УДК 004.8; 330.46
DOI 10.26118/2782-4586.2026.94.41.052

Лыков Артем Валерьевич
Кубанский государственный технологический университет

Оценка экономической эффективности решений на основе RAG и Fine-tuning для корпоративной аналитики на базе больших языковых моделей

Аннотация. Исследована формализованная модель оценки экономической эффективности решений корпоративной аналитики на базе больших языковых моделей, реализованных в архитектурах RAG (Retrieval-Augmented Generation, генерация, дополненная поиском) и Fine-tuning (дообучение модели). Рассмотрена структура затрат, возникающих при эксплуатации указанных архитектур в корпоративных информационных системах, включая стоимость инференса, обновления знаний и периодического переобучения моделей. Разработан имитационный стенд, учитывающий интенсивность пользовательских запросов и динамику обновления базы знаний. На основе моделирования TCO (Total Cost of Ownership, совокупная стоимость владения) проведено сравнение экономических характеристик рассматриваемых архитектур при различных сценариях нагрузки. Полученные результаты позволили определить границы рационального применения каждой архитектуры. Показано, что при низкой интенсивности запросов более экономически эффективным является использование RAG-подхода, тогда как при высокой нагрузке и большом количестве обращений преимущество получает архитектура Fine-tuning за счёт эффекта масштабирования затрат.

Ключевые слова: большие языковые модели, LLM, корпоративная аналитика, RAG, retrieval-augmented generation, fine-tuning, экономическая эффективность, TCO, совокупная стоимость владения, имитационное моделирование, динамика знаний, интенсивность запросов, информационные системы, архитектура ИИ, корпоративные информационные системы.

Lykov Artem Valerievich
Kuban State Technological University

Assessment of the economic efficiency of RAG and Fine-tuning solutions for corporate analytics based on Large Language Models

Abstract. A formalized model for evaluating the economic efficiency of corporate analytics solutions based on large language models is presented. The study considers systems implemented using RAG (Retrieval-Augmented Generation) and Fine-tuning architectures. The cost structure associated with the deployment of these approaches in corporate information systems is analyzed, including inference costs, knowledge updates, and periodic model retraining. A simulation framework was developed to account for request intensity and the dynamics of knowledge base updates. Based on TCO (Total Cost of Ownership) modeling, the economic characteristics of the considered architectures were compared under different workload scenarios. The results allowed the identification of rational application boundaries for each architecture. It is shown that the RAG approach is more economically efficient under low request intensity, whereas Fine-tuning becomes advantageous at higher workloads due to the scaling effect of operational costs.

Key words: large language models, LLM, corporate analytics, RAG, retrieval-augmented generation, fine-tuning, economic efficiency, TCO, total cost of ownership, simulation modeling, knowledge dynamics, request intensity, information systems, AI architecture, corporate information systems.

Введение

В последние годы LLM (Large Language Models, большие языковые модели) активно внедряются в корпоративные информационные системы, включая инструменты управленческой и финансовой аналитики [4, 9]. Использование LLM позволяет автоматизировать обработку внутренних регламентов, отчётности и аналитических запросов, повышая скорость принятия управленческих решений. Однако при проектировании корпоративных решений возникает принципиальный выбор архитектуры: использование RAG, основанной на динамическом извлечении данных из внешнего индекса знаний, либо Fine-tuning, предполагающего дообучение модели на корпоративном массиве данных [1, 9].

Несмотря на широкое распространение обеих архитектур, их сравнение преимущественно носит технологический характер [9, 11]. В ряде работ анализируются архитектурные особенности RAG и механизмы адаптации моделей [10, 14], однако экономические аспекты – полная стоимость владения, зависимость затрат от интенсивности запросов и динамики обновления знаний – остаются недостаточно формализованными. В экономической литературе вопросы ценообразования и оптимального распределения ресурсов LLM рассматриваются в более общем контексте [13], что не позволяет напрямую выбрать архитектуру для корпоративной аналитики.

Целью настоящего исследования является разработка формализованной модели оценки экономической эффективности RAG и Fine-tuning решений для корпоративной аналитики. Предложена математическая структура затрат, учитывающая интенсивность запросов, динамику обновления базы знаний и периодическое переобучение модели. На основе имитационного моделирования определены границы рационального применения каждой архитектуры и рассчитаны точки экономической безубыточности.

Научная новизна исследования заключается в формализации зависимости полной стоимости владения LLM-решений от параметров нагрузки и обновляемости знаний, а также в разработке воспроизводимого экспериментального стенда, позволяющего количественно сопоставлять архитектурные подходы в корпоративной среде.

Теоретическая модель экономической оценки

Архитектурная структура затрат RAG и Fine-tuning в корпоративной аналитике

Экономическая оценка архитектур на базе больших языковых моделей должна исходить не из алгоритмических различий, а из структуры формирования затрат в условиях корпоративной эксплуатации. В прикладных системах управленческой аналитики языковая модель выступает не как автономный генератор текста, а как компонент цифровой инфраструктуры, встроенный в контур обработки регламентов, отчётности и аналитических запросов.

Архитектура RAG предполагает разделение хранения знаний и механизма генерации [9]. Корпоративные данные сохраняются во внешнем индексе (векторном или гибридном), а при поступлении запроса осуществляется поиск релевантных фрагментов, которые включаются в контекст обращения к модели [6]. Таким образом, модель выполняет функцию интерпретатора извлечённых данных, а не их хранилища.

С экономической точки зрения данная архитектура характеризуется:

1. Переменными издержками на каждый запрос (стоимость токенов увеличивается за счёт расширенного контекста);
2. Затратами на поддержку и обновление индекса знаний;
3. Относительной независимостью качества ответа от частоты обновлений при условии своевременного индексирования.

Fine-tuning, напротив, переносит корпоративные знания непосредственно в параметры модели посредством дополнительного обучения [12, 14]. После завершения цикла дообучения система способна отвечать на специализированные запросы без обращения к внешнему источнику данных.

Структура затрат в этом случае иная:

1. Значительные фиксированные издержки на процедуру retraining;
2. Сниженные переменные издержки инференса (отсутствует расширенный контекст);
3. Зависимость качества ответа от интервала между циклами переобучения.

Ключевое различие между RAG и Fine-tuning состоит в соотношении фиксированных и переменных затрат. В терминах теории издержек RAG ближе к модели с высокой долей переменных затрат и низким порогом входа, тогда как Fine-tuning представляет собой стратегию с существенными первоначальными инвестициями и эффектом масштаба при росте нагрузки.

Для корпоративной аналитики это различие приобретает особую значимость, поскольку интенсивность запросов и динамика обновления знаний могут существенно варьироваться в зависимости от отрасли, организационной структуры и уровня цифровизации процессов. Следовательно, выбор архитектуры должен основываться на формализованной оценке совокупных затрат с учётом параметров эксплуатации, а не на технологических предпочтениях.

Формализация экономической модели полной стоимости владения

Для формализации экономической оценки введём систему параметров, отражающих эксплуатационные условия корпоративного LLM-решения.

Пусть:

- λ - средняя интенсивность запросов (запросов в день);
- T - горизонт анализа (дней);
- $N = \lambda T$ - общее число запросов за период;
- $U \in [0,1]$ - доля обновления базы знаний за анализируемый период;
- C_{train} - стоимость одного цикла дообучения модели;
- $C_{index}(U)$ - совокупная стоимость обновления и поддержания индекса знаний;
- p_{in}, p_{out} - стоимость входных и выходных токенов;
- t_{in}, t_{out} - среднее число входных и выходных токенов на один запрос.

Переменная стоимость одного запроса определяется токеновой моделью ценообразования:

$$c = p_{in} \cdot t_{in} + p_{out} \cdot t_{out}.$$

В архитектуре RAG величина t_{in}увеличивается за счёт включения в контекст извлечённых фрагментов документов, поэтому:

$$c_{rag} = p_{in} \cdot (t_{base} + t_{ctx}) + p_{out} \cdot t_{out},$$

где t_{ctx} - дополнительное число токенов, обусловленное механизмом retrieval.

Для Fine-tuning контекст, как правило, короче, и переменная стоимость запроса может быть представлена как:

$$c_{ft} = p_{in} \cdot t_{base} + p_{out} \cdot t_{out}.$$

Предполагается, что: $c_{rag} > c_{ft}$, что отражает большую токеновую нагрузку при использовании внешнего контекста.

Полная стоимость владения (Total Cost of Ownership) за период Tдля RAG определяется как:

$$TCO_{RAG}(\lambda, U) = C_{index}(U) + N \cdot c_{rag}.$$

Для Fine-tuning:

$$TCO_{FT}(\lambda, U) = C_{train} + N \cdot c_{ft}.$$

В данной модели фиксированные затраты имеют различную природу. Для RAG они зависят от динамики знаний:

$$C_{index}(U) = \alpha U,$$

где α – коэффициент, отражающий стоимость обработки единицы обновления данных (включая перерасчёт векторных представлений и поддержание инфраструктуры хранения).

