УДК 004.4

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПРОДУКТИВНОСТІ ВЕЛИКИХ ТА МАЛИХ МОВНИХ МОДЕЛЕЙ В УМОВАХ ОБМЕЖЕНИХ ОБЧИСЛЮВАЛЬНИХ РЕСУРСІВ

Сарієв Г.В.;[1] Рудніченко М.Д., к.т.н., доц.[2]
[1] Міжрегіональна Академія управління персоналом, [2] Національний університет «Одеська політехніка»
[1] Україна, Київ, [2] Україна, Одеса,
sarievgeo1@gmail.com; nickolay.rud@gmail.com

Анотація. *У статті розглянуто підхід до описання та побудови системи бенчмаркінгу мовних моделей (LLM) у середовищах із обмеженими ресурсами. Пропонується концептуальна модель оцінювання, що поєднує критерії якості генерації тексту, продуктивності інференсу та ефективності використання апаратних ресурсів. Описано ключові принципи вимірювання, методи квантизації та способи інтеграції локальних моделей у споживчі пристрої. Робота узагальнює сучасні тенденції у напрямі створення малих мовних моделей (SLM) та пропонує методологічні засади для порівняльного аналізу їхньої ефективності.*

Ключові слова: *мовна модель; бенчмаркінг; квантизація; ефективність; локальні моделі; оптимізація.*

Вступ. Зі зростанням складності архітектур великих мовних моделей (LLM) значно зросли їхні апаратні вимоги — як щодо пам’яті, так і енергоспоживання [1]. Це створює виклик для їхнього розгортання на пристроях користувачів, де ресурси обмежені. У відповідь на це формується нова дослідницька парадигма, спрямована на створення компактних моделей (Small Language Models, SLM), які зберігають функціональність більших аналогів, але працюють із мінімальним навантаженням [2].

Для оцінювання доцільності використання таких моделей необхідно мати єдину систему показників, що враховує не лише точність або логічну повноту відповіді, а й швидкодію, споживання пам’яті, стабільність та здатність до роботи в офлайн-середовищі [3]. Саме тому важливим стає розроблення узагальненого підходу до бенчмаркінгу LLM, який не залежить від конкретного заліза, але враховує загальні закономірності між архітектурою моделі та ефективністю виконання.

Метою роботи є створення бенчмаркінгу, сутність якого полягає не в абсолютному вимірюванні продуктивності, а у встановленні пропорцій між якістю генерації та витратами ресурсів. Тобто ефективна модель — це не обов’язково найрозумніша, а та, яка забезпечує оптимальний баланс між обчислювальними витратами та якістю вихідного тексту [4].

Основна частина. Побудова такої системи спирається на принципи відносної порівняльності, багатокритеріальності та масштабної інваріантності. Тобто кожна модель повинна оцінюватися не у відриві від інших, а у межах спільного тестового поля. Якість генерації тексту не може бути єдиним показником. Необхідно враховувати час реакції, стабільність роботи, використання ресурсів і навіть поведінку при довгих запитах або неочікуваних структурах [5]. Методика повинна залишатися застосованою як для великих LLM, так і для квантизованих моделей, що працюють на CPU або мобільних GPU.

Бенчмаркінг мовних моделей можна розглядати як багатошаровий процес, що охоплює кілька рівнів вимірювань, кожен з яких має специфічні метрики та методики оцінки. На когнітивному рівні основна увага приділяється якості текстової генерації: логічній послідовності, когерентності, семантичній насиченості та здатності моделі розуміти контекст запиту. Для цього застосовуються різноманітні показники, такі як BLEU, ROUGE, METEOR, а також cosine similarity із еталонними відповідями [6]. Оцінка на цьому рівні дозволяє виявити, наскільки модель адекватно відтворює зміст, зберігає смислові зв'язки та адаптується до різних стилів тексту, включно з науковим, технічним та повсякденним мовленням.

На обчислювальному рівні вимірюється продуктивність моделі, що включає затримку інференсу (response latency), кількість токенів, оброблених за секунду, а також використання ресурсів — процесора, оперативної пам'яті та відеопам'яті. Ці показники дозволяють порівнювати моделі різного масштабу з точки зору швидкодії та ефективності обробки даних. Додатково проводиться аналіз поведінки моделі при тривалих сесіях, великих обсягах тексту та складних запитах, що дозволяє визначити потенційні вузькі місця та обмеження у продуктивності.

