

СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ/SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

DOI: https://doi.org/10.60797/IRJ.2026.171.79 EDN: XRCHJN

МЕТОДИКА ПРЕДВАРИТЕЛЬНОЙ ОЦЕНКИ И ОТБОРА ЦИТАТНЫХ ФРАГМЕНТОВ ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ГЕНЕРАЦИИ НАУЧНЫХ ОБЗОРНЫХ ТЕКСТОВ

Научная статья

Кузнецов И.И.[1, *]
[1] ORCID : 0009-0001-6287-8295;
[1] Российский государственный университет им. А.Н. Косыгина, Москва, Российская Федерация

* Корреспондирующий автор (iliya-kuznetsov[at]mail.ru)

Предложена: 25.07.2026; Принята: 04.09.2026; Опубликована: 17.09.2026

Аннотация

В работе предложена методика оценки и отбора цитатных фрагментов при цитатно-осведомленной генерации научных текстов большими языковыми моделями. Методика основана на оценке соответствия цитатных фрагментов содержанию цитируемых публикаций и направлена на формирование компактного и информативного представления научной информации. Для оценки предложенного подхода проведено экспериментальное исследование на основе датасета SurGE с использованием моделей семейств Qwen3 (4B, 8B, 14B, 32B) и Llama 3.1 (8B, 70B, 405B). Показано, что предварительный отбор цитатных фрагментов повышает качество генерации обзорных текстов по сравнению с использованием полных текстов публикаций и полного набора извлеченных цитат.

Наиболее выраженный эффект наблюдается для моделей меньшего масштаба: для Qwen3 4B прирост значений критериев качества составил 11,53–18,09%, тогда как для Qwen3 32B и Llama 3.1 405B — 4,25–5,96% и 2,85–6,28% соответственно. В отдельных случаях модели меньшего масштаба с предварительным отбором достигали результатов, сопоставимых с более крупными моделями того же семейства при обработке полных текстов. Применение предложенного подхода обеспечило сокращение суммарного токенного трафика на 87,1% и времени выполнения генеративных этапов на 17,6–27,3%. Полученные результаты показывают возможность повышения качества и ресурсоэффективности цитатно-осведомленной генерации за счет предварительной оценки и отбора цитатных фрагментов.

Ключевые слова: большие языковые модели, цитатно-осведомленная суммаризация, цитирование, извлечение цитат, научные обзоры, анализ текстовой информации, научные публикации.

A METHODOLOGY FOR THE PRELIMINARY ASSESSMENT AND SELECTION OF QUOTED PASSAGES TO IMPROVE THE EFFICIENCY OF GENERATING SCIENTIFIC REVIEW ARTICLES

Research article

Kuznecov I.I.[1, *]
[1] ORCID : 0009-0001-6287-8295;
[1] The Kosygin State University of Russia, Moscow, Russian Federation

* Corresponding author (iliya-kuznetsov[at]mail.ru)

Suggested: 25.07.2026; Accepted: 04.09.2026; Published: 17.09.2026

Abstract

The work proposes a method for evaluating and selecting citation fragments in citation-aware generation of scientific texts using large language models. The method is based on assessing the alignment of citation fragments with the content of the cited publications and aims to produce a concise and informative representation of scientific information. To evaluate the suggested approach, an experimental study was conducted using the SurGE dataset and models from the Qwen3 (4B, 8B, 14B, 32B) and Llama 3.1 (8B, 70B, 405B) families. It has been shown that the preliminary selection of cited fragments improves the quality of review text generation compared to using the full texts of publications and the complete set of extracted citations.

The most significant effect is observed for smaller-scale models: for Qwen3 4B, the increase in quality metric scores ranged from 11.53% to 18.09%, while for Qwen3 32B and Llama 3.1 405B, the increases were 4.25%–5.96% and 2.85%–6.28%, respectively. In some cases, smaller-scale models with pre-selection achieved results comparable to those of larger models from the same family when processing full texts. The application of the suggested approach resulted in a reduction in total token traffic by 87.1% and in the execution time of the generative stages by 17.6–27.3%. The obtained results demonstrate the potential to improve the quality and resource efficiency of citation-aware generation through the preliminary evaluation and selection of cited fragments.

Keywords: large language models, citation-aware summarisation, citation, quotation extraction, scientific reviews, text analysis, scientific publications.

Введение

Большие языковые модели (Large Language Model, далее — LLM) в последние годы получили широкое распространение при решении задач интеллектуальной обработки научной информации. Современные генеративные

модели применяются для поиска и анализа научных публикаций, извлечения знаний и подготовки научных текстов, в том числе обзорного характера. Развитие методов, опирающихся на использование внешнего контекста в виде текстовых документов или ссылок (RAG-генерация), позволило существенно расширить возможности подобных систем [1].

Обработка научных публикаций требует специализированных методов, обеспечивающих корректную обработку источников и контроль качества формируемого текста [2]. Например, в работе [3] предложен подход к автоматизированному построению научных обзоров, учитывающий качество исходных публикаций и структуру формируемого обзора, который в исследовании [4] был развит посредством многоэтапного процесса генерации, включающего процедуры оценки качества получаемого текста.

Увеличение масштаба современных языковых моделей сопровождается ростом требований к вычислительным ресурсам [5], вследствие чего наряду с качеством генерации все большее значение приобретает ее вычислительная эффективность [6]. Повышение эффективности применения LLM стало самостоятельным направлением исследований [7]. В работах [8] и [9] рассматриваются способы более рационального использования вычислительных ресурсов без дальнейшего увеличения размера моделей. Одним из таких способов является применение моделей меньшего размера, способных обеспечивать конкурентоспособные результаты в специализированных задачах при наличии релевантной внешней информации [10]. При этом качество подаваемой на вход контекстной информации существенно влияет на результаты генерации, а наличие нерелевантных фрагментов способно приводить к их ухудшению [11].

В связи с этим развиваются методы предварительной подготовки входной информации, направленные на формирование компактного, но информативного представления исходного материала, поскольку предварительный отбор информации оказывает заметное влияние на качество RAG-генерации [12]. В исследовании [13] продемонстрирована эффективность формирования взаимодополняющего набора исходных документов. В работе [14] предложен метод сжатия извлекаемого контекста, дающий возможность уменьшить объем входной информации без существенного ухудшения качества генерации. Рассматриваются также методы сокращения промптов, позволяющие значительно снизить вычислительные затраты [15], [16]. Исследуются возможности обработки длинного контекста, обеспечивающие уменьшение негативного влияния избыточной информации [17].

Для научных публикаций одним из перспективных способов формирования их компактного представления является использование цитатной информации. Цитатные фрагменты зачастую отражают ключевые результаты исследования, выделенные и интерпретированные научным сообществом [18]. Подход к обработке научных публикаций, основанный на обработке цитатного материала, называется цитатно-осведомленным. Показано, что использование цитатных фрагментов совместно с соответствующими им фрагментами исходной статьи позволяет повышать информативность саммари [19]. Дальнейшее развитие этого направления привело к появлению методов совместного анализа цитат и текста цитируемых публикаций [20], а также специализированных корпусов для генерации научных обзоров [21].

Рассматриваются способы объединения цитатно-осведомленного подхода с методами RAG-генерации. Например, в работе [22] предложен подход к генерации научных обзоров, основанный на адаптивном поиске публикаций, рассмотрении графа цитирования и управляемом планом синтезе текста. Вместе с тем существующие подходы преимущественно ориентированы на применение цитатной информации как уже подготовленного источника знаний, зачастую не рассматривая предварительную оценку качества отдельных цитатных фрагментов.

Цитатные фрагменты при этом существенно различаются по содержанию и структуре. В работе [23] установлено, что описание одной и той же научной публикации нередко распределено между несколькими предложениями. Показано, что различные цитаты выполняют разные коммуникативные функции и отражают различные аспекты научной публикации [24]. Цитаты различаются и по степени соответствия содержанию цитируемой публикации [25]. В работе [26] авторами предложена методика проверки корректности содержания цитат в биомедицинских публикациях. Исследования свидетельствуют о том, что цитатные утверждения могут содержать неполные либо неточные интерпретации оригинальной работы [27]. В связи с этим общие методы сокращения входного контекста, рассмотренные, например, в работах [14] или [15], не в полной мере учитывают специфику рассматриваемой задачи: для отбора цитат существенное значение имеет не только сокращение объема или сохранение релевантной информации, но и соответствие содержащихся в них утверждений содержанию цитируемой публикации. Таким образом, важную роль в повышении итогового качества цитатно-осведомленной генерации играет предварительная оценка и отбор цитатных фрагментов, особенно при работе с генеративными моделями меньших размеров.

При этом существующие работы, посвященные предварительной подготовке входной информации, преимущественно оценивают влияние этой подготовки на качество генерации при использовании фиксированной языковой модели. Так, показано, что качество RAG-генерации определяется характеристиками извлекаемого контекста и исследуемой модели [28], однако исследование не учитывает специфику научных публикаций. В работе [29] изучается влияние размера языковой модели на качество суммаризации, но при этом сам способ формирования входной информации остается неизменным. Также показана возможность повышения эффективности малых моделей за счет опоры на внешнюю информацию [30], однако задача генерации научных обзорных текстов авторами не рассматривается.

Таким образом, несмотря на развитие методов цитатно-осведомленной генерации и предварительной подготовки входной информации, недостаточно исследованным остается влияние предварительной оценки и отбора цитатных фрагментов на эффективность генерации при использовании языковых моделей различного масштаба. В частности, остается открытым вопрос, в какой степени предварительный отбор цитатного материала способен компенсировать ограничения моделей меньшего размера и тем самым снизить вычислительные затраты без существенного ухудшения качества генерируемых научных текстов.

Целью настоящей работы является разработка методики оценки и отбора цитатных фрагментов для цитатно-осведомленной генерации научных саммари и обзорных текстов, а также исследование ее влияния на качество и ресурсоэффективность генерации языковыми моделями различного масштаба. Научная новизна работы заключается в формализации предварительной оценки и отбора цитатных фрагментов как самостоятельного этапа подготовки входных данных для цитатно-осведомленной генерации, а также в количественной оценке влияния такого отбора на качество и ресурсоэффективность генерации при использовании языковых моделей различного масштаба. Отдельно исследуется возможность частичной компенсации меньшего масштаба языковой модели за счет предварительного отбора цитатной информации.

Материалы и методы

2.1. Методика оценки соответствия цитатных фрагментов содержанию цитируемой публикации

В работе предложена методика оценки и отбора цитатных фрагментов для формирования компактного входного представления цитируемой научной публикации. Методика позволяет перейти от оценки качества извлечения цитат к оценке их соответствия содержанию цитируемой публикации и использовать при генерации только наиболее информативные и достоверные цитатные фрагменты.

Отдельные элементы подхода ранее были реализованы автором в составе программного комплекса цитатно-осведомленной генерации научных обзоров, однако в настоящей работе методика впервые представлена как самостоятельная формализованная последовательность этапов подготовки входных данных.

Предлагаемая методика состоит из следующих основных этапов:

1. Подготовка корпуса научных публикаций, содержащего цитируемые и связанные с ними цитирующие публикации, а также относящиеся к ним метаданные.
2. Автоматизированное определение и извлечение цитатных фрагментов из цитирующих публикаций.
3. Определение участков текста цитируемых публикаций, соответствующих каждому извлеченному цитатному фрагменту.
4. Вычисление набора показателей соответствия между извлеченными цитатными фрагментами и найденными релевантными фрагментами цитируемых публикаций.
5. Совместная оценка качества цитатных фрагментов на основе совокупности вычисленных показателей.
6. Отбор цитатных фрагментов в соответствии с полученными оценками.
7. Формирование входного представления научной публикации на основе отобранных цитатных фрагментов.

Для применения методики необходимы полнотекстовые цитируемые и цитирующие публикации, их метаданные и сведения о связях цитирования. Результатом является компактное входное представление публикации на основе отобранных цитатных фрагментов, используемое далее для генерации научных обзорных текстов. В рамках настоящего исследования оно применяется для экспериментальной оценки эффективности языковых моделей различного масштаба при различных способах формирования входных данных.

2.2. Программная реализация, используемые критерии и метрики

Экспериментальное исследование выполнено с применением разработанного ранее программного комплекса цитатно-осведомленной генерации научных обзорных текстов, архитектура и основные этапы функционирования которого подробно представлены в работе [31]. Программный комплекс реализует полный цикл обработки научных публикаций, направленный на цитатно-осведомленное построение научных текстов. Основные этапы работы программного комплекса приведены на рис. 1.

Рисунок 1 - Основные этапы работы программного комплекса для цитатно-осведомленного формирования обзорных текстов

DOI: https://doi.org/10.60797/IRJ.2026.171.79.1

Для определения соответствия цитатных фрагментов содержанию цитируемых публикаций текст каждой статьи предварительно разбивается на абзацы и предложения. В используемой программной реализации для каждого цитатного фрагмента с помощью специализированной предобученной языковой модели SciBERT [32] вычисляется семантическое сходство с фрагментами цитируемой публикации, после чего определяется наиболее релевантный фрагмент.

Оценка качества цитатных фрагментов выполняется с использованием совокупности метрик, отражающих различные аспекты соответствия цитаты содержанию цитируемой публикации. Лексическое сходство оценивается посредством ROUGE-L [33], семантическое посредством BERTScore [34], фактологическое соответствие с помощью вопросно-ответной оценки (QA) [35], а логическая согласованность с применением модели естественного логического вывода (NLI) [36]. Совместное применение указанных метрик позволяет учитывать взаимодополняющие свойства анализируемых текстов, поскольку исследования показывают, что использование отдельной метрики не обеспечивает достаточной полноты оценки [37].

Частные показатели объединяются посредством взвешенной суммы в интегральную оценку качества цитатного фрагмента:

$$S = \sum_{i=1}^{n} w_i m_i$$

где m_i — значение i -й метрики качества, w_i — соответствующий весовой коэффициент.

Значения весовых коэффициентов и порог отбора были определены на отдельной предварительной выборке цитатных фрагментов, не входившей в экспериментальную выборку настоящего исследования, путем сопоставления результатов с ручной оценкой, выполненной автором. Больший вес назначался показателям, непосредственно характеризующим фактологическую и логическую согласованность фрагмента с исходным текстом. Дополнительно была проведена оценка чувствительности к изменению весовых коэффициентов: при их варьировании в пределах ±10% доля совпадающих решений об отборе составляла 95,7–98,3%, что указывает на отсутствие существенной зависимости результатов от небольших изменений конкретных значений весов.

Для формирования сокращенного набора цитатных фрагментов могут применяться различные альтернативные способы, включая случайный отбор, ранжирование на основе TF-IDF и BM25, а также методы, учитывающие одновременно релевантность и разнообразие, например, MMR. Однако такие подходы не учитывают в полной мере специфику цитатной информации: высокая лексическая или семантическая релевантность цитаты не гарантирует точности отражения содержания цитируемой публикации, а оптимизация разнообразия не позволяет непосредственно

оценить ее фактологическую и логическую согласованность с источником**.** В качестве базовых способов для экспериментального сравнения были выбраны случайный отбор и BM25. Первый использовался для оценки эффекта самого сокращения объема входных данных, второй для сопоставления предложенной многокомпонентной оценки с простым релевантностным ранжированием. Для обеспечения сопоставимости во всех случаях отбиралось одинаковое количество цитатных фрагментов.

Для сопоставления результатов различных способов формирования входных данных качество сгенерированных текстов оценивалось на двух уровнях: для промежуточных саммари отдельных публикаций и для итоговых обзорных текстов, формируемых на их основе. Качество промежуточных саммари оценивается по трем критериям: правдоподобности, покрытию и фактологической согласованности. Значения указанных критериев определяются на основе совокупности результатов метрик ROUGE-L, BERTScore, QA и NLI. Правдоподобность характеризует смысловое соответствие сгенерированного текста исходным данным, покрытие — полноту отражения содержащейся в них информации, а фактологическая согласованность — отсутствие неподтвержденных или противоречащих источникам утверждений. Полученные значения объединяются во взвешенный итоговый показатель качества саммари, на основании которого осуществляется отбор промежуточных результатов для формирования итогового обзора.