Для Fine-tuning фиксированные затраты связаны с периодическим retraining. Если модель переобучается каждые τ дней, то число циклов дообучения за период T равно (T/τ), а совокупные фиксированные затраты составляют:

$$C_{train}^{tot} = \left\lceil \frac{T}{\tau} \right\rceil C_{train}.$$

Тогда итоговое выражение принимает вид:

$$TCO_{FT} = \left\lceil \frac{T}{\tau} \right\rceil C_{train} + \lambda T \cdot c_{ft}.$$

В итоге получены две линейные функции по λ с различными свободными членами и коэффициентами наклона. Экономическая интерпретация модели сводится к анализу влияния параметров λ, U и τ на относительную величину совокупных затрат.

Формализованная структура позволяет перейти к исследованию чувствительности системы к изменению нагрузки и динамики знаний, что является необходимым условием для определения границы рационального применения каждой архитектуры.

Модель динамики знаний и влияние обновляемости данных на затраты и качество

В корпоративной аналитике база знаний не является статичной. Регламенты, отчётные показатели, внутренние процедуры и нормативные документы регулярно изменяются. Для формализации этого процесса введём параметр $U \in [0,1]$, отражающий долю обновления массива знаний за анализируемый период T.

В архитектуре RAG обновление данных приводит к необходимости пересчёта векторных представлений и актуализации поискового индекса. При допущении линейной зависимости издержек от объёма обновлений совокупная стоимость поддержки индекса может быть представлена в виде

$$C_{index}(U) = \alpha U,$$

где α характеризует стоимость обработки единицы обновлённого массива данных. В этом случае производная по U положительна:

$$\frac{\partial C_{index}}{\partial U} = \alpha > 0.$$

Следовательно, рост динамики знаний прямо увеличивает совокупные издержки RAG.

Для Fine-tuning влияние обновлений имеет иной характер. Новые данные не вызывают немедленных затрат, однако приводят к постепенному расхождению между текущим состоянием базы знаний и параметрами модели. Пусть модель переобучается каждые τ дней. Тогда за интервал между циклами retraining доля накопленных обновлений составляет приблизительно

$$U_\tau = U \cdot \frac{\tau}{T}.$$

Рост U_τ связан со снижением релевантности ответов. В упрощённой форме ожидаемое качество можно представить как убывающую функцию накопленного обновления:

$$Q_{ft} = Q_0 - \beta U_\tau,$$

где Q_0 является базовым уровнем качества, а $\beta > 0$ отражает чувствительность модели к устареванию знаний.

В результате динамика знаний воздействует на архитектуры по-разному. В RAG обновления трансформируются в дополнительные прямые затраты, но качество ответа остаётся устойчивым при условии своевременной индексации. В Fine-tuning издержки носят дискретный характер, однако качество снижается между циклами переобучения.

Это различие создаёт дополнительный критерий выбора архитектуры. При высокой изменчивости данных увеличение частоты retraining снижает деградацию качества, но одновременно увеличивает фиксированные издержки:

$$C_{train}^{tot} = \left\lceil \frac{T}{\tau} \right\rceil C_{train}.$$

Следовательно, параметр τ выступает регулятором компромисса между качеством и затратами. Малое значение τ повышает устойчивость модели, но увеличивает TCO. Большое значение τ снижает расходы, но повышает риск потери актуальности знаний.

Подобные эффекты деградации качества при накоплении новых данных без переобучения отмечаются и в прикладных исследованиях по адаптации LLM [12, 15].

Динамика знаний через параметры U и τ формирует дополнительное измерение экономической эффективности, которое должно учитываться совместно с интенсивностью запросов λ.

Определение границы экономической целесообразности и сравнительная статика модели

Сопоставление архитектур требует определения значения нагрузки, при котором их совокупные затраты равны. Пусть горизонт анализа фиксирован и равен T. Условие равенства полной стоимости владения имеет вид

$$C_{index}(U) + \lambda T \cdot c_{rag} = \left\lceil \frac{T}{\tau} \right\rceil C_{train} + \lambda T \cdot c_{ft}.$$

Обозначим

$$C_{train}^{tot} = \left\lceil \frac{T}{\tau} \right\rceil C_{train}.$$

Тогда при непрерывной аппроксимации по λточка безразличия λ^* определяется из выражения

$$\lambda^* = \frac{C_{train}^{tot} - C_{index}(U)}{T \cdot (c_{rag} - c_{ft})}.$$

Соответствующее количество запросов за период

$$N^* = \lambda^* T = \frac{C_{train}^{tot} - C_{index}(U)}{c_{rag} - c_{ft}}.$$

Экономический смысл полученного выражения согласуется с результатами исследований, посвящённых оптимальному распределению токеновых ресурсов и затрат на fine-tuning [13] и заключается в следующем. Если разность переменных издержек $c_{rag} - c_{ft}$ положительна, то при росте числа запросов суммарные затраты RAG увеличиваются быстрее. В этом случае при $N > N^*$ предпочтение следует отдавать Fine-tuning. Если же нагрузка ниже порогового значения, рациональным остаётся использование RAG.

