УДК 004.8
DOI: 10.26102/2310-6018/2026.57.6.020

Математическая модель процесса автоматизированного построения обзорных текстов, основанного на использовании цитатно-осведомленной суммаризации

И.И. Кузнецов✉

Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство), Москва, Российская Федерация

Резюме. В работе рассматривается задача автоматизированного построения научных обзорных текстов на основе анализа корпуса научных публикаций. Разработана математическая модель, описывающая подход, основанный на использовании цитатно-осведомленной суммаризации с предварительным отбором цитатных фрагментов, и формализующая полный цикл обработки данных от отбора публикаций и извлечения цитирующих фрагментов до генерации частных саммари и итогового обзорного текста. Модель задает единое формальное описание последовательности операторов преобразования данных и включает систему критериев качества, обеспечивающих контроль правдоподобности, покрытия и фактологической согласованности результатов на всех этапах обработки. На основе разработанной модели реализована программная система в виде модульного конвейера. С использованием разработанной модели экспериментальные исследования на датасете SurGE, включающем 114 тематик, более 7 тыс. цитируемых и свыше 73 тыс. цитирующих публикаций. Показано, что использование цитатно-осведомленного подхода с предварительным отбором фрагментов обеспечивает улучшение качества генерации саммари по сравнению с альтернативными методами. Для итоговых обзорных текстов достигнуты следующие значения критериев качества: правдоподобность – 0,8744, покрытие – 0,9356, фактологическая достоверность – 0,9713 и LLM-оценка – 0,9232, что превосходит результаты генерации на основе полного текста источников (на 7,67 % для правдоподобности, 4,63 % для покрытия, 9,22 % для фактологической согласованности и 4,42 % для LLM-оценки). Полученные результаты подтверждают эффективность предложенного подхода и разработанной модели, и их применимость для задач автоматизированного формирования достоверных научных обзоров.

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

Благодарности. Автор благодарит профессора кафедры искусственного интеллекта, прикладной математики и программирования РГУ им. А.Н. Косыгина Новикова Олега Пантелеевича и доцента кафедры цифровых технологий обработки данных РТУ МИРЭА Ильина Дмитрия Юрьевича за помощь в обсуждении методов и результатов.

Для цитирования: Кузнецов И.И. Математическая модель процесса автоматизированного построения обзорных текстов, основанного на использовании цитатно-осведомленной суммаризации. *Моделирование, оптимизация и информационные технологии.* 2026;14(6). URL: https://moitvivt.ru/ru/journal/article?id=2375 DOI: 10.26102/2310-6018/2026.57.6.020

A mathematical model of the process of automated construction of review texts based on the use of quotation-aware summarization

I.I. Kuznetsov✉

The Kosygin State University of Russia, Moscow, the Russian Federation

Abstract. This paper examines the automated generation of scientific review texts based on the analysis of a corpus of scientific publications. A mathematical model has been developed that describes an approach based on the use of citation-aware summarization with preliminary selection of citation fragments and formalizes the full data processing cycle, from publication selection and extraction of citing fragments to the generation of individual summaries and the final review text. The model defines a unified formal description of the sequence of data transformation operators and includes a system of quality criteria that ensure control of the plausibility, coverage, and factual consistency of results at all stages of processing. A software system in the form of a modular pipeline is implemented based on the developed model. Experimental studies using the developed model were conducted on the SurGE dataset, which includes 114 topics, over 7,000 cited and over 73,000 citing publications. It is shown that the use of a citation-aware approach with preliminary fragment selection improves the quality of summary generation compared to alternative methods. The following quality criteria were achieved for the resulting review texts: credibility – 0.8744, coverage – 0.9356, factual reliability – 0.9713, and LLM score – 0.9232, which outperforms the results of generation based on the full text of sources (by 7.67 % for credibility, 4.63 % for coverage, 9.22 % for factual consistency, and 4.42 % for LLM score). The obtained results confirm the effectiveness of the proposed approach and the developed model and their applicability for the automated generation of reliable scientific reviews.

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

Acknowledgements: The author thanks O.P. Novikov, Professor of the Department of Artificial Intelligence, Applied Mathematics and Programming of the Institute of Information Technology and Digital Transformation at The Kosygin State University of Russia, and D.Yu Ilin, Associate Professor at the department of Data Processing Digital Technologies, Institute of Cybersecurity and Digital Technologies, MIREA – Russian Technological University, for their assistance in discussing the methods and results.

For citation: Kuznetsov I.I. A mathematical model of the process of automated construction of review texts based on the use of quotation-aware summarization. *Modeling, Optimization and Information Technology.* 2026;14(6). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2375 DOI: 10.26102/2310-6018/2026.57.6.020

Введение

В последние годы наблюдается рост объема научных публикаций, что существенно усложняет задачу построения научных обзоров. Подготовка такого обзора требует анализа большого числа статей, выявления ключевых результатов и формирования целостного представления об исследуемом направлении. Выполнение этих задач вручную является трудоемким и плохо масштабируется. В связи с этим возрастает актуальность разработки автоматизированных методов обзорно-аналитической обработки научных текстов [1, 2].

Большие языковые модели (Large Language Model, далее – LLM) активно применяются для автоматизированной обработки научных публикаций, включая генерацию аннотаций и сравнительных обзоров [3]. Показана возможность использования LLM для полного цикла формирования обзоров, включающего поиск источников, извлечение ключевой информации и генерацию текста на ее основе [4]. В [5] предложена система генерации обзоров с многоуровневым контролем за качеством генерации, а в [6] предложен фреймворк, ориентированный на декомпозицию обзора на структурные элементы. В [7] предлагается метод, основанный на промежуточной оценке и итеративной доработке генерируемых текстов, а в [8] рассматривается модульная система с контролем генерации на различных этапах. Предлагаются бенчмарки и датасеты, предназначенные для оценки качества обзоров, сгенерированных при помощи LLM [9, 10].

Однако использование LLM в задачах обработки научной литературы сопряжено с рядом существенных ограничений, таких, как склонность моделей к галлюцинациям [11]. При генерации обзоров это проявляется в искажении фактов, добавлении недостоверной информации и нарушении логической связи с источником и т. д. [12, 13]. Даже современные модели демонстрируют систематические ошибки и требуют специализированных методов их оценки [14, 15]. LLM также склонны воспроизводить распространенные, но некорректные утверждения [16]. Практические исследования в научных и клинических задачах также указывают на риски использования LLM без дополнительной верификации [17]. Указывается, что модели способны генерировать правдоподобные, но логически необоснованные выводы, создавая иллюзию достоверности текста [18].

Как следствие, активно развиваются подходы, направленные на повышение достоверности и обоснованности генерируемого текста, например, интеграция механизмов поиска внешней информации в процесс генерации [19, 20]. В работе [21] предложена архитектура Retrieval-Augmented Generation (RAG), ориентированная на повышение достоверности генерации за счет опоры на корпус источников. Рассматриваются способы повышения качества отбора исходных статей для обзора, для снижения количества «ложных» ссылок [22, 23]. В исследовании [24] предлагается система генерации со встроенным контролем качества, опирающимся на ряд критериев. В работе [25] предложена многоагентная схема для подобных систем с отдельными компонентами для основных этапов генерации. Исследуются методы повышения логической согласованности генерации, включая использование набора вопросов для анализа содержательности текста [26].

Одно из направлений повышения качеств генерируемых текстов связано с использованием научного цитирования, то есть цитатно-осведомленной генерации [27]. Это позволяет формировать более информативные и обоснованные представления, поскольку цитаты часто содержат ключевые идеи и вклад публикации в контексте научного дискурса [28]. Предлагаются подходы, в которых тексты цитирования используются непосредственно как источник данных для обучения моделей суммаризации [29] или как опорные фрагменты для генерации [30]. В работе [31] предлагается система генерации обзоров, в котором информация о цитатных взаимосвязях статей используется для структуризации исходных статей по тематикам. Рассматриваются возможности построения графа цитирующих работ для повышения качества обзорных текстов [32]. В ряде работ рассматривается эффективность дополнения RAG-архитектуры цитатной информацией [33, 34].

Существующие проблемы и ограничения указывают на то, что при построении систем генерации научных обзоров требуется обеспечение контроля качества на всех ключевых стадиях [35]. В связи с этим рассматриваются возможности формализовать процесс автоматизированной обработки научной литературы в виде математических моделей. Так, в исследовании [36] формализуется задача объединения информации из множества научных публикаций на основе цитатных связей. В [37] предложена модель суммаризации, формализующая объединение текста статьи и структуры цитирующих работ. В исследовании [38] рассматривается модель сравнительной суммаризации, основанная на формальном отображении цитат для сопоставления результатов различных работ. В работе [39] процесс обработки научной литературы формализуется в виде прикладного модульного конвейера.

Однако такие подходы, как правило, носят частный характер и не обеспечивают целостного формального описания всего процесса. Современные обзоры указывают на отсутствие единой формализованной схемы, объединяющей этапы извлечения, оценки и генерации научных знаний в рамках одной модели [40]. В связи с этим актуальной

задачей является разработка обобщенной математической модели процесса построения обзорных текстов, позволяющей обеспечить согласованное управление качеством и структурой формируемого результата.

Таким образом, анализ существующих подходов показывает, что, несмотря на значительный прогресс в области обработки научных текстов с использованием LLM и применения методов, учитывающих научное цитирование, решение задачи автоматизированного построения обзорных текстов остается фрагментарным и сталкивается с рядом ограничений. Современные методы и модели либо ориентированы на решение частных задач, таких как суммаризация отдельных документов или извлечение информации, либо реализуются в виде прикладных модульных конвейеров без единой формальной основы. При этом отсутствует целостная модель, позволяющая согласованно описать процесс отбора, оценки и агрегации содержательных фрагментов научных публикаций с последующим формированием обзорного представления. В связи с этим целью настоящей работы является разработка подхода к автоматизированному построению обзорных текстов, основанного на использовании цитатно-осведомленной суммаризации, а также формализация данного процесса в виде обобщенной математической модели. Предлагаемый подход направлен на обеспечение согласованной интеграции этапов извлечения, оценки и агрегирования информации из множества научных публикаций и создание теоретической основы для построения систем генерации достоверных и структурированных научных обзоров.

Материалы и методы

Модель цитатно-осведомленной генерации обзорных текстов на основании научных публикаций. Задача построения обзорного текста по заданной тематике с опорой на цитирующие текстовые фрагменты формализуется как построение отображения:

$$F: R \to \Theta, \tag{1}$$

где Θ – заданная тематика обзора, а R – итоговый обзорный текст.

В общем случае это отображение реализуется как композиция нескольких операторов:

$$R = F_1 \circ F_2 \circ F_3 \circ F_4 \circ F_5 \circ F_6 \circ F_7(\Theta), \tag{2}$$

где F_1 – отбор исходных публикаций по теме; F_2 – построение связанного корпуса цитирующих текстов, относящихся к выбранным публикациям; F_3 – выделение кандидатных цитирующих текстовых фрагментов; F_4 – оценка и отбор содержательно значимых фрагментов; F_5 – построение промежуточных представлений содержания; F_6 – генерация частных саммари; F_7 – генерация итогового обзорного текста.

Обозначим множество научных публикаций, доступных системе, как:

$$P = \{p_i\}_{i=1}^{N^P}. \tag{3}$$

Для каждой публикации $p_i \in P$ заданы ее метаданные M_i, ее текстовое содержание T_i и множество исходных цитирующих текстов C_i, связанных с данной публикацией.

Текст публикации T_i представим в виде множества элементарных фрагментов:

$$T_i = \{t_{iq}\}_{q=1}^{N_i^T}, \tag{4}$$

где t_{iq} может быть предложением, окном из нескольких предложений или иной единицей анализа.

Далее в формулах используется следующая индексация:

- i – индекс публикации;
- j – индекс цитирующего текста, относящегося к публикации p_i;
- k – индекс элементарного или кандидатного фрагмента внутри текста c_{ij};
- l – индекс смысловой единицы, относящейся к публикации p_i.

На первом этапе формируется множество публикаций, релевантных тематике Θ. Для каждой публикации вычисляется вектор признаков:

$$v_i^P = f_P(p_i, \Theta) \in \mathbb{R}^{d_P}, \tag{5}$$

где $f_P: P \times \Theta \to \mathbb{R}^{d_P}$ – отображение, сопоставляющее публикации p_i при заданной тематике Θ вектор признаков фиксированной размерности; $\mathbb{R}$ – множество действительных чисел; d_P – размерность пространства признаков публикации.

Компоненты вектора v_i^P описывают признаки, значимые для отбора, такие как полнота и качество доступных данных, качество источника и т. д. На основе этого вектора задается функция полезности публикации:

$$Q_P(p_i) = \Psi_P(v_i^P), \tag{6}$$

где Ψ_P – функция агрегированной оценки качества цитируемой публикации.

Тогда множество отобранных публикаций определяется как:

$$\tilde{P} = \{p_i \in P | Q_P(p_i) \geq \delta_P\}, \tag{7}$$

где δ_P – порог отбора публикаций.

Для каждой публикации $p_i \in \tilde{P}$ формируется подмножество цитирующих текстов из множества C_i, пригодных для дальнейшего анализа.

Для каждого цитирующего текста $c_{ij} \in C_i$ вычисляется вектор признаков:

$$v_{ij}^C = f_C(c_{ij}, p_i, \Theta) \in \mathbb{R}^{d_C}, \tag{8}$$

где $f_C: C_i \times P \times \Theta \to \mathbb{R}^{d_C}$ – отображение признаков цитирующего текста; d_C – размерность пространства признаков цитирующих текстов. На основе этого вектора задается функция полезности публикации, отражающая, насколько текст подходит для извлечения цитирующих фрагментов:

$$Q_C(c_{ij}) = \Psi_C(v_{ij}^C), \tag{9}$$

где Ψ_C – функция агрегированной оценки качества цитирующей публикации.

Тогда множество выбранных связанных текстов определяется как:

$$\widetilde{C_\iota} = \{c_{ij} \in C_i | Q_C(c_{ij}) \geq \delta_C\}, \tag{10}$$

где δ_C – порог отбора связанных цитирующих текстов.

Каждый связанный текст $c_{ij} \in \widetilde{C_\iota}$ разбивается на последовательность элементарных фрагментов:

$$U_{ij} = \{u_{ijk}\}_{k=1}^{N_{ij}^U}. \tag{11}$$

Для каждого элементарного фрагмента $u_{ijk} \in U_{ij}$ вычисляется степень его связи с публикацией p_i:

$$\gamma(u_{ijk}, p_i) \in [0,1]. \tag{12}$$

Величина $\gamma(u_{ijk}, p_i)$ отражает степень уверенности в том, что цитирующий текстовый фрагмент действительно относится к анализируемой цитируемой

публикации. На основе этого формируется множество фрагментов, потенциально связанных с публикацией p_i:

$$\widetilde{U_i} = \{u_{ijk} | c_{ij} \in C_i, \gamma(u_{ijk}, p_i) \geq \delta_U\}, \quad (13)$$

где δ_U – порог связи фрагмента с публикацией.

При необходимости элементарный цитирующий фрагмент может быть расширен до контекстного цитирующего фрагмента с помощью оператора

$$E_U : u \to x, \quad (14)$$

где E_U строит текстовый отрезок, включающий соседние предложения. Тогда множество кандидатных цитирующих текстовых фрагментов для публикации p_i определяется как:

$$X_i = \{x = E_U(u) | u \in \widetilde{U_i}\}. \quad (15)$$

Далее под фрагментом x понимается именно кандидатный фрагмент $x \in X_i$, то есть фрагмент после применения оператора E_U либо после тождественного преобразования, если расширение не выполняется.

Для того чтобы оценить, насколько кандидатный фрагмент соответствует содержанию публикации, каждой публикации $p_i \in \tilde{P}$ ставится в соответствие множество опорных фрагментов

$$S_i \in G_S(t_i), \quad (16)$$

где G_S – оператор построения опорного содержательного представления публикации на основе ее текстовых фрагментов t_i. Множество S_i представляет собой содержательное представление публикации, с которым сравниваются извлеченные фрагменты. Для оценки конкретного фрагмента $x \in X_i$ может использоваться подмножество наиболее близких опорных фрагментов:

$$S_i^{(K)}(x) \in S_i, \quad (17)$$

где $S_i^{(K)}(x)$ определяется на основе функции близости $p_i : X_i \times S_i \to \mathbb{R}$ и правила отбора K наиболее близких элементов множества S_i.

Для каждого фрагмента $x \in X_i$ вычисляется вектор частных показателей:

$$m_X(x) = \left(m_{X,1}(x), m_{X,2}(x), \dots, m_{X,r_X}(x)\right), \quad (18)$$

где каждая компонента $m_{X,r}(x)$ отражает один из критериев качества фрагмента. Эти критерии могут характеризовать смысловую близость, текстовое сходство, логическую согласованность и иные свойства.

Тогда общая функция качества фрагмента задается как:

$$Q_X(x) = \Psi_X\left(m_X(x)\right), \quad (19)$$

где Ψ_X – функция объединения частных критериев качества.

Множество фрагментов, прошедших первичный отбор, определяется как:

$$Y_i = \{x \in X_i | Q_X(x) \geq \delta_X\}, \quad (20)$$

где δ_X – порог качества кандидатного фрагмента.

После первичного отбора множество может оставаться слишком большим или содержать повторяющиеся по смыслу элементы. Поэтому вводится задача построения компактного множества $Z_i \subseteq Y_i$, которое используется в дальнейших этапах. Пусть $Q_X(x)$ – качество отдельного фрагмента $x \in Y_i$; $rel_\Theta(x)$ – релевантность фрагмента тематике обзора Θ; $sim(x, x')$ – близость двух фрагментов, отражающая степень их

смысловой избыточности. Тогда задача отбора компактного множества может быть записана в виде:

$$Z_i = \underset{Z \subseteq Y_i}{\operatorname{argmax}}[\lambda_1 \sum_{x \in Z} Q_X(x) + \lambda_2 \sum_{x \in Z} rel_\Theta(x) - \lambda_3 \sum_{x,x' \in Z} sim_X(x, x')], \quad (20)$$

где λ_1, λ_2, λ_3 – весовые коэффициенты, удовлетворяющие условию $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Первый и второй члены целевой функции отвечают за качество и тематическую полезность фрагментов, а третий уменьшает избыточность итогового множества.

На основе множества Z_i строится промежуточное представление содержания публикации. В общем случае оно задается как множество смысловых единиц:

$$E_i = \{e_{il}\}_{l=1}^{N_i^E} = G_E(Z_i), \quad (22)$$

где G_E – оператор построения смысловых единиц, а каждая единица e_{il} представляет отдельное положение, результат, характеристику метода, или иной смысловой элемент. Каждой единице e_{il} ставится в соответствие множество поддерживающих фрагментов:

$$B_i(e_{il}) = \{x \in Z_i | \rho(x, e_{il}) = 1\}, \quad (23)$$

где $\rho: Z_i \times E_i \to \{0,1\}$ – функция, определяющая соответствие фрагмента x смысловой единице e_{il}. Степень поддержки смысловой единицы определяется функцией:

$$Q_E(Q_{il}) = \Psi_E(B_i(e_{il})), \quad (24)$$

где Ψ_E – функция, агрегирующая сведения о поддерживающих фрагментах. Тогда множество допустимых смысловых единиц задаётся как:

$$\widetilde{E_i} = \{e \in E_i | Q_E(e) \geq \delta_E\}, \quad (25)$$

где δ_E – порог поддержки смысловой единицы.

Для каждой публикации $p_i \in \tilde{P}$ на основе множества смысловых единиц $\widetilde{E_i}$ формируется частное саммари ω_i. Формально частное саммари определяется как решение задачи:

$$\omega_i = \underset{\omega \in \Omega_i}{\operatorname{argmax}} Q_\Omega(\omega | \widetilde{E_i}, \Theta), \quad (26)$$

где Ω_i – множество допустимых текстов длины не более $L_{\Omega,i}$, а Q_Ω – функция качества частного саммари. Функция качества частного саммари определяется на основе набора частных показателей. Для каждого частного саммари ω_i вводится вектор показателей качества:

$$m_\Omega(\omega | Z_i, \Theta) = \left(m_{\Omega,1}(\omega | Z_i, \Theta), m_{\Omega,2}(\omega | Z_i, \Theta), \dots, m_{\Omega,n_\Omega}(\omega | Z_i, \Theta)\right), \quad (27)$$

где каждая компонента $m_{\Omega,n}(\omega | Z_i, \Theta))$ отражает отдельный аспект качества текста, связанный с его содержательной, структурной или тематической характеристикой. На основе этих показателей определяется функция качества частного саммари:

$$Q_\Omega(\omega | Z_i, \Theta) = \Psi_\Omega\left(m_\Omega(\omega | Z_i, \Theta)\right), \quad (28)$$

где Ψ_Ω – функция агрегирования частных показателей в итоговую оценку. Тогда множество саммари, допущенных к финальной агрегации, определяется как:

$$\widetilde{\Omega} = \{\omega_i | Q_\Omega(\omega | Z_i, \Theta) \geq \delta_\Omega\}, \quad (29)$$

где δ_Ω – порог допуска частного саммари к финальной агрегации.

На основе $\widetilde{\Omega}$ формируется итоговый обзорный текст R. Формально итоговый обзор определяется как решение задачи:

$$R = \underset{r \in R_{L_R}}{\operatorname{argmax}}\, Q_{\mathrm{R}}\left(r\middle|\widetilde{\Omega}, \Theta\right), \tag{30}$$

где R_{L_R} – множество допустимых обзорных текстов длины не более L_R, а Q_R – функция качества итогового обзора. Функция Q_R определяется на основе набора частных показателей. Для каждого текста r вводится вектор показателей качества:

$$m_{\mathrm{R}}\left(r|\widetilde{\Omega}, \Theta\right) = \left(m_{R,1}\left(r|\widetilde{\Omega}, \Theta\right), m_{R,2}\left(r|\widetilde{\Omega}, \Theta\right), \dots, m_{R,a_R}\left(r|\widetilde{\Omega}, \Theta\right)\right), \tag{31}$$

где каждая компонента $m_{R,a}\left(r|\widetilde{\Omega}, \Theta\right)$) отражает отдельный аспект качества итогового обзора, связанный с его содержательной, структурной или тематической характеристикой. На основе этих показателей определяется функция качества сгенерированного обзора:

$$Q_R\left(r|\widetilde{\Omega}, \Theta\right) = \Psi_{\mathrm{R}}(m_{\mathrm{R}}(r|\widetilde{\Omega}, \Theta)), \tag{32}$$

где Ψ_{R} – функция агрегирования частных показателей в итоговую оценку. Таким образом, итоговый обзор рассматривается как текст, оптимальный по совокупности частных критериев качества, определяемых его соответствием входным саммари и заданной тематике.

Для оценки качества итогового обзора может использоваться эталонный текст R', подготовленный экспертом или сформированный иным доверенным способом. Тогда вводится вектор эталонных оценок:

$$m_R^{ref}(R, R') = \left(m_{R,1}^{ref}(R, R'), m_{R,2}^{ref}(R, R'), \dots, m_{R,b_R}^{ref}(R, R')\right), \tag{33}$$

где каждая компонента $m_{R,b}^{ref}(R, R')$ отражает отдельный аспект сходства, соответствия или качества итогового текста по отношению к эталону. На основе этого вектора определяется эталонная оценка итогового обзора:

$$Q_R^{ref}(R, R') = \Psi_R^{ref}(m_R^{ref}(R, R'), \tag{34}$$

где Ψ_R^{ref} – функция агрегирования показателей сравнения с эталоном.

Поскольку итоговое качество обзора зависит от всех этапов построения, введем общий функционал качества системы, в котором учитываются качество отобранных содержательных фрагментов, качество частных саммари и внешняя оценка итогового обзора.

$$J = J\left(\{Z_i\}, \widetilde{\Omega}, R, R'\right). \tag{35}$$

В простейшем случае общий функционал можно записать в виде:

$$J = \mu_1 Q_Z + \mu_2 Q_\Omega^{agg} + \mu_3 Q_R^{ref}(R, R'), \tag{36}$$

где μ_1, μ_2, μ_3 – весовые коэффициенты, удовлетворяющие условию $\mu_1 + \mu_2 + \mu_3 = 1$.

Здесь

$$Q_Z = \frac{1}{|\tilde{P}|}\sum_{p_i \in \tilde{P}}\left(\frac{1}{|Z_i|}\sum_{x \in Z_i} Q_X(x)\right) \tag{37}$$

– агрегированная оценка качества отобранных содержательных фрагментов;

$$Q_\Omega^{agg} = \frac{1}{|\widetilde{\Omega}|}\sum_{\omega_i \in \widetilde{\Omega}} Q_\Omega(\omega_i) \tag{38}$$

– агрегированная оценка качества частных саммари, допущенных к финальной агрегации.

Тогда задача функционирования системы сводится к максимизации функционала при заданных ограничениях на объем данных, число объектов, длину генерируемых текстов и допустимые значения промежуточных оценок.

Программная реализация модели. Разработанная математическая модель задает формальное описание отображения от заданной тематики Θ к итоговому обзорному тексту R, согласно (2). Экспериментальная реализация данной модели выполнена в виде программы на языке Python.

В основе реализации лежит модульный конвейер, в котором этапы обработки данных соответствуют основным элементам модели. Это обеспечивает воспроизводимость всех введенных в модели преобразований и возможность их вычислительной реализации. Схема модульного устройства программы приведена на Рисунке 1. Обозначения входных и выходных данных для модулей даны в соответствии с описанной выше моделью, где P и C – исходные множества цитируемых и цитирующих статей, $\tilde{P}$ и $\tilde{C}$ – итоговые множества цитируемых и цитирующих статей, X – множество извлеченных цитатных фрагментов, Z – компактное множество отобранных цитатных фрагментов, Ω – множество сгенерированных саммари, $\tilde{\Omega}$ – множество саммари, прошедших отбор, R – сгенерированный обзор, R' – эталонный обзор.

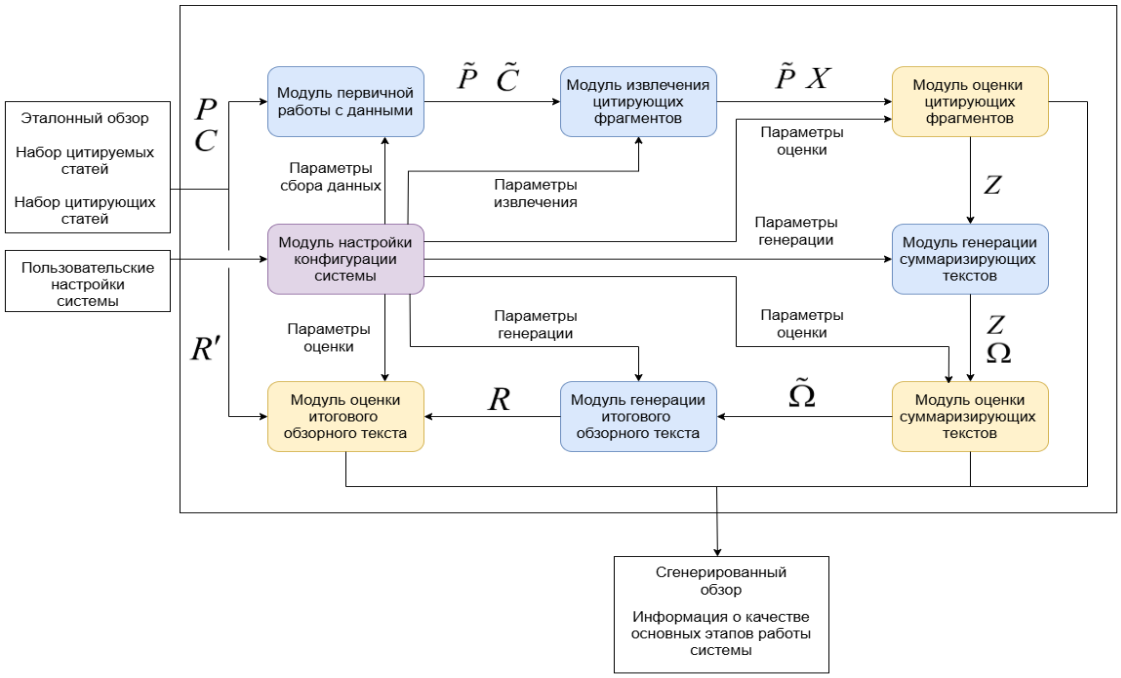

Рисунок 1 – Общая схема модульного устройства программы
Figure 1 – General diagram of the modular structure of the program

Модуль настройки конфигурации системы позволяет пользователю задавать параметры, используемые на том или ином этапе работы системы. Параметрами могут быть тематика обзора, пороги для функций отбора, веса для весовых функций и другие значения, зависящие от особенностей решаемой задачи.

Модуль первичной работы с данными предназначен для формирования исходного набора данных, содержащего в себе множество цитируемых публикаций в формате PDF-документов, соответствующих теме обзора, а также метаданные и множество исходных цитирующих текстов, связанных с каждой публикацией.

В рамках эксперимента модуль ориентирован на работу с электронной библиотекой Semantic Scholar в качестве открытого источника данных. С помощью API библиотеки осуществляется выгрузка необходимых данных. Тематика обзора формируется в виде запроса на естественном языке, при этом отбор наиболее релевантных публикаций осуществляются за счет рекомендательных механизмов самой библиотеки. Параметры подбора публикаций задаются пользователем через модуль конфигурации. В рамках модуля происходит формирование итогового набора данных, что соответствует формированию множества $\tilde{P}$ (7) на основании вектора признаков публикации v^p (5) и функции полезности публикации Q_P (6). Аналогично для каждой цитируемой публикации выполняется итоговый отбор цитирующих публикаций $\widetilde{C_l}$ (10). На Рисунке 2 приведен обобщенный алгоритм функционирования модуля.

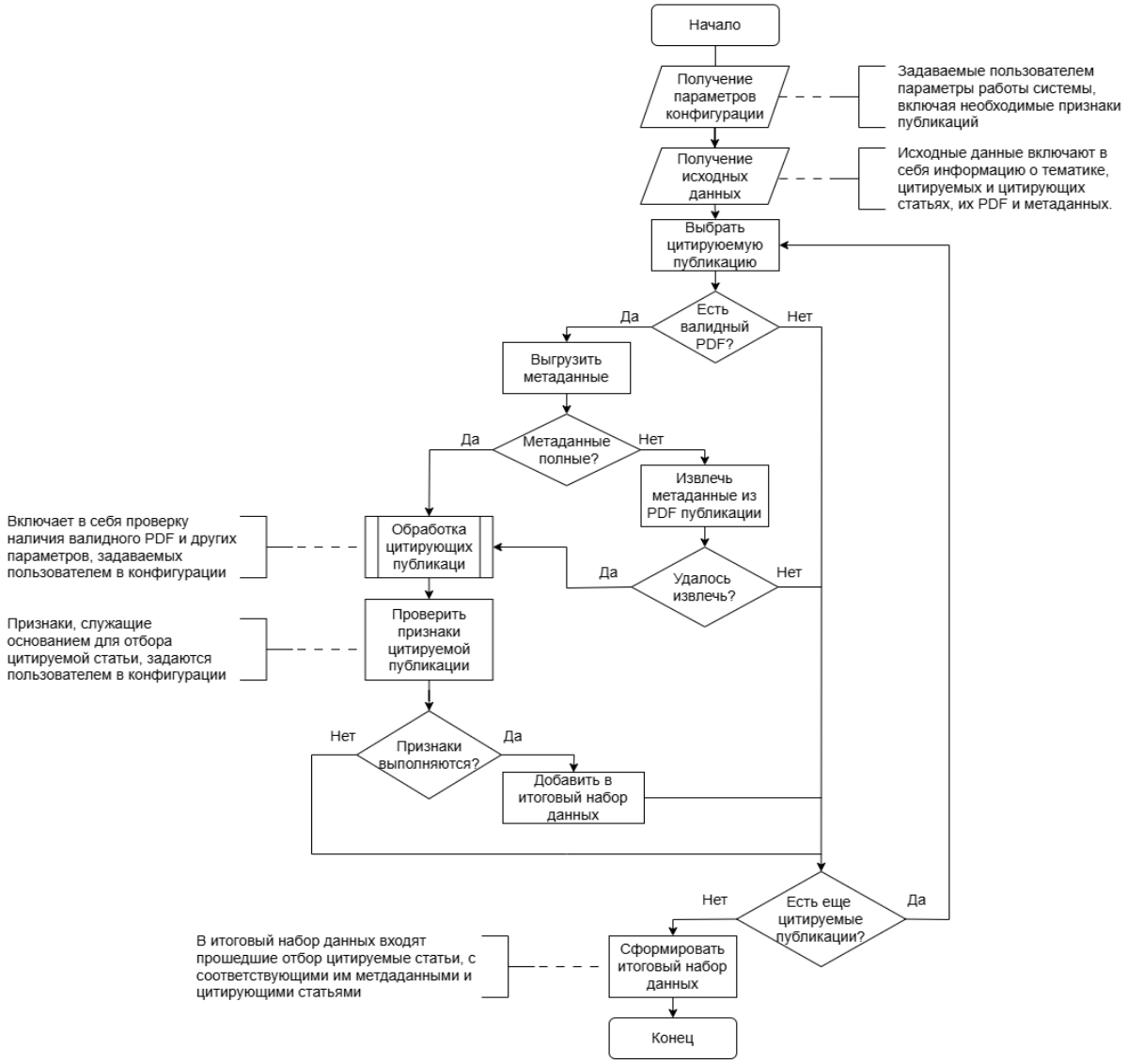

Рисунок 2 – Обобщенный алгоритм функционирования модуля первичной обработки данных
Figure 2 – Generalized algorithm for the operation of the primary data processing module

Также может использоваться готовый датасет, содержащий набор эталонных обзоров R'. Такой датасет обязательно должен содержать полный список использованных источников для каждого эталонного обзора. В таком случае вместо формирования запроса по тематике к электронной библиотеке и получения с помощью

API набора релевантных цитируемых публикаций, используется тематика из датасета, а в качестве P выступает множество работ, использованных для формирования эталонного обзора по данной тематике. Дальнейший анализ и отбор работ происходит, как и в описанном выше общем случае. При этом необходимо, чтобы все используемые для R' источники удовлетворяли заданным требованиям к цитируемым работам. То есть, конкретный эталон R' может использоваться для оценки только в случае, если $\tilde{P}$ идентично P.

В модуле извлечения цитирующих фрагментов процесс выполняется в несколько этапов. На Рисунке 3 приведен обобщенный алгоритм функционирования модуля.

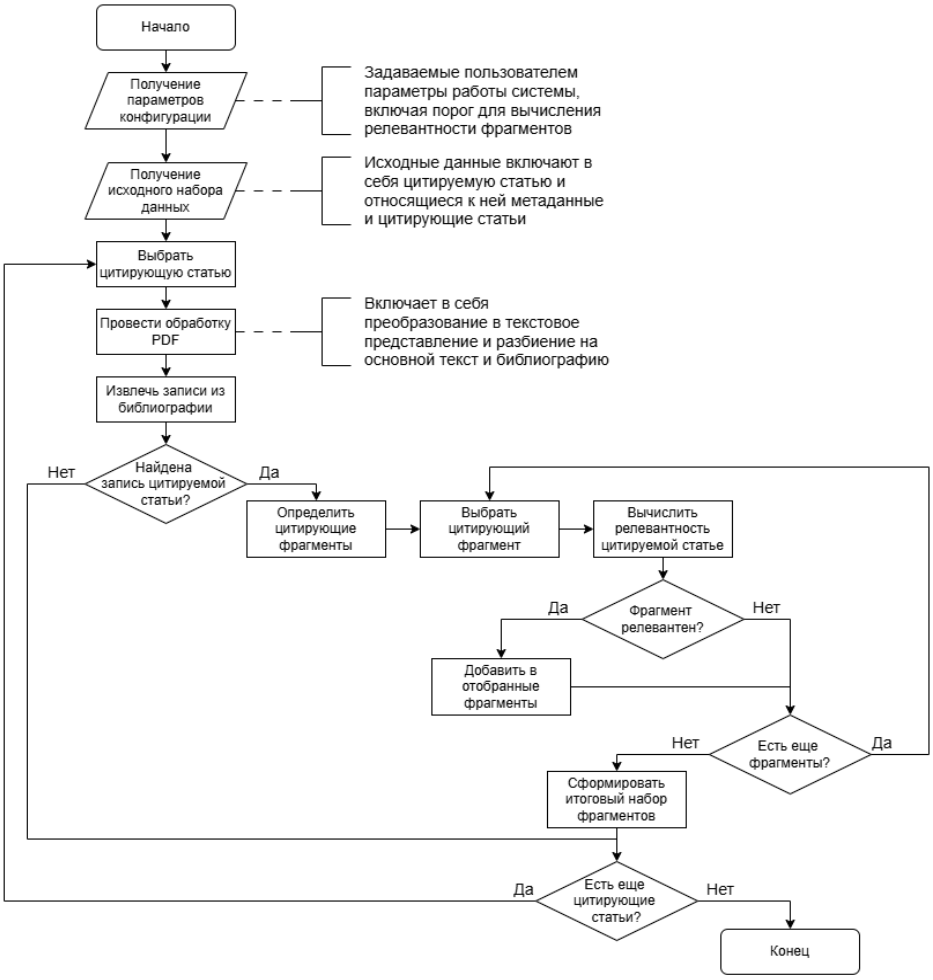

Рисунок 3 – Обобщенный алгоритм функционирования модуля извлечения цитирующих фрагментов

Figure 3 – Generalized algorithm for the operation of the quoting fragment extraction module

Для проверки семантической релевантности извлекаемых фрагментов применяется предобученная модель SciBERT [41], представляющая собой специализированную модификацию BERT, обученную на корпусе научных документов. С применением SciBERT формируются векторные представления названий и аннотаций цитируемой статьи, и цитатного фрагмента, после чего вычисляется косинусное сходство между ними для проверки семантической близости. Исходя из этого, формируется величина $\gamma(u_{ijk}, p_i)$ (12), после чего для каждой статьи производится отбор множества цитирующих фрагментов X_i (15).

После формирования X_i для каждой цитируемой статьи производится оценка качества фрагментов в соответствующем модуле. Под качеством подразумевается соответствие содержания цитирующих фрагментов содержанию цитируемой статьи. Для оценки применяются четыре метрики: ROUGE, BERTScore, QAEval и NLI, описанные в подразделе «Используемые метрики и критерии оценки». Для вычисления показателей этих метрик осуществляется подбор наиболее близких семантически фрагментов текста цитируемой публикации, которые используются как $S_i^{(K)}(x)$, согласно (17). Показатели этих метрик выступают в качестве компонент вектора (18) для каждого цитирующего фрагмента. Для функции оценки фрагмента (19) в качестве Ψ_X выступает функция взвешенной суммы, где веса задаются пользователем, как и порог итогового отбора цитатного фрагмента. В рамках этого же модуля осуществляется формирование итогового компактного множества Z_i цитирующих фрагментов (21).

В рамках модуля генерации суммаризирующих текстов для каждой цитируемой публикации производится генерация саммари. Для генерации используется одна из коммерческих LLM. Выбор модели, используемой для генерации саммари, осуществляется пользователем. Передаваемый моделям промпт унифицирован для минификации отличий и искажений между моделями на этапе их интерпретации входных данных. Промпт для генерации саммари предписывает использовать только заданные утверждения без внесения новой информации, вводятся ограничения на объем текста, требования к научному стилю изложения, информационной плотности и обязательному охвату ключевых аспектов. При ограниченности входной информации предполагается сокращение длины текста без добавления обобщающих формулировок. Обобщенный алгоритм функционирования модуля приведен на Рисунке 4.

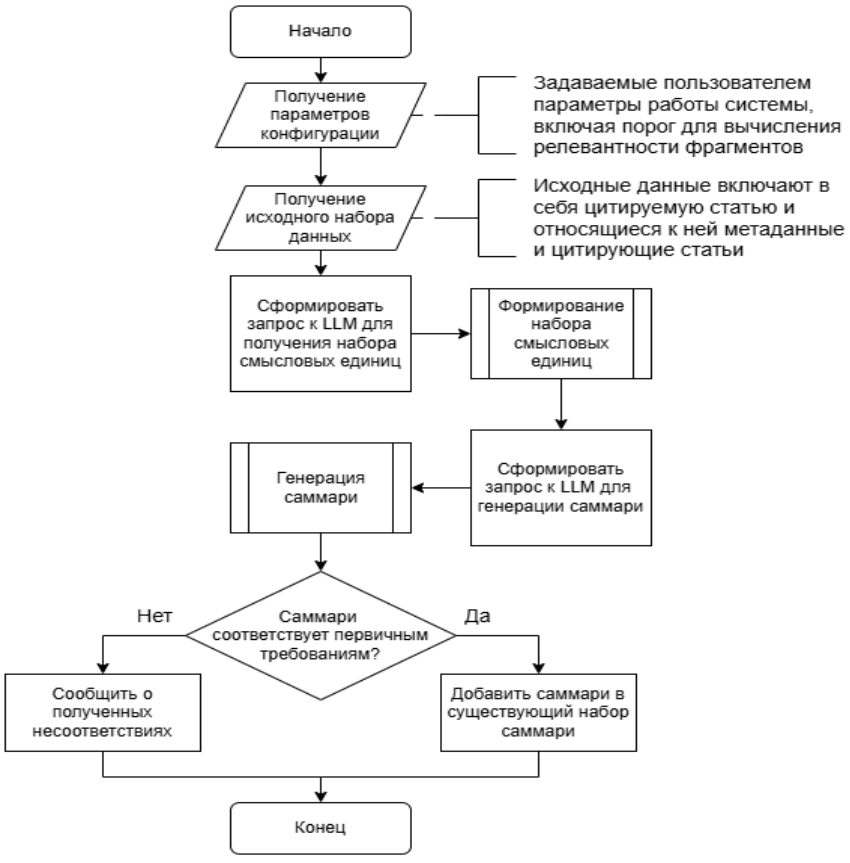

Рисунок 4 – Обобщенный алгоритм функционирования модуля генерации саммари
Figure 4 – Generalized algorithm for the functioning of the summary generation module

После генерации саммари производится их оценка согласно нескольким критериям в соответствующем модуле. Используются 3 критерия качества: правдоподобность, покрытие и фактологическая достоверность, выступающие в качестве компонентов $m_\Omega(\omega|\widetilde{E}_\iota, \Theta)$ (27). Критерии подробно описаны в подразделе «Используемые метрики и критерии оценки». Вычисляется показатель итогового качества сгенерированного саммари согласно (28), где в качестве Ψ_Ω выступает функция взвешенной суммы. Веса, как и порог отбора, задаются пользователем. На основании этого производится итоговый отбор множества саммари $\widetilde{\Omega}$ (29). На Рисунке 5 показан обобщенный алгоритм функционирования модуля.

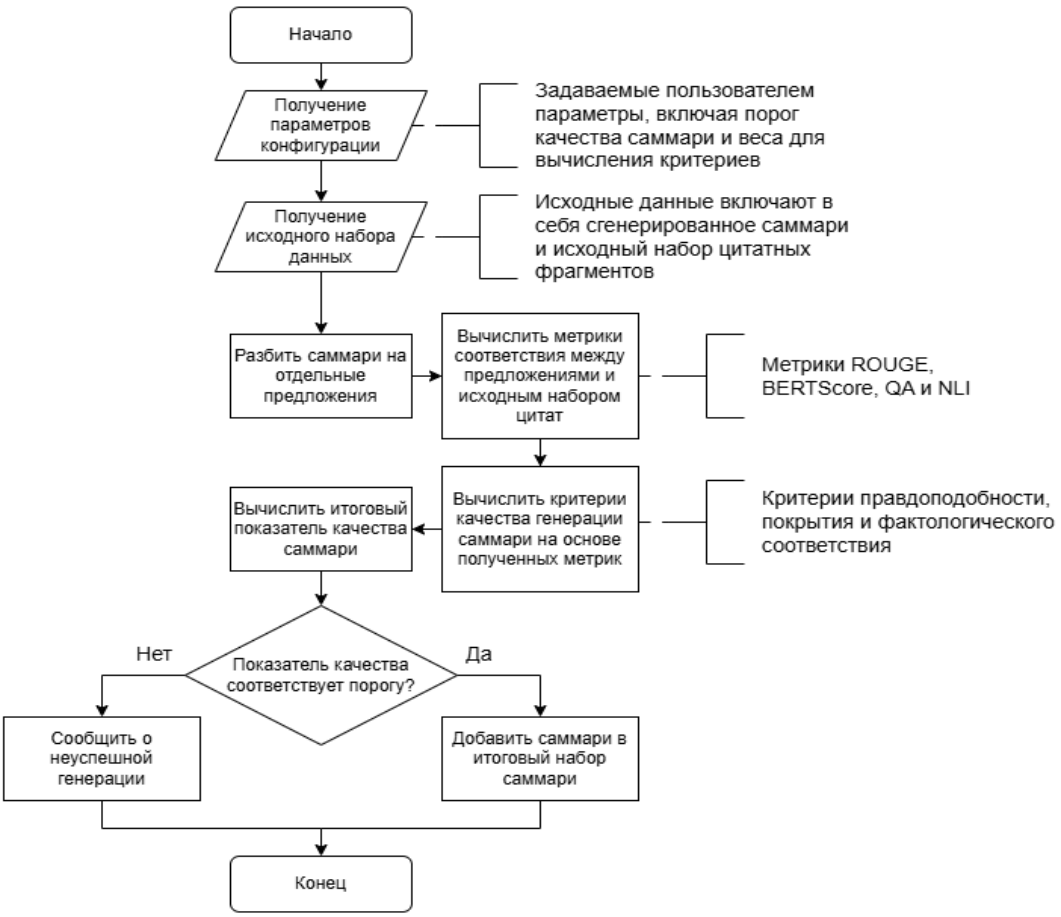

Рисунок 5 – Обобщенный алгоритм функционирования модуля оценки и отбора сгенерированных саммари

Figure 5 – Generalized algorithm for the functioning of the module for evaluating and selecting generated summaries

В рамках модуля генерации обзорного текста производится генерация на основании множества отобранных саммари. Генерация производится с использованием LLM. Промпт ограничивает модель использованием исключительно предоставленных саммари и запрещает добавление внешней информации. Промпт регламентирует требования к структуре и стилю результирующего текста: формирование связного научного изложения, обеспечение логической связности, охват тематических направлений, представленных во входных данных. Задаются ограничения на избыточность и повторяемость информации, а также требование к высокой плотности содержания. При недостаточности входных данных предполагается сокращение итогового текста без введения обобщающих утверждений.

После генерации обзорного текста производится оценка результата генерации. Реализована возможность оценки путем сравнения с эталонным обзором R', составленным экспертом. В качестве основных критериев используются правдоподобность, покрытие, и фактологическая согласованность, а также LLM-оценка качества. Эти критерии выступают в качестве компонентов $m_{R,b}^{ref}(R, R')$ (33). Критерии описаны в подразделе «Используемые метрики и критерии оценки».

Итоговая оценка сгенерированного текста производится согласно (34). В качестве Ψ_R^{ref} выступает функция взвешенной суммы, веса задаются пользователем. Это значение рассматривается в качестве итогового показателя качества сгенерированного текста.

Используемые метрики и критерии оценки. Для оценки соответствия извлеченных цитатных фрагментов содержанию цитируемой статьи применяются четыре метрики: ROUGE [42], BERTScore [43], QA [44] и NLI [45]. Совместное использование этих метрик позволяет провести комплексную оценку соответствия цитаты, где ROUGE отражает поверхностное текстовое совпадение, BERTScore – семантическую близость выражений, QAEval – фактологическую точность передачи содержания, а NLI – логическую непротиворечивость между цитатой и оригиналом.

Оценка качества сгенерированных текстов осуществляется на основе трех критериев: правдоподобности, покрытия и фактологической согласованности. Критерии определяется через комбинацию метрик ROUGE, BERTScore, QA и NLI. Для NLI используются вероятности «следует» $p_{ent}(T, C)$ и «противоречит» $p_{con}(T, C)$.

Пусть $S = \{s_j\}_{j=1}^{|S|}$ – множество предложений сгенерированного текста, а $X = \{x_i\}_{i=1}^{|X|}$ – множество единиц сравнения. В задаче генерации саммари X соответствует либо множеству цитирующих фрагментов, либо набору наиболее релевантных предложений исходной статьи. В задаче генерации обзорного текста X соответствует множеству предложений эталонного экспертного обзора.

Критерий правдоподобности определяется как степень логической согласованности утверждений сгенерированного текста с исходными данными. Для каждой пары (x_i, s_j) вычисляется оценка логического соответствия:

$$NLI(x_i, s_j) = \max\left(0, p_{ent}(x_i, s_j) - p_{con}(x_i, s_j)\right). \quad (39)$$

Далее для каждого предложения определяется агрегированная оценка:

$$f_{faith}(s_j, X) = \max_{x_i \in X}(0{,}5 \cdot NLI(x_i, s_j) + 0{,}3 \cdot BERTScore(x_i, s_j) + 0{,}2 \cdot ROUGE(x_i, s_j)). \quad (40)$$

Для этого и последующих критериев веса, используемые в рамках эксперимента, были определены экспертным способом с помощью анализа выборки из 200 сгенерированных саммари и 50 сгенерированных обзоров. Исходя из агрегированной оценки, правдоподобность определяется как:

$$Faith(S, X) = \frac{1}{|S|}\sum_{j=1}^{|S|} f_{faith}(s_j, X). \quad (41)$$

Критерий покрытия характеризует степень охвата информации из X сгенерированным текстом S. Для каждого элемента x_i определяется степень его отображения:

$$f_{cover}(x_i, S) = \max_{s_j \in S}\left(0{,}4 \cdot BERTScore(s_j, x_i) + 0{,}4 \cdot QA(s_j, x_i) + 0{,}2 \cdot ROUGE(s_j, x_i)\right). \quad (42)$$

Тогда покрытие определяется как:

$$Cover(S,X) = \frac{1}{|X|}\sum_{i=1}^{|X|} f_{cover}(x_i, S). \quad (43)$$

Критерий фактологической согласованности отражает степень подтверждения конкретных фактов сгенерированного текста исходными данными. Он является, с некоторой погрешностью, обратным показателю галлюцинаций, поскольку показывает, какая доля фактов была передана точно, а не придумана LLM. Для каждого предложения s_j определяется:

$$f_{fact}(s_j, X) = \max_{x_i \in X}(0{,}6 \cdot QA(x_i, s_j) + 0{,}4 \cdot NLI(x_i, s_j)). \quad (44)$$

Тогда фактологическая согласованность определяется как:

$$Fact(S,X) = \frac{1}{|S|}\sum_{j=1}^{|S|} f_{fact}(s_j, X). \quad (45)$$

Итоговая оценка качества сгенерированного текста определяется как взвешенная сумма критериев:

$$Q(S,X) = \alpha_1 \cdot Faith(S,X) + \alpha_2 \cdot Cover(S,X) + \alpha_3 \cdot Fact(S,X). \quad (46)$$

Здесь выбор весовых коэффициентов определяется приоритетами конкретной задачи.

Дополнительно для оценки обзорного текста используется подход на основе LLM, выступающей в роли эксперта (LLM-judge), позволяющий получить общую оценку факторов, трудно формализуемых с помощью автоматических метрик. Использование LLM-оценки позволяет учитывать глобальные свойства текста (связность, логичность изложения, корректность обобщений), которые не полностью отражаются в локальных метриках соответствия, и тем самым дополняет формальную систему критериев [46]. Модель сравнивает сгенерированный и эталонный обзоры с учетом связности, полноты, корректности передачи фактов, отсутствия искажений и противоречий, и возвращает нормированную числовую оценку в заданном диапазоне.

Такой набор критериев позволяет достаточно эффективно оценивать качество генерации обзора, в том числе, уровень точности передачи фактов и галлюцинаций, что до сих пор является одним из проблемных аспектов задачи генерации научных текстов с использованием LLM.

Результаты

Исходные данные и конфигурация программы. Для эксперимента использовался датасет SurGE [9], предназначенный для задач генерации научных обзорных текстов. Датасет включает 205 тематик, каждая из которых содержит корпус публикаций и эталонный обзор, подготовленный экспертами. Выбор датасета обусловлен его ориентированностью на задачу генерации научных обзоров и наличием явного соответствия между набором источников и эталонным текстом.

В Таблице 1 приведены основные параметры, используемые в эксперименте. Заданные веса для критериев оценки качества генерации заданы исходя из ориентации на повышение достоверности передачи информации из исходных статей.

Для эксперимента был произведен дополнительный отбор эталонных обзоров из датасета, чтобы для каждого эталонного обзора все цитируемые статьи удовлетворяли требованиям: доступ к PDF-документу и наличие не менее 10 цитирующих работ с доступом PDF-документу. В итоговую выборку вошло 114 обзоров, которым соответствовало суммарно 7 103 цитируемых работы, а также суммарно 73 514 цитирующих статей.

Таблица 1 – Основные параметры конфигурации системы при проведении эксперимента
Table 1 – Main system configuration parameters during the experiment

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 фрагментов	0,4	Вес фактологической достоверности для отбора саммари	0,4
Порог отбора цитатных фрагментов	0,63	Порог отбора саммари	0,78
Длина саммари, слов	от 80 до 100	Длина обзорного текста, слов	от 1500 до 3000

Этапы эксперимента и полученные результаты. Эксперимент с использованием программной реализации модели генерации обзорных текстов состоял из 3 этапов. Два этапа были связаны с цитатно-осведомленной генерацией саммари, как с одной из ключевых составляющих процесса, третий с генерацией обзорного текста. На первом этапе было проведено сравнение 4 коммерческих LLM применительно к задаче цитатно-осведомленной генерации саммари. На втором этапе было проведено сравнение нескольких подходов к генерации саммари для проверки утверждения, что описанный в модели подход показывает достаточную эффективность. На третьем этапе проводилась генерация обзорного текста на основании сгенерированных саммари.

На первом этапе для задачи генерации саммари сравнивались четыре коммерческих LLM: GPT-5-Mini, Claude-Sonnet-4-5-20250929, DeepSeek-V3 и Gemini-2.5-Flash. Промпт был унифицирован для всех моделей. После генерации были получены оценки качества саммари по трем критериям: правдоподобность, покрытие, фактологическая согласованность. На Рисунке 6 приведена визуализация результатов.

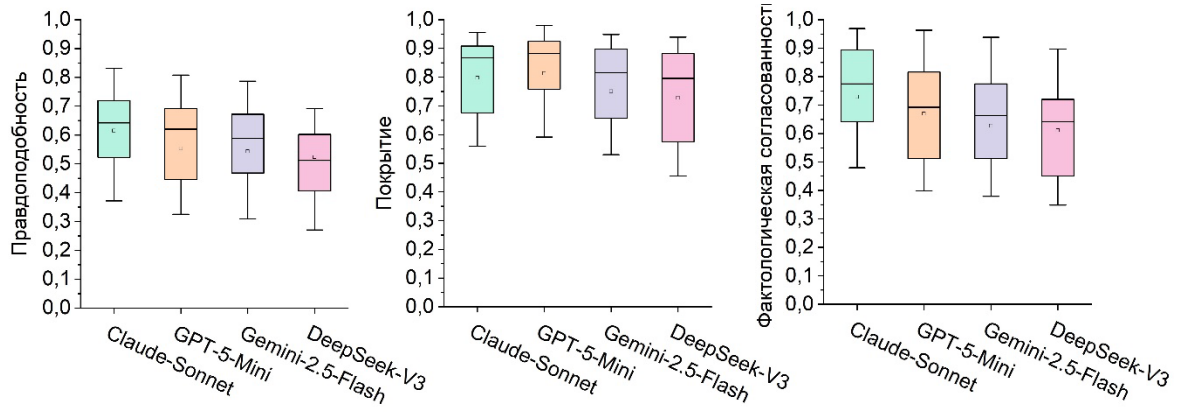

Рисунок 6 – Значения критериев качества генерации саммари для сравниваемых моделей
Figure 6 – Values of the summary generation quality criteria for the compared models

Анализ результатов первого этапа показывает, что лучшие значения всех критериев достигаются для Claude-Sonnet-4-5-20250929. GPT-5-Mini сопоставима по качеству, при этом при повышении покрытия наблюдается снижение правдоподобности и информативности. Это может быть обусловлено большей вариативностью генерируемых формулировок у этой модели. Gemini-2.5-Flash и DeepSeek-V3 демонстрируют более низкие значения критериев качества, при этом для DeepSeek-V3 характерно наибольшее снижение информативности, что свидетельствует о недостаточной способности передаче конкретных фактов и склонности к обобщённым описаниям.

В Таблице 2 приведены показатели медианных значений критериев по результатам первого этапа эксперимента.

Таблица 2 – Медианные значения критериев качества для сравниваемых моделей
Table 2 – Median values of quality criteria for the compared models

Модель	Критерий правдоподобности	Критерий покрытия	Критерий фактологической достоверности
Claude-Sonnet-4-5-20250929	0,6429	0,8674	0,775
GPT-5-Mini	0,6205	0,8825	0,6927
Gemini-2.5-Flash	0,5893	0,8168	0,6641
DeepSeek-V3	0,5123	0,7953	0,6426

На втором этапе эксперимента были рассмотрены три подхода к генерации саммари, различающиеся используемыми входными данными для генеративной модели. Первый подход (далее подход 1) предполагает генерацию саммари на основе полного текