УДК 004.8

Захаров А.С.

магистрант 2 курса

Санкт-Петербургский государственный университет

аэрокосмического приборостроения

Российская Федерация, г. Санкт-Петербург

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЛОКАЛЬНЫХ И ОБЛАЧНЫХ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Аннотация: *В статье рассматриваются особенности локального и облачного развёртывания моделей искусственного интеллекта. На основе научных исследований сопоставлены качество обработки запросов, скорость формирования ответов, вычислительные затраты и подходы к обеспечению конфиденциальности информации. Особое внимание уделено условиям использования интеллектуальных систем в организациях, работающих с чувствительными данными. Определены преимущества, ограничения и факторы выбора подходящей модели развёртывания.*

Ключевые слова: *искусственный интеллект; языковые модели; облачные вычисления; локальные вычисления; производительность; информационная безопасность; экономическая эффективность.*

Zakharov A.S.

second-year master's student

Saint Petersburg State University of Aerospace Instrumentation

Russian Federation, Saint Petersburg

COMPARATIVE ANALYSIS OF LOCAL AND CLOUD-BASED ARTIFICIAL INTELLIGENCE MODELS

Abstract: *The article examines the characteristics of local and cloud-based artificial intelligence model deployment. Based on scientific research, it compares response quality, inference speed, computational costs, and*

approaches to data confidentiality. Particular attention is paid to the use of intelligent systems in organizations processing sensitive information. The advantages, limitations, and key factors influencing the choice of deployment model are identified.

Keywords: *artificial intelligence; language models; cloud computing; local computing; performance; information security; economic efficiency.*

Распространение искусственного интеллекта (ИИ) сопровождается развитием различных способов использования интеллектуальных моделей. Наиболее привычным для широкого круга пользователей является облачный подход, при котором обработка запросов осуществляется на удалённых серверах поставщика услуги. Одновременно увеличивается доступность моделей, которые можно запускать непосредственно на персональных компьютерах или собственной вычислительной инфраструктуре организации. Оба варианта позволяют решать задачи обработки текста, анализа информации и автоматизации отдельных операций, однако различаются по техническим требованиям, стоимости эксплуатации и условиям работы с данными. Цель настоящей статьи заключается в сравнительном анализе локальных и облачных моделей ИИ с точки зрения качества результатов, скорости обработки запросов, экономических затрат и информационной безопасности. Исследование основано на сопоставлении опубликованных научных работ без проведения самостоятельного вычислительного эксперимента.

Под локальным использованием ИИ понимается выполнение вычислений на устройстве пользователя либо на серверах, находящихся под контролем организации. Облачный подход предполагает обращение к удалённой инфраструктуре через пользовательский интерфейс или программный интерфейс приложения. При этом необходимо различать саму модель и место её выполнения. Одна и та же модель при наличии

необходимых технических условий может использоваться как локально, так и на удалённом сервере. Следовательно, различия в качестве ответов не всегда определяются способом развёртывания: существенное значение имеют архитектура модели, параметры её настройки и доступные вычислительные ресурсы.

Главным преимуществом облачных решений является доступность производительной инфраструктуры без необходимости её самостоятельного приобретения и обслуживания. Пользователь получает возможность работать с крупными моделями через обычный браузер, не располагая дорогостоящим компьютерным оборудованием. Поставщик сервиса обеспечивает размещение модели, распределение вычислительной нагрузки и техническую поддержку. Для частных пользователей и небольших организаций подобная модель взаимодействия особенно удобна, поскольку позволяет начать использование ИИ практически без первоначальных капитальных затрат. Однако доступность сервиса зависит от интернет-соединения, условий поставщика и установленных ограничений использования.

Локальное развёртывание предоставляет больше возможностей контроля над программной средой. Пользователь самостоятельно определяет используемую модель, её версию и параметры выполнения. При полностью автономной конфигурации обработка запросов возможна без подключения к внешнему сервису. Это может быть полезно при работе с внутренними документами организации, ограниченном доступе к интернету или необходимости сохранять неизменность программной конфигурации. Вместе с тем локальное выполнение предъявляет требования к объёму оперативной и видеопамяти, производительности вычислительных компонентов и организации программного окружения. Поэтому доступность такого подхода зависит от размера модели и возможностей имеющегося оборудования.

Вопрос качества результатов требует отдельного рассмотрения. Распространённое представление о безусловном превосходстве облачных систем связано с тем, что удалённая инфраструктура позволяет предоставлять доступ к наиболее ресурсоёмким моделям. Однако небольшие локальные модели также способны эффективно выполнять ограниченные задачи, например классифицировать документы, формировать краткие описания и обрабатывать структурированные запросы. В исследовании В. Л. Чанга и соавторов, посвящённом платформе Chatbot Arena, показана возможность сравнительной оценки языковых моделей на основе предпочтений пользователей при решении различных задач [1]. Подобные исследования подтверждают необходимость оценивать качество конкретных моделей, а не делать выводы только по месту их размещения.

При выборе решения для организации важно учитывать характер предполагаемой нагрузки. Крупная универсальная модель может демонстрировать высокое качество на сложных аналитических заданиях, но её использование для простых повторяющихся операций не всегда экономически оправданно. В то же время компактная модель может обеспечивать достаточную точность при обработке типовых документов, особенно если задача имеет чётко определённые требования. Поэтому практическое сравнение целесообразно проводить на одинаковых запросах, оценивая правильность ответа, полноту результата и необходимость последующей проверки человеком. Универсальные рейтинги являются полезным ориентиром, но не заменяют оценку применимости модели к конкретному рабочему процессу.

Скорость обработки запросов также зависит от нескольких факторов. При облачном использовании общая задержка включает передачу данных по сети, ожидание обработки запроса и непосредственное формирование ответа. Локальная система не испытывает задержки обращения к внешнему

серверу, однако ограничивается производительностью доступного оборудования. Поэтому небольшая локальная модель способна быстро отвечать на простые вопросы, тогда как более крупная может работать медленно при недостатке видеопамяти. Облачная инфраструктура, напротив, позволяет распределять запросы между вычислительными устройствами, хотя при высокой загрузке сервиса также возможно увеличение времени ожидания.

Значение оптимизации вычислительной инфраструктуры рассматривается в исследовании В. Квона и соавторов, посвящённом технологии PagedAttention и системе vLLM. Авторы показали, что более эффективное управление памятью при обработке запросов к большим языковым моделям позволяет увеличивать пропускную способность сервера. В проведённых экспериментах vLLM обеспечивала повышение производительности обработки запросов в 2–4 раза по сравнению с рассмотренными альтернативными системами при сопоставимой задержке [2]. При этом указанный результат относится к сравнению программных механизмов обслуживания моделей, а не непосредственно к сопоставлению локальной и облачной инфраструктуры. Он показывает, насколько существенное влияние на скорость оказывает организация вычислений.

Экономические различия между подходами проявляются прежде всего в структуре затрат. Облачные сервисы обычно предполагают оплату подписки либо фактически использованных вычислительных ресурсов. Такой вариант позволяет избежать крупных первоначальных вложений, однако расходы могут увеличиваться при росте количества пользователей и объёма запросов. Локальное решение требует приобретения оборудования или использования существующих вычислительных мощностей. Дополнительно возникают затраты на электроэнергию, техническое обслуживание, обновление программного обеспечения и работу

специалистов. Следовательно, отсутствие платы за каждый отдельный локальный запрос не означает бесплатности эксплуатации модели.

В исследовании Г. Пэна и Х. Вана предложена модель сопоставления экономических затрат на локальное размещение больших языковых моделей и использование коммерческих облачных сервисов [3]. Авторы показывают, что экономическая целесообразность собственной инфраструктуры существенно зависит от размера модели, интенсивности использования и требований организации. При небольшой или нерегулярной нагрузке облачный подход может оказаться выгоднее благодаря отсутствию капитальных затрат. При постоянном значительном объёме вычислений локальное размещение в некоторых условиях становится конкурентоспособным. Таким образом, выбор должен основываться на совокупной стоимости владения, а не исключительно на цене оборудования или облачной подписки.

Особое значение имеет вопрос конфиденциальности данных. При использовании внешнего облачного сервиса организация передаёт запросы и, возможно, внутренние документы за пределы собственной инфраструктуры. Это требует оценки условий обработки информации, политики хранения, разграничения доступа и договорных обязательств поставщика. Для компаний, работающих с коммерческой тайной, персональными данными и другой чувствительной информацией, такие обстоятельства могут становиться определяющими. Однако само облачное размещение не свидетельствует об отсутствии безопасности. Корпоративные сервисы способны предусматривать специальные механизмы изоляции данных и контроля доступа.