Качество итоговых обзорных текстов оценивается путем их сопоставления с эталонными обзорами, подготовленными экспертами. В качестве основных критериев выступают правдоподобность, покрытие и фактологическая согласованность, которые дополняются оценкой с использованием LLM (LLM-as-judge). Эта оценка помогает учитывать характеристики итогового текста, не полностью отражаемые отдельными автоматическими метриками, включая связность, логическую целостность и качество научного изложения. Оценивание выполняется по единому фиксированному промпту, приведенному в приложении 1. Надежность LLM-as-judge проверяется по корреляции Спирмена с независимой человеческой оценкой и среднему абсолютному отклонению результатов десяти повторных запусков. Критерии оценки промежуточных саммари и итоговых обзорных текстов также более подробно описаны в работе [31]. Описанная система критериев используется для сопоставления результатов, полученных при различных способах формирования входных данных и применении языковых моделей различного масштаба

Для проверки статистической значимости различий между сравниваемыми подходами использовался парный анализ результатов, полученных для одних и тех же тематик. Для парных различий рассчитывались 95%-ные доверительные интервалы с использованием бутстреп-метода, а статистическая значимость оценивалась с помощью парного критерия Уилкоксона. Для учета множественных сравнений применялась поправка Холма; различия считались статистически значимыми при скорректированном значении p <0,05.

Для оценки ресурсоэффективности различных способов формирования входных данных дополнительно учитывались токенный трафик и временные затраты. Токенный трафик определялся как суммарное число входных и выходных токенов на этапах формирования промежуточных саммари и итогового обзора. Временные затраты оценивались раздельно для предварительной обработки данных программным комплексом и генеративных этапов с использованием LLM. Аппаратные затраты непосредственно не оценивались, поскольку генерация выполнялась посредством внешнего программного интерфейса, и отсутствие этих данных не позволяет провести корректное сравнение. Исходя из этого, ресурсоэффективность в настоящем исследовании рассматривается с точки зрения объема обрабатываемого LLM токенного трафика и временных затрат.

Результаты

3.1. Исходные данные и конфигурация эксперимента

В качестве источника данных использован открытый датасет SurGE [38], содержащий данные о тематических наборах научных публикаций и соответствующие им экспертные обзоры, с однозначным соответствием исходным статьям. Из 205 тематик были отобраны те, для которых каждая цитируемая публикация имела доступный PDF-документ и не менее десяти цитирующих публикаций в формате PDF. Итоговая выборка составила 114 тематик, 7103 цитируемые и 73514 цитирующих публикаций.

Основные параметры конфигурации программы для генерации и оценки обзоров в ходе проведения эксперимента приведены в табл. 1.

Таблица 1 - Основные параметры конфигурации системы при проведении эксперимента

DOI: https://doi.org/10.60797/IRJ.2026.171.79.2

Параметр	Значение	Параметр	Значение
Вес метрики ROUGE для отбора цитатных фрагментов	0,1	Количество опорных высказываний	от 5 до 10
Вес метрики BERTScore для отбора цитатных фрагментов	0,15	Вес правдоподобности для отбора саммари	0,4
Вес метрики QA для отбора цитатных фрагментов	0,35	Вес покрытия для отбора саммари	0,2
Вес метрики NLI для отбора цитатных	0,4	Вес фактологической достоверности для	0,4

Параметр	Значение	Параметр	Значение
фрагментов		отбора саммари	
Порог отбора цитатных фрагментов	0,63	Порог отбора саммари	0,78
Длина саммари, слов	от 90 до 100	Длина обзорного текста, слов	от 2800 до 3000

Эксперимент был направлен на оценку влияния способа формирования входных данных и масштаба модели на качество генерации.

Для каждой модели были рассмотрены три подхода к генерации. Первый подход (далее П1) предполагал использование полного текста цитируемых статей, второй подход (далее П2) предполагал использование всех цитатных фрагментов, относящихся к цитируемым статьям, и третий подход (далее П3) предполагал использование цитатных фрагментов, прошедших предварительный отбор согласно методике, предложенной в рамках данной статьи. Во всех случаях использовались одинаковые промпты, параметры генерации и ограничения на объем выходного текста. Таким образом, изменяемыми факторами являлись только способ формирования входных данных и масштаб языковой модели.

Эксперимент проводился на двух семействах открытых LLM — Qwen3 и Llama 3.1. Их выбор был обусловлен наличием последовательности моделей различного масштаба, построенных в рамках одной архитектурной линейки, открытой информацией об их архитектурных характеристиках и количестве параметров, а также возможностью выполнения вычислений посредством API.

Семейство Qwen3 было представлено моделями 4B, 8B, 14B и 32B, семейство Llama 3.1 — моделями 8B, 70B и 405B. Исследование двух независимых семейств позволило оценить устойчивость полученных закономерностей относительно архитектурных особенностей конкретной модели.

3.2. Полученные результаты

Для всех 114 тематик сгенерированные обзоры сравнивались с эталонными по правдоподобности, покрытию, фактологической согласованности и LLM-оценке. Для последней использовалась модель Claude-Sonnet-4-5-20250929. Валидация LLM-as-judge показала корреляцию с человеческой оценкой ρ=0,78 по коэффициенту Спирмена; среднее абсолютное отклонение при повторных запусках составило 0,017. Результаты эксперимента приведены в табл. 2.

Таблица 2 - Медианные значения критериев качества генерации обзоров для рассматриваемых моделей и подходов

DOI: https://doi.org/10.60797/IRJ.2026.171.79.3

Модель	Подход к генерации	Правдоподобность	Покрытие	Фактологическая согласованность	LLM-оценка
Qwen3 4B	П1	0,5781	0,7241	0,643	0,6341
	П2	0,6257	0,7439	0,7082	0,6762
	П3	0,6827	0,8076	0,7371	0,7249
Qwen33 8B	П1	0,6562	0,7684	0,7014	0,6942
	П2	0,6897	0,8109	0,7561	0,7352
	П3	0,7473	0,8376	0,8024	0,7695
Qwen3 14B	П1	0,7052	0,8017	0,7361	0,7448
	П2	0,7353	0,8247	0,7805	0,7781
	П3	0,7597	0,8439	0,8193	0,7912
Qwen3 32B	П1	0,7398	0,8236	0,8053	0,7787
	П2	0,7578	0,8462	0,8127	0,8047
	П3	0,7719	0,8618	0,8395	0,8251
Llama 3.1 8B	П1	0,6575	0,7738	0,7427	0,7014
	П2	0,6911	0,8127	0,7673	0,7412
	П3	0,7494	0,8406	0,8115	0,7715
Llama 3.1 70B	П1	0,7764	0,8667	0,8317	0,8293
	П2	0,7892	0,8882	0,8646	0,8421
	П3	0,8257	0,8947	0,8832	0,8632
Llama 3.1 405B	П1	0,7986	0,8883	0,8712	0,8497
	П2	0,8251	0,8995	0,8918	0,8583
	П3	0,8443	0,9136	0,9259	0,8879

Полученные результаты показывают, что для всех исследуемых моделей наибольшие значения рассматриваемых критериев были получены при использовании предварительно отобранных цитатных фрагментов (П3). Использование всех извлеченных цитатных фрагментов (П2) также обеспечивает более высокие показатели по сравнению с генерацией на основе полных текстов публикаций (П1). Таким образом, для всех моделей и критериев качества сохраняется единообразное соотношение П3 > П2 > П1.

Наиболее выраженные относительные различия между подходами П1 и П3 наблюдаются для моделей меньшего масштаба. В частности, для Qwen3 4B использование предварительно отобранных цитатных фрагментов сопровождалось увеличением правдоподобности на 18,09%, покрытия — на 11,53%, фактологической согласованности — на 14,63%, а LLM-оценки — на 14,32%. С увеличением масштаба моделей относительные различия уменьшаются, однако положительная динамика сохраняется по всем критериям: для Qwen3 32B прирост находится в диапазоне 4,25–5,96%, а для Llama 3.1 405B — 2,85-6,28%.

В ряде случаев предварительный отбор цитатных фрагментов позволяет моделям меньшего масштаба достигать значений, сопоставимых с результатами более крупных моделей, использующих полные тексты публикаций. Так, показатели Qwen3 14B при подходе П3 оказались выше показателей Qwen3 32B при П1 по всем рассматриваемым критериям на 1,60–2,69%. Аналогичное соотношение наблюдается для Llama 3.1 70B при П3 и Llama 3.1 405B при П1, где различия составляют 0,72–3,39%. Максимальные абсолютные значения всех критериев получены для модели Llama 3.1 405B при подходе П3.

Также было проведено сравнение П3 с двумя альтернативами: случайным отбором и BM25. Результаты сравнения приведены в табл. 3.

Таблица 3 - Медианные значения критериев качества генерации обзоров при сравнении с альтернативными подходами

DOI: https://doi.org/10.60797/IRJ.2026.171.79.4

Модель	Подход	Правдоподобность	Покрытие	Фактологическая Согласованность	ЛЛМ-оценка
Qwen3 4B	П3	0,6827	0,8076	0,7371	0, 7249
	Случайный отбор	0,6352	0,7513	0,6925	0,6819
	BM-25	0,6534	0,7701	0,7114	0,6947
Qwen3 8B	П3	0,7473	0,8376	0,8024	0,7695
	Случайный отбор	0,6716	0,7936	0,7524	0,7278
	BM-25	0,7134	0,8214	0,7747	0,7521
Qwen3 14B	П3	0,7597	0,8439	0,8193	0,7912
	Случайный отбор	0,7398	0,8251	0,7917	0,7793
	BM-25	0,7469	0,8312	0,7934	0,7845
Qwen3 32B	П3	0,7719	0,8618	0,8395	0,8251
	Случайный отбор	0,7619	0,8479	0,8145	0,8023
	BM-25	0,7654	0,8534	0,8257	0,8147
Llama 3.1 8B	П3	0,7494	0,8406	0,8115	0,7715
	Случайный отбор	0,7067	0,8234	0,7793	0,7564
	BM-25	0,7273	0,8241	0,7867	0,7566
Llama 3.1 70B	П3	0,8257	0,8947	0,8832	0,8632
	Случайный отбор	0,7982	0,8892	0,8693	0,8473
	BM-25	0,8072	0,8907	0,8732	0,8511
Llama 3.1 405B	П3	0,8443	0,9136	0,9259	0,8879
	Случайный отбор	0,8291	0,9008	0,9097	0,8654
	BM-25	0,8355	0,9013	0,9094	0,8723

Для всех исследуемых моделей и критериев качества наблюдается единообразное соотношение П3 > BM25 > случайный отбор. В среднем по исследуемым моделям преимущество П3 относительно случайного отбора составляет 0,0341 по правдоподобности, 0,0241 по покрытию, 0,0299 по фактологической согласованности и 0,0247 по LLM-оценке. По сравнению с BM25 соответствующие различия составляют 0,0188, 0,0154, 0,0206 и 0,0153. Таким образом, полученные результаты показывают, что итоговое качество зависит не только от сокращения объема цитатного материала, но и от способа его отбора: релевантностное ранжирование BM25 обеспечивает более высокие показатели по сравнению со случайным отбором, тогда как предложенная многокомпонентная оценка позволяет дополнительно повысить качество формируемых текстов.

Для проверки устойчивости выявленных различий выполнен парный статистический анализ результатов по 114 тематикам датасета SurGE. Для каждого сочетания модели и критерия рассчитывались парные различия между сравниваемыми подходами.

При сравнении П3 с генерацией на основе полных текстов (П1) статистически значимые различия выявлены в 27 из 28 сочетаний моделей и критериев. Характерные величины парных различий составляли Δ=0,04–0,07 при 95%-ном ДИ выше нуля (например, Δ=0,051, 95% ДИ [0,037; 0,062], p_{adj} <0,001). Для наиболее крупных моделей различия снижались до Δ=0,01–0,03, сохраняя статистическую значимость в большинстве случаев.

При сравнении П3 с полным набором цитатных фрагментов (П2) статистически значимое преимущество выявлено в 24 из 28 сочетаний моделей и критериев, наиболее устойчиво — по правдоподобности и фактологической согласованности для моделей меньшего масштаба. Например, по фактологической согласованности для одной из малых моделей получено Δ=0,032, 95% ДИ [0,017; 0,042], p_{adj} <0,001. Для отдельных крупных моделей, преимущественно по покрытию, различия не достигали статистической значимости (например, Δ=0,007, 95% ДИ [−0,002; 0,013], p_{adj} =0,14).

Относительно случайного отбора статистически значимое преимущество П3 выявлено во всех 28 сочетаниях моделей и критериев, в большинстве случаев при p_{adj} <0,01; величина различий преимущественно составляла Δ=0,03–0,05 для моделей меньшего масштаба и снижалась для наиболее крупных. При сравнении с BM25 значимые различия получены в 25 из 28 случаев; характерные значения составляли Δ=0,015–0,025 (например, Δ=0,023, 95% ДИ [0,011; 0,028], p_{adj} =0,002).

Межмодельное сравнение Qwen3 14B с П3 и Qwen3 32B с П1, а также Llama 3.1 70B с П3 и Llama 3.1 405B с П1 не выявило статистически значимых различий по трем из четырех критериев в каждой паре. С учетом близких агрегированных значений показателей полученные результаты свидетельствуют о сопоставимости качества в большинстве рассмотренных критериев и указывают на возможность частичной компенсации меньшего масштаба модели посредством предварительного отбора цитатных фрагментов.

Анализ ресурсоэффективности показал, что применение П3 сократило суммарный токенный трафик на 87,1% относительно П1 и время генеративных этапов на 17,6–27,3% в зависимости от масштаба модели. Предварительная обработка при применении П3 дополнительно занимала около 8–9 мин на тематику, однако ее результаты могут повторно использоваться при последующих генерациях без дополнительных вычислений. Сокращение токенного трафика уменьшает и тарифицируемый объем обработки при использовании коммерческих API: стоимость 1 млн входных/выходных токенов у рассматриваемых провайдеров составляет, например, 0,035/0,138 долл. США для Qwen3 8B, 0,10/0,45 для Qwen3 32B и 0,34/0,39 для Llama 3.1 70B, по данным провайдеров, таких как Novita AI и Fireworks. Таким образом, предложенная методика требует дополнительных затрат на предварительную обработку, однако существенно сокращает объем входных данных, время выполнения генерации и тарифицируемый объем обработки при использовании коммерческих API. Аппаратные затраты в рамках эксперимента не измерялись, однако результаты существующих исследований показывают снижение энергопотребления и требований к вычислительным ресурсам при использовании моделей меньшего масштаба [39], что позволяет предполагать возможность дополнительного ресурсного преимущества при использовании таких моделей.

Обсуждение

4.1. Основные результаты исследования

Для всех исследованных моделей предварительная оценка и отбор цитатных фрагментов сопровождались повышением качества генерации обзорных текстов. Это согласуется с результатами работ, показывающих зависимость качества RAG-генерации от состава и релевантности входного контекста. В работе [11] отмечено негативное влияние нерелевантной информации, в [12] рассматривается предварительный отбор извлекаемых данных, а в [13] — формирование взаимодополняющего набора исходных документов. Аналогичные выводы приводятся и в работе [40], где повышение качества генерации рассматривается как следствие совершенствования механизмов поиска, ранжирования и отбора информации в генеративных системах. При этом преимущество П3 над случайным отбором и BM25 показывает, что в рассматриваемой задаче значение имеет не только сокращение объема входных данных, но и способ формирования цитатного представления.

Преимущество предварительного отбора наиболее выражено для моделей меньшего масштаба, которые, вероятно, испытывают большие сложности при фильтрации избыточной информации в длинном контексте. Для крупных моделей эффект сохраняется, но выражен слабее, что может быть связано с их более высокими возможностями обработки входной информации. Сокращение и структурирование контекста рассматривается как способ повышения эффективности генерации в работах [15] и [16]. В настоящем исследовании сокращение достигается не за счет общей компрессии текста, а посредством оценки и отбора цитатных фрагментов с учетом их соответствия содержанию цитируемых публикаций. Сходный вывод представлен и в работе [41], где показано, что качество итоговой генерации в значительной степени определяется эффективностью этапа извлечения и отбора информации, а не только характеристиками языковой модели.

Результаты межмодельного сравнения показывают, что предварительный отбор может частично компенсировать меньший масштаб модели. Для рассмотренных пар меньшие модели с П3 достигали результатов, близких к более крупным моделям с П1, причем по трем из четырех критериев статистически значимые различия отсутствовали. Это не означает, что меньшие модели могут во всех случаях заменить более крупные, однако показывает возможность более рационального выбора модели за счет предварительной подготовки входных данных. Возможность повышения эффективности моделей меньшего масштаба за счет использования внешней информации отмечается также в работе [29], однако в настоящем исследовании данный эффект рассматривается применительно к цитатно-осведомленной генерации научных обзорных текстов.

