вихідним текстом або спотворюють його зміст.

Традиційні метрики, засновані переважно на лексичному збігу з еталонним узагальненням, не дають змоги повною мірою оцінити якість результатів генеративних моделей. Лексично відмінне формулювання може правильно передавати зміст оригіналу, тоді як текст із високим рівнем формального збігу не завжди є логічним, повним або

фактично узгодженим. Це зумовлює необхідність поєднання лексичних, семантичних і фактологічних метрик з оцінюванням зв'язності, зрозумілості та загальної якості сформованого тексту.

Практичний вибір великої мовної моделі також не може ґрунтуватися лише на показниках якості. Важливими є швидкість генерації, вартість використання, обчислювальні вимоги та можливість локального розгортання. Модель із найвищою якістю узагальнення може бути недоцільною для систем, що обробляють значні обсяги інформації або працюють у режимі реального часу. Крім того, ефективність моделей залежить від типу вихідних текстів і заданої довжини узагальнення, тому результати оцінювання мають розглядатися з урахуванням конкретного сценарію застосування.

Отже, актуальність дослідження визначається необхідністю розроблення багатовимірного підходу до порівняння великих мовних моделей, який одночасно враховує якість, фактичну узгодженість, семантичну відповідність, лексичну подібність та практичну ефективність. Такий підхід дає змогу об'єктивніше визначати сильні й слабкі сторони моделей та формувати обґрунтовані рекомендації щодо їх використання для автоматичного узагальнення текстів різних типів.

Мета дослідження полягає в розробленні та апробації методики комплексного порівняння великих мовних моделей у задачах автоматичного узагальнення текстів із використанням дворівневої системи оцінювання. Запропонована система має враховувати не лише формальну подібність згенерованого узагальнення до еталонного тексту, а й його семантичну відповідність, фактичну узгодженість, зв'язність, читабельність, повноту та лаконічність. Окрему увагу приділено практичним характеристикам моделей, зокрема швидкості генерації, вартості обробки запитів, обчислювальним вимогам і можливості локального використання.

Проблема автоматичного узагальнення текстів є однією з ключових задач сучасної обробки природної мови (Natural Language Processing, NLP). З розвитком цифрових технологій і зростанням обсягів текстових даних дослідження у цій сфері активно розвиваються, що зумовило появу великої кількості підходів, моделей та методів оцінювання якості узагальнення. У сучасній науковій літературі простежується еволюція підходів до автоматичного реферування від класичних статистичних методів до сучасних нейронних моделей і великих мовних моделей.

Ранні системи автоматичного узагальнення тексту переважно базувалися на екстрактивних методах, які вибирали найбільш інформативні речення з вихідного документа на основі статистичних характеристик тексту [1]. З розвитком глибокого навчання почали активно використовуватися абстрактні моделі, здатні генерувати новий текст, який передає зміст оригінального документа. Сучасні дослідження показують, що використання трансформерних архітектур і попередньо навчених мовних моделей значно підвищило якість автоматичного узагальнення та дозволило створювати більш зв'язні та семантично узгоджені тексти.

Важливим напрямом сучасних досліджень є розроблення методів оцінювання якості узагальнення текстів. Традиційно для цього використовуються автоматичні метрики, зокрема ROUGE, яка вимірює ступінь лексичної подібності між згенерованим узагальненням і еталонним текстом шляхом порівняння n-грам. Проте численні дослідження показують, що такі метрики мають обмеження, оскільки вони не враховують семантичну відповідність, стилістичну якість або фактичну точність узагальнення [2].

З метою подолання цих обмежень у сучасних роботах пропонуються семантичні метрики оцінювання, такі як BERTScore, які використовують контекстні векторні представлення слів для визначення семантичної подібності текстів. Крім того, дослідники розробляють спеціалізовані підходи для оцінювання фактичної узгодженості тексту, оскільки генеративні моделі можуть створювати інформацію, яка не відповідає змісту джерела [3].

Останніми роками значна увага приділяється застосуванню великих мовних моделей як інструментів оцінювання якості тексту. У такому підході моделі використовуються як

експертні системи, які оцінюють узагальнення за певними критеріями (релевантність, зв'язність, фактична точність тощо). Дослідження показують, що LLM-орієнтовані методи оцінювання можуть краще узгоджуватися з людськими оцінками якості тексту порівняно з традиційними автоматичними метриками [4].

Проблема багатовимірного оцінювання якості текстової генерації вже активно досліджується. Зокрема, у межах SummEval було здійснено систематичне переоцінювання широкого набору автоматичних метрик і зіставлення їх із експертними та краудсорсинговими оцінками якості узагальнень [5]. Підхід UniEval передбачає використання єдиного оцінювача для аналізу декількох пояснюваних вимірів якості, зокрема зв'язності, узгодженості та інших характеристик, що дозволяє адаптувати оцінювання до різних завдань генерації тексту [6]. У роботі, присвяченій G-Eval, запропоновано використання великих мовних моделей як оцінювачів якості з формалізованими критеріями та послідовністю міркувань [7]. Окремий напрям представлений фреймворком HELM, орієнтованим на стандартизоване та відтворюване оцінювання мовних моделей за широким набором сценаріїв і характеристик, включаючи не лише якість, а й ефективність, надійність та інші аспекти їх використання [8].

Крім того, у сучасних роботах активно проводяться порівняльні дослідження ефективності різних мовних моделей у задачах узагальнення тексту. Такі дослідження зазвичай використовують стандартні набори даних і комбінують декілька метрик оцінювання, що дозволяє виявити сильні та слабкі сторони різних моделей у різних прикладних умовах. При цьому результати показують, що жодна модель не демонструє стабільної переваги у всіх задачах, а ефективність значною мірою залежить від типу тексту, доменної специфіки та обраних метрик оцінювання [9].

Отже, аналіз сучасних досліджень свідчить про активний розвиток методів автоматичного узагальнення текстів і зростання ролі великих мовних моделей у цій сфері. Водночас залишається відкритою проблема створення комплексних методів оцінювання, які б одночасно враховували різні аспекти якості узагальнення, зокрема семантичну відповідність, фактичну точність, читабельність і ефективність моделей. Це зумовлює необхідність подальших досліджень, спрямованих на розроблення багатовимірних підходів до оцінювання сучасних систем автоматичного узагальнення текстів.

У дослідженні проведено експериментальне порівняння сучасних великих мовних моделей у задачах автоматичного узагальнення тексту з використанням багатовимірної системи оцінювання. Методологія дослідження передбачала комплексний аналіз якості згенерованих узагальнень за декількома незалежними критеріями, а також оцінювання ефективності моделей з точки зору швидкодії та вартості використання.

Для експериментального аналізу було відібрано 4 великі мовні моделі, що представляють різні підходи до реалізації LLM, зокрема комерційні API-моделі та моделі з відкритим кодом. До вибірки увійшли моделі від провідних розробників, зокрема OpenAI, Google, Anthropic, а також відкрита модель LLaMA, що запускається локально. Це дозволило порівняти як пропрієтарні рішення, так і моделі з відкритим доступом. У Таблиці 1 наведено список моделей, що були використані в дослідженні.

