3

Проблемы искусственного интеллекта. 2026. № 3 (42). С.122-134
Problems of Artificial Intelligence. 2026; 3(42): 122-134
Системный анализ, управление и обработка информации, статистика
Научная статья

УДК 004.8:001.18:519.25 **DOI** 10.24412/2413-7383-2026-3-42-122-134

А. В. Звягинцева[1], А. Р. Бражников[2]
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Донецкий государственный университет»
283001, г. Донецк, ул. Университетская, д. 24.
[1]a.zvyagintseva@donnu.ru; ORCID 0000-0003-2218-9230
[2]yanox123321@gmail.com; ORCID 0009-0005-4420-9206

К ВОПРОСУ О ФЕНОМЕНОЛОГИЧЕСКИХ ЗАКОНОМЕРНОСТЯХ РАЗВИТИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ВОЗМОЖЕН ЛИ АНАЛОГ ЗАКОНА МУРА?

A. V. Zviagintseva, A. R. Brazhnikov
Federal State Budgetary Educational Institution of Higher Education “Donetsk State University”
283001, Donetsk, Universitetskaya str. 24

ON THE QUESTION OF PHENOMENOLOGICAL REGULARITIES IN THE DEVELOPMENT OF LARGE LANGUAGE MODELS: IS AN ANALOGUE OF MOORE'S LAW POSSIBLE?

Систематизированы количественные характеристики современных больших языковых моделей и поставлен вопрос о феноменологической закономерности, аналогичной закону Мура. Предложено пятимерное пространство параметров, сводимое к двухфакторному уравнению состояния. На базе десяти моделей 2019–2025 гг. (зарубежных и российских) показано, что динамика качества (MMLU) описывается насыщающейся логистической кривой, обучающий бюджет растёт экспоненциально, а удельная стоимость вывода – затраты на единицу качества – устойчиво снижается. Отдельно рассмотрена применимость к моделям самостоятельного развёртывания при квантизации и разреженных Mixture-of-Experts архитектурах.
Ключевые слова: большие языковые модели; закон Мура; законы масштабирования; MMLU; MERA; GigaChat; YandexGPT; DeepSeek; Claude; self-hosted модели; квантизация.

This paper systematizes the quantitative characteristics of modern large language models and raises the question of a phenomenological regularity analogous to Moore's law. A five-dimensional parameter space is proposed and reduced to a two-factor equation of state. On ten models for 2019–2025 (foreign and Russian) it is shown that the quality dynamics (MMLU) follow a saturating logistic curve, the training compute budget grows exponentially, and the unit cost of inference – the cost per unit of quality – steadily decreases. The applicability to self-hosted models under quantization and sparse Mixture-of-Experts architectures is considered separately.
Key words: large language models; Moore’s law; scaling laws; MMLU; MERA; GigaChat; YandexGPT; DeepSeek; Claude; self-hosted models; quantization.

Введение

Закон Мура [1] более пяти десятилетий определял планирование полупроводниковой отрасли, задавая ориентир для инвестиций, стандартов и дорожных карт. Закономерен вопрос: существует ли аналогичное обобщение для моделей искусственного интеллекта (ИИ)? Актуальность определяется тремя обстоятельствами: инвестиции в ИИ-инфраструктуру требуют устойчивых количественных ориентиров; технологический форсайт нуждается в операциональных метриках прогресса [2], [3]; сравнительный анализ моделей методологически неустоявшийся [4]. Это особенно значимо для отечественной отрасли с проприетарными моделями GigaChat [5] и YandexGPT [6], верифицируемыми бенчмарком MERA [7].

Цель работы – систематизация количественных характеристик моделей ИИ, анализ их динамики и формулировка исследовательской программы по построению закономерностей, аналогичных закону Мура, с разделением моделей на закрытые проприетарные и доступные для самостоятельного развёртывания.

Постановка задачи: на основе систематизации количественных характеристик больших языковых моделей (БЯМ) из открытых данных бенчмарков [4], [7], [8] и техотчётов [5], [6], [9-14] за 2019-2025 гг., а также сопоставительного регрессионного анализа временных рядов [15] провести оценку динамики БЯМ и определить возможность существования феноменологической закономерности, аналогичной закону Мура. Для чего выделить многомерное пространство параметров, рассмотреть применимость выявленных тенденций к моделям с открытыми весами (self-hosted) и сформулировать проверяемые гипотезы о динамике качества, вычислительного бюджета и удельной стоимости вывода.

Количественное пространство состояний моделей ИИ

На основе открытых источников [3], [4], [8], [16] выделим пятимерное пространство количественных характеристик современных больших языковых моделей (в соответствии с таблицей 1).

Таблица 1 – Пятимерное пространство количественных характеристик моделей ИИ

Координата	Характеристика	Ед. измерения	Источник
X_1	Общее число параметров модели	млрд	Технические отчёты [5], [9-12]
X_2	Объём обучающей выборки	трлн токенов	Технические отчёты [5], [9-12]
X_3	Качество решения задач (MMLU)	%	HELM [4]; MERA [7], [5], [6]
X_4	Стоимость вывода (inference)	USD / млн ток.	Прайс-листы разработчиков [9-14]; обзор [16]
X_5	Качество генерации кода (HumanEval)	%	Open LLM Leaderboard [8]

Здесь и далее координата качества обозначается символом X_3; наименование MMLU сохраняется лишь для самого бенчмарка (и подписей осей и столбцов), которым измеряется X_3. Представленное пространство охватывает морфологические (X_1, X_2), функциональные (X_3, X_5) и экономические (X_4) характеристики. Его принципиальное отличие от одномерного закона Мура [1] состоит в необходимости совместного описания нескольких координат, что ставит задачу редукции многомерной динамики к аналитически разрешимым закономерностям [17].

Формализация уравнения состояния. Координаты X_1–X_5 не независимы, а связаны системой соотношений, понижающей размерность пространства состояний. Качество X_3 определяется обучающим бюджетом $C \approx 6 \cdot X_1 \cdot X_2$ (стандартная оценка для плотных моделей) через насыщающуюся функцию $X_3 = Q(C)$, где $Q: \mathbb{R}^+ \to [0{,}100]$ отражает ограниченность метрики сверху; качество генерации кода X_5 коррелирует с X_3 как другая проекция общей способности. Стоимость вывода X_4 определяется не полным, а активируемым числом параметров (для MoE существенно меньшим полного X_1), что ослабляет прямую связь X_4 с X_1. Тем самым состояние модели задаётся эффективно двумя управляющими величинами – обучающим бюджетом C и долей активируемых параметров, – а прочие координаты являются их функциями.

Ключевым эмпирическим индикатором динамики выступает метрика X_3 (MMLU). Данные по десяти моделям за 2019–2025 гг. (в соответствии с таблицей 2) указывают на устойчивый рост качества решения задач: более чем в два раза за пятилетний период, что по темпу сопоставимо с исходным законом Мура [1].

Таблица 2 – Динамика ключевых характеристик моделей ИИ (2019–2025 гг.)

Модель	Год	Архитектура	X_1, млрд	MMLU, %	HumanEval, %
GPT-3 (OpenAI) [9]	2020	dense	175	43,9	≈5
GPT-4 (OpenAI) [10]	2023	MoE (зак.)	≈1800	86,4	67,0
DeepSeek-V2 [11]	2024	MoE	236 / 21	77,8	73,8
Qwen 2.5-72B [14]	2024	dense	72,7	86,1	≈86
Llama 3.1-70B [13]	2024	dense	70	83,6	80,5
GigaChat 2 MAX [5]	2025	MoE	н/д (зак.)	86,0	87,1
Claude Sonnet 4 [18]	2025	dense (зак.)	н/д (зак.)	84,8	≈93
Claude Opus 4 [18]	2025	dense (зак.)	н/д (зак.)	87,4	≈96
YandexGPT 5 Pro [6]	2025	dense (зак.)	н/д	83,0	–
DeepSeek-R1 [12]	2025	MoE, RL	671 / 37	90,8	≈96

Сокращение «зак.» обозначает закрытые проприетарные данные. Составлено по данным [4-6], [8-14], [18].

Оценка $X_1 \approx 1800$ млрд для GPT-4 основана на неофициальных публикациях и официально не подтверждена. Значения MMLU получены по протоколам разработчиков: GPT-3, GPT-4, Llama 3.1-70B и Qwen 2.5-72B – 5-shot; DeepSeek-V2, DeepSeek-R1, Claude Opus 4 и Claude Sonnet 4 – с цепочкой рассуждений (CoT); GigaChat 2 MAX и YandexGPT 5 Pro – по данным MERA; поэтому значения строго не сопоставимы.

Из таблицы 2 следуют четыре наблюдения. За пятилетний период (2020–2025 гг.) значение метрики X_3 возросло с 43,9% до 90,8%, прирост 46,9 процентных пункта (п.п., ≈2,1×). Лидирующие модели 2024–2025 гг. (DeepSeek-R1, Claude Opus 4) приближаются к уровню эксперта-человека (≈89,8%) [19], что свидетельствует о насыщении бенчмарка. Начиная с 2024 г. наблюдается переход к разреженным архитектурам Mixture-of-Experts (MoE): DeepSeek-V2 и DeepSeek-R1 [11], [12], GigaChat 2 [5] – при которых число активируемых параметров существенно меньше общего, что радикально изменяет соотношение «качество – стоимость вывода».

Предварительный регрессионный анализ. Аппроксимация динамики X_3 (MMLU) по выборке из десяти моделей (в соответствии с таблицей 2, рисунком 1) линейной зависимостью от года даёт средний прирост около 7,7 п.п./год (коэффициент детерминации $R^2 \approx 0{,}81$); однако оценка определяется преимущественно ранней точкой (GPT-3, 2020 г.). При её исключении на интервале 2023–2025 гг. наклон падает до ≈1,2 п.п./год ($R^2 \approx 0{,}06$), а значения X_3 концентрируются в диапазоне 78–91%. Ввиду малого объёма и неравномерности выборки (по одной модели на 2020 и 2023 гг., восемь –

на 2024–2025 гг.) обнаруженное замедление на интервале 2023–2025 гг. следует трактовать не как подтверждённое насыщение, а как предварительный сигнал – исследовательскую гипотезу, требующую верификации на расширенном корпусе (100+ моделей) с адаптивными бенчмарками (MMLU-Pro [20], MATH-500, Arena-Hard).

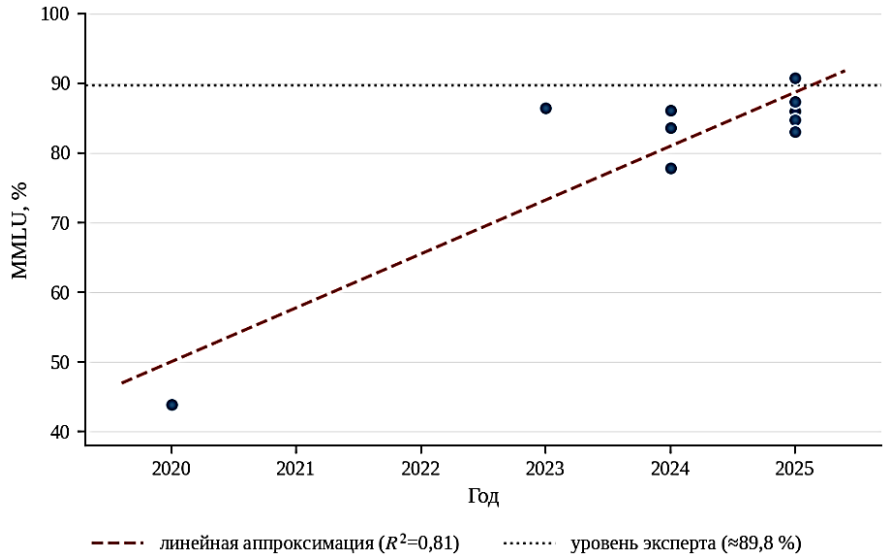


Рисунок 1 – Динамика метрики MMLU по годам (десять моделей, в соответствии с таблицей 2): линейная аппроксимация и уровень эксперта-человека. Составлено авторами

Существующие аналоги закона Мура

Обобщённый закон Мура и его статистические модели для развития вычислительных технологий ранее рассматривались в работах [21], [22]. Наиболее строгим из существующих аналогов являются законы масштабирования. Дж. Капланом и соавторами [23] показано, что функция потерь L убывает по степенному закону при росте N, D и C: $L(N, D) \approx A \cdot N^{-\alpha} + B \cdot D^{-\beta} + L_\infty$ (здесь N – число параметров, D – объём обучающих данных, A, B – масштабные коэффициенты, α, β – показатели степени, L_∞ – неустранимый уровень потерь). В работе Дж. Хоффмана и соавторов («Chinchilla») [24] получено оптимальное соотношение $N \propto C^{0,5}$ (против $N \propto C^{0,73}$ у Каплана [23]). Расхождение в полтора раза свидетельствует о чувствительности к условиям обучения; тем не менее эти результаты приняты в отрасли в качестве стандарта при проектировании архитектур. При согласованном расписании («Chinchilla») данные и параметры должны масштабироваться примерно одинаково – около 20 обучающих токенов на параметр, $X_2 \approx 20 \cdot X_1$. Это означает, что при компьютер-оптимальном обучении X_1 и X_2 связаны, что дополнительно понижает число независимых координат пространства состояний. Epoch AI [3] зафиксировал экспоненциальный рост обучающего вычислительного бюджета: с 2012 по 2023 г. рост более 3×10^5 раз.

Параллельно с ростом возможностей наблюдается рост эффективности на единицу вычислений. DeepSeek-V2 [12] при общем числе параметров 236 млрд активирует в режиме вывода лишь 21 млрд (MoE), обеспечивая качество LLaMA 3 70B при стоимости вывода приблизительно на два порядка ниже. Семейство GigaChat [5] использует аналогичный принцип: открытая модель GigaChat-A3B при 20 млрд общих параметров активирует около 3,3 млрд, демонстрируя качество 7-8-миллиардных dense-моделей.

Применимость к моделям самостоятельного развёртывания

Самостоятельное развёртывание (self-hosted inference) моделей с открытыми весами в 2024–2025 гг. приобрело массовый характер: открытые семейства Llama, Qwen, DeepSeek, GigaChat-A3B и YandexGPT 5 Lite доступны для локального запуска через типовые рантаймы [16]. Аппаратная граница self-hosted определяется требованием размещения весов модели в видеопамяти ускорителя: $V = N \cdot b/8$, где V – требуемый объём видеопамяти, N – число параметров, b – битность представления весов. Это даёт прямую связь между размерностью модели и доступностью её для конечного пользователя (в соответствии с таблицей 3).

Таблица 3 – Соотношение размерности модели и аппаратных требований к её самостоятельному развёртыванию. Составлено авторами по данным [5], [6], [11–14], [16]

Модель	Параметры	VRAM (BF16)	VRAM (INT4)	Аппаратная конфигурация
DeepSeek-R1-Distill 1,5B	1,5 млрд	≈3 ГБ	<1 ГБ	ноутбук без GPU
Qwen 2.5-7B; LLaMA 3.1-8B; YandexGPT 5 Lite	7–8 млрд	≈16 ГБ	≈4–6 ГБ	GPU ≥ 8 ГБ; Apple Silicon ≥ 16 ГБ
DeepSeek-R1-Distill 14B; Qwen-Coder 32B	14–32 млрд	≈28–64 ГБ	≈9–20 ГБ	RTX 3090/4090 (24 ГБ)
Llama 3.1-70B; Qwen 2.5-72B	70–73 млрд	≈140 ГБ	≈40 ГБ	2 × RTX 4090 либо A100 80 ГБ
DeepSeek-R1 (полная)	671/37	≈1,3 ТБ	≈400 ГБ	кластер 6–8 × A100 80 ГБ (INT4)

Анализ таблицы 3 позволяет сформулировать три тезиса. Во-первых, экспоненциальный рост frontier-моделей по X_1 делает полные версии недоступными пользователю: DeepSeek-R1 (671 млрд) при 4-битной квантизации требует ≈400 ГБ VRAM – кластер 6–8 × A100 80 ГБ [12], [16]. Во-вторых, здесь преобладает иной механизм прогресса – не рост числа параметров, а дистилляция и архитектурная оптимизация. В-третьих, квантизация (BF16 → INT4) играет роль, аналогичную закону Мура: при неизменной размерности она снижает ресурсы вчетверо; для Qwen 2.5–72B [14] и Llama 3.1-70B [13] это переход от ЦОД к двум RTX 4090. Таким образом, прогресс в self-hosted определяется двумя совместными переменными – размерностью модели (X_1) и эффективной битностью представления весов (новая координата X_6). Эмпирическую иллюстрацию «закона квантизации» даёт таблица 4: переход FP16 → INT4 сокращает память весов вчетверо и повышает пропускную способность на ~35–50% при потере качества 1–3%.

Таблица 4 – Влияние квантизации INT4 против FP16 (ориентировочные значения по сводным бенчмаркам MLPerf Inference [25] и данным о методе квантизации AWQ [26])

Показатель	FP16	INT4 (AWQ/GPTQ)
Память весов, 70B	≈140 ГБ	≈35 ГБ
GPU A100 80 ГБ для 70B	4×	1×
Относительная пропускная способность	1,0	1,35–1,5×
Потеря качества (MMLU)	–	1–3 п.п.
Память весов, 8B	≈16 ГБ	≈4 ГБ

Методологические ограничения

Прямая аналогия с законом Мура [1] наталкивается на ряд принципиальных ограничений. Закон Мура описывал одну переменную – плотность транзисторов, тогда как прогресс ИИ разворачивается в пространстве X_1–X_5 (с дополнительной координатой X_6 в self-hosted сегменте), и совместная динамика этих координат не сводится к скалярному показателю без существенных упрощений. В работе Дж. Вэя и соавторов [27] зафиксированы качественные скачки способностей при пересечении пороговых значений масштаба (эмерджентные способности), несовместимые с гладкими степенными законами [23], [24]. Вместе с тем трактовка эмерджентности как подлинного фазового перехода оспаривается: Р. Шеффер и соавторы [28]. При этом критика Шеффера не отменяет наблюдений для задач, требующих композиции подзадач (compositional reasoning). Прирост X_3 на стандартном MMLU [4], [8] практически остановился в диапазоне 86–91% для ведущих моделей начиная с 2024 г.; для долгосрочного мониторинга требуются адаптивные бенчмарки (MMLU-Pro, MERA [7]). Для коммерческих моделей GPT-4 [10], Claude Opus 4 и Claude Sonnet 4 [20], GigaChat 2 MAX [5], YandexGPT 5 Pro [6] значения X_1 и X_2 частично или полностью не раскрываются разработчиками [29], что ограничивает эмпирическую верификацию в отличие от открытых семейств [11-14]. Кроме того, ограниченная интерпретируемость нейросемантических моделей затрудняет содержательное объяснение наблюдаемых эффектов [30]. Переход к RLHF и обучению с подкреплением в режиме чистой обратной связи [12], использование синтетических данных на этапе постобучения [5] потенциально изменяет форму законов масштабирования.

Сопоставимость значений X_3 между моделями ограничена разнородностью протоколов (0/5-shot, chain-of-thought, правдоподобие против генерации), поэтому значения X_3 строго не сопоставимы и характеризуют порядковую динамику. Ключевые ограничения интерпретации: 1. Потолок метрики: приближение X_3 к границе MMLU может означать, что тест стал слишком лёгким, а не предел возможностей – нужны адаптивные бенчмарки (MMLU-Pro [20], MATH-500, Arena-Hard). 2. Смещение выжившего: в выборку попадают лишь успешные публичные модели, что завышает наблюдаемый темп прогресса. 3. Аппаратная гетерогенность: оценка через FLOP не учитывает алгоритмическую эффективность реализаций (Flash Attention 2, FP8 [16]), дающую выигрыш без изменения N и D, поэтому положение точек по оси вычислений имеет погрешность. 4. Данные постобучения: RL, SFT и синтетические данные смещают кривую Каплана, делая её неприменимой к постобученным моделям (средний сдвиг DeepSeek-R1 выше кривой в соответствии с рисунком 2).