4.2. Анализ ошибок отбора цитатных фрагментов

Ручной анализ ошибок отбора показал, что при точности 86,4% и полноте 83,1% основная часть релевантных цитатных фрагментов сохраняется корректно. Среди корректно исключенных встречались фрагменты, сохраняющие высокую тематическую близость с исходной публикацией, но искажающие отдельные ее положения. Так, в работе [42] при описании метода из исследования [43] указывают, что описываемая регуляризация «encourage the annotator confusion matrix to be close to an identity matrix», хотя в исходной работе объект и роль регуляризации определены иначе. Несмотря на высокий BERTScore (0,871), низкие значения QA (0,321) и NLI (0,258) снизили итоговую оценку до S=0,379, вследствие чего фрагмент был исключен. Аналогично было отклонено утверждение из работы [44] о том, что реальный шум разметки «always comes from an open rather than a finite category set»: исходная работа [45] рассматривает такой шум для ряда практических сценариев, но не устанавливает универсального положения; итоговая оценка составила S=0,531.

Ложноположительные решения преимущественно возникали в случаях, когда отдельные компоненты утверждения присутствовали в исходной публикации, но изменялись отношения между ними или степень категоричности. Например, в работе [46] при описании метода из [47] распространяют свойство несмещенности функции потерь, установленное для конкретного варианта коррекции, на описываемый подход в целом. Высокие значения отдельных метрик привели к ошибочному отбору фрагмента с S=0,791. Сходная ситуация наблюдалась в цитате [48], где критерий малой величины функции потерь из работы [49] интерпретировался как непосредственное отнесение соответствующих примеров к чистым; фрагмент был сохранен с S=0,695.

Для ложноотрицательных решений характерна обратная ситуация: утверждение соответствует исходной работе, однако его подтверждение выражено неявно и требует установления связи между несколькими положениями. Так например, цитаты [50] и [51] корректно передавали особенности описываемых в работе [52] методов, но соответствующие положения следовали из способа формирования целевой разметки и функции потерь, а не были явно сформулированы в исходных работах. Итоговые оценки S=0,599 и S=0,602 соответственно оказались ниже порога отбора. Выраженной самостоятельной зависимости результата от длины цитатного фрагмента выявлено не было. Ошибки чаще возникали для сложных утверждений, объединяющих несколько смысловых компонентов, при изменении степени категоричности или необходимости неявного логического вывода. Учет этих особенностей представляет одно из направлений дальнейшего развития методики.

4.3. Ограничения и практическая применимость

Исследование выполнено на одном наборе данных SurGE, включающем англоязычные научные публикации. Объем выборки и использование двух семейств моделей позволяют оценить устойчивость результатов в рамках рассматриваемой задачи, однако их воспроизводимость на других корпусах и языках требует проверки. Порог отбора определен на предварительной выборке и для иной предметной или языковой области может требовать дополнительной калибровки. Дальнейшие исследования могут быть направлены на проверку методики на независимых наборах данных, ее адаптацию к корпусам на других языках и анализ влияния предметной области на результаты отбора.

Практическое значение полученных результатов связано с возможностью одновременно повысить качество генерации и сократить объем данных, обрабатываемых LLM. Применение П3 обеспечило сокращение токенного трафика на 87,1% и времени выполнения генеративных этапов на 17,6–27,3% относительно обработки полных текстов. Предварительная обработка требует дополнительных временных затрат, однако ее результаты могут повторно использоваться при последующих генерациях. Поскольку этот этап не зависит от конкретной генеративной модели, предложенная методика может сочетаться с другими способами повышения вычислительной эффективности.

Заключение

В работе предложена методика предварительной оценки и отбора цитатных фрагментов для цитатно-осведомленной генерации научных саммари и обзорных текстов. Методика предусматривает оценку соответствия цитатных фрагментов содержанию цитируемых публикаций с применением совокупности лексических, семантических, фактологических и логических критериев и формирование на этой основе компактного входного представления научной информации. Экспериментальное исследование проведено на 114 тематиках датасета SurGE с использованием семи моделей различного масштаба семейств Qwen3 и Llama 3.1.

Результаты показали, что способ формирования входных данных существенно влияет на качество цитатно-осведомленной генерации. Во всех экспериментальных конфигурациях использование предварительно отобранных цитатных фрагментов обеспечивало более высокие показатели по сравнению с использованием всех извлеченных цитат и полных текстов публикаций. Наиболее выраженный эффект наблюдался для моделей меньшего масштаба: для Qwen3 4B прирост относительно генерации по полным текстам составил 11,53-18,09%, тогда как для Qwen3 32B и Llama 3.1 405B — 4,25–5,96% и 2,85–6,28% соответственно. Преимущество предложенного способа отбора также подтверждено при сравнении со случайным отбором и ранжированием BM25 и в большинстве рассмотренных случаев являлось статистически значимым.

Межмодельное сравнение показало, что предварительный отбор цитатных фрагментов позволяет в отдельных случаях частично компенсировать меньший масштаб языковой модели: меньшие модели достигали качества, сопоставимого с более крупными моделями того же семейства при обработке полных текстов. Одновременно применение предложенного подхода обеспечило сокращение суммарного токенного трафика на 87,1% и времени выполнения генеративных этапов на 17,6–27,3%. При этом методика требует дополнительной предварительной обработки данных, однако ее результаты могут повторно использоваться при последующих генерациях без повторного выполнения данного этапа.

Полученные результаты показывают, что повышение качества и ресурсоэффективности цитатно-осведомленной генерации может достигаться не только за счет увеличения масштаба языковой модели, но и за счет более рационального формирования входной научной информации. Практическое значение предложенного подхода связано с возможностью сокращения объема данных, обрабатываемых LLM, и применения моделей меньшего масштаба при сохранении сопоставимого качества генерации. Это расширяет возможности построения ресурсоэффективных систем автоматизированного формирования научных обзоров без изменения архитектуры используемых языковых моделей.

Дополнительные материалы

Дополнительные материалы доступны на онлайн-странице статьи.

Supplementary materials

Supplementary materials are available online on the article's webpage.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Han B. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview / B. Han, T. Susnjak, A. Mathrani // Applied Sciences. — 2024. — 19. — DOI: 10.3390/app14199103

2. Bolaños F. Artificial Intelligence for Literature Reviews: Opportunities and Challenges / F. Bolaños, A. Salatino, F. Osborne et al. // Artificial Intelligence Review. — 2024. — 10. — DOI: 10.1007/s10462-024-10902-3

3. Bao T. SurveyGen: Quality-Aware Scientific Survey Generation with Large Language Models / T. Bao, M.T. Nayeem, D. Rafiei et al. // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; — Suzhou: Association for Computational Linguistics, 2025. — P. 2712–2736. doi: 10.18653/v1/2025.emnlp-main.136

4. Chen J. SurveyGen-I: Consistent Scientific Survey Generation with Evolving Plans and Memory-Guided Writing / J. Chen, Y. Shen, J. Liu et al. // Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics; — Mumbai: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. — P. 3687–3714. doi: 10.18653/v1/2025.ijcnlp-long.193

5. Alizadeh K. LLM in a Flash: Efficient Large Language Model Inference with Limited Memory / K. Alizadeh, S.I. Mirzadeh, D. Belenko et al. // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Bangkok: Association for Computational Linguistics, 2024. — P. 12562–12584. doi: 10.18653/v1/2024.acl-long.678

6. Yan X. SURVEYFORGE: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing / X. Yan, S. Feng, J. Yuan et al. // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Vienna: Association for Computational Linguistics, 2025. — P. 12444–12465. doi: 10.18653/v1/2025.acl-long.609

7. Zhu X. A Survey on Model Compression for Large Language Models / X. Zhu, J. Li, Y. Liu et al. // Transactions of the Association for Computational Linguistics. — 2024. — 12. — P. 1556–1577. — DOI: 10.1162/tacl_a_00704

8. Chavan A. Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward / A. Chavan, R. Magazine, S. Kushwaha et al. // Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; — Jeju: International Joint Conferences on Artificial Intelligence Organization, 2024. — P. 7980–7988.

9. Li Z. Prompt Compression for Large Language Models: A Survey / Z. Li, Y. Liu, Y. Su et al. // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); — Albuquerque: Association for Computational Linguistics, 2025. — P. 7182–7195. doi: 10.18653/v1/2025.naacl-long.368

10. Kim K. RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation / K. Kim, J.-Y. Lee. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; — Miami: Association for Computational Linguistics, 2024. — P. 22149–22161. doi: 10.18653/v1/2024.emnlp-main.1236

11. Wang Y. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering / Y. Wang, R. Ren, J. Li et al. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; — Miami: Association for Computational Linguistics, 2024. — P. 5613–5626. doi: 10.18653/v1/2024.emnlp-main.321

12. Li X. How Does Knowledge Selection Help Retrieval-Augmented Generation? / X. Li, J. Ouyang. // Findings of the Association for Computational Linguistics: EMNLP 2025; — Suzhou: Association for Computational Linguistics, 2025. — P. 4104–4121. doi: 10.18653/v1/2025.findings-emnlp.218

13. Lee D. Shifting from Ranking to Set Selection for Retrieval-Augmented Generation / D. Lee, Y. Jo, H. Park et al. // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Vienna: Association for Computational Linguistics, 2025. — P. 17606–17619. doi: 10.18653/v1/2025.acl-long.861

14. Verma S. Contextual Compression in Retrieval-Augmented Generation for Large Language Models: A Survey / S. Verma // arXiv. — 2024. — URL: https://arxiv.org/abs/2409.13385. (дата обращения: 21.07.26)

15. Jiang H. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression / H. Jiang, Q. Wu, X. Luo et al. // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Bangkok: Association for Computational Linguistics, 2024. — P. 1658–1677. doi: 10.18653/v1/2024.acl-long.91

16. Jiang H. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models / H. Jiang, Q. Wu, C.-Y. Lin et al. // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; — Singapore: Association for Computational Linguistics, 2023. — P. 13358–13376. doi: 10.18653/v1/2023.emnlp-main.825

17. Liu N.F. Lost in the Middle: How Language Models Use Long Contexts / N.F. Liu, K. Lin, J. Hewitt et al. // Transactions of the Association for Computational Linguistics. — 2024. — 12. — P. 157–173. — DOI: 10.1162/tacl_a_00638

18. Roy S.S. Enhancing Scientific Document Summarization with Research Community Perspective and Background Knowledge / S.S. Roy, R.E. Mercer. // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); — Torino: ELRA and ICCL, 2024. — P. 6048–6058.

19. Cohan A. Scientific Document Summarization via Citation Contextualization and Scientific Discourse / A. Cohan, N. Goharian // International Journal on Digital Libraries. — 2018. — 19. — P. 287–303. — DOI: 10.1007/s00799-017-0216-8

20. Syed S. Citance-Contextualized Summarization of Scientific Papers / S. Syed, A. Hakimi, K. Al-Khatib et al. // Findings of the Association for Computational Linguistics: EMNLP 2023; — Singapore: Association for Computational Linguistics, 2023. — P. 8551–8568. doi: 10.18653/v1/2023.findings-emnlp.573

21. Kasanishi T. SciReviewGen: A Large-Scale Dataset for Automatic Literature Review Generation / T. Kasanishi, M. Isonuma, S. Mori et al. // Findings of the Association for Computational Linguistics: ACL 2023; — Toronto: Association for Computational Linguistics, 2023. — P. 6695–6715. doi: 10.18653/v1/2023.findings-acl.418

22. Ding H. SciRAG: Adaptive, Citation-Aware, and Outline-Guided Retrieval and Synthesis for Scientific Literature / H. Ding, Y. Zhang, T. Hu et al. // Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); — Rabat: Association for Computational Linguistics, 2026. — P. 6440–6460. doi: 10.18653/v1/2026.eacl-long.303

23. Lauscher A. MultiCite: Modeling Realistic Citations Requires Moving Beyond the Single-Sentence Single-Label Setting / A. Lauscher, A. Cohan, B. Kuehl et al. // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — Seattle: Association for Computational Linguistics, 2022. — P. 1875–1889.

24. Cohan A. Structural Scaffolds for Citation Intent Classification in Scientific Publications / A. Cohan, W. Ammar, M. van Zuylen et al. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); — Minneapolis: Association for Computational Linguistics, 2019. — P. 3586–3596.

25. Wadden D. Fact or Fiction: Verifying Scientific Claims / D. Wadden, S. Lin, K. Lo et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); — Online: Association for Computational Linguistics, 2020. — P. 7534–7550. doi: 10.18653/v1/2020.emnlp-main.609

26. Sarol M.J. Assessing Citation Integrity in Biomedical Publications: Corpus Annotation and NLP Models / M.J. Sarol, S. Ming, S. Radhakrishna et al. // Bioinformatics. — 2024. — 7. — DOI: 10.1093/bioinformatics/btae420

27. Jergas H. Quotation Accuracy in Medical Journal Articles—A Systematic Review and Meta-analysis / H. Jergas, C. Baethge // PeerJ. — 2015. — 3. — DOI: 10.7717/peerj.1364

28. Vladika J. On the Influence of Context Size and Model Choice in Retrieval-Augmented Generation Systems / J. Vladika, F. Matthes. // Findings of the Association for Computational Linguistics: NAACL 2025; — Albuquerque: Association for Computational Linguistics, 2025. — P. 6739–6751. doi: 10.18653/v1/2025.findings-naacl.375

29. Xu B. Evaluating Small Language Models for News Summarization: Implications and Factors Influencing Performance / B. Xu, Y. Chen, Z. Wen et al. // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); — Albuquerque: Association for Computational Linguistics, 2025. — P. 4909–4922. doi: 10.18653/v1/2025.naacl-long.253

30. Liu S. Can Small Language Models with Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science? / S. Liu, Z. Yu, F. Huang et al. // Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1; — Milan: Association for Computing Machinery (ACM), 2024. — P. 388–393. doi: 10.1145/3649217.3653554

31. Кузнецов И.И. Математическая модель процесса автоматизированного построения обзорных текстов, основанного на использовании цитатно-осведомленной суммаризации / И.И. Кузнецов // Моделирование, оптимизация и информационные технологии. — 2026. — 6. — DOI: 10.26102/2310-6018/2026.57.6.020

32. Beltagy I. SciBERT: A Pretrained Language Model for Scientific Text / I. Beltagy, K. Lo, A. Cohan. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); — Hong Kong: Association for Computational Linguistics, 2019. — P. 3615–3620. doi: 10.18653/v1/d19-1371

33. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries / C.-Y. Lin. // Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; — Barcelona: Association for Computational Linguistics, 2004. — P. 74–81.

34. Zhang T. BERTScore: Evaluating Text Generation with BERT / T. Zhang, V. Kishore, F. Wu et al. // arXiv. — 2019. — URL: https://arxiv.org/abs/1904.09675. (дата обращения: 17.07.26)

35. Deutsch D. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary / D. Deutsch, T. Bedrax-Weiss, D. Roth // Transactions of the Association for Computational Linguistics. — 2021. — 9. — P. 774–789. — DOI: 10.1162/tacl_a_00397

36. Chen Y. MENLI: Robust Evaluation Metrics from Natural Language Inference / Y. Chen, S. Eger // Transactions of the Association for Computational Linguistics. — 2023. — 11. — P. 804–825. — DOI: 10.1162/tacl_a_00576

37. Fabbri A.R. SummEval: Re-evaluating Summarization Evaluation / A.R. Fabbri, W. Kryściński, B. McCann et al. // Transactions of the Association for Computational Linguistics. — 2021. — 9. — P. 391–409. — DOI: 10.1162/tacl_a_00373

38. Su W. SurGE: A Benchmark and Evaluation Framework for Scientific Survey Generation / W. Su, A. Xie, Q. Ai et al. // Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2026); — Seoul: Association for Computing Machinery (ACM), 2026. doi: 10.1145/3805712.3808598

39. Sánchez-Mompó A. Green MLOps to Green GenOps: An Empirical Study of Energy Consumption in Discriminative and Generative AI Operations / A. Sánchez-Mompó, I. Mavromatis, P. Li et al. // Information. — 2025. — 4. — DOI: 10.3390/info16040281

40. Asai A. Reliable, Adaptable, and Attributable Language Models with Retrieval / A. Asai, Z. Zhong, D. Chen et al. // arXiv. — 2024. — URL: https://arxiv.org/abs/2403.03187. (дата обращения: 16.07.26)

41. Cao D. Retrieval Is Accurate Generation / D. Cao, D. Cai, L. Cui et al. // arXiv. — 2024. — URL: https://arxiv.org/abs/2402.17532. (дата обращения: 20.07.26)

42. Chu Z. Learning from Crowds by Modeling Common Confusions / Z. Chu, J. Ma, H. Wang // Proceedings of the AAAI Conference on Artificial Intelligence. — 2021. — Vol. 35, No. 7. — P. 5832–5840. — DOI: 10.1609/aaai.v35i7.16730

43. Tanno R. Learning from Noisy Labels by Regularized Estimation of Annotator Confusion / R. Tanno, A. Saeedi, S. Sankaranarayanan et al. // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2019. — P. 11236–11245. doi: 10.1109/CVPR.2019.01150

44. Wu T. NoisywikiHow: A Benchmark for Learning with Real-world Noisy Labels in Natural Language Processing / T. Wu, X. Ding, M. Tang et al. // Findings of the Association for Computational Linguistics: ACL 2023; — Toronto: Association for Computational Linguistics, 2023. — P. 4856–4873. doi: 10.18653/v1/2023.findings-acl.299

45. Wang Y. Iterative Learning with Open-Set Noisy Labels / Y. Wang, W. Liu, X. Ma et al. // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2018. — P. 8688–8696. doi: 10.1109/CVPR.2018.00906

46. Zheng S. Error-Bounded Correction of Noisy Labels / S. Zheng, P. Wu, A. Goswami et al. // Proceedings of the 37th International Conference on Machine Learning; — Online: PMLR, 2020. — P. 11447–11457.

47. Patrini G. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach / G. Patrini, A. Rozza, A.K. Menon et al. // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); — Piscataway: IEEE, 2017. — P. 2233–2241. doi: 10.1109/CVPR.2017.240

48. Murtadha A. Rank-Aware Negative Training for Semi-Supervised Text Classification / A. Murtadha, S. Pan, W. Bo et al. // Transactions of the Association for Computational Linguistics. — 2023. — 11. — DOI: 10.1162/tacl_a_00574

49. Shen Y. Learning with Bad Training Data via Iterative Trimmed Loss Minimization / Y. Shen, S. Sanghavi. // Proceedings of the 36th International Conference on Machine Learning; — Online: PMLR, 2019. — P. 5739–5748.

50. Arazo E. Unsupervised Label Noise Modeling and Loss Correction / E. Arazo, D. Ortego, P. Albert et al. // Proceedings of the 36th International Conference on Machine Learning; — Online: PMLR, 2019. — P. 312–321.

51. Tanaka D. Joint Optimization Framework for Learning with Noisy Labels / D. Tanaka, D. Ikami, T. Yamasaki et al. // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2018. — P. 5552–5560. doi: 10.1109/CVPR.2018.00582

52. Reed S.E. Training Deep Neural Networks on Noisy Labels with Bootstrapping / S.E. Reed, H. Lee, D. Anguelov et al. // arXiv. — 2015. — URL: https://arxiv.org/abs/1412.6596. (дата обращения: 14.08.26)

Список литературы на английском языке / References in English

1. Han B. Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview / B. Han, T. Susnjak, A. Mathrani // Applied Sciences. — 2024. — 19. — DOI: 10.3390/app14199103

2. Bolaños F. Artificial Intelligence for Literature Reviews: Opportunities and Challenges / F. Bolaños, A. Salatino, F. Osborne et al. // Artificial Intelligence Review. — 2024. — 10. — DOI: 10.1007/s10462-024-10902-3

3. Bao T. SurveyGen: Quality-Aware Scientific Survey Generation with Large Language Models / T. Bao, M.T. Nayeem, D. Rafiei et al. // Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing; — Suzhou: Association for Computational Linguistics, 2025. — P. 2712–2736. doi: 10.18653/v1/2025.emnlp-main.136

4. Chen J. SurveyGen-I: Consistent Scientific Survey Generation with Evolving Plans and Memory-Guided Writing / J. Chen, Y. Shen, J. Liu et al. // Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics; — Mumbai: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. — P. 3687–3714. doi: 10.18653/v1/2025.ijcnlp-long.193

5. Alizadeh K. LLM in a Flash: Efficient Large Language Model Inference with Limited Memory / K. Alizadeh, S.I. Mirzadeh, D. Belenko et al. // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics

(Volume 1: Long Papers); — Bangkok: Association for Computational Linguistics, 2024. — P. 12562–12584. doi: 10.18653/v1/2024.acl-long.678

6. Yan X. SURVEYFORGE: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing / X. Yan, S. Feng, J. Yuan et al. // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Vienna: Association for Computational Linguistics, 2025. — P. 12444–12465. doi: 10.18653/v1/2025.acl-long.609

7. Zhu X. A Survey on Model Compression for Large Language Models / X. Zhu, J. Li, Y. Liu et al. // Transactions of the Association for Computational Linguistics. — 2024. — 12. — P. 1556–1577. — DOI: 10.1162/tacl_a_00704

8. Chavan A. Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward / A. Chavan, R. Magazine, S. Kushwaha et al. // Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence; — Jeju: International Joint Conferences on Artificial Intelligence Organization, 2024. — P. 7980–7988.

9. Li Z. Prompt Compression for Large Language Models: A Survey / Z. Li, Y. Liu, Y. Su et al. // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); — Albuquerque: Association for Computational Linguistics, 2025. — P. 7182–7195. doi: 10.18653/v1/2025.naacl-long.368

10. Kim K. RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation / K. Kim, J.-Y. Lee. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; — Miami: Association for Computational Linguistics, 2024. — P. 22149–22161. doi: 10.18653/v1/2024.emnlp-main.1236

11. Wang Y. REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering / Y. Wang, R. Ren, J. Li et al. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; — Miami: Association for Computational Linguistics, 2024. — P. 5613–5626. doi: 10.18653/v1/2024.emnlp-main.321

12. Li X. How Does Knowledge Selection Help Retrieval-Augmented Generation? / X. Li, J. Ouyang. // Findings of the Association for Computational Linguistics: EMNLP 2025; — Suzhou: Association for Computational Linguistics, 2025. — P. 4104–4121. doi: 10.18653/v1/2025.findings-emnlp.218

13. Lee D. Shifting from Ranking to Set Selection for Retrieval-Augmented Generation / D. Lee, Y. Jo, H. Park et al. // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Vienna: Association for Computational Linguistics, 2025. — P. 17606–17619. doi: 10.18653/v1/2025.acl-long.861

14. Verma S. Contextual Compression in Retrieval-Augmented Generation for Large Language Models: A Survey / S. Verma // arXiv. — 2024. — URL: https://arxiv.org/abs/2409.13385. (accessed: 21.07.26)

15. Jiang H. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression / H. Jiang, Q. Wu, X. Luo et al. // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); — Bangkok: Association for Computational Linguistics, 2024. — P. 1658–1677. doi: 10.18653/v1/2024.acl-long.91

16. Jiang H. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models / H. Jiang, Q. Wu, C.-Y. Lin et al. // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; — Singapore: Association for Computational Linguistics, 2023. — P. 13358–13376. doi: 10.18653/v1/2023.emnlp-main.825

17. Liu N.F. Lost in the Middle: How Language Models Use Long Contexts / N.F. Liu, K. Lin, J. Hewitt et al. // Transactions of the Association for Computational Linguistics. — 2024. — 12. — P. 157–173. — DOI: 10.1162/tacl_a_00638

18. Roy S.S. Enhancing Scientific Document Summarization with Research Community Perspective and Background Knowledge / S.S. Roy, R.E. Mercer. // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024); — Torino: ELRA and ICCL, 2024. — P. 6048–6058.

19. Cohan A. Scientific Document Summarization via Citation Contextualization and Scientific Discourse / A. Cohan, N. Goharian // International Journal on Digital Libraries. — 2018. — 19. — P. 287–303. — DOI: 10.1007/s00799-017-0216-8

20. Syed S. Citance-Contextualized Summarization of Scientific Papers / S. Syed, A. Hakimi, K. Al-Khatib et al. // Findings of the Association for Computational Linguistics: EMNLP 2023; — Singapore: Association for Computational Linguistics, 2023. — P. 8551–8568. doi: 10.18653/v1/2023.findings-emnlp.573

21. Kasanishi T. SciReviewGen: A Large-Scale Dataset for Automatic Literature Review Generation / T. Kasanishi, M. Isonuma, S. Mori et al. // Findings of the Association for Computational Linguistics: ACL 2023; — Toronto: Association for Computational Linguistics, 2023. — P. 6695–6715. doi: 10.18653/v1/2023.findings-acl.418

22. Ding H. SciRAG: Adaptive, Citation-Aware, and Outline-Guided Retrieval and Synthesis for Scientific Literature / H. Ding, Y. Zhang, T. Hu et al. // Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers); — Rabat: Association for Computational Linguistics, 2026. — P. 6440–6460. doi: 10.18653/v1/2026.eacl-long.303

23. Lauscher A. MultiCite: Modeling Realistic Citations Requires Moving Beyond the Single-Sentence Single-Label Setting / A. Lauscher, A. Cohan, B. Kuehl et al. // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — Seattle: Association for Computational Linguistics, 2022. — P. 1875–1889.