Таблиця 1. **Список досліджуваних моделей та їх особливості**

Розробник	Модель	Тип моделі	Метод доступу
OpenAI	GPT-5.2	Комерційна	API
Anthropic	Claude-4.5-Sonnet	Комерційна	API
Google	Gemini-2.5-Flash	Комерційна	API
Open Source	LLaMA-4-Maverick	Відкритий код	Локальний запуск

Джерело: **розроблено авторами**

Експериментальне оцінювання проводилося на наборах даних, що представляють різні типи текстових джерел. Художні тексти відібрано з корпусу SQuALITY, сформованого на основі англомовних творів художньої літератури з Project Gutenberg; корпус містить довгі

тексти обсягом переважно 4-6 тис. слів та підготовлені людьми еталонні узагальнення. Нехудожню літературу, наукові публікації та технічну документацію відібрано з відповідних англомовних корпусів із наявними еталонними узагальненнями. Зокрема, для наукових текстів використано SumSurvey, сформований на основі оглядових публікацій arXiv: основний текст статті використовується як вхідний документ, а авторська анотація – як еталонне узагальнення. Усі використані тексти є англомовними. Для кожного набору даних випадковим чином відбиралося по 20 документів, що достатньо для контрольованого порівняння моделей за однакових умов.

Для кожного документа генерувалися узагальнення різної довжини – 50, 100 та 150 токенів, що дало можливість дослідити вплив довжини резюме на якість узагальнення. Усі моделі використовували однакові мінімальні інструкції (prompts) та однакові параметри генерації, що дозволило мінімізувати випадковість результатів і забезпечити коректність порівняння. Значення temperature = 0,1; значення top-p = 1.0, щоб не обмежувати вибір через цей параметр і контролювати випадковість температурою; seed = 1234.

Оцінювання якості узагальнень здійснювалося за допомогою декількох груп метрик, які охоплюють різні аспекти якості тексту [10], [11].

1. Лексична подібність. Для вимірювання лексичного збігу між згенерованим узагальненням і еталонним текстом використовувалася метрика ROUGE, зокрема показники ROUGE-1, ROUGE-2 та ROUGE-L. Ці метрики визначають ступінь збігу n-грам між текстами та широко використовуються у задачах автоматичного реферування.

2. Семантична відповідність. Для оцінювання змістовної подібності між текстами використовувалася метрика BERTScore, яка базується на контекстних векторних представленнях слів. На відміну від метрик на основі n-грам, BERTScore дозволяє враховувати семантичні зв’язки між словами навіть у випадках, коли вони не мають прямого лексичного збігу.

3. Фактична узгодженість. Для перевірки відповідності згенерованого узагальнення змісту вихідного документа застосовувався метод SummaC, що використовує моделі природно-мовного висновування (NLI) для виявлення суперечностей між текстами.

4. Оцінювання якості тексту. Додатково використовувалося оцінювання за допомогою мовних моделей, яке імітує людське сприйняття тексту. У цьому підході моделі-оцінювачі аналізували узагальнення за такими критеріями, як релевантність, логічність, фактична точність, лаконічність і загальна якість. Для цього використовувалися GPT-5.4 та Claude-4.6-Sonnet. Для уникнення ефекту самопереваги оцінювач обирався таким чином, щоб він не збігався з моделлю, результат якої оцінювався. Оцінювання проводилося за п’ятьма критеріями: релевантність, зв’язність, фактологічна узгодженість, стислість та загальна якість. Кожен критерій оцінювався за шкалою від 1 до 5; для зменшення стохастичності використовувалося temperature = 0,1.

Окрім метрик якості, у дослідженні також враховувалися метрики ефективності, зокрема середній час генерації одного узагальнення, обчислювальні витрати, вартість обробки одного запиту при використанні API-моделей. Під час оцінювання ефективності для хмарних моделей вимірювався повний час від надсилання API-запиту до отримання відповіді, включно з мережевою затримкою. Локальна модель LLaMA запускалася на комп’ютері з процесором Apple M4 Max, для неї враховувався повний час локального інференса. Вартість використання API-моделей розраховувалася за фактичною кількістю вхідних і вихідних токенів відповідно до офіційних тарифів провайдерів на момент проведення експерименту. Для локальної LLaMA прямі API-витрати відсутні, а показник вартості визначався на основі оцінки обчислювальних витрат локального запуску.