Дальнейший анализ позволяет оценить влияние параметров модели на положение границы эффективности. Рассмотрим частные производные по ключевым переменным.

1. Влияние стоимости переобучения:

$$\frac{\partial N^*}{\partial C_{train}^{tot}} = \frac{1}{c_{rag} - c_{ft}} > 0.$$

Рост совокупных затрат retraining сдвигает границу вправо, что делает Fine-tuning экономически оправданным только при большей нагрузке.

2. Влияние динамики знаний через $C_{index}(U)$:

$$\frac{\partial N^*}{\partial U} = -\frac{\alpha}{c_{rag} - c_{ft}}.$$

Если $\alpha > 0$ и $c_{rag} > c_{ft}$, то производная отрицательна. Увеличение обновляемости базы знаний снижает порог N^*, поскольку RAG несёт возрастающие издержки индексирования. При высокой динамике данных Fine-tuning становится конкурентоспособным при меньшей нагрузке.

3. Влияние разницы переменных издержек:

$$\frac{\partial N^*}{\partial (c_{rag} - c_{ft})} = -\frac{C_{train}^{tot} - C_{index}(U)}{(c_{rag} - c_{ft})^2}.$$

Чем больше разрыв в стоимости одного запроса, тем быстрее достигается экономическое преимущество архитектуры с меньшими переменными издержками.

Рассмотрим предельные режимы.

При $\lambda \to 0$ совокупные затраты приближаются к фиксированной составляющей. В этом случае выбор определяется соотношением $C_{index}(U)$ и C_{train}^{tot}. Если стоимость

retraining существенно выше затрат на поддержание индекса, использование Fine-tuning нецелесообразно.

При $\lambda \to \infty$ доминирует переменная составляющая. Тогда архитектура с меньшим значением cстановится предпочтительной независимо от фиксированных затрат.

Полученные зависимости показывают, что экономическая эффективность RAG и Fine-tuning определяется совместным действием трёх параметров: интенсивности запросов λ, динамики обновления знаний U и периода переобучения τ. Формализованная модель задаёт аналитическую основу для последующей экспериментальной проверки и количественной оценки границ рационального выбора архитектуры.

Практическая реализация и экспериментальная апробация

Архитектура имитационного стенда

Для проверки теоретической модели разработан программный стенд имитационного моделирования, воспроизводящий эксплуатационные условия корпоративной аналитической системы на базе LLM. Целью моделирования являлась количественная оценка полной стоимости владения RAG и Fine-tuning решений при варьировании интенсивности запросов и динамики обновления знаний.

В рамках эксперимента сформирован синтетический корпус из 5000 документов, отражающих типовые элементы корпоративной базы знаний: управленческие отчёты, регламенты, показатели деятельности, план-фактные значения. На основе корпуса сгенерирован набор аналитических запросов, охватывающих извлечение значений показателей, проверку лимитов и расчёт отклонений.

Нагрузка моделировалась как пуассоновский поток с параметром λ, что соответствует случайному характеру обращений пользователей в реальной корпоративной среде. Горизонт анализа установлен равным 30 дням.

Для архитектуры RAG реализован механизм поиска релевантных документов и формирования расширенного контекста запроса. Для Fine-tuning использована модель с периодическим переобучением через фиксированный интервал $\tau = 7$ дней. Обновляемость базы знаний задавалась параметром U, распределённым равномерно по дням периода. Подход соответствует распространённой практике корпоративного применения RAG для управления знаниями [9, 10].

Стоимость инференса рассчитывалась на основе токеновой модели ценообразования с раздельным учётом входных и выходных токенов. Число токенов аппроксимировалось через длину текстового контекста. Полная нагрузка масштабировалась при необходимости, что позволяло сохранять точность оценки при ограничении объёма симуляции.

Стенд воспроизводит ключевые экономические параметры теоретической модели: λ, U, τ, структуру фиксированных и переменных издержек, а также зависимость качества от устаревания данных.

Параметры сценариев и экспериментальный дизайн

Экспериментальная апробация проводилась при варьировании двух ключевых параметров модели: интенсивности запросов λ и динамики обновления знаний U. Горизонт анализа во всех сценариях фиксирован и равен $T = 30$ дней.

Для оценки влияния нагрузки рассматривались три уровня интенсивности запросов:

$$\lambda \in \{100,\ 1000,\ 10000\} \text{ запросов в день.}$$

Такая градация отражает различные масштабы корпоративной эксплуатации: от локального аналитического сервиса до системы, интегрированной в контур массовых бизнес-процессов.

Динамика обновления базы знаний моделировалась параметром $U \in \{0.01,\ 0.05,\ 0.20\}$, что соответствует обновлению 1%, 5% и 20% массива данных за анализируемый период. Значения выбраны таким образом, чтобы охватить диапазон от относительно стабильной нормативной базы до высокодинамичной управленческой среды.

Период переобучения модели в архитектуре Fine-tuning установлен равным $\tau = 7$ дней, что отражает типичную практику регулярного обновления корпоративных моделей. Соответственно, за период анализа выполнялось несколько циклов retraining, формирующих фиксированную компоненту затрат.

Для каждого сочетания параметров (λ, U) рассчитывались следующие показатели:

- полная стоимость владения TCO_{RAG} и TCO_{FT};
- средняя стоимость одного запроса;
- распределение задержек инференса;
- доля корректных ответов как прокси-показатель качества.

С целью обеспечения статистической устойчивости моделирование выполнялось с фиксированным генератором случайных чисел. При высоких значениях λиспользовалась процедура масштабирования переменной составляющей затрат, что позволяло сохранить корректность расчётов без избыточного увеличения времени симуляции.

Сформированный экспериментальный дизайн обеспечивает сопоставимость архитектур в идентичных условиях нагрузки и обновляемости данных, что позволяет интерпретировать различия в результатах как следствие именно экономической структуры моделей.

Результаты моделирования

Графические результаты расчётов представлены на Рисунках 1–3. Общие результаты содержатся в файле summary_pretty.xlsx репозитория [19]. Графический анализ позволяет выявить закономерности изменения совокупных затрат и качества в зависимости от интенсивности запросов и динамики обновления знаний.

На Рисунке 1 показана зависимость полной стоимости владения от интенсивности запросов при различных значениях U. Для повышения наглядности ось ординат представлена в логарифмической шкале. Во всех сценариях функции TCO_{RAG} и TCO_{FT} имеют линейный характер по λ, что соответствует теоретической модели. При низкой нагрузке архитектура RAG демонстрирует существенно меньшие совокупные затраты вследствие отсутствия значительных фиксированных издержек переобучения. При росте интенсивности запросов наблюдается пересечение кривых, и при высоких значениях λ Fine-tuning становится экономически предпочтительным за счёт более низкой переменной стоимости запроса.

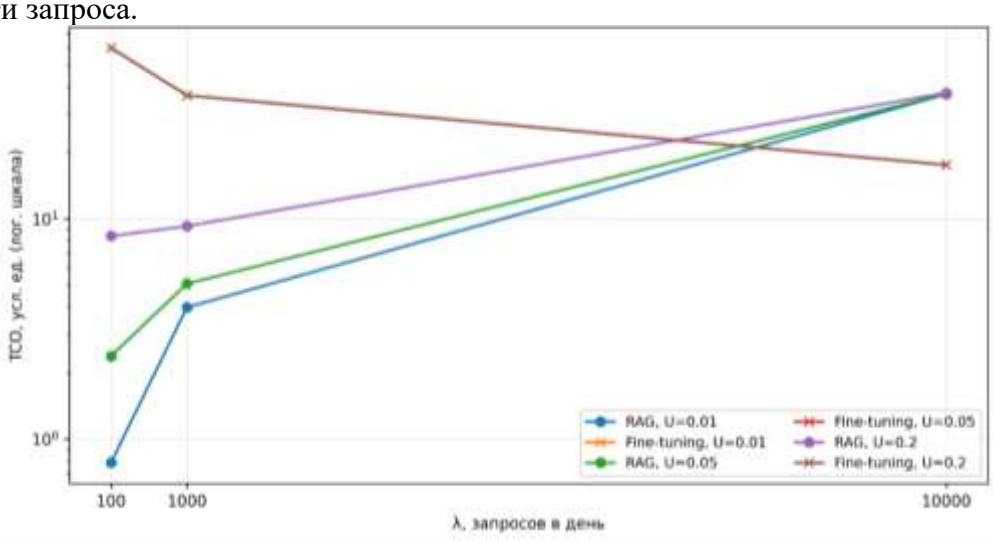

Рисунок 1 – Зависимость TCO от интенсивности запросов при различных U [19].

На Рисунке 2 представлена зависимость средней стоимости одного запроса от интенсивности нагрузки. Ось ординат также приведена в логарифмической шкале, что позволяет корректно сопоставить сценарии с различающимися порядками величин затрат. В архитектуре RAG средняя стоимость запроса остаётся относительно стабильной и определяется расширенным контекстом. В случае Fine-tuning наблюдается выраженный эффект масштаба: при увеличении числа запросов фиксированные затраты retraining

распределяются на больший объём обращений, что приводит к снижению средней стоимости.

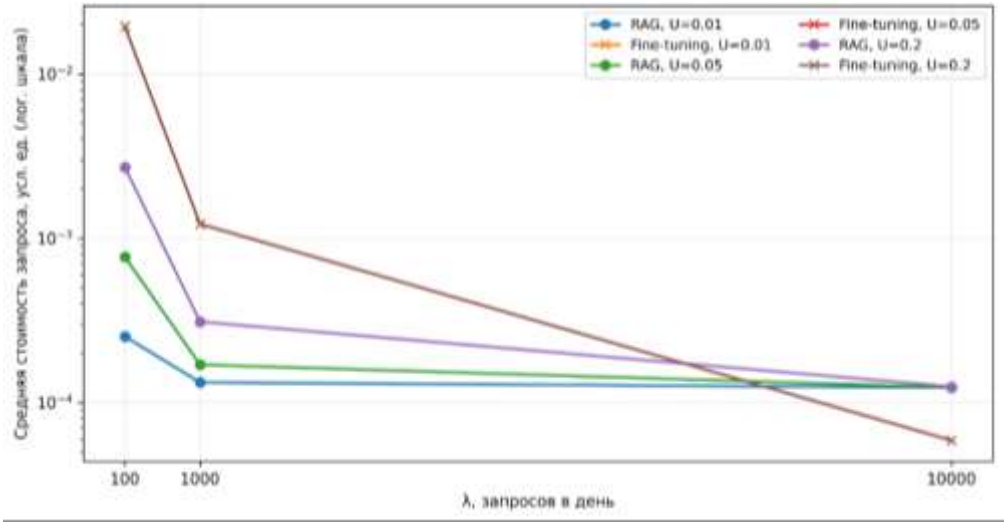

Рисунок 2 – Средняя стоимость запроса в зависимости от λ [19].

На Рисунке 3 показана разница долей корректных ответов $\Delta accuracy = accuracy_{RAG} - accuracy_{FT}$ в зависимости от параметра U для различных значений λ. Представление в виде разности позволяет непосредственно оценить сравнительное преимущество архитектур. При всех уровнях нагрузки RAG демонстрирует более высокую устойчивость качества при росте динамики обновления базы знаний. Увеличение U приводит к расширению разрыва в качестве, что отражает влияние устаревания параметров модели в архитектуре Fine-tuning при фиксированном интервале переобучения.

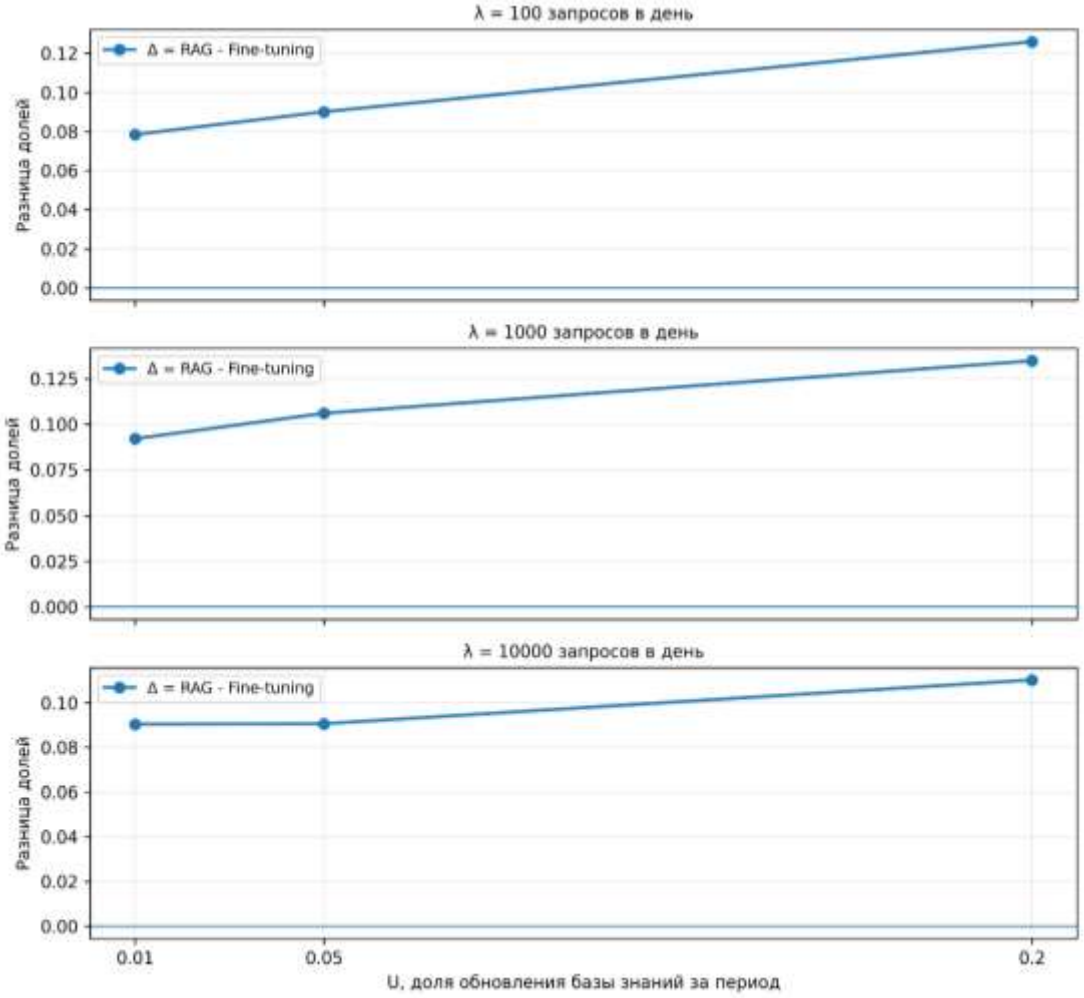

Рисунок 3 – Разница долей корректных ответов (Δaccuracy) в зависимости от U [19].

Совокупный анализ результатов подтверждает выводы теоретического раздела. Экономическая эффективность архитектур определяется совместным действием интенсивности запросов и динамики обновления данных. При низкой нагрузке рациональным остаётся использование RAG, тогда как при высокой нагрузке Fine-tuning обеспечивает снижение совокупных затрат. Рост обновляемости базы знаний усиливает преимущества RAG с точки зрения устойчивости качества.

Интерпретация результатов и практические выводы

Полученные результаты позволяют перейти от формального сопоставления затрат к выработке прикладных рекомендаций для корпоративных систем.

При низкой интенсивности запросов доминирующее влияние оказывают фиксированные издержки. В рассматриваемых сценариях стоимость retraining существенно превышает затраты на поддержание индекса, поэтому внедрение Fine-tuning экономически нецелесообразно для сервисов с ограниченным числом обращений. В таких условиях RAG обеспечивает более низкий порог входа и гибкость масштабирования.

При росте нагрузки наблюдается эффект распределения фиксированных затрат Fine-tuning на увеличивающийся объём запросов. Это приводит к снижению средней стоимости инференса и формирует экономическое преимущество при массовой эксплуатации системы. Для крупных организаций с высокой частотой аналитических обращений данный фактор становится определяющим.

Динамика обновления базы знаний оказывает разнонаправленное воздействие на архитектуры. В RAG увеличение объёма обновлений приводит к росту инфраструктурных затрат, однако качество ответа остаётся стабильным при условии регулярной индексации. В Fine-tuning высокая изменчивость данных требует сокращения интервала между циклами retraining, что увеличивает совокупные издержки. При неизменном периоде переобучения качество ответа снижается.

Следовательно, выбор архитектуры зависит от сочетания трёх параметров: интенсивности запросов, динамики обновления знаний и допустимого уровня качества. При стабильной нормативной базе и высокой нагрузке оправдано инвестирование в Fine-tuning. При высокой изменчивости данных или умеренной нагрузке предпочтительным остаётся RAG-подход.

Полученные выводы согласуются с аналитическими зависимостями теоретической модели и подтверждают применимость формализованного подхода к задачам экономического обоснования архитектурных решений в корпоративной аналитике [8, 16].

Заключение

В работе разработана формализованная модель оценки экономической эффективности архитектур RAG и Fine-tuning в корпоративной аналитике на базе больших языковых моделей. Предложена структура полной стоимости владения, учитывающая интенсивность запросов, динамику обновления базы знаний и периодичность переобучения модели. Выведена аналитическая формула границы экономической целесообразности, позволяющая определить порог нагрузки, при котором одна архитектура становится предпочтительной по сравнению с другой.

Результаты имитационного моделирования подтвердили теоретические выводы. При низкой интенсивности обращений доминирующее значение имеют фиксированные издержки, что делает RAG экономически более оправданным. При высокой нагрузке переменные затраты становятся определяющими, и Fine-tuning демонстрирует эффект масштаба.

Разработанный имитационный стенд обеспечивает воспроизводимость расчётов и может быть адаптирован под конкретные параметры организации.

Ограничением модели является использование синтетического корпуса и аппроксимация токеновой нагрузки. Дальнейшие исследования могут быть направлены на

применение методики к реальным корпоративным данным и расширение модели с учётом дополнительных факторов инфраструктурных затрат [13].

