УДК 004.83:004.032.26 DOI: 10.31673/2412-4338.2025.048926

Бондар Віталій Володимирович
Черкаський державний технологічний університет, м. Черкаси
ORCID 0009-0002-3230-5979

Бабенко Віра Григорівна
Черкаський державний технологічний університет, м. Черкаси
ORCID 0000-0003-2039-2841

Козлов Дмитро Євгенович
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0007-1454-9036

МАСШТАБУВАННЯ ОБЧИСЛЕНЬ ПІД ЧАС ГЕНЕРАЦІЇ ЯК УНІВЕРСАЛЬНИЙ ПРИНЦИП ДЛЯ ГЕНЕРАТИВНИХ МОДЕЛЕЙ

Анотація: *Протягом останнього десятиліття покращення продуктивності нейронних мереж досягалося переважно через масштабування обчислювальних ресурсів під час навчання. Проте вичерпання доступних даних та зростання енергетичних витрат створюють фундаментальні обмеження для цієї парадигми. Дослідження спрямоване на виявлення спільних принципів альтернативного підходу – масштабування обчислень безпосередньо під час генерації, коли додаткові ресурси виділяються на етапі використання моделі. Роботу присвячено аналізу концептуальних аналогій між ітеративними процесами уточнення в різних архітектурах генеративних моделей. У великих мовних моделях масштабування реалізується через ланцюжок міркувань, де проміжні токени послідовно уточнюють представлення задачі. Дифузійні моделі досягають аналогічного ефекту через багаторазові кроки розшумлення, трансформуючи шум у структуровані дані. Моделі узгодження потоків використовують контроль точності інтегрування траєкторій між розподілами. Всі підходи об'єднує спільний принцип: виділення додаткових обчислень для послідовного уточнення трансформації імовірнісних розподілів. Дослідження встановлює, що за фіксованого бюджету, або обмеженості ресурсів, компактні моделі з додатковими обчисленнями під час генерації можуть перевершувати архітектури на порядок більші. Така стратегія забезпечує адаптивне виділення ресурсів залежно від складності запиту – властивість недосяжна при статичному масштабуванні. Аналіз обчислювальних графів виявляє лінійне масштабування витрат з кількістю ітерацій при степеневому зростанні якості. Результати формують теоретичну основу для розуміння масштабування під час генерації як універсального механізму підвищення продуктивності генеративних систем. Практичне значення полягає у зміщенні парадигми оптимізації від збільшення розміру моделей до інтелектуального розподілу обчислень, що відкриває шляхи для створення ефективніших систем в умовах обмежених ресурсів.*

Ключові слова: *машинне навчання, генеративні моделі, ланцюжок міркувань, дифузійні моделі, узгодження потоків, великі мовні моделі, масштабування, оптимальний розподіл ресурсів, обчислювальна ефективність.*

Bondar Vitalii
Cherkasy State Technological University, Cherkasy
ORCID 0009-0002-3230-5979

Babenko Vira
Cherkasy State Technological University, Cherkasy
ORCID 0000-0003-2039-2841

Kozlov Dmytro
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0009-0007-1454-9036

INFERENCE-TIME COMPUTATIONAL SCALING CALCULATIONS AS A UNIVERSAL PRINCIPLE FOR GENERATIVE MODELS

Abstract: *Over the past decade, neural network performance improvements have been achieved primarily through scaling computational resources during training. However, exhaustion of available data and rising energy costs create fundamental limitations for this paradigm. This study identifies common principles of an alternative approach – scaling computation directly during generation, where additional resources are allocated at the deployment stage. The work analyzes conceptual analogies between iterative refinement processes across different generative model architectures. In*

large language models, scaling is realized through chain-of-thought techniques, where intermediate tokens sequentially refine task representations. Diffusion models achieve analogous effects through multiple denoising steps, transforming noise into structured data. Flow matching models utilize control over integration precision of trajectories between distributions. All approaches share a common principle: allocating additional computation for sequential refinement of probability distribution transformations. The study establishes that under fixed budgets, or limited resources, compact models with additional inference-time computation can outperform architectures an order of magnitude larger. This strategy enables adaptive resource allocation depending on query complexity – a property unattainable with static scaling. Computational graph analysis reveals linear cost scaling with iteration count alongside power-law quality growth. The findings form a theoretical foundation for understanding inference-time scaling as a universal mechanism for enhancing generative system performance. The practical significance lies in shifting optimization paradigms from increasing model size to intelligent computation distribution, opening pathways for more efficient systems under resource constraints.