Гипотезы и исследовательская программа

Сформулированы четыре гипотезы. H1 (единая гипотеза о динамике качества): значение метрики X_3 во времени описывается насыщающейся логистической кривой $X_3 \propto t^{\gamma}$ (быстрый рост; где t – время, γ – показатель степени), а поздний (2023–2025 гг.) – приближением к асимптоте. H2: рост вычислительного бюджета за 2012–2024 гг. подчиняется экспоненциальному закону $C(t) \propto e^{\lambda t}$ (где λ – показатель роста) с периодом удвоения около пяти месяцев (для фронтирных моделей с 2020 г.) [21] – в четыре–пять раз быстрее закона Мура [1]. H3 (предположение о текущем режиме): по имеющимся данным динамика X_3 вошла в поздний (насыщающийся) участок: прирост в подвыборке 2023–2025 гг. замедляется, тогда как X_4 убывает, а доля self-hosted моделей растёт; фронт прогресса смещается от масштаба к эффективности [2], [3], [15]. H4 (эффективность вывода): удельная стоимость вывода – затраты на достижение 1%

MMLU в одном запросе, формализуемая координатой $E = X_4/X_3$ (Cost/ Quality; альтернативно – удельное качество $P = X_3/N_{акт}$ на активируемый параметр) – убывает во времени экспоненциально: $E(t) \propto e^{-\lambda_E t}$, где $\lambda_E > 0$ – показатель убывания; снижение опережает удешевление оборудования. Эта закономерность является inference-аналогом закона масштабирования Хоффмана [24] и проверяется регрессией logE по годам (ожидаемый $R^2 > 0{,}9$ по экспоненте) для моделей с раскрытой стоимостью вывода (DeepSeek-V2, GigaChat-A3B, Llama). Программа предполагает три этапа: I – сбор данных по 80–100 моделям за 2018–2025 гг.; II – регрессионный анализ с подгонкой степенной и экспоненциальной форм, тест Чоу на структурный сдвиг для H3, а также проверку H4 (регрессия log E по годам).

Вычислительный эксперимент

Для количественной проверки аналога закона Мура оценён обучающий бюджет моделей в числе операций с плавающей точкой (FLOP): $C \approx 6 \cdot N \cdot D$, где N – активируемое число параметров, D – объём обучающей выборки в токенах (значения взяты из технических отчётов [9], [11–14]); для закрытой модели GPT-4 принята оценка Epoch AI $C \approx 2 \cdot 10^{25}$ FLOP [21]. Зависимость X_3 от $\log_{10} C$ (в соответствии с рисунком 2) имеет выраженно насыщающийся характер. Логистическая аппроксимация $X_3 = Q(\log_{10} C)$ даёт асимптоту $Q_\infty \approx 87\,\%$ при коэффициенте детерминации $R^2 \approx 0{,}98$.

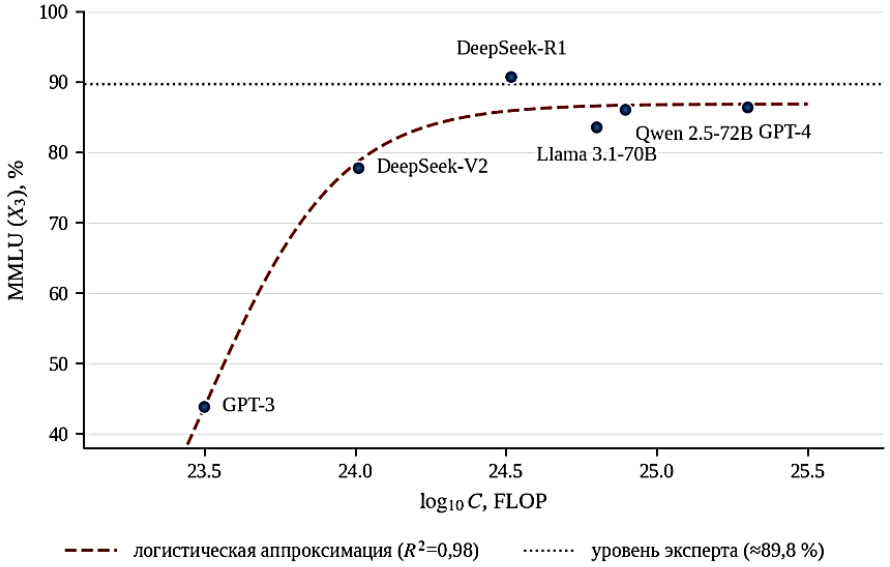


Рисунок 2 – Зависимость MMLU от логарифма обучающего бюджета (FLOP) для шести моделей с раскрытыми параметрами: логистическая аппроксимация и уровень эксперта-человека. Составлено авторами

Существенно, что рассуждающая модель DeepSeek-R1, дообученная с подкреплением (reinforcement learning, RL), достигает $X_3 = 90{,}8\,\%$ при бюджете меньшем, чем у GPT-4 и Qwen 2.5, располагаясь на ≈5 п.п. выше кривой масштабирования. Это согласуется с оценками Epoch AI о более высокой отдаче дообучения с подкреплением на единицу вычислений [3] и показывает, что качество определяется не только обучающим бюджетом, но и типом постобучения. Обучающий бюджет ведущих моделей возрастает примерно в пять раз за год, тогда как по закону Мура характеристики удваиваются за то же время (в соответствии с рисунком 3).

Многомерность прогресса иллюстрирует фронт «качество-стоимость» (в соответствии с рисунком 4): MoE-модели (DeepSeek-R1, GigaChat-A3B) достигают уровня dense-моделей при на порядок меньшем числе активируемых параметров, т.е. меньшей стоимости вывода X_4.

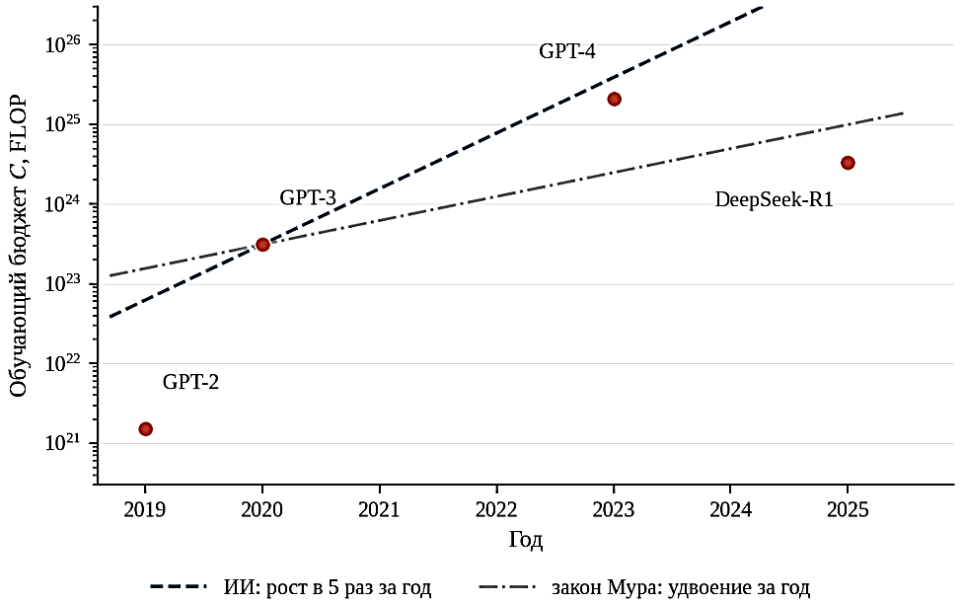


Рисунок 3 – Рост обучающего бюджета (FLOP) по годам: ИИ (×5/год, Epoch [3]) против закона Мура (×2/год); точки – GPT-2, GPT-3, GPT-4, DeepSeek-R1. Составлено авторами

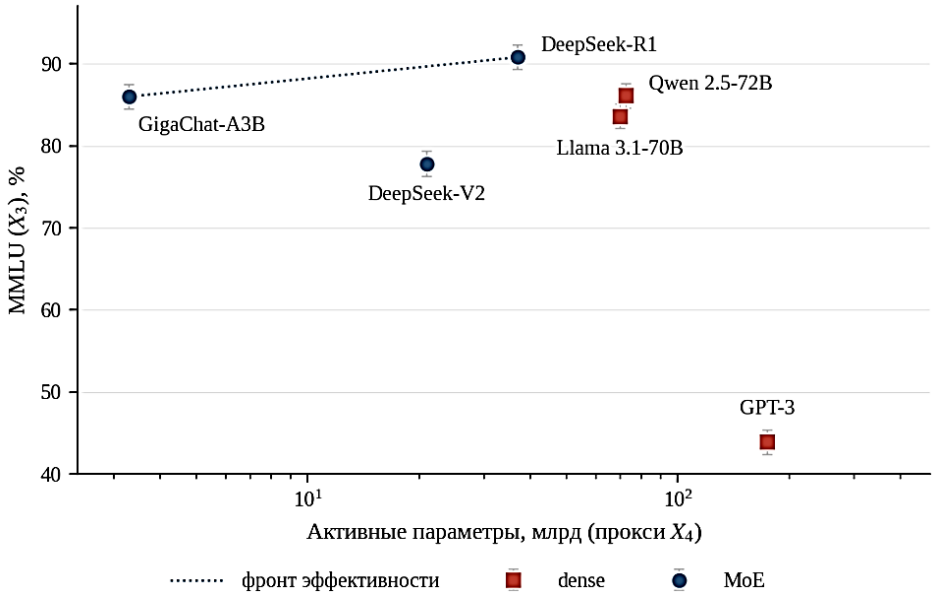


Рисунок 4 – Качество X_3 (MMLU) в зависимости от активируемых параметров (прокси стоимости вывода X_4) для dense- и MoE-моделей; пунктир – фронт эффективности. Планки – разброс из-за протокола оценки. Составлено авторами

Заключение

В работе для больших языковых моделей предложено пятимерное пространство состояний X_1–X_5 (с координатой X_6 для self-hosted) и процедура его анализа: сведение к двухфакторному уравнению состояния и сопоставительный регрессионный анализ

динамики качества по обучающему бюджету. Оценка выполнена на десяти зарубежных и российских моделях 2019–2025 гг. по открытым бенчмаркам (HELM, Open LLM Leaderboard, MERA) и техническим отчётам, с фиксацией методологических ограничений: разнородность протоколов оценки, потолок метрики, смещение выжившего и влияние постобучения.

Научная новизна: впервые для БЯМ 2019–2025 гг. предложено пятимерное пространство X1–X5, редуцируемое к двухфакторному уравнению состояния, где архитектурные инновации трактуются как преобразование координат. Единый регрессионный анализ MMLU от обучающего бюджета для зарубежных и российских моделей выявляет общий кластер масштабирования, переход от степенного роста к логистическому насыщению и снижение удельной стоимости вывода. Введена координата X6 (эффективная битность), раскрывающая квантизацию как драйвер self-hosted прогресса.

Сопоставление с законом Мура показывает неприменимость прямой одномерной аналогии: в пространстве X_1–X_5 сосуществуют степенная (по качеству, до насыщения), экспоненциальная (по бюджету) и структурно неустойчивая (эмерджентные эффекты) закономерности, не сводимые к единому скалярному показателю. Прогностическая сила аппарата ограничена малым объёмом и неоднородностью выборки, и эффектом потолка бенчмарков и требует проверки на расширенном корпусе (100+ моделей) с адаптивными тестами. Вероятные источники ошибки в горизонте двух лет: смена доминирующего бенчмарка (MMLU-Pro/MATH-500), опережающий рост эффективности за счёт постобучения и рассуждающих моделей (H4), исчерпание общедоступных данных [3]. Практическая значимость – поддержка стратегий разработки, мониторинга отрасли и выбора моделей для самостоятельного развёртывания.

Список источников

1. Moore G. E. Cramming More Components onto Integrated Circuits // Electronics. 1965. Vol. 38, No. 8. P. 114–117.
2. Douglas R., Verstyuk S. Long-run Progress in AI: A Quantitative Analysis [Electronic resource]. 2025. URL: https://arxiv.org/pdf/2501.17894 (accessed: 20.06.2026).
3. Epoch AI. Trends in Artificial Intelligence [Electronic resource]. 2025. URL: https://epoch.ai/trends (accessed: 20.06.2026).
4. Liang P., Bommasani R., Lee T. et al. Holistic Evaluation of Language Models (HELM). Stanford CRFM, 2022. arXiv:2211.09110 [Electronic resource]. URL: https://arxiv.org/abs/2211.09110 (accessed: 20.06.2026).
5. GigaChat Team. GigaChat Family: Efficient Russian Language Modeling Through Mixture of Experts Architecture [Electronic resource]. SaluteDevices, 2025. URL: https://arxiv.org/abs/2506.09440 (accessed: 20.06.2026).
6. Бут А. и др. YandexGPT 5: технологический отчёт [Электронный ресурс] // Хабр. Яндекс, 2025. URL: https://habr.com/ru/companies/yandex/articles/885218/ (дата обращения: 20.06.2026).
7. Fenogenova A., Chervyakov A., Martynov N. et al. MERA: A Comprehensive LLM Evaluation in Russian [Electronic resource] // Proceedings of the 62nd Annual Meeting of the ACL. Bangkok, 2024. P. 9920–9948. arXiv:2401.04531. URL: https://arxiv.org/abs/2401.04531 (accessed:20.06.2026).
8. Open LLM Leaderboard [Electronic resource] // Hugging Face. URL: https://huggingface.co/spaces/open-llm-leaderboard (accessed:20.06.2026).
9. Brown T. B., Mann B., Ryder N. et al. Language Models are Few-Shot Learners [Electronic resource]. OpenAI, 2020. arXiv:2005.14165. URL: https://arxiv.org/abs/2005.14165 (accessed: 20.06.2026).
10. OpenAI. GPT-4 Technical Report [Electronic resource]. 2023. URL: https://arxiv.org/abs/2303.08774 (accessed: 20.06.2026).
11. DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model [Electronic resource]. 2024. arXiv:2405.04434. URL: https://arxiv.org/abs/2405.04434 (accessed: 20.06.2026).
12. DeepSeek-AI, Guo D., Yang D. et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Electronic resource]. 2025. arXiv:2501.12948. URL: https://arxiv.org/abs/2501.12948 (accessed: 20.06.2026).

13. Grattafiori A., Dubey A., Jauhri A. et al. The Llama 3 Herd of Models [Electronic resource]. Meta AI, 2024. arXiv:2407.21783. URL: https://arxiv.org/abs/2407.21783 (accessed: 20.06.2026).
14. Qwen Team. Qwen2.5 Technical Report [Electronic resource]. Alibaba Group, 2024. arXiv:2412.15115. URL: https://arxiv.org/abs/2412.15115 (accessed: 20.06.2026).
15. Звягинцева, А.В. Анализ моделей классификации для распознавания прецедентных событий в технологических процессах добычи нефти и газа / А.В. Звягинцева, И.Ю. Ковалев // Проблемы искусственного интеллекта. 2025. № 2 (37). С. 66–78. DOI 10.24412/2413-7383-2025-2-37-66-78. EDN GUKVXJ.
16. Zhou Z., Ning X., Hong K. et al. A Survey on Efficient Inference for Large Language Models [Electronic resource]. 2024. arXiv:2404.14294. URL: https://arxiv.org/abs/2404.14294 (accessed: 20.06.2026).
17. Аверин, Г.В. Феноменологический подход к комплексной оценке сложности многомерных систем / Г.В. Аверин, А.В. Звягинцева, А.Р. Бражников // Вестник Донецкого национального университета. Серия Г: Технические науки. 2026. № 1. С. 124–140. DOI 10.5281/zenodo.19201881. EDN WSFTVO.
18. Anthropic. Claude 4 System Card: Opus 4 and Sonnet 4 [Electronic resource]. 2025. URL: https://www.anthropic.com/claude/opus. (accessed: 20.06.2026).
19. Hendrycks D., Burns C., Basart S. et al. Measuring Massive Multitask Language Understanding [Electronic resource] // International Conference on Learning Representations (ICLR). 2021. arXiv:2009.03300. URL: https://arxiv.org/abs/2009.03300 (accessed: 20.06.2026).
20. Wang Y., Ma X., Zhang G. et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark [Electronic resource] // Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Datasets and Benchmarks Track. 2024. P. 95266–95290. arXiv: 2406.01574. URL: https://arxiv.org/ abs/2406.01574 (accessed: 20.06.2026).
21. Аноприенко А.Я. Закономерности развития компьютерных технологий и обобщённый закон Мура // Вестник Донецкого национального технического университета. 2016. № 2 (2). С. 3–17. EDN XEEEDV.
22. Аноприенко А.Я. Представление закона Мура статистическими распределениями временной последовательности событий / А.Я. Аноприенко, Г.В. Аверин // Системный анализ и информационные технологии в науках о природе и обществе. 2018. № 1(14)-2(15). С. 63–72.
23. Kaplan J., McCandlish S., Henighan T. et al. Scaling Laws for Neural Language Models. OpenAI [Electronic resource]. 2020. arXiv: 2001.08361. URL: https://arxiv.org/abs/2001.08361 (accessed: 20.06.2026).
24. Hoffmann J., Borgeaud S., Mensch A. et al. Training Compute-Optimal Large Language Models [Electronic resource]. DeepMind, 2022. arXiv:2203.15556. URL: https://arxiv.org/abs/2203.15556 (accessed: 20.06.2026).
25. Reddi V. J., Cheng C., Kanter D. et al. MLPerf Inference Benchmark [Electronic resource] // ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). 2020. P. 446–459. arXiv:1911.02549. URL: https://arxiv.org/abs/1911.02549 (accessed: 20.06.2026).
26. Lin J., Tang J., Tang H. et al. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration [Electronic resource] // Proceedings of Machine Learning and Systems (MLSys). 2024. arXiv:2306.00978. URL: https://arxiv.org/abs/2306.00978 (accessed: 20.06.2026).
27. Wei J., Tay Y., Bommasani R. et al. Emergent Abilities of Large Language Models [Electronic resource] // Transactions on Machine Learning Research. 2022. URL: https://arxiv.org/abs/2206.07682 (accessed: 20.06.2026).
28. Schaeffer R., Miranda B., Koyejo S. Are Emergent Abilities of Large Language Models a Mirage? [Electronic resource] // Advances in Neural Information Processing Systems (NeurIPS). 2023. arXiv:2304.15004. URL: https://arxiv.org/abs/2304.15004 (accessed: 20.06.2026).
29. Bom Electronic resource masani R., Klyman K., Longpre S. et al. The Foundation Model Transparency Index [Электронный ресурс]. Stanford CRFM, 2023. arXiv:2310.12941. URL: https://arxiv.org/abs/2310.12941 (accessed: 20.06.2026).
30. Никитенко, К.А. Интерпретируемость нейросемантических моделей при их применении в прикладных областях / К.А. Никитенко, А.В. Звягинцева // Проблемы искусственного интеллекта. 2025. № 2 (37). С. 79–90. DOI 10.24412/2413-7383-2025-2-37-79-90. EDN VFRDWH.

References

1. Moore G. E. Cramming More Components onto Integrated Circuits // Electronics. – 1965. – Vol. 38, No. 8. – P. 114–117.
2. Douglas R., Verstyuk S. Long-run Progress in AI: A Quantitative Analysis [Electronic resource]. – 2025. – URL: https://arxiv.org/pdf/2501.17894 (accessed: 20.06.2026).
3. Epoch AI. Trends in Artificial Intelligence [Electronic resource]. – 2025. – URL: https://epoch.ai/trends (accessed: 20.06.2026).
4. Liang P., Bommasani R., Lee T. et al. Holistic Evaluation of Language Models (HELM) [Electronic resource] // Stanford CRFM. – 2022. – arXiv: 2211.09110. – URL: https://arxiv.org/abs/2211.09110
5. GigaChat Team. GigaChat Family: Efficient Russian Language Modeling Through Mixture of Experts Architecture [Electronic resource]. – SaluteDevices, 2025. – URL: https://arxiv.org/abs/2506.09440 (accessed: 20.06.2026).

6. But A. et al. YandexGPT 5: Technical Report [Electronic resource] // Habr. Yandex. – 2025. – URL: https://habr.com/ru/ companies/yandex/articles/885218/ (accessed: 20.06.2026).
7. Fenogenova A., Chervyakov A., Martynov N. et al. MERA: A Comprehensive LLM Evaluation in Russian [Electronic resource] // Proceedings of the 62nd Annual Meeting of the ACL. – Bangkok, 2024. – P. 9920–9948. – arXiv:2401.04531. – URL: https://arxiv.org/abs/2401.04531 (accessed: 20.06.2026).
8. Open LLM Leaderboard [Electronic resource] // Hugging Face. – URL: https://huggingface.co/spaces/open-llm-leaderboard (accessed: 20.06.2026).
9. Brown T.B., Mann B., Ryder N. et al. Language Models are Few-Shot Learners [Electronic resource]. – OpenAI, 2020. – arXiv:2005.14165. – URL: https://arxiv.org/abs/2005.14165 (accessed: 20.06.2026).
10. OpenAI. GPT-4 Technical Report [Electronic resource]. – 2023. – URL: https://arxiv.org/abs/2303.08774 (accessed: 20.06.2026).
11. DeepSeek-AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model [Electronic resource]. – 2024. – arXiv:2405.04434. URL: https://arxiv.org/abs/2405.04434 (accessed: 20.06.2026).
12. DeepSeek-AI, Guo D., Yang D. et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Electronic resource]. – 2025. – arXiv:2501.12948. – URL: https://arxiv.org/abs/2501.12948 (accessed: 20.06.2026).
13. Grattafiori A., Dubey A., Jauhri A. et al. The Llama 3 Herd of Models [Electronic resource]. – Meta AI, 2024. – arXiv:2407.21783. – URL: https://arxiv.org/abs/2407.21783 (accessed: 20.06.2026).
14. Qwen Team. Qwen2.5 Technical Report. Alibaba Group [Electronic resource]. – 2024. – arXiv:2412.15115. – URL: https://arxiv.org/abs/2412.15115 (accessed: 20.06.2026).
15. Zviagintseva A. V., Kovalev I. Yu. Analysis of Classification Models for the Recognition of Precedent Events in Oil and Gas Production Processes // Problems of Artificial Intelligence. – 2025. – No. 2 (37). – P. 66–78. – DOI 10.24412/2413-7383-2025-2-37-66-78. – EDN GUKVXJ.
16. Zhou Z., Ning X., Hong K. et al. A Survey on Efficient Inference for Large Language Models [Electronic resource]. – 2024. – arXiv:2404.14294. – URL: https://arxiv.org/abs/2404.14294 (accessed: 20.06.2026).
17. Averin G.V., Zviagintseva A.V., Brazhnikov A.R. Phenomenological Approach to the Integrated Assessment of the Complexity of Multidimensional Systems // Vestnik of Donetsk National University. Series G: Technical Sciences. – 2026. – No. 1. – P. 124–140. – DOI 10.5281/zenodo.19201881. – EDN WSFTVO.
18. Anthropic. Claude 4 System Card: Opus 4 and Sonnet 4 [Electronic resource]. – 2025. – URL: https://www.anthropic.com/claude/opus (accessed: 20.06.2026).
19. Hendrycks D., Burns C., Basart S. et al. Measuring Massive Multitask Language Understanding [Electronic resource] // International Conference on Learning Representations (ICLR). – 2021. – arXiv:2009.03300. – URL: https://arxiv.org/abs/2009.03300 (accessed: 20.06.2026).
20. Wang Y., Ma X., Zhang G. et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark [Electronic resource] // Advances in Neural Information Processing Systems 37 (NeurIPS 2024). Datasets and Benchmarks Track. – 2024. – P. 95266–95290. – arXiv:2406.01574. – URL: https://arxiv.org/ abs/2406.01574 (accessed: 20.06.2026).
21. Anoprienko A.Y. Patterns of Computer Technology Development and the Generalized Moore's Law // Vestnik of Donetsk National Technical University. – 2016. – No. 2(2). – P. 3–17. – EDN XEEEDV.
22. Anoprienko A.Y., Averin G.V. Representation of Moore's Law by Statistical Distributions of a Temporal Sequence of Events // System Analysis and Information Technologies in the Sciences of Nature and Society. – 2018. – No.1(14) – 2(15). – P. 63–72.
23. Kaplan J., McCandlish S., Henighan T. et al. Scaling Laws for Neural Language Models [Electronic resource]. – OpenAI, 2020. – arXiv:2001.08361. – URL: https://arxiv.org/abs/2001.08361 (accessed: 20.06.2026).
24. Hoffmann J., Borgeaud S., Mensch A. et al. Training Compute-Optimal Large Language Models [Electronic resource]. – DeepMind, 2022. – arXiv:2203.15556. – URL: https://arxiv.org/abs/2203.15556 (accessed: 20.06.2026).
25. Reddi V.J., Cheng C., Kanter D. et al. MLPerf Inference Benchmark [Electronic resource] // ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA). – 2020. – P. 446–459. – arXiv: 1911.02549. – URL: https://arxiv.org/abs/1911.02549 (accessed: 20.06.2026).
26. Lin J., Tang J., Tang H. et al. AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration [Electronic resource] // Proceedings of Machine Learning and Systems (MLSys). – 2024. – arXiv:2306.00978. – URL: https://arxiv.org/abs/2306.00978 (accessed: 20.06.2026).
27. Wei J., Tay Y., Bommasani R. et al. Emergent Abilities of Large Language Models // Transactions on Machine Learning Research. 2022. – URL: https://arxiv.org/abs/2206.07682 (accessed: 20.06.2026).
28. Schaeffer R., Miranda B., Koyejo S. Are Emergent Abilities of Large Language Models a Mirage? [Electronic resource] // Advances in Neural Information Processing Systems (NeurIPS). – 2023. – arXiv:2304.15004. – URL: https://arxiv.org/abs/2304.15004 (accessed: 20.06.2026).
29. Bommasani R., Klyman K., Longpre S. et al. The Foundation Model Transparency Index [Electronic resource]. – Stanford CRFM, 2023. – arXiv:2310.12941. – URL: https://arxiv.org/abs/2310.12941
30. Nikitenko K.A., Zviagintseva A.V. Interpretability of Neuro-Semantic Models in Their Application to Applied Domains // Problems of Artificial Intelligence. – 2025. – No. 2 (37). – P. 79–90. – DOI 10.24412/2413-7383-2025-2-37-79-90. – EDN VFRDWH.

RESUME

A. V. Zviagintseva, A. R. Brazhnikov

On the question of phenomenological regularities in the development of large language models: is an analogue of Moore's law possible?

Introduction. Moore's Law has served as an empirical benchmark for the semiconductor industry for over five decades, and with the rapid development of large language models, there is a question about the existence of a similar phenomenological pattern that can predict AI progress. The relevance of this research is driven by the need for numerical benchmarks for investment, technology foresight, and model comparisons, particularly in the domestic segment (GigaChat, YandexGPT).