Список источников

1. Шмат А. В. Применение больших языковых моделей и технологий Retrieval-Augmented Generation для корпоративных ассистентов // Вестник цифровых технологий. — 2024. — № 3. — С. 45–58.
2. Иванов Д. С., Петрова Е. Н. Экономическая оценка внедрения интеллектуальных информационных систем в корпоративной среде // Экономика и управление. — 2023. — № 12. — С. 67–75.
3. Кузнецов М. А. Имитационное моделирование информационных систем предприятия. — М.: Инфра-М, 2022. — 256 с.
4. Сидоров А. П., Белова Н. И. Цифровая трансформация корпоративной аналитики на основе технологий искусственного интеллекта // Управленческие науки. — 2024. — Т. 14, № 2. — С. 89–101.
5. Григорьев В. Л. Экономика информационных технологий. — СПб.: Питер, 2021. — 304 с.
6. Архитектура Retrieval-Augmented Generation: обзор и применение [Электронный ресурс] // Habr. — 2025. — Режим доступа: https://habr.com/ru/articles/931396 (дата обращения: 27.02.2026).
7. RAG vs Fine-tuning: что выбрать бизнесу и разработчикам в 2025 году [Электронный ресурс] // ServerFlow. — 2025. — Режим доступа: https://serverflow.ru/blog/stati/rag-vs-fine-tuning-chto-vybrat-dlya-biznesa-i-razrabotchikov-v-2025-godu (дата обращения: 27.02.2026).
8. RAG или Fine-tuning — как выбрать метод для LLM-задач [Электронный ресурс] // Napoleon IT. — 2025. — Режим доступа: https://napoleonit.ru/blog/rag-ili-fine-tuning-kak-vybrat-pravilnyy-metod-dlya-nastroyki-llm (дата обращения: 27.02.2026).
9. Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey // arXiv preprint. — 2023. — Режим доступа: https://arxiv.org/abs/2312.10997 (дата обращения: 28.02.2026).
10. Karakurt E., Akbulut A. Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) for Enterprise Knowledge Management and Document Automation: A Systematic Literature Review // Applied Sciences. — 2026. — Vol. 16, No. 1. — Article 368.
11. Shen M., Gupta U., Zhang Y. et al. Towards Understanding Systems Trade-offs in Retrieval-Augmented Generation Model Inference // arXiv preprint. — 2024. — Режим доступа: https://arxiv.org/abs/2412.11854 (дата обращения: 28.02.2026).
12. Devine P. ALoFTRAG: Automatic Local Fine Tuning for Retrieval Augmented Generation // arXiv preprint. — 2025. — Режим доступа: https://arxiv.org/abs/2501.11929 (дата обращения: 28.02.2026).
13. Bergemann D., Bonatti A., Smolin A. The Economics of Large Language Models: Token Allocation, Fine-Tuning, and Optimal Pricing // arXiv preprint. — 2025. — Режим доступа: https://arxiv.org/abs/2502.07736 (дата обращения: 28.02.2026).
14. Ren R., Li Q., Zhang T. Adaptive Two-stage Retrieval Augmented Fine-Tuning Method // Expert Systems with Applications. — 2025. — Vol. 244.
15. Robust Fine-Tuning for Retrieval Augmented Generation // Proceedings of the ACM Conference on Information and Knowledge Management. — 2025.
16. RAG vs. Fine-Tuning: Comparative Analysis [Электронный ресурс] // Monte Carlo Data. — 2025. — Режим доступа: https://www.montecarlodata.com/blog-rag-vs-fine-tuning (дата обращения: 01.03.2026).
17. Should You Fine-Tune Your Large Language Models or Let RAG Do the Heavy Lifting [Электронный ресурс] // Centific. — 2025. — Режим доступа:

https://www.centific.com/blog/should-you-fine-tune-your-large-language-models-or-let-rag-do-the-heavy-lifting (дата обращения: 01.03.2026).

18. Fine-Tuning vs RAG Trade-offs in Large Language Models for Domain-Specific Tasks // Journal of Medical Internet Research. — 2026.

19. Лыков А.В. Economic Evaluation of RAG and Fine-Tuning Architectures [Электронный ресурс]. — Режим доступа: https:// https:// https://github.com/MrMixaDj32/rag-ft-economic-evaluation (дата обращения: 02.03.2026).

Сведения об авторе

Лыков Артем Валерьевич, студент направления подготовки «Программная инженерия», Кубанский государственный технологический университет, Краснодар, Россия,

Научный руководитель

Волик Александр Георгиевич, старший преподаватель кафедры информационных систем и программирования, Кубанский государственный технологический университет, Краснодар, Россия

Information about the author

Lykov Artem Valerievich, student of the training program "Software Engineering", Kuban State Technological University, Krasnodar, Russia

Scientific supervisor

Volik Alexander Georgievich, Senior Lecturer, Department of Information Systems and Programming, Kuban State Technological University, Krasnodar, Russia