Keywords: *machine learning, generative models, chain-of-thought reasoning, diffusion models, flow matching, large language models, scaling, optimal resource allocation, computational efficiency.*

Постановка проблеми

Протягом останнього десятиліття основним шляхом покращення продуктивності нейронних мереж було масштабування обчислювальних ресурсів під час навчання. Дослідження встановили чіткі степеневі закони, що описують залежність якості моделі від її розміру, обсягу навчальних даних та обчислювальних витрат на тренування [1, 2]. Цей підхід забезпечив вражаючі результати у різних областях машинного навчання, від обробки природної мови до комп'ютерного зору, дозволивши створити моделі з мільярдами параметрів, здатні розв'язувати складні задачі. Проте сучасні дослідження виявляють фундаментальні обмеження такої парадигми: вичерпання доступних навчальних даних, зростання енергетичних витрат на тренування наднових моделей та практичні складнощі розгортання моделей гігантського розміру.

Альтернативою традиційному підходу стає концепція масштабування обчислень під час генерації, що передбачає виділення додаткових обчислювальних ресурсів безпосередньо в процесі генерації відповіді. У великих мовних моделях цей підхід реалізується через техніки ланцюжка міркувань, де модель генерує проміжні кроки обґрунтування перед фінальною відповіддю [3], або через паралельне семплювання з подальшим вибором найкращого результату [4]. Дослідження демонструють, що невеликі моделі з додатковими обчисленнями під час генерації можуть перевершувати моделі у десятки разів більші за розміром, що відкриває нову вимірність для масштабування без необхідності повторного навчання. Водночас аналогічні механізми почали з'являтися і в інших типах генеративних моделей: дифузійні моделі та узгодження потоків використовують багаторазове обчислення задля покращення якості [5].

Незважаючи на успіхи окремих досліджень, залишається відкритим питання про універсальність принципів масштабування під час генерації для різних архітектур генеративних моделей. Поточні роботи здебільшого фокусуються на специфічних підходах для конкретних типів моделей, не розглядаючи спільні концептуальні основи цього явища. Відсутність єдиної теоретичної бази ускладнює розуміння фундаментальних механізмів, що забезпечують ефективність масштабування під час генерації, та перешкоджає розробці універсальних методів оптимізації. Це створює потребу в систематичному аналізі спільних принципів масштабування обчислень під час генерації для різних класів генеративних моделей – від авторегресивних мовних моделей до моделей на основі дифузії та узгодження потоків.

Аналіз останніх досліджень і публікацій

Фундаментальні дослідження закономірностей масштабування нейронних мереж встановили емпіричні степеневі залежності між продуктивністю моделей та обчислювальними ресурсами, витраченими на навчання. Результати дослідження [1] показали, що втрати мовних моделей масштабуються як степеневі функції від розміру моделі, обсягу даних та загальної кількості обчислень, при цьому ці залежності зберігаються на сім порядків величини. У роботі [2] дослідили закони масштабування та уточнили оптимальне співвідношення між розміром моделі та обсягом навчальних даних для фіксованого обчислювального бюджету. Ці роботи створили теоретичну основу для розуміння масштабування під час навчання, проте не розглядали можливості масштабування на етапі використання моделі.