У дослідженні застосовано дворівневу систему оцінювання, яка дозволяє поєднати детальний аналіз окремих характеристик якості узагальнення з узагальненим рейтингом моделей, придатним для практичного використання. На першому рівні для отримання узагальненого результату різним метрикам було надано різну вагу залежно від їх значущості: показник фактичної узгодженості – 35 %, семантична подібність – 25 %, оцінювання якості

тексту – 25 %, лексичний збіг – 15 %. Аналогічно формується зведений показник ефективності, що поєднує час виконання та вартість обробки.

На першому рівні фактична узгодженість має вагу 35 %, оскільки непідтверджені або суперечливі твердження є найбільш критичним ризиком, особливо у медичних, юридичних та інших відповідальних застосуваннях. Семантична подібність отримує 25 %, оскільки оцінює збереження змісту за можливого перефразування. LLM-оцінювання якості також має 25 %, оскільки відображає релевантність, зв'язність, читабельність, лаконічність і загальне сприйняття тексту. Лексичний збіг має вагу 15 %: ROUGE забезпечує порівнянність із традиційними дослідженнями, але є чутливим до формулювання та не повністю враховує семантичну еквівалентність.

Для забезпечення порівнянності всі показники нормалізувалися до діапазону [0;1] за методом min-max. Для метрик, де більше значення є кращим, використовувалося пряме нормалізоване значення, а для часу обробки та вартості застосовувалася інверсія, щоб у всіх випадках більше значення відповідало кращому результату. Після цього нормалізовані показники використовувалися для розрахунку зважених інтегральних оцінок.

На другому рівні відбувається ранжування моделей – результати першого рівня перетворюються у ранги моделей за окремими компонентами: загальна якість, фактична узгодженість (як окремий критично важливий показник), «людиноподібна» якість, ефективність, економічна ефективність.

Кожна модель отримує позицію (ранг) у кожній категорії (1 – найкращий результат). Після цього формується підсумковий рейтинг як зважена комбінація рангів: загальна якість – 30% через важливість читабельності й логічності, фактична узгодженість – 25 % через критичність достовірності, оцінювання якості тексту – 20 % через важливість читабельності й логічності, ефективність – 15 % як характеристика швидкодії, ефективність витрат – 10 % як показник, що найбільше залежить від тарифів і середовища розгортання.

Такий підхід дозволяє уникнути домінування окремих метрик, врахувати різні аспекти якості, забезпечити баланс між точністю та практичністю використання моделей.

Таким чином, дворівнева система оцінювання забезпечує як детальний аналіз окремих характеристик моделей, так і формування інтегрального рейтингу, що відображає їхню загальну придатність для практичного використання у задачах автоматичного узагальнення текстів. Отримані результати порівняння наведено на Рис. 1.

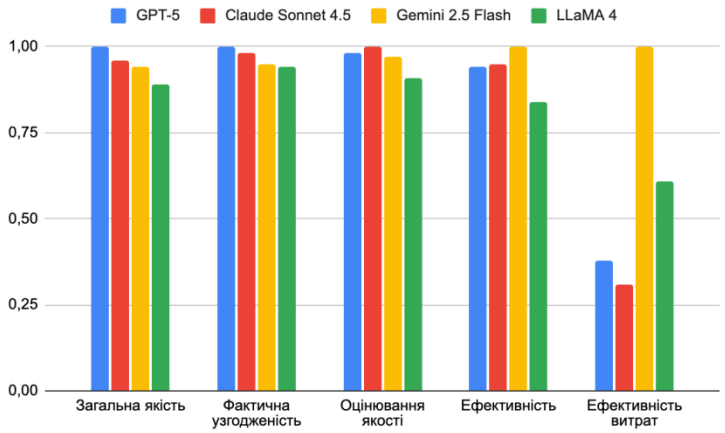

Рис. 1. **Порівняльний аналіз моделей за всіма метриками, нормалізований до діапазону 0-1 для забезпечення коректного порівняння**

Джерело: **розроблено авторами**

Висновки. У роботі проведено комплексне дослідження можливостей сучасних великих мовних моделей у задачах автоматичного узагальнення текстів із використанням багатовимірної системи оцінювання. Запропонований підхід дозволив поєднати різні аспекти якості узагальнення – фактичну узгодженість, семантичну відповідність, лексичну подібність

та якість – із показниками ефективності, зокрема швидкістю обробки та вартістю використання моделей.

У ході дослідження підтверджено наявність компромісу між фактичною точністю та сприйняттям якості узагальнення: коротші тексти характеризуються вищою фактичною узгодженістю, тоді як довші узагальнення краще оцінюються за критеріями повноти та зрозумілості. Крім того, встановлено, що ефективність моделей суттєво залежить від типу тексту: найкращі результати досягаються для художніх текстів, тоді як технічні та наукові тексти залишаються більш складними для коректного узагальнення.

Результати дослідження показали, що моделі демонструють різні сильні сторони залежно від обраних критеріїв оцінювання. Модель GPT забезпечує збалансовані результати за більшістю метрик, поєднуючи достатньо високу якість узагальнення з прийнятною ефективністю. Модель Claude відзначається високим рівнем якості тексту, забезпечуючи кращу зв'язність і читабельність узагальнень. Модель Gemini демонструє високу швидкодію та економічну ефективність, що робить її придатною для задач, пов'язаних із обробкою великих обсягів даних у реальному часі. Водночас модель LLaMA, попри нижчі загальні показники якості, залишається важливою завдяки можливості локального використання та гнучкості налаштування. Жодна з досліджених моделей не є універсально оптимальною для всіх сценаріїв використання. Вибір моделі має здійснюватися з урахуванням пріоритетів конкретної задачі.

Конфлікт інтересів: Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті

Фінансування: Дослідження проводилося без фінансової підтримки

Використання засобів штучного інтелекту (ШІ): Автори підтверджують, що ШІ не використовувався для генерації тексту, формул і таблиць, формування наукових положень, отримання результатів або формулювання висновків. Наукові положення, результати та висновки є власним інтелектуальним внеском авторів.

СПИСОК ЛІТЕРАТУРИ

1. Törnberg, P. "How to use large language models for text analysis". *SAGE Publications.* 2024. DOI: https://doi.org/10.4135/9781529683707.

2. El-Kassas, W. S., Salama, C. R., Rafea, A. A. & Mohamed, H. K. "Automatic text summarization: A comprehensive survey". *Expert Systems with Applications.* 2021; 165: 113679. DOI: https://doi.org/10.1016/j.eswa.2020.113679.

3. Koh, H. Y., Ju, J., Liu, M. & Pan, S. "An empirical survey on long document summarization: Datasets, models, and metrics". *ACM Computing Surveys.* 2022; 55 (8): 1–35. DOI: https://doi.org/10.1145/3545176.

4. Zhang, H., Yu, P. S. & Zhang, J. "Text summarization comparative analysis". *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2406.11289.

5. Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R. & Radev, D. "SummEval: Re-evaluating summarization evaluation". *Transactions of the Association for Computational Linguistics*. 2021; 9: 391–409. DOI: https://doi.org/10.1162/tacl_a_00373.

6. Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H. & Han, J. "Towards a unified multi-dimensional evaluator for text generation". *Proceedings of EMNLP* 2022. p. 2023–2038. DOI: https://doi.org/10.18653/v1/2022.emnlp-main.131.

7. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. & Zhu, C. "G-Eval: NLG evaluation using GPT-4 with better human alignment". *Proceedings of EMNLP*. 2023. p. 2511–2522. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.153.

8. Liang, P., Bommasani, R., Lee, T., et al. "Holistic evaluation of language models. transactions on machine learning research". *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2211.09110.

9. Nguyen, H., Chen, H., Pobbathi, L. & Ding, J. "A comparative study of quality evaluation methods for text summarization". *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2407.00747.

