12. Becker G.S. Human Capital. New York: Columbia University Press, 1964. 534 p.
13. Kotler P., Turner R.E. Marketing Management: Analysis, Planning, Implementation and Control. Scarborough, Ont.: Prentice Hall Canada, 1998. 896 p. URL: https://archive.org/details/marketingmanagem0000kotl/ mode/ 2up (дата обращения: 03.02.2025).

УДК 004.85:338.3

БОЙКОВА Анна Викторовна – профессор, д. э. н., доцент кафедры экономики и управления производством ТвГТУ, Тверь (alexmario@mail.ru)
КУЗНЕЦОВ Кирилл Сергеевич – магистрант кафедры экономики и управления производством ТвГТУ, Тверь (kirill.kuznetsov.95@yandex.ru)

ФАКТОРЫ, ВЛИЯЮЩИЕ НА СТОИМОСТЬ ОБУЧЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Аннотация. Указано, что процесс создания больших языковых моделей сопряжен с трудоемкими операциями, значительными затратами и высокими техническими требованиями. Отмечены проблемы развития данных моделей. Перечислены ключевые факторы, обусловливающие итоговую стоимость модели.

Ключевые слова: затраты, модели обучения, искусственный интеллект, обучение моделей, большие языковые модели, капитальные затраты, эксплуатационные затраты, облачные технологии.

Boikova A.V. – Professor, Doctor of Economics, Associate Professor of the Department of Economics and Production Management of TvSTU, Tver (alexmario@mail.ru)
Kuznetsov K.S. – Graduate Student of the Department of Economics and Production Management of TvSTU, Tver (kirill.kuznetsov.95@yandex.ru)

FACTORS AFFECTING THE COST OF TRAINING LARGE LANGUAGE MODELS

Abstract. It is stated that the process of creating large language models involves time-consuming operations, significant costs and high technical requirements. The problems of the development of these models are noted. The key factors determining the final cost of the model are listed.

Keywords: costs, AI learning models, artificial intelligence, AI model training, large language models, capital costs, operating costs, cloud technology.

Модели типа Gemini и GPT-4 произвели революцию во взаимодействии людей с технологиями. С помощью таких моделей можно генерировать текст, улучшать поиск и решать творческие задачи. Данные модели облегчают работу клиентских служб.

Большие языковые модели (Large language models, или LLMs) представляют собой сложные системы, функционирующие на основе глубоких нейронных сетей, создающие и распознающие различные тексты. Они обучаются на огромных объемах данных, включающих сотни миллиардов предложений из открытых источников (интернета и т. п.).

Концепция языковых моделей базируется на прогнозировании вероятной последовательности слов, благодаря чему значительно улучшается способность машин понимать контекст и смысл. Технологии эволюционировали постепенно: сначала они представляли собой системы, основанные на правилах, затем превратились в статистические модели, а потом стали современными нейросетями. Сегодня доминируют трансформеры, которые благодаря глубокому обучению могут производить невероятно точные анализ и генерацию языка [2].

Большие языковые модели, о чем мы вскользь говорили выше, обучаются на значительных объемах данных из разных источников, что позволяет им глубже «понимать» языковые структуры, грамматику и особенности использования языка. Благодаря этому они выполняют широкий спектр задач, имитируя человеческое восприятие. Последнее делает взаимодействие индивида с технологиями более естественным и удобным, чем когда-либо [2].

В 2024 году Стэндфордским университетом была опубликована седьмая редакция отчета о ключевых тенденциях и достигнутых результатах в области искусственного интеллекта AI Index Report 2024 [1]. В нем, в частности, среди проблем развития LLMs эксперты отмечают недостаток информации для их обучения. За последние несколько лет чат-боты, функционирующие на базе искусственного интеллекта, стали крайне прогрессивными. Это стало возможным во многом благодаря тому, что LLMs обучались на все возрастающем количестве источников данных, таких как книги, статьи и т. д. Однако усиливающаяся зависимость моделей искусственного интеллекта от информации привела к возникновению опасений, что будущие поколения исследователей будут испытывать нехватку сведений для дальнейшего масштабирования и совершенствования своих систем [4].

По заявлениям компании Epoch AI, опубликованным в 2022 году, высококачественные языковые данные могут быть полностью исчерпаны к 2024 году, а резерв низкокачественных, согласно прогнозам, будет полностью использован в течение следующих двух десятилетий, ресурсы изображений и графической информации будут потрачены к концу 2030-х или середине 2040-х годов. Одним из решений указанной проблемы может стать обучение LLMs на синтетических данных, которые модели создают самостоятельно. По мнению исследователей из Стэнфордского университета, такой подход не только поможет справиться с возможным истощением данных, но и позволит генерировать информацию там, где ее с самого начала недостаточно [4]. Однако некоторые ученые отмечают, что существуют ограничения, связанные с обучением моделей на синтетических данных. Например, в определенный момент такие LLMs «теряют способность запоминать истинные распределения данных и начинают выдавать узкий диапазон результатов» [4].

Обучение больших языковых моделей, характеризующихся миллионами параметров, требует значительных вычислительных мощностей. В первую очередь последнее связано с необходимостью обрабатывать огромные массивы данных и оптимизировать параметры модели для повышения точности прогнозов. Вычислительные затраты на обучение модели зависят от ряда факторов:

1) объема данных (значительный объем информации, необходимый для обучения, может перегружать вычислительные ресурсы);

2) ограниченности ресурсов (нехватка памяти, дефицит графических процессоров и даже высокие тарифы на электроэнергию могут замедлить обучение);

3) качества данных (является ключевым моментом; плохие или предвзятые сведения могут привести к ошибочным результатам) [6].

Размеры нейронных сетей, используемых в LLMs, растут в геометрической прогрессии. Например, если ранние модели, такие как GPT-2, содержали около 1,5 млрд параметров, что соответствует мозгу небольшого животного (например, медоносной пчелы), то последние модели (допустим, GPT-4) имеют около триллиона параметров, что приближает их по масштабу к человеческому мозгу. Экспоненциальный рост размеров и сложности моделей привели к увеличению затрат на обучение. Так, для моделей типа GPT-3 и PaLM нужен примерно 1 терабайт памяти. С усилением спроса на вычислительные мощности несоответствие между требованиями искусственного интеллекта и возможностями оборудования становится все более очевидным [6].

Увеличение возможностей моделей ведет к росту затрат на их обучение и дальнейшую эксплуатацию [4]. Это вызывает интерес у теоретиков и беспокойство у практиков. Согласно закону Мура,

производительность моделей может повышаться только за счет роста вычислительных мощностей. Однако в то время как потребность в вычислительных ресурсах увеличивается в 10 раз каждый год, производительность оборудования – лишь в 3 раза каждые два года. Этот разрыв приводит к необходимости использовать больше машин, что еще сильнее повышает стоимость обучения, особенно LLMs. Для решения указанной проблемы необходимо искать более экономичные подходы [6].

По мнению директора лаборатории искусственного интеллекта в компании Anthropic Д. Амодея, в ближайшие два года расходы на обучение моделей вышеназванного интеллекта могут достичь примерно 10 млрд долл. [4; 5]. Такие расходы будут по карману только крупным организациям. Если эта тенденция сохранится, то к 2027 году стоимость обучения моделей искусственного интеллекта, согласно прогнозам специалистов Epoch AI, превысит 1 млрд долл. [8]. По их наблюдениям, ежегодные затраты увеличиваются в среднем в 2,4 раза [2]. В связи с этим имеет смысл выделить основные факторы, которые влияют на этот процесс. К таким факторам относятся:

1. Запуск больших языковых моделей (требует огромных вычислительных мощностей; большая часть (47–67 %) совокупных затрат приходится на аппаратное обеспечение (из них на серверные компоненты – 15–22 %; межсоединения на уровне кластера – 9–13 %); 29–49 % составляют расходы на персонал; 2–6 % – затраты на электроэнергию [8]).

2. Использование для обучения LLMs графических процессоров высшего класса, приобретение или аренда которых обходится довольно дорого. По оценке Д. Хуанга, генерального директора NVIDIA, для обучения языковой модели Generative Pre-trained Transformer Mixture of Experts потребовалось привлечь в течение 3–5 месяцев 25 тыс. графических процессоров на базе Ampere [8].

3. Повышение интеллектуальности модели (часто связано с применением крайне сложных архитектур или крупных моделей (например, масштабирование с 7 до 300 млрд параметров) или с одновременным задействованием большего числа экспертов (допустим, в модели Mixture of Experts, или MoE). Очевидно, что такие модели стоят дороже [3]).

4. Количество токенов, которое поступает на вход модели и которое получает пользователь на выходе (оно влияет на время обработки и требуемые вычислительные ресурсы, расход энергии для обработки. Все это в конечном счете вызывает рост затрат [3]).

5. Тип отрабатываемых моделью данных: текст, графика, аудио или видео (он влияет на стоимость; для обработки аудио и видео обычно требуется затратить больше ресурсов, чем для обработки текста [3]).

На данный момент для оценки затрат на обучение моделей искусственного интеллекта, в том числе и больших языковых моделей, используются, как правило, два подхода. В рамках первого из них обучение проводится на собственном или арендованном оборудовании; второго – с применением технологий облачных вычислений. Некоторые ученые пришли к выводу, что величина затрат, определенная согласно второму подходу, обычно намного выше, чем установленная по первому [4].

Таким образом, был рассмотрен инструментарий, позволяющий выяснить, какой из вариантов (локальный вычислительный кластер или облачная инфраструктура) является более экономичным решением для реализации обучения больших языковых моделей.

Библиографический список

1. AI Index Report 2024. URL: https://hai.stanford.edu/ai-index (дата обращения: 14.01.2025).
2. Epoch AI: официальный сайт. URL: https://epoch.ai/ (дата обращения: 14.01.2025).
3. Benram G. Understanding the cost of Large Language Models (LLMs). URL: https://www.tensorops.ai/post/understanding-the-cost-of-large-language-models-llms (дата обращения: 14.01.2025).
4. The rising costs of training frontier AI models / B. Cottier, R. Rahman, L. Fattorini, N. Maslej, D. Owen. URL: https://arxiv.org/html/2405.21015v1#abstract (дата обращения: 14.01.2025).
5. Grinkevičius P. The cost of training AI models is rising exponentially. URL: https://cybernews.com/tech/rising-cost-of-training-ai-/ (дата обращения: 14.01.2025).
6. Juhasz Z. Quantitative cost comparison of on-premise and cloud infrastructure based EEG data processing. URL: https://www.researchgate.net/publication/342545255_Quantitative_cost_comparison_of_on-premise_and_cloud_infrastructure_based_EEG_data_processing (дата обращения: 14.01.2025).
7. What if Dario Amodei Is Right about A.I.? URL: https://www.nytimes.com/2024/04/12/opinion/ezra-klein-podcast-dario-amodei.html?showTranscript=1 (дата обращения: 15.01.2025).
8. Luccioni A.S., Viguier S., Ligozat A.-L. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. URL: https://www.researchgate.net/publication/365081230_Estimating_the_Carbon_Footprint_of_BLOOM_a_176B_Parameter_Language_Model (дата обращения: 15.01.2025).