Масштабування обчислень під час генерації у великих мовних моделях отримало розвиток після впровадження техніки ланцюжка міркувань. У роботі [3] автори продемонстрували, що генерація проміжних кроків обґрунтування перед фінальною відповіддю суттєво покращує здатність моделей розв'язувати складні задачі багатокрокового міркування. Цей ефект проявляється як властивість

моделей масштабу близько ста мільярдів параметрів, причому додаткові обчислення виділяються безпосередньо під час генерації, а не закладаються в архітектуру на етапі навчання.

Систематичне дослідження оптимального розподілу обчислень показало суттєве покращення ефективності ресурсів. У роботі [4] дослідники встановили, що обчислювально оптимальна стратегія може підвищити ефективність масштабування вчетверо порівняно з підходом best-of-N. Ключовим висновком стало спостереження, що ефективність стратегій критично залежить від складності запиту, мотивуючи адаптивний розподіл ресурсів. В умовах фіксованого бюджету невелика модель з додатковими обчисленнями може перевершити модель у чотирнадцять разів більшу. У роботі [5] автори розширили це розуміння через аналіз паралельного семплювання, показавши, що покриття задач масштабується з кількістю семплів на чотири порядки величини. Це демонструє фундаментальну властивість стохастичної генерації: збільшення спроб дозволяє досліджувати ширший простір рішень, створюючи потребу в ефективних механізмах верифікації результатів.

Всебічний аналіз масштабування для складних задач виявив нюанси та обмеження цього підходу. У науковій праці [6] дослідники встановили, що переваги варіюються залежно від типу задачі та зменшуються зі зростанням складності.

Дифузійні моделі демонструють природну можливість варіювання обчислювальних витрат через регулювання кроків денойзингу. У [7] автори встановили базовий механізм ітеративного видалення шуму через навчання послідовності ймовірнісних моделей, використовуючи зв'язок між дифузійними моделями та узгодження оцінки розшумлення (denoising score matching). Збільшення кроків традиційно покращує якість, проте ефект швидко згасає через накопичення похибок апроксимації та дискретизації. Надалі запропонували навчання через оцінку градієнтів розподілу на різних рівнях шуму [8], а подальше узагальнення через стохастичні диференціальні рівняння об'єднало різні підходи в єдину структуру [9].

Сучасні дослідження запропонували принципово новий підхід через пошук кращих початкових умов. У дослідження поданому в [10] продемонстровано, що виділення ресурсів на пошук оптимальних шумів суттєво покращує якість набагато більше за просте збільшення кроків розшумлення. Замість залучення всіх обчислень для здійснення обрахунків під час розшумлення, частина ресурсів спрямовується на дослідження простору початкових шумів. Цей підхід демонструє концептуальну схожість з методами пошуку в мовних моделях.

Моделі узгодження потоків пропонують альтернативний підхід через регресію векторних полів умовних траєкторій ймовірності. У роботі [11] автори представили метод навчання безперервних нормалізуючих потоків, що узагальнює дифузійні шляхи та дозволяє використовувати інтерпретацію з теорії оптимального транспорту. Узгодження потоків забезпечує пряміші траєкторії від шуму до даних, полегшуючи масштабування обчислень під час генерації. Застосування підходу для задач моделювання демонструє масштабованість до складних наукових задач [12]. Дослідження масштабування обчислень під час генерації для зовнішніх моделей винагороди у навчанні з підкріпленням розширює концепцію на задачі оцінки якості, показуючи, що правильні методи навчання забезпечують більш ефективну масштабованість порівняно з масштабуванням самих моделей [13].

Теоретичні роботи формують єдине бачення різних типів моделей через призму трансформації імовірнісних розподілів. У роботі [14] автори розглядають генеративні моделі як функції трансформації ймовірності, надаючи концептуальну основу для розуміння спільних принципів. У цьому контексті масштабування під час генерації інтерпретується як виділення обчислень для точнішої апроксимації трансформації через ітеративні уточнення. Дослідження застосування генеративних моделей підкреслюють практичну важливість розуміння компромісів між підходами [15].

Проте поточні роботи фокусуються на специфічних техніках масштабування під час генерації – ланцюжок роздумів, різонінг та інші для великих мовних моделей, пошук шумів для дифузійних моделей, варіювання траєкторій для моделей узгодження потоків – не розглядаючи фундаментальні концептуальні аналогії.

Мета і задачі дослідження

Метою даного дослідження є виявлення та систематизація спільних принципів масштабування обчислень під час генерації для різних архітектур генеративних моделей, що дозволяє розглядати цей підхід як універсальний механізм підвищення продуктивності, а не специфічну техніку окремих типів моделей. Дослідження спрямоване на встановлення концептуальних аналогій між ітеративними процесами уточнення в авторегресивних мовних моделях, дифузійних моделях та моделях узгодження потоків, обґрунтовуючи можливість застосування єдиних принципів оптимізації обчислювальних ресурсів під час генерації.

Результати дослідження

Генеративні моделі різних архітектур можна концептуально об'єднати через призму ітеративного уточнення трансформації імовірнісних розподілів. Сучасні роботи розглядають глибокі генеративні моделі як функції трансформації ймовірності, що перетворюють простий базовий розподіл у складний цільовий розподіл даних [15]. У цьому контексті масштабування під час генерації природно інтерпретується як виділення додаткових обчислювальних ресурсів для послідовного уточнення цієї трансформації, незалежно від специфічної архітектури моделі. Ключова гіпотеза полягає в тому, що ефективність такого масштабування є фундаментальною властивістю процесу ітеративного наближення, а не артефактом конкретної реалізації.

У великих мовних моделях масштабування під час генерації реалізується через генерацію проміжних кроків міркування перед фінальною відповіддю. Техніка ланцюжка міркувань дозволяє моделі розкладати складну задачу на послідовність простіших підзадач, де кожен проміжний токен виконує роль уточнення внутрішнього представлення проблеми [3]. Формально, якщо стандартна авторегресійна генерація моделює розподіл $P(y|x)$ безпосередньо, то з використанням ланцюжка міркувань модель обчислює $P(y|x) = \sum_k P(y|r,x)P(r,x)$, де r представляє послідовність проміжних кроків обґрунтування довжиною k токенів. Збільшення k відповідає залученню більших обчислювальних ресурсів під час генерації та емпірично демонструє покращення якості для складних задач багатокрокового міркування [4].

Дифузійні моделі демонструють структурно аналогічний механізм через послідовність кроків розшумлення. Процес генерації розпочинається з випадкового шуму $x_T \sim N(0, I)$ та ітеративно уточнює наближення до цільового розподілу через послідовність $x_T \to x_{T-1} \to \cdots \to x_1 \to x_0$, де кожен крок t моделює зворотній дифузійний перехід $p_\theta(x_{t-1}|x_t)$ (Ho 2020). Збільшення кількості кроків T є прямим аналогом збільшення довжини ланцюжка міркувань у мовних моделях – обидва підходи виділяють додаткові обчислення для послідовного уточнення результату. Важливо, що сучасні дослідження виходять за межі простого збільшення кроків розшумлення, пропонуючи виділення ресурсів на пошук оптимальних початкових шумів, що концептуально відповідає стратегіям пошуку в мовних моделях [10].

Формальна подібність проявляється в структурі обчислювальних графів обох типів моделей. Позначимо через $C(n)$ обчислювальну складність генерації з n ітеративними уточненнями. Для мовних моделей маємо $c_{LLM}(n) = n \cdot C_{forward}$, де $C_{forward}$ – вартість одного прямого проходу через трансформер для генерації токена різонінгу. Для дифузійних моделей аналогічно $C_{diff}(n) = n \cdot C_{denoise}$, де $C_{denoise}$ – вартість одного кроку розшумлення через U-Net або іншу архітектуру розшумлення. В обох випадках загальні обчислювальні витрати масштабуються лінійно з кількістю ітерацій уточнення, а якість результату *Q(n)* демонструє логарифмічне або степеневе зростання, створюючи характерний компроміс між точністю та обчислювальними ресурсами.

Концептуальну схожість також можна проілюструвати через подібність послідовностей обчислень. Процес генерації в авторегресивних мовних моделях: початковий запит *x* проходить через послідовність блоків різонінгу $[R_1, R_2, \dots, R_k]$, де кожен блок R_i є обчисленням мережі на основі початково запиту та вже обрахованих токенів різонінгу, поступово наближаючись до фінальної відповіді *y*. Дифузійна генерація: початковий шум x_T послідовно трансформується через блоки розшумлення $[D_T, D_{T-1}, \dots, D_1]$, де кожен блок D_t застосовує навчену нейронну мережу для передбачення та видалення шуму, поступово наближаючись до чистих даних x_0. Критично, що в обох випадках кожен наступний крок залежить від результату попереднього, створюючи каскадну структуру послідовного уточнення, а збільшення кількості кроків відповідає залученню додаткових обчислювальних ресурсів у процес генерації.

Моделі узгодження потоків розширюють ці принципи через безперервну інтерполяцію між розподілами. Узгодження потоків навчає векторне поле $v_t(x)$, що визначає траєкторію трансформації від базового розподілу до цільового через звичайне диференціальне рівняння $\frac{dx}{dt} = v_t(x)$ [11]. Масштабування під час генерації реалізується через контроль точності чисельного інтегрування цього рівняння – збільшення кількості кроків інтегратора відповідає виділенню більших обчислювальних ресурсів для точнішого наближення траєкторії. Такий підхід забезпечує пряміші шляхи трансформації порівняно з дифузійними моделями, проте фундаментальний принцип залишається незмінним: якість результату покращується через виділення додаткових обчислень на послідовне уточнення процесу генерації [12].

Універсальність масштабування під час генерації має важливі практичні наслідки для оптимального розподілу обчислювальних ресурсів у генеративних системах. Дослідження показують,

що за фіксованого обчислювального бюджету стратегія з меншою моделлю та більшою кількістю ітеративних уточнень може перевершувати підхід з великою моделлю без додаткових обчислень під час генерації [4]. Це спостереження справедливе як для мовних моделей, де невелика модель з різонінг може перевершити модель у чотирнадцять разів більшу, так і для дифузійних моделей, де компактна архітектура з оптимізованим семплюванням досягає якості значно більших систем [10]. Ефективність стратегії критично залежить від складності задачі та якості механізмів верифікації проміжних результатів, проте сама можливість такого компромісу демонструє фундаментальну властивість генеративних систем незалежно від їхньої архітектури.

Аналіз компромісів між масштабуванням моделей та масштабуванням під час генерації виявляє зміщення парадигми оптимізації генеративних моделей. Традиційний підхід фокусувався на максимізації якості через збільшення розміру моделі та обсягу навчальних даних, що описується степеневими законами масштабування [1, 2]. Проте в умовах вичерпання доступних даних та зростання вартості навчання надвеликих моделей, масштабування під час генерації пропонує альтернативну вісь оптимізації. Ключова перевага полягає в можливості адаптивного виділення обчислень залежно від складності конкретного запиту – прості задачі вирішуються швидко з мінімальними ітераціями, тоді як складні отримують додаткові обчислювальні ресурси для досягнення необхідної якості. Така гнучкість недосяжна в парадигмі масштабування моделей, де розмір моделі фіксується на етапі навчання та не може змінюватися залежно від запиту.

Висновки і перспективи подальших досліджень

Масштабування обчислень під час генерації є універсальним принципом для генеративних моделей різних архітектур, а не специфічною технікою окремих класів систем. Дослідження виявило концептуальні аналогії між ітеративними процесами уточнення в авторегресивних мовних моделях, дифузійних моделях та моделях узгодження потоків, де збільшення кількості обчислювальних кроків забезпечує послідовне покращення якості результату. У всіх розглянутих архітектурах додаткові обчислювальні ресурси використовуються для уточнення апроксимації трансформації імовірнісних розподілів через каскадні структури послідовного уточнення.