10. Ampel, B. M., Yang, C.-H., Hu, J. & Chen, H. “Large language models for conducting advanced text analytics information systems research”. *ACM Transactions on Management Information Systems.* 2025. DOI: https://doi.org/10.1145/3682069.

11. Rehman, T., Ghosh, S., Das, K., Bhattacharjee, S., Sanyal, D. K. & Chattopadhyay, S. “Evaluating LLMs and pre-trained models for text summarization across diverse datasets”. *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2502.19339.

Отримано 28.07.2026
Отримано після перегляду 28.08.2026
Прийнято 04.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.10
UDC 004.451:502/504

Multi-dimensional evaluation of large language models for text summarization

Oleh Y. Yehorov[1)]
ORCID: https://orcid.org/0000-0002-8260-9463; egoroffoleg@ukr.net
Volodymyr O. Borshchenko[1)]
ORCID: https://orcid.org/0009-0007-1089-5580; v.o.borshchenko@ust.edu.ua
[1)] Ukrainian State University of Science and Technologies, 2, Lazariana St. Dnipro, 49010, Ukraine

ABSTRACT

The rapid growth of digital textual data across various domains necessitates effective methods for automatic text summarization. Large language models have demonstrated strong capabilities in generating coherent and informative summaries; however, comprehensive evaluation of their performance remains a challenging task. Existing approaches often rely on a limited number of evaluation measures and fail to capture multiple dimensions of summary quality and practical deployment factors. This study aims to conduct a comparative analysis of leading families of large language models in text summarization tasks using a multi-dimensional evaluation framework. The methodology includes experimental evaluation across diverse datasets, with summaries generated at multiple lengths. The models are assessed using a combination of evaluation measures, including lexical overlap, semantic similarity, factual consistency, and evaluation by large language models for human-like quality. Additionally, efficiency measures such as processing time and computational cost are incorporated into the analysis. The results reveal significant differences in model performance across evaluation dimensions. A key finding is the trade-off between factual consistency and perceived quality: shorter summaries tend to be more factually accurate, whereas longer summaries are rated higher in terms of completeness and readability. The proposed multi-dimensional evaluation approach provides a more comprehensive understanding of the capabilities of large language models and supports informed model selection for specific applications. The findings can be applied in real-world scenarios requiring automated processing of large-scale textual data and highlight directions for future research in improving evaluation methodologies and domain-specific adaptation of large language models.

Keywords: Information technologies; neural networks; artificial intelligence; large language models; natural language processing; automatic text summarization

ІНФОРМАЦІЯ ПРО АВТОРІВ

Єгоров Олег Йосипович - кандидат технічних наук, доцент кафедри Електронних обчислювальних машин. Український державний університет науки і технологій, вул. Лазаряна, 2. Дніпро, 49010, Україна
***Галузь досліджень*:** автоматизація роботи сортувальних станцій; імітаційне моделювання технологічних процесів

Oleh Y. Yehorov - Candidate (PhD) of Engineering Sciences, Associate Professor, Department of Computer Engineering. Ukrainian State University of Science and Technologies, 2, Lazariana St. Dnipro, 49010, Ukraine
ORCID: https://orcid.org/0000-0002-8260-9463; egoroffoleg@ukr.net
Research field: automation of sorting stations; simulation of technological processes

Борщенко Володимир Олександрович - аспірант кафедри Електронних обчислювальних машин. Український державний університет науки і технологій, вул. Лазаряна, 2. Дніпро, 49010, Україна
***Галузь досліджень*:** програмна інженерія

Volodymyr O. Borshchenko - postgraduate student, Department of Computer Engineering. Ukrainian State University of Science and Technologies, 2, Lazariana St. Dnipro, 49010, Ukraine
ORCID: https://orcid.org/0009-0007-1089-5580; v.o.borshchenko@ust.edu.ua
Research field: software engineering