УДК 004.85:620.9 DOI: 10.18372/2073-4751.85.21093

Дорош О.І.,
orcid.org/0000-0003-2488-0500,
oleh.dorosh@npp.kai.edu.ua,
Гузій М.М., к.т.н.,
orcid.org/0000-0003-4807-8862,
mykola.huzii@npp.kai.edu.ua

МЕТОДИ ОЦІНЮВАННЯ ЕНЕРГОЕФЕКТИВНОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Державний університет «Київський авіаційний інститут»

Вступ

В процесі еволюції штучного інтелекту концепція *"Red AI"* (досягнення точності великих мовних моделей шляхом експоненціального зростання обчислювальних витрат) змінюється на концепцію *"Green AI"* (ефективність моделі оцінюється по критерію "інтелектуальна щільність" отриманих результатів). У визначенні *"Green AI"* метрика якості моделі – це енергоефективність інференсу (*Efficiency*) при достатній когнітивній спроможності (*Accuracy*) мовної моделі [1].

Перехід до технології "*Green AI*" дозволяє отримати результати з прийнятною точністю моделі без суттєвого збільшення обчислювальних витрат за метрикою якості "результат/одиниця затраченої енергії". Критерії для оцінки якості моделей:

- Вартість навчання моделі: показники точність результатів; кількість операцій з плаваючою комою (*FPO*s); спожита електроенергія в кВт-год.
- Доступність (*Inclusivity*): "*Red AI*" доступний великим корпораціям (наприклад *Google*, *OpenAI, etc*). Технології *"Green AI"* дозволяють дослідникам з обмеженими ресурсами створювати легкі "розумні" моделі.
- Ефективність: створення легких мовних моделей з меншою кількостю параметрів. Енергетична ефективність має стати одним з основних критеріїв оцінки роботи великих мовних моделей [1].

Великі мовні моделі (*LLM — Large Language Models)* представляють собою підмножину моделей глибокого навчання, призначених для обробки та генерації тексту на основі аналізу великих обсягів даних. Стрімкий розвиток технологій *LLM* суттєво розширює можливості створення прикладних автономних інтелектуальних систем. Побудовані на архітектурі трансформерів сімейства моделей *LLM*, зокрема *GPT, Gemini, Llama, Qwen*, продемонстрували здатність до генерації тексту, перекладу, інтелектуального аналізу даних та виконання інших мультимодальних завдань.

Постановка задачі дослідження

Однією з важливих проблем практичного використання *LLM* є зростання обчислювальної складності алгоритмів при збільшенні довжини послідовності токенів, що зумовлює значні обчислювальні та енергетичні витрати інференсу, а зростання розмірів авторегресійних моделей збільшує енергетичний та екологічний вплив на екосистему. Великі обчислювальні ресурси, необхідні для навчання *LLM,* споживають значний об’єм електроенергії, складні алгоритми та тривалий процес глибокого машинного навчання моделей також збільшує загальні витрати електроенергії. Підвищення енергоефективності *LLM*

має економічне, екологічне та соціальне значення.

Завдання дослідження - оцінити енергоефективність відкритих мовних моделей (на прикладі *Cogito, Phi4, Mistral, RNJ-1),* провести порівняльний аналіз отриманих результатів, розробити практичні алгоритми підвищення енергоефективності *LLM*.

Аналіз останніх досліджень та публікацій

В роботі [2] дослідники *E. Strubell, A. Ganesh, A. McCallum,* у роботі *«Energy and Policy Considerations for Deep Learning in NLP»* першими звернули увагу на екологічний та ресурсний вплив великих моделей глибокого навчання. Автори проаналізували енергетичні витрати та вуглецевий слід (викиди CO_2) при тренуванні сучасних *NLP*-моделей і показали, що масштабні моделі можуть мати значний вуглецевий слід, що стимулює перегляд підходів до моделювання та оптимізації мовних моделей.

У технічному звіті *NVIDIA (2024)* «*Energy Efficiency Trends in AI Inference»* компанія розглядає тенденції оптимізації інференсу на сучасних прискорювачах *(GPU/TPU*), зокрема:

- важливість *CUDA*-оптимізації;
- апаратні засоби енергозбереження;
- алгоритмічні підходи для зменшення споживаної потужності при інференсі.

Основний висновок - інференс, а не тренування, стає визначальним фактором загального енергоспоживання LLM, оскільки запити до моделей виконуються мільйонами у продуктивних системах [3]. Автори направляють дискусію з академічної площини до інженерної, визначають конкретні практичні напрями оптимізації (керування пам'яттю, тонкі настройки *GPU*-ядра, прискорення конвеєрів операцій).

В роботі *«LoRA: Low-Rank Adaptation of Large Language Models»* автори запропонували метод *LoRA — низькорозмірну адаптацію параметрів*, що дозволяє зменшувати пам'яткову та обчислювальну вартість моделей при збереженні продуктивності [4]. В роботі безпосередньо не вимірюється енергоспоживання, вплив низькорозмірної адаптації параметрів моделі на енергоефективності непрямий - моделі з *LoRA* потребують менше обчислень для тонкого налаштування та інференсу, що зменшує енергоспоживання при використанні в продуктивних системах, зокрема це важливо для промислового *IoT.* Дослідження активізує напрям оптимізації параметрів та моделей при збереженні їх продуктивності.

Концепція *"Green AI"* є основою методології, яку використовує Optimum-Benchmark. Сучасні підходи до масштабованого та реплікабельного бенчмаркінгу моделей для оцінювання продуктивності *LLM* наведені в публікаціях репозиторію *Hugging Face Optimum-Benchmark.* Тестування моделей фокусується здебільшого на їх продуктивності *(latency, throughput, memory)*, без оцінки енергоефективності *LLM*, а інструментарії для оцінювання часто не включають пряме вимірювання енергоспоживання *GPU/CPU* [5].

Проблеми енергоспоживання розподілених систем є одним важливих напрямів підвищення енергоефективності моделей. У роботі *«Distributed Inference of Large Language Models: Challenges and Opportunities»* аналізуються проблеми розподіленого *inference LLM* у кластерних та хмарних середовищах [6].

***Мета дослідження*:** провести оцінювання енергоспоживання великих мовних моделей та розробити методи підвищення їх енергоефективності на протязі життєвого циклу *LLM*.

Основна частина дослідження

Системний аналіз публікацій показав, що для дослідження сучасних

великих мовних моделей практично не використовують алгоритми моделювання з відкритим кодом. Набір бенчмарків *Hugging Face Optimum-Benchmark* надає гнучкий *Python*-орієнтований інструментарій для оцінювання продуктивності трансформерних моделей у різних середовищах (*PyTorch, ONNX Runtime, TensorRT*) [7]. Його основною перевагою є кросплатформенність та відтворюваність результатів — вимірювання пропускної здатності, затримки та використання пам'яті за контрольованих умов.

У даному дослідженні автори дотримуються енергоцентричного підходу, пріоритетом є зменшення кількості спожитої електроенергії на один запит.

Методика порівняння моделей за енергоцентричним підходом базується на оцінці енерго-інтелектуальної ефективності. Інтегральний показник *E-IQ (Energy-Intelligence Quotient)* обчислюється як відношення інтелектуальної потужності моделі до її енергетичного сліду:

$$S_{eff} = A * W / E_{req} \quad (1),$$

де:

A (Accuracy): показник точності моделі за стандартизованим бенчмарком.

W (Weighting Factor): ваговий коефіцієнт складності.

E_{req} *(Energy per Request):* загальна енергія, витрачена на генерацію відповіді.

Набір бенчмарків *Optimum-Benchmark* дозволяє інтегрувати телеметричні плагіни, проте його головний акцент залишається на продуктивності. Інтеграція відбору потужності *GPU* (через *NVML*) могла б стати природним наступним кроком для об'єднання цих методологій. Крім того *Optimum Benchmark* має закритий код, кількість моделей дуже обмежена.

В даній роботі на основі показників функціонування обладнання на протязі життєвого циклу *LLM* запропоновано метод та реалізовано алгоритм аналізу енергоефективності, створено універсальний застосунок для оцінювання енергоефективності *LLM*-моделі довільної архітектури, що дозволяє прогнозувати характеристики перспективних авторегресійних *LLM* на етапі розробки та тестування.

Методика експерименту.

В роботі запропоновано метод аналізу енергоефективності *LLM* на основі показників функціонування апаратних засобів шляхом вимірювання споживаної потужності на графічному процесорі *NVIDIA RTX 3070 Ti* з використанням прискорення *CUDA* для відкритих мовних моделей *Cogito, Phi4, Mistral, RNJ-1.*

Запити кожній моделі надсилаються через *API*, вимірюються основні системні параметри (час роботи, навантаження *CPU* та *GPU*). Алгоритм достатньо універсальний, дозволяє задавати архітектуру моделі, кількість запитів до неї, а також структуру самих запитів. В результаті система видає *CSV* файл, який збирає статистику по використанню системних ресурсів за кожен запит а також обчислює середньостатистичне енергоспоживання моделі. Мовні моделі видають результат в різних форматах даних, алгоритм адаптовано до коректної обробки даних різних форматів. Розроблений застосунок має зручний інтуїтивно зрозумілий інтерфейс користувача, дозволяє легко змінювати параметри моделі в процесі тестування.

Результати експериментального дослідження енергоспоживання LLM

Експериментальне середовище для оцінювання енергоспоживання великих мовних моделей розроблено з урахуванням варіацій їх розміру, архітектурних особливостей, апаратних засобів для функціонування *LLM*. Методика дослідження передбачає

багаторазове виконання однакових запитів для кожної моделі з подальшим вимірюванням основних параметрів функціонування моделі. Відповідно до концепції "Green AI" енергоефективність *LLM* оцінюється через "інтелектуальну щільність" (кількість операцій та спожиту енергію / на одиницю результату) замість показника "токен /за секунду". Результати тестування наведені в таблиці 1.

Таблиця 1 Показники енергоефективності мовних моделей

Модель	Розмір моделі (параметри)	Об'єм пам'-яті (VRAM Q4)	Розмір MLP	Енергія за запит (Wh)	Кількість запитів на 1 kWh	Інтелектуа-льна щіль-ність (IPJ)
RNJ-1	3.8B	~2.4 GB	8,960	0.017	~58 823	4.03
Cogito	7B	~4.5 GB	11,008	0.022	~45 454	3.86
Mistral	7B	~4.8 GB	14,336	0.058	~17 241	1.39
Phi-4	14B	~7.9 GB	16,384	0.070	~14 285	1.16

На рис. 1 представлено результати моделювання енергоефективності *LLM.*

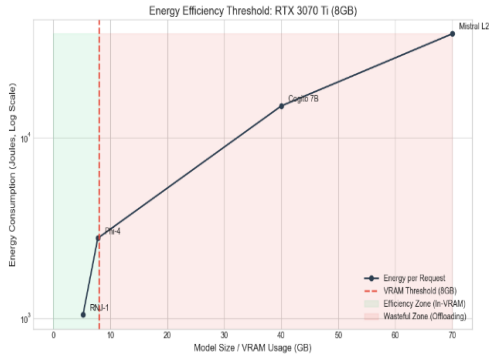

Рис. 1. Залежність енергоспоживання моделей *LLM* від об'єму *VRAM*

В зеленій зоні (зліва від пунктирної лінії) знаходяться моделі *RNU-1* та Phi-4, які повністю вміщуються у 8GB відеопам'яті, що дозволяє мінімізувати енерговитрати на обмін даними. В червоній зоні (справа від пунктирної лінії) знаходяться моделі *Cogito 7B* та *Mistral L2*, які зберігають дані в *VRAM* та частину даних в оперативній пам'яті *RAM*. Вихід моделі за межі *VRAM* приводить до значних енергозатрат.

Середні значення енергоспоживання моделей наведені на рис. 2. Результати порівняльного аналізу моделей одного класу показали, що моделі *Cogito* та *Phi-4* демонструють вищу енергоефективність порівняно з моделями *Mistral* та іншими аналогами при близькій кількості параметрів (порядку 7 млрд). Енергоспоживання визначається розміром моделі, а також особливостями архітектури, наявними апаратними ресурсами для реалізації моделі.

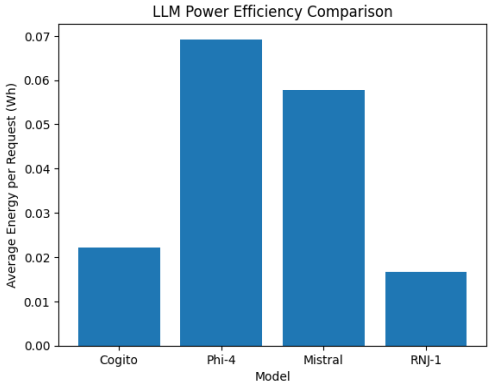

Рис. 2. Середні показники енергоспоживання *LLM.*

Ефективність функціонування *LLM* також суттєво залежить від додаткових факторів, зокрема методу токенізації, способу зберігання параметрів, простору представлення моделей.

Перспектива подальших досліджень.

Для зменшення споживання *LLM* перспективними є наступні напрями досліджень:

- Квантування (*Quantization*): зменшення точності ваг моделі, що суттєво знижує навантаження на пам'ять.
- Дистиляція (*Distillation*): навчання легкої моделі на основі знань великої мовної моделі.
- Оптимізація механізмів уваги (*Sparse Attention*): модель обробляє зв'язки в тексті по ранжируванню.

Подальші дослідження доцільно зосередити на аналізі енергоефективності великих мовних моделей (кількість параметрів до 100В), які потребують більших обчислювальних ресурсів, передусім оперативної пам’яті та серверного обладнання. Для підвищення енергоефективності *LLM* доцільно обирати метод квантування для зменшення ваги моделі до 90% об'єму доступної *VRAM*.

Одним із перспективних напрямів оптимізації є зниження напруги живлення (undervolting) під час фази декодування, при цьому *CUDA*-сумісні пристрої можуть працювати зі зниженим енергоспоживанням без втрати продуктивності. Аналогічні механізми доцільно реалізовувати також на програмному рівні.

Іншим перспективним підходом є маршрутизація запитів між різними моделями залежно від складності запиту. Запити, що не потребують глибокого семантичного аналізу, можуть оброблятися легкими моделями (3–7 млрд параметрів), а складні завдання доцільно передавати більш потужним моделям. Враховуючи, що навіть “розширені” варіанти моделей споживають на порядок більше ресурсів у режимах інтенсивного міркування, диференціація запитів дозволяє суттєво зменшити загальне енергоспоживання системи.

Для реалізації даного підходу необхідно розробити механізми класифікації запитів. Більшість запитів до *LLM* не потребує складного аналізу, використання легких моделей для їх обробки є ефективною стратегією з точки зору енергоспоживання. Розробка спеціалізованих моделей для конкретної предметної області, у поєднанні з API-маршрутизацією, дозволяє досягти подальшого зниження енерговитрат.

Висновки

Аналіз функціонування великих мовних моделей підтверджує енергоефективність концепції "*Green AI*". Дослідження показали, що *LLM*-моделі з близькими показниками точності та "інтелектуальності" результатів обчислень можуть суттєво відрізнятися за енергоефективністю.

Запропоновано метод оцінювання енергоефективності *LLM* по критерію балансу між потужністю інтелекту та фізичними обмеженнями апаратних засобів. За наявності відповідних апаратних ресурсів даний підхід можна масштабувати на великі моделі (порядку 10^{11} параметрів) для оцінювання ефективності нових *LLM*.

Впровадження технологій "*Green AI*" дозволить зменшити енергоспоживання прикладних систем штучного інтелекту, отримати значний економічний та екологічний ефект.

Література

1. Green AI / R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni. Communications of the ACM. 2020. Vol. 63, No. 12. P. 54–63.

2. E. Strubell, A. Ganesh, A. McCallum, “Energy and Policy Considerations for Deep Learning in NLP,” ACL 2019. https://aclanthology.org/P19-1355

3. NVIDIA Corp., “Energy Efficiency Trends in AI Inference,” NVIDIA Whitepaper, 2024. https://developer.nvidia.com

4. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv preprint, 2021. https://arxiv.org/abs/2106.09685

5. Hugging Face, “Optimum-Benchmark GitHub Repository,” 2025. https://github.com/huggingface/optimum-benchmark

6. Zhang et al., "Distributed Inference of Large Language Models: Challenges and Opportunities," IEEE TPDS, 2024.

7. Li et al., "Adaptive Energy-Aware Scheduling for Distributed Transformer Inference," ACM SoCC, 2024.

8. ThUnderVolt: Enabling Aggressive Voltage Underscaling and Timing Error Resilience for Energy Efficient Deep Neural Network Accelerators https://arxiv.org/abs/1802.03806

9."FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance" https://openreview.net/forum?id=XUZ2S0JVJP

Дорош О.І., Гузій М.М.

МЕТОДИ ОЦІНКИ ЕНЕРГОЕФЕКТИВНОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

У статті розглядаються методи оцінки енергоефективності авторегресивних великих мовних моделей, побудованих на архітектурі трансформерів, зокрема представників сімейств Cogito, Phi-4, Mistral та RNJ-1. З огляду на стрімке зростання обчислювальної складності механізмів уваги та відповідних енергетичних витрат під час використання моделей, дослідження зосереджується на експериментальному вимірюванні споживаної потужності моделей на споживчому графічному процесорі NVIDIA RTX 3070 Ti із використанням прискорення CUDA. Запропонований підхід дозволяє кількісно оцінити середні, мінімальні та максимальні показники енергоспоживання, а також визначити відносну енергоефективність різних моделей у типових сценаріях генерації тексту. Отримані результати доцільно використовувати для подальших досліджень енергоощадного розгортання систем штучного інтелекту та підкреслюють промислову й екологічну важливість оптимізації енергоспоживання сучасних LLM. Крім того, у статті наведено ряд інших підходів щодо покращення енергоефективності LLM таких як маршрутизація запитів а також динамічна зміна потужності при розшифруванні запиту. Комплексне використання різних методологій оптимізації є важливим фактором в розробці та впровадженні нейромереж LLM.

***Ключові слова**: Green AI, великі мовні моделі; енергоефективність; LLM; benchmarking.*

Dorosh O.I., Huzii M.M.

METHODS FOR EVALUATING THE ENERGY EFFICIENCY OF LARGE LANGUAGE MODELS

This paper examines methods for evaluating the energy efficiency of autoregressive large language models based on the transformer architecture, focusing on representatives of the Cogito, Phi-4, Mistral, and RNJ-1 families. Given the rapidly increasing computational complexity of attention mechanisms and the associated power demands during inference, the study emphasizes experimental measurement of model power consumption on a consumer-grade NVIDIA RTX 3070 Ti GPU using CUDA acceleration. The proposed approach enables quantitative assessment of average, minimum, and maximum power draw, as well as comparative analysis of relative energy efficiency across models in typical text-generation scenarios. The obtained results provide a baseline for further research on energy-efficient deployment of artificial intelligence systems and highlight the industrial and societal importance of reducing the energy footprint of modern LLMs. In addition, the article presented a number of other approaches to improving the energy efficiency of LLM, such as query routing and dynamic power adjustment when working with query decryption. The comprehensive use of various optimization methodologies is an important factor in the development and implementation of neural networks LLM.

***Keywords:** Green AI, large language models; energy efficiency; LLM; benchmarking.*

Стаття подана до редакції: 25/03/2026
Стаття прийнята до опублікування: 30/03/2026
Стаття опублікована: 27/04/2026