Практичне значення цього принципу полягає у зміщенні парадигми оптимізації від масштабування моделей до адаптивного розподілу обчислень. За фіксованого бюджету компактні моделі з додатковими обчисленнями під час генерації можуть конкурувати з архітектурами на порядок більшого розміру, причому гнучкість виділення ресурсів залежно від складності запиту забезпечує ефективність недосяжну для статичних рішень. Це відкриває шляхи для розробки генеративних систем з оптимізованим співвідношенням якості та обчислювальних витрат.

Перспективним напрямком є дослідження оптимальних стратегій розподілу ресурсів між параметрами моделі та обчисленнями під час генерації для конкретних класів задач. Потребують вивчення механізми автоматичної адаптації кількості ітерацій залежно від складності запиту, методи верифікації проміжних результатів для ефективного пошуку, а також теоретичне обґрунтування степеневих законів масштабування для обчислень під час генерації. Окремий інтерес становить узагальнення принципів масштабування на мультимодальні генеративні системи та дослідження компромісів між латентністю і якістю в умовах реального часу.

Список використаної літератури

1. Kaplan J., McCandlish S., Henighan T. et al. Scaling laws for neural language models. arXiv preprint. 2020. arXiv:2001.08361. URL: https://arxiv.org/abs/2001.08361

2. Hoffmann J., Borgeaud S., Mensch A et al. Training compute-optimal large language models. Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 30016–30030. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html

3. Wei J., Wang X., Schuurmans D. et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

4. Snell C., Lee J., Xu K. et al. Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint. 2024. arXiv:2408.03314. URL: https://arxiv.org/abs/2408.03314

5. Brown B., Juravsky J., Ehrlich R. et al. Large language monkeys: Scaling inference compute with repeated sampling. arXiv preprint. 2024. arXiv:2407.21787. URL: https://arxiv.org/abs/2407.21787

6. Balachandran V., Chen J., Chen L. et al. Inference-time scaling for complex tasks: Where we stand and what lies ahead. arXiv preprint. 2025. arXiv:2504.00294. URL: https://arxiv.org/abs/2504.00294

7. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 6840–6851. URL: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

8. Song Y., Ermon S. Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems. 2019. Vol. 32. P. 11895–11907. URL: https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html

9. Song Y., Sohl-Dickstein J., Kingma D. P. et al. Score-based generative modeling through stochastic differential equations. arXiv preprint. 2020. arXiv:2011.13456. URL: https://arxiv.org/abs/2011.13456

10. Ma N., Tong S., Jia H. et al. Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint. 2025. arXiv:2501.09732. URL: https://arxiv.org/abs/2501.09732

11. Lipman Y., Chen R. T. Q., Ben-Hamu H. et al. Flow matching for generative modeling. arXiv preprint. 2022. arXiv:2210.02747. URL: https://arxiv.org/abs/2210.02747

12. Wildberger J., Dax M., Buchholz S. et al. Flow matching for scalable simulation-based inference. Advances in Neural Information Processing Systems. 2023. Vol. 36. P. 16837–16864. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/3663ae53ec078860bb0b9c6606e092a0-Abstract-Conference.html

13. Liu Z., Wang P., Xu R. et al. Inference-time scaling for generalist reward modeling. arXiv preprint. 2025. arXiv:2504.02495. URL: https://arxiv.org/abs/2504.02495

14. Bondar V., Babenko V., Trembovetskyi R. et al. Deep generative models as the probability transformation functions. arXiv preprint. 2025. arXiv:2506.17171. URL: https://arxiv.org/abs/2506.17171

15. Bondar V., Babenko V. Investigation of the main aspects of the generative models applications in machine learning. Herald of Khmelnytskyi National University. Technical sciences. 2025. Vol. 347, No. 1. P. 69–72. URL: https://doi.org/10.31891/2307-5732-2025-347-8