24. Cohan A. Structural Scaffolds for Citation Intent Classification in Scientific Publications / A. Cohan, W. Ammar, M. van Zuylen et al. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); — Minneapolis: Association for Computational Linguistics, 2019. — P. 3586–3596.

25. Wadden D. Fact or Fiction: Verifying Scientific Claims / D. Wadden, S. Lin, K. Lo et al. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); — Online: Association for Computational Linguistics, 2020. — P. 7534–7550. doi: 10.18653/v1/2020.emnlp-main.609

26. Sarol M.J. Assessing Citation Integrity in Biomedical Publications: Corpus Annotation and NLP Models / M.J. Sarol, S. Ming, S. Radhakrishna et al. // Bioinformatics. — 2024. — 7. — DOI: 10.1093/bioinformatics/btae420

27. Jergas H. Quotation Accuracy in Medical Journal Articles—A Systematic Review and Meta-analysis / H. Jergas, C. Baethge // PeerJ. — 2015. — 3. — DOI: 10.7717/peerj.1364

28. Vladika J. On the Influence of Context Size and Model Choice in Retrieval-Augmented Generation Systems / J. Vladika, F. Matthes. // Findings of the Association for Computational Linguistics: NAACL 2025; — Albuquerque: Association for Computational Linguistics, 2025. — P. 6739–6751. doi: 10.18653/v1/2025.findings-naacl.375

29. Xu B. Evaluating Small Language Models for News Summarization: Implications and Factors Influencing Performance / B. Xu, Y. Chen, Z. Wen et al. // Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); — Albuquerque: Association for Computational Linguistics, 2025. — P. 4909–4922. doi: 10.18653/v1/2025.naacl-long.253

30. Liu S. Can Small Language Models with Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science? / S. Liu, Z. Yu, F. Huang et al. // Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1; — Milan: Association for Computing Machinery (ACM), 2024. — P. 388–393. doi: 10.1145/3649217.3653554

31. Kuzneczov I.I. Matematicheskaya model' processa avtomatizirovannogo postroeniya obzorny'x tekstov, osnovannogo na ispol'zovanii citatno-osvedomlennoj summarizacii [A mathematical model of the process of automated construction of review texts based on the use of quotation-aware summarization] / I.I. Kuzneczov // Modeling, Optimization and Information Technology. — 2026. — 6. — DOI: 10.26102/2310-6018/2026.57.6.020 [in Russian]

32. Beltagy I. SciBERT: A Pretrained Language Model for Scientific Text / I. Beltagy, K. Lo, A. Cohan. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); — Hong Kong: Association for Computational Linguistics, 2019. — P. 3615–3620. doi: 10.18653/v1/d19-1371

33. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries / C.-Y. Lin. // Text Summarization Branches Out: Proceedings of the ACL-04 Workshop; — Barcelona: Association for Computational Linguistics, 2004. — P. 74–81.

34. Zhang T. BERTScore: Evaluating Text Generation with BERT / T. Zhang, V. Kishore, F. Wu et al. // arXiv. — 2019. — URL: https://arxiv.org/abs/1904.09675. (accessed: 17.07.26)

35. Deutsch D. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary / D. Deutsch, T. Bedrax-Weiss, D. Roth // Transactions of the Association for Computational Linguistics. — 2021. — 9. — P. 774–789. — DOI: 10.1162/tacl_a_00397

36. Chen Y. MENLI: Robust Evaluation Metrics from Natural Language Inference / Y. Chen, S. Eger // Transactions of the Association for Computational Linguistics. — 2023. — 11. — P. 804–825. — DOI: 10.1162/tacl_a_00576

37. Fabbri A.R. SummEval: Re-evaluating Summarization Evaluation / A.R. Fabbri, W. Kryściński, B. McCann et al. // Transactions of the Association for Computational Linguistics. — 2021. — 9. — P. 391–409. — DOI: 10.1162/tacl_a_00373

38. Su W. SurGE: A Benchmark and Evaluation Framework for Scientific Survey Generation / W. Su, A. Xie, Q. Ai et al. // Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2026); — Seoul: Association for Computing Machinery (ACM), 2026. doi: 10.1145/3805712.3808598

39. Sánchez-Mompó A. Green MLOps to Green GenOps: An Empirical Study of Energy Consumption in Discriminative and Generative AI Operations / A. Sánchez-Mompó, I. Mavromatis, P. Li et al. // Information. — 2025. — 4. — DOI: 10.3390/info16040281

40. Asai A. Reliable, Adaptable, and Attributable Language Models with Retrieval / A. Asai, Z. Zhong, D. Chen et al. // arXiv. — 2024. — URL: https://arxiv.org/abs/2403.03187. (accessed: 16.07.26)

41. Cao D. Retrieval Is Accurate Generation / D. Cao, D. Cai, L. Cui et al. // arXiv. — 2024. — URL: https://arxiv.org/abs/2402.17532. (accessed: 20.07.26)

42. Chu Z. Learning from Crowds by Modeling Common Confusions / Z. Chu, J. Ma, H. Wang // Proceedings of the AAAI Conference on Artificial Intelligence. — 2021. — Vol. 35, No. 7. — P. 5832–5840. — DOI: 10.1609/aaai.v35i7.16730

43. Tanno R. Learning from Noisy Labels by Regularized Estimation of Annotator Confusion / R. Tanno, A. Saeedi, S. Sankaranarayanan et al. // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2019. — P. 11236–11245. doi: 10.1109/CVPR.2019.01150

44. Wu T. NoisywikiHow: A Benchmark for Learning with Real-world Noisy Labels in Natural Language Processing / T. Wu, X. Ding, M. Tang et al. // Findings of the Association for Computational Linguistics: ACL 2023; — Toronto: Association for Computational Linguistics, 2023. — P. 4856–4873. doi: 10.18653/v1/2023.findings-acl.299

45. Wang Y. Iterative Learning with Open-Set Noisy Labels / Y. Wang, W. Liu, X. Ma et al. // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2018. — P. 8688–8696. doi: 10.1109/CVPR.2018.00906

46. Zheng S. Error-Bounded Correction of Noisy Labels / S. Zheng, P. Wu, A. Goswami et al. // Proceedings of the 37th International Conference on Machine Learning; — Online: PMLR, 2020. — P. 11447–11457.

47. Patrini G. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach / G. Patrini, A. Rozza, A.K. Menon et al. // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); — Piscataway: IEEE, 2017. — P. 2233–2241. doi: 10.1109/CVPR.2017.240

48. Murtadha A. Rank-Aware Negative Training for Semi-Supervised Text Classification / A. Murtadha, S. Pan, W. Bo et al. // Transactions of the Association for Computational Linguistics. — 2023. — 11. — DOI: 10.1162/tacl_a_00574

49. Shen Y. Learning with Bad Training Data via Iterative Trimmed Loss Minimization / Y. Shen, S. Sanghavi. // Proceedings of the 36th International Conference on Machine Learning; — Online: PMLR, 2019. — P. 5739–5748.

50. Arazo E. Unsupervised Label Noise Modeling and Loss Correction / E. Arazo, D. Ortego, P. Albert et al. // Proceedings of the 36th International Conference on Machine Learning; — Online: PMLR, 2019. — P. 312–321.

51. Tanaka D. Joint Optimization Framework for Learning with Noisy Labels / D. Tanaka, D. Ikami, T. Yamasaki et al. // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); — Los Alamitos: IEEE Computer Society, 2018. — P. 5552–5560. doi: 10.1109/CVPR.2018.00582

52. Reed S.E. Training Deep Neural Networks on Noisy Labels with Bootstrapping / S.E. Reed, H. Lee, D. Anguelov et al. // arXiv. — 2015. — URL: https://arxiv.org/abs/1412.6596. (accessed: 14.08.26)