В частности, М. Храпек и соавторы исследовали возможности конфиденциального выполнения больших языковых моделей в защищённых вычислительных средах на центральных и графических процессорах [4]. Результаты показывают, что применение аппаратных

механизмов защиты способно повышать конфиденциальность вычислений при ограниченных дополнительных затратах производительности. Следовательно, защита информации определяется не только физическим расположением вычислительного оборудования, но и используемыми техническими и организационными мерами.

Локальное развёртывание позволяет уменьшить зависимость от внешнего поставщика при обработке данных, однако также не гарантирует полной безопасности. Информация может сохраняться в журналах работы приложения, временных файлах и оперативной памяти, а неправильно настроенные сетевые интерфейсы способны создавать дополнительные риски. В исследовании Ю. Х. З. Ибрагима и соавторов рассмотрены потенциальные нарушения конфиденциальности в программных системах локального выполнения языковых моделей. Авторы обнаружили проблемы хранения данных и разграничения доступа в исследованных конфигурациях [5]. Данная работа показывает, что безопасность локальной системы требует самостоятельной настройки и регулярной проверки. Результаты конкретного исследования не следует автоматически распространять на все локальные модели.

Для организаций целесообразность того или иного подхода определяется сочетанием нескольких критериев. Приоритет облачного решения может быть оправдан при необходимости быстрого внедрения, переменной нагрузке и отсутствии собственной вычислительной инфраструктуры. Локальная система представляет интерес при постоянном использовании, наличии технических специалистов и повышенных требованиях к контролю информации. Возможен и смешанный подход, при котором обработка конфиденциальных данных выполняется внутри организации, а для менее чувствительных задач применяются внешние сервисы. Подобное распределение позволяет использовать преимущества

разных способов развёртывания, хотя требует дополнительных механизмов управления доступом и обменом информацией.

Таким образом, локальные и облачные модели искусственного интеллекта обладают различными преимуществами и ограничениями, которые не позволяют выделить универсально предпочтительное решение. Облачная инфраструктура обеспечивает удобный доступ к вычислительным ресурсам и снижает первоначальные затраты, тогда как локальное развёртывание предоставляет больший контроль над программной средой и обработкой информации. При этом качество ответов зависит преимущественно от характеристик конкретной модели, скорость – от вычислительных ресурсов и организации обработки запросов, а экономическая эффективность – от интенсивности использования. Для предприятий существенное значение приобретает обеспечение конфиденциальности данных, но его нельзя связывать исключительно с физическим расположением модели. Наиболее обоснованным является выбор способа развёртывания на основе требований конкретных задач, совокупных затрат и оценки информационных рисков.

Использованные источники:

1. Chiang W.-L., Zheng L., Sheng Y. et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference // Proceedings of Machine Learning Research. — 2024. — Vol. 235. — P. 8359–8388. — URL: https://proceedings.mlr.press/v235/chiang24b.html (дата обращения: 14.11.2026).

2. Kwon W., Li Z., Zhuang S. et al. Efficient Memory Management for Large Language Model Serving with PagedAttention // Proceedings of the 29th Symposium on Operating Systems Principles. — 2023. — P. 611–626. — DOI: 10.1145/3600006.3613165.

3. Pan G., Wang H. A Cost-Benefit Analysis of On-Premise Large Language Model Deployment: Breaking Even with Commercial LLM Services // 2025 IEEE International Conference on Big Data (BigData). — 2025. — DOI: 10.1109/BigData66926.2025.11400957.
4. Chrapek M., Copik M., Mettaz E., Hoefler T. Confidential LLM Inference: Performance and Cost Across CPU and GPU TEEs // 2025 IEEE International Symposium on Workload Characterization (IISWC). — 2025. — DOI: 10.1109/IISWC66894.2025.00017.
5. Ibrahim Y. H. Z., Ikram M., Salama M. K. The Illusion of Local Privacy: Confidentiality Boundary Failures in Consumer LLM Serving Systems // arXiv. — 2026. — arXiv:2609.18526. — URL: https://arxiv.org/abs/2609.18526 (дата обращения: 14.11.2026).