Materials and methods. The quantitative characteristics of ten foreign and Russian models for 2019–2025 (GPT-3, GPT-4, DeepSeek-V2, DeepSeek-R1, Llama 3.1-70B, Qwen 2.5-72B, Claude Sonnet 4, Claude Opus 4, GigaChat 2 MAX, YandexGPT 5 Pro) have been systematized. The source data is extracted from open benchmarks (HELM, Open LLM Leaderboard, MERA) and technical reports. A five-dimensional parameter space (number of parameters, data volume, MMLU quality, HumanEval quality, and inference cost) is constructed, supplemented with an effective weight bitness coordinate for the self-hosted segment. The analysis is performed using time series regression analysis, logistic approximation, and computational budget estimation.

Results. It is shown that the quality dynamics of X_3 (MMLU) increased from 43.9% to 90.8% over a five-year period, but when the early point (GPT-3, 2020) is excluded, the growth slows down to ≈1.2 p.p./year in the interval 2023–2025, which indicates that the benchmark is saturated. The training budget of leading models grows exponentially with a doubling period of about 5 months, which is 4–5 times faster than the Moore's law. For self-hosted models, the key driver of progress is quantization (the transition from BF16 to INT4 reduces memory by a factor of four) and sparse MoE architectures, which allow for the same quality of dense models at a fraction of the cost. The unit cost of inference (the cost per 1% MMLU) is steadily decreasing, outpacing the cost reduction of hardware.

Conclusion. A direct one-dimensional analogy with Moore's law is not applicable: AI progress unfolds in a multidimensional space where saturating logistic dynamics of quality, exponential growth of budget, and structural shifts due to architectural innovations coexist. The proposed framework allows for numerical assessment of model states and can serve as a basis for industry monitoring, self-deployment model selection, and strategic planning. Further research requires expanding the sample (100+ models) and transitioning to adaptive benchmarks (MMLU-Pro, MATH-500).

РЕЗЮМЕ

А. В. Звягинцева, А. Р. Бражников

К вопросу о феноменологических закономерностях развития больших языковых моделей: возможен ли аналог закона Мура?

Введение. Закон Мура более пяти десятилетий служил эмпирическим ориентиром для полупроводниковой отрасли; с бурным развитием больших языковых моделей возникает вопрос о существовании аналогичной феноменологической закономерности, позволяющей прогнозировать прогресс ИИ. Актуальность исследования обусловлена потребностью в численных ориентирах для инвестиций, технологического форсайта и сравнения моделей, особенно в отечественном сегменте (GigaChat, YandexGPT).

Материалы и методы. Систематизированы количественные характеристики десяти зарубежных и российских моделей за 2019–2025 гг. (GPT-3, GPT-4, DeepSeek-V2, DeepSeek-R1, Llama 3.1-70B, Qwen 2.5-72B, Claude Sonnet 4, Claude Opus 4, GigaChat 2 MAX, YandexGPT 5 Pro). Исходные данные извлечены из открытых бенчмарков (HELM, Open LLM Leaderboard, MERA) и технических отчётов. Построено пятимерное

пространство параметров (число параметров, объём данных, качество MMLU, качество HumanEval, стоимость вывода), дополненное координатой эффективной битности весов для self-hosted сегмента. Анализ выполнен с использованием регрессионного анализа временных рядов, логистической аппроксимации и оценки вычислительного бюджета.

Результаты. Показано, что динамика качества X_3 (MMLU) за пятилетний период возросла с 43,9% до 90,8%, однако при исключении ранней точки (GPT-3, 2020 г.) прирост на интервале 2023–2025 гг. замедляется до ≈1,2 п.п./год, что свидетельствует о насыщении бенчмарка. Обучающий бюджет ведущих моделей растёт экспоненциально с периодом удвоения около 5 месяцев, что в 4–5 раз быстрее закона Мура. Для self-hosted моделей ключевой драйвер прогресса – квантизация (переход BF16→INT4 сокращает память вчетверо) и разреженные MoE-архитектуры, позволяющие достигать качества dense-моделей при меньшей стоимости вывода на порядок. Удельная стоимость вывода (затраты на 1% MMLU) устойчиво снижается, опережая удешевление оборудования.

Заключение. Прямая одномерная аналогия с законом Мура неприменима: прогресс ИИ разворачивается в многомерном пространстве, где сосуществуют насыщающаяся логистическая динамика качества, экспоненциальный рост бюджета и структурные сдвиги за счёт архитектурных инноваций. Предложенный аппарат позволяет численно оценивать состояние моделей и может служить основой для мониторинга отрасли, выбора моделей для самостоятельного развёртывания и стратегического планирования. Дальнейшие исследования требуют расширения выборки (100+ моделей) и перехода к адаптивным бенчмаркам (MMLU-Pro, MATH-500).

Звягинцева Анна Викторовна – д.т.н., доцент, профессор кафедры компьютерных технологий ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, a.zvyagintseva@donnu.ru. *Область научных интересов*: системный анализ, событийная и комплексная оценка; безопасность и управление социально-экономическими и техногенными системами, социофизика; информационно-аналитические системы и технологии в прикладных областях; обработка и анализ многомерных и больших данных. ORCID 0000-0003-2218-9230

Бражников Андрей Романович – магистрант, ФГБОУ ВО «ДонГУ», 283001, Донецк, ул. Университетская, 24, телефон +7(949) 343-1659, yanox123321@gmail.com. *Область научных интересо*в: методы анализа и обработки больших данных; проектирование и оптимизация архитектуры распределённых веб-систем; автоматизация процессов разработки и развёртывания программного обеспечения. ORCID 0009-0005-4420-9206

Zviagintseva Anna Viktorovna – Dr. of Engineering, Associate Professor, Professor of the Department of Computer Technologies, Donetsk State University, 24 Universitetskaya Street, Donetsk, 283001, a.zvyagintseva@donnu.ru. Research interests: system analysis, event-based and comprehensive assessment; security and management of socio-economic and technogenic systems, sociophysics; information and analytical systems and technologies in applied fields; processing and analysis of multidimensional and big data. ORCID 0000-0003-2218-923

Brazhnikov Andrey Romanovich – Master's student, Donetsk State University, 283001, Donetsk, Universitetskaya str., 24, phone +7(949) 343-1659, yanox123321@gmail.com. Research interests: methods of analysis and processing of big data; design and optimization of architecture of distributed web systems; automation of software development and deployment processes. ORCID 0009-0005-4420-9206

Статья поступила в редакцию 02.06.2026;
одобрена после рецензирования 03.07.2026; принята к публикации 02.09.2026
The article was submitted 02.06.2026; approved after reviewing 03.07.2026;
accepted for publication 02.09.2026.