Обзор современных больших языковых моделей и бенчмарков для работы на локальных устройствах

В. А. Слепов[1], Д. Е. Чикрин[2]

Казанский федеральный университет

E-mail: [1]inf.vlad.s@bk.ru, [2]dmitry.kfu@ya.ru

***Аннотация*. В работе рассматриваются современные большие языковые модели, их основные характеристики и подходы к оценке с позиции локального развертывания на устройствах ограниченной производительности. Приводится обзор моделей, различающихся по степени открытости, размеру контекстного окна и количеству параметров. Рассматриваются распространенные бенчмарки, применяемые для сравнения качества языковых моделей. Рассмотрено утверждение, что при выборе модели для локального применения необходимо учитывать не только результаты стандартных тестов, но и требования к вычислительным ресурсам, возможности квантования и совместимость с локальными средствами запуска.**

Ключевые слова: большие языковые модели, локальное развертывание, бенчмарки, контекстное окно, квантование, одноплатные устройства, ограниченные вычислительные ресурсы, модели класса 1B

I. Введение

Большие языковые модели играют немаловажную роль в современном развитии искусственного интеллекта. Существенный этап формирования данной области связан с публикацией работы Attention Is All You Need, в которой была предложена архитектура трансформера, ставшая основой для последующего развития систем обработки естественного языка [1]. В дальнейшем появились модели нового поколения, включая BERT и GPT, а после выхода GPT-3 большие языковые модели стали рассматриваться как универсальный инструмент для решения широкого круга прикладных задач [2, 3].

Для сравнения моделей используются специализированные бенчмарки, позволяющие оценивать знания, способность к рассуждению, следование инструкциям и фактическую точность. Однако рост возможностей БЯМ сопровождается увеличением требований к вычислительным ресурсам, что важно при запуске на локальных устройствах. В этих условиях требуется учитывать не только результаты бенчмаркинга, но и ограничения, связанные с объемом памяти, вычислительной мощностью и доступными средствами развертывания [4, 5].

II. Анализ современных больших языковых моделей

Современные большие языковые модели различаются по нескольким ключевым параметрам. Прежде всего, следует выделить степень открытости модели. На практике существуют закрытые, полуоткрытые и полностью открытые решения.

- Закрытые модели ориентированы преимущественно на облачное использование и не предполагают полноценного локального развертывания;
- Полуоткрытые модели допускают использование готовых весов, но могут иметь ограничения, связанные с лицензированием или доступностью отдельных артефактов;
- Полностью открытые модели представляют особый интерес для исследовательской и инженерной практики, поскольку позволяют запускать и адаптировать модель под конкретные условия применения [6].

Сравнение современных больших языковых моделей по их ключевым параметрам приведено в табл. 1.

ТАБЛИЦА I. Сравнение ключевых параметров моделей

Модель	Открытая/Закрытая	Размер контекста	Количество токенов для обучения	Максимальный размер параметров
GTP 4	Закрытая	-	-	-
Llama 3	Полуоткрытая	128K	15T	405B
Granite 4	Полуоткрытая	128K	22T	32B
GLM 4.5	Полуоткрытая	128K	23T	355B
Kimi K2	Полуоткрытая	128K	15,5T	1T
Gemma 3	Полуоткрытая	128K	-	27B
OLMo 2	Полностью открытая	4K	6,5T	32B
Apertus	Полностью открытая	64K	15T	70B
Phi4	Полуоткрытая	16K	10T	14B

Как видно из табл. 1, важными параметрами являются размер контекстного окна, количество токенов для обучения и максимальный размер параметров. Современные модели поддерживают контекст от нескольких тысяч до сотен тысяч токенов. Это расширяет возможности работы с объемными документами и сложными сценариями взаимодействия, однако одновременно увеличивает нагрузку на память и вычислительную составляющую устройства.

Модели, представленные в табл. 1 – это актуальные решения, которые отражают разные подходы к развитию языковых моделей: от закрытых облачных решений до открытых моделей, ориентированных на более гибкое использование.

Развитие БЯМ сопровождается расширением их возможностей, что также влияет на необходимость создания тестов для правильной оценки работоспособности и эффективности модели.

III. Бенчмарки оценки больших языковых моделей

Для анализа качества больших языковых моделей применяются бенчмарки – это стандартизированные наборы задач, позволяющие сравнивать модели по различным направлениям [6]. Их использование необходимо как для научного сопоставления моделей, так и для предварительного выбора решений под конкретные прикладные сценарии.

К самым популярным бенчмаркам можно отнести следующие: MMLU, BBH, DROP, GSM8K, IFEval, MATH. Бенчмарк MMLU предназначен для оценки многозадачного понимания по широкому кругу дисциплин [7]. Для проверки сложного рассуждения используется BIG-Bench Hard, включающий задачи повышенной трудности [8]. В задачах понимания текста и извлечения информации применяется DROP, ориентированный на дискретное рассуждение по абзацам текста [9].

Отдельную группу образуют математические бенчмарки. К ним относятся GSM8K, включающий задачи школьного уровня, и MATH, ориентированный на более сложные математические задания [10, 11]. Для анализа следования инструкциям используется IFEval, а для оценки безопасности и фактической достоверности применяются специализированные наборы, связанные с безопасностью модели [12].

Такой набор тестов позволяет анализировать модель сразу по нескольким направлениям: общему уровню знаний, рассуждению, математическим способностям, следованию инструкциям, безопасности и фактологической устойчивости. Для локального развертывания это особенно важно, поскольку высокая точность по одному тесту не гарантирует практической пригодности модели в реальных условиях эксплуатации.

При выборе модели для локального развертывания результатов стандартных тестов недостаточно, поскольку высокие показатели на бенчмарках не гарантируют эффективной работы в условиях ограниченных вычислительных ресурсов. Важно также учитывать особенности самих моделей для эффективного развертывания на устройстве.

IV. Особенности выбора моделей для локального развертывания

При локальном использовании больших языковых моделей логика их оценки меняется. Если в облачной среде основное внимание чаще уделяется максимальному качеству ответа и запоминанию информации, то при локальном запуске особенно важным становится баланс между качеством и ресурсоемкостью. На устройствах ограниченной производительности необходимо учитывать объем оперативной памяти, вычислительные возможности процессора, доступный объем хранилища и общую устойчивость модели при длительной нагрузке.

Одним из ключевых практических факторов выступает возможность квантования модели. Квантование позволяет уменьшить объем памяти, занимаемый весами, и сделать запуск модели более доступным для локальной среды. Существенное значение имеет и реальный размер модели на устройстве, поскольку даже открытая модель может оказаться непригодной для практического использования из-за высоких требований к памяти и вычислительным ресурсам.

С точки зрения практического применения особый интерес представляют модели малого класса, в частности модели от 1 до 8 миллиардов параметров. Они не снимают полностью проблему ограниченных ресурсов, однако делают локальное использование более реалистичным по сравнению с более крупными системами. К данному классу можно отнести Llama 3.2, Gemma 3 и OLMo 2. Эти модели имеют современную архитектуру БЯМ и необходимое количество параметров для удобного локального развертывания [13, 14, 15].

На рис. 1 представлено сравнение средних и максимальных значений потребления оперативной памяти моделями Llama 3.2 1B, Gemma 3 и OLMo2 полученных на одноплатном устройстве Khadas VIM 3.

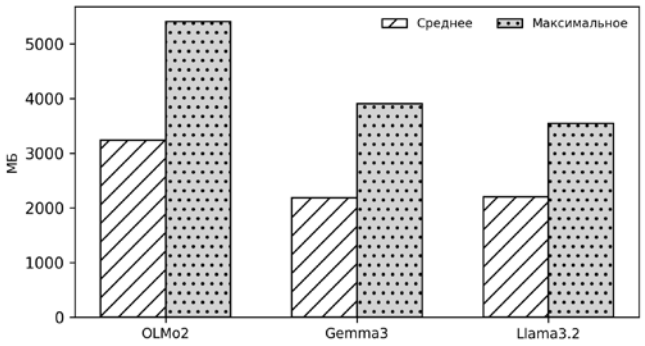

Рис. 1. Среднее и максимальное потребление оперативной памяти

V. Заключение

Современные большие языковые модели применяются в широком круге задач, а для оценки их качества используется набор специализированных бенчмарков, позволяющих анализировать знания модели, способность к рассуждению, следование инструкциям, безопасность и фактическую достоверность.

При выборе модели для локального развертывания необходимо учитывать не только результаты стандартных тестов, но и такие характеристики, как

степень открытости, размер контекстного окна, количество параметров, возможность квантования и совместимость со средствами локального запуска.

Таким образом, проведенный обзор показывает, что для локального развертывания подходят модели малого масштаба, которые могут служить как отправной точкой для практического применения на устройствах с ограниченными вычислительными возможностями, так и хорошо справляться с задачами на устройствах средних мощностей.

Список литературы

[1] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30.

[2] Koroteev M.V. BERT: A Review of Applications in Natural Language Processing and Understanding [Электронный ресурс]. 2021.URL:https://www.researchgate.net/publication/350287107_BERT_A_Review_of_Applications_in_Natural_Language_Processing_and_Understanding (Дата обращения: 21.03.2026).

[3] Назаров Д.М., Бадаев Ф.И. Применение больших языковых моделей в сфере здравоохранения // Менеджер здравоохранения. 2025. № 5. С. 142–154.

[4] Ni S., Chen G., Li S., Chen X., Li S., Wang B., Wang Q., Wang X., Zhang Y., Fan L., Li C., Xu R., Sun L., Yang M. A Survey on Large Language Model Benchmarks [Электронный ресурс]. 2025. URL: https://arxiv.org/html/2508.15361v1 (Дата обращения: 23.03.2026).

[5] Muhammad Usman Hadi, Qasem Al-Tashi, Abbas Shah, Rizwan Qureshi Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects [Электронный ресурс]. 2025. URL: https://www.researchgate.net/publication/383058502_Large_Language_Models_A_Comprehensive_Survey_of_its_Applications_Challenges_Limitations_and_Future_Prospects (Дата обращения: 27.03.2026).

[6] Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. Measuring Massive Multitask Language Understanding // Proceedings of the International Conference on Learning Representations. 2021.

[7] Suzgun M., Scales N., Schärli N., Gehrmann S., Tay Y., Chung H. W., Chowdhery A., Le Q. V., Chi E. H., Zhou D., Wei J. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them // Findings of the Association for Computational Linguistics: ACL 2023 P. 13003-13051. DOI: 10.18653/v1/2023.findings-acl.824.

[8] Dua D., Wang Y., Dasigi P., Stanovsky G., Singh S., Gardner M. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. P. 2368–2378. DOI: 10.18653/v1/N19-1246.

[9] Cobbe K., Kosaraju V., Bavarian M., Chen M., Jun H., Kaiser Ł., Plappert M., Tworek J., Hilton J., Nakano R., Hesse C., Schulman J. Training Verifiers to Solve Math Word Problems [Электронный ресурс]. 2021. URL: https://arxiv.org/abs/2110.14168 (Дата обращения: 07.04.2026).

[10] Zhou J., Lu T., Mishra S., Brahma S., Basu S., Luan Y., Zhou D., Hou L. Instruction-Following Evaluation for Large Language Models [Электронный ресурс]. 2023. URL: https://arxiv.org/abs/2311.07911 (Дата обращения: 07.04.2026).

[11] Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E., Song D., Steinhardt J. Measuring Mathematical Problem Solving With the MATH Dataset // Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2021. Vol. 1. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html (Дата обращения: 07.04.2026).

[12] The Llama 3 Herd of Models [Электронный ресурс]. 2024. URL: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ (Дата обращения: 07.04.2026).

[13] Gemma: Open Models Based on Gemini Research and Technology [Электронный ресурс] / Gemma Team, Google DeepMind. 2024. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf (Дата обращения: 07.04.2026).

[14] OLMo Team. OLMo 2 Furious // Conference on Language Modeling. 2025. URL: https://openreview.net/forum?id=2ezugTT9kU (Дата обращения: 07.04.2026)