На системному рівні акцент робиться на стабільності роботи, відмовостійкості та енергоспоживанні під час тривалих експлуатаційних сесій [7]. Для вбудованих або мобільних систем цей аспект набуває особливого значення, адже саме енергоефективність визначає реальну придатність моделі для автономної роботи в умовах обмежених ресурсів. Крім того, системний рівень передбачає моніторинг інтеграції моделі з апаратними компонентами, управління кешем, контекстом та оптимізацію взаємодії з іншими сервісами. Такий багаторівневий підхід забезпечує комплексну оцінку як когнітивних, так і технічних характеристик моделей, формуючи повне уявлення про їхню придатність до реального використання [8].

Архітектура системи оцінювання мовних моделей базується на поєднанні кількох взаємопов'язаних компонентів, які забезпечують повний цикл тестування — від генерації вхідних запитів до отримання узагальнених показників ефективності. Її основна мета полягає у створенні відтворюваного середовища, здатного об'єктивно порівнювати різні моделі за якістю відповідей, швидкістю обробки та ефективністю використання ресурсів в умовах обмежених обчислювальних потужностей. Важливим аспектом є можливість інтегрувати різноманітні моделі, включаючи великі трансформери та легковагові квантизовані варіанти, у єдину тестову платформу, що дозволяє формувати порівнянні метрики для різних типів завдань.

На початковому етапі функціонує модуль формування запитів, який відповідає за створення стандартизованих тестових наборів. Ці запити охоплюють широкий спектр завдань — від простих питань природною мовою до більш складних структур, таких як логічні послідовності, порівняння текстових патернів, генерація відповідей на основі контексту чи базові математичні обчислення [9]. Окрему увагу приділено різноманітності запитів, щоб оцінка враховувала поведінку моделей при нетипових запитах, довгих діалогах або наявності неоднозначних даних.

Далі в процесі бере участь модуль вимірювання продуктивності, який фіксує часові та ресурсні характеристики виконання кожного тесту. Саме на цьому етапі відбувається збір таких показників, як середній час відгуку, використання оперативної пам'яті, навантаження центрального процесора, активність GPU та споживання енергії. У реальних реалізаціях цей компонент може ґрунтуватися на інструментах системного моніторингу — наприклад, бібліотеках psutil, nvidia-smi або вбудованих API середовища виконання, що дозволяють

знімати метрики безпосередньо під час інференсу. Додатково модуль може здійснювати логування помилок, перевантажень та непередбачуваної поведінки моделі, що забезпечує глибший аналіз стабільності та надійності системи. Таке розширене вимірювання дозволяє отримати комплексне уявлення про продуктивність моделі не лише у стандартних сценаріях, а й у стресових умовах, що критично важливо для оцінки придатності моделей для реальних користувацьких пристроїв.

Отримані дані передаються до аналітичного модуля, який виконує обробку, нормалізацію та інтеграцію результатів у зведену оцінку. Цей модуль формує агреговані показники якості, будує графіки продуктивності та обчислює інтегральні індекси ефективності. Для кількісного вираження узагальненого результату часто використовується співвідношення:

$$E = \frac{Q}{t \cdot R}$$

де Q — середній показник якості відповіді, t — середній час відгуку, а R — відносне ресурсоспоживання системи [10]. Така формула дозволяє інтерпретувати поняття «ефективності» не лише як показник точності або швидкості, а як баланс між інтелектуальною ємністю моделі та її економічністю. Завдяки цьому архітектура системи оцінювання набуває універсальності, дозволяючи порівнювати як потужні серверні моделі, так і легковагові варіанти, оптимізовані для роботи на пристроях кінцевих користувачів.

На продуктивність моделей у середовищах із обмеженими ресурсами впливають кілька ключових факторів. Перш за все — архітектурна глибина моделі, тобто кількість шарів трансформера та обсяг параметрів. Вона визначає потенціал якості, але водночас прямо впливає на затримку відповіді.

Другий фактор — стратегія квантизації, що перетворює ваги моделі у менш розрядний формат (int8, int4). Особлива увага приділяється методам квантизації та оптимізації моделей, адже вони безпосередньо впливають на швидкодію та ефективність використання апаратних ресурсів. Квантизація передбачає перетворення чисел з високою точністю (float32 або float16) у менш розрядні формати, такі як int8 або int4, що дозволяє значно зменшити обсяг пам’яті та прискорити обчислення на CPU та GPU. Проте цей процес потребує тонкого налаштування, щоб уникнути втрати семантичної точності або деградації якості генерації тексту. Серед популярних методів — посттренувальна квантизація, що застосовується після завершення навчання, та квантизація під час навчання, яка дозволяє моделі адаптуватися до обмежених розрядних форматів. Крім того, важливим елементом оптимізації є структурне спрощення моделей: видалення зайвих параметрів, об’єднання схожих вагових матриць та застосування спеціалізованих алгоритмів скорочення контексту для довгих запитів. Використання таких підходів дозволяє отримувати легковагові варіанти моделей із мінімальними втратами якості, забезпечуючи при цьому можливість швидкої інтеграції на споживчих пристроях і вбудованих системах. Додатково застосовуються методи апаратної оптимізації, зокрема векторизація операцій, пріоритизація кешу та багатопоточність, що дозволяє ефективніше використовувати ресурси сучасних процесорів і графічних адаптерів.

Третій аспект — механізми кешування та контекстного скорочення, що дозволяють підтримувати довгі діалоги без повторного обчислення всіх шарів. І нарешті, ефективність токенайзера, яка визначає кількість одиниць, необхідних для представлення тексту, суттєво впливає на загальну швидкість генерації. У сукупності ці фактори визначають, наскільки модель здатна функціонувати у середовищах із обмеженою енергією, наприклад у мобільних телефонах, дронах або автономних агентах.

Висновки. Системи оцінювання мовних моделей відіграють ключову роль у пошуку балансу між інтелектуальною складністю та практичною доцільністю їх застосування. Запропонований описовий підхід дозволяє формувати універсальні принципи бенчмаркінгу, незалежні від конкретного обладнання чи середовища. Розуміння взаємозв'язку між архітектурою, квантизацією, обчислювальними обмеженнями та якістю генерації формує основу для створення більш адаптивних і доступних моделей майбутнього.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] OpenAI (2024). *GPT-4 Technical Report*. arXiv:2303.08774.

[2] Abdin, M. et al. (2024). *Phi-3 Technical Report: Lightweight Language Models at Scale*. Microsoft Research.

[3] Touvron, H. et al. (2024). *LLaMA 3: Open and Efficient Foundation Models*. Meta AI.

[4] Sun, J. & Chen, Q. (2025). *Comparative Evaluation of Small Language Models for On-Device Inference*. *Journal of AI Systems*, 12(3), 120–134.

[5] Jiang, H. et al. (2023). *Multidimensional Metrics for Evaluating LLM Efficiency*. *IEEE Transactions on Neural Networks*, 34(4), 855–869.

[6] Papineni, K. et al. (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation*. *ACL Conference*.

[7] Zhu, L. & Wang, X. (2024). Energy-Aware Optimization for AI Edge Inference. ACM Computing Surveys, 56(7).

[8] Gholami, A. Et al. (2021). A Survey on Model Compression and Acceleration. arXiv:2106.08962.

[9] Lin, C. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. *ACL Workshop*.

[10] Dettmers, T. et al. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. *arXiv:2305.14314*.

Sariev G., Rudnichenko M.

Comparative Analysis of the Performance of Large and Small Language Models under Limited Computing Resources

Abstract. *The article discusses an approach to describing and building a language model benchmarking (LLM) system in resource-constrained environments. A conceptual evaluation model is proposed, combining the criteria of text generation quality, inference performance and efficiency in the use of hardware resources. Key measurement principles, quantization methods, and ways to integrate local models into consumer devices are described. The paper summarizes current trends in the direction of creating small language models (SLM) and offers methodological foundations for a comparative analysis of their effectiveness.*

Keywords: *language model; Benchmarking; quantization; effectiveness; local models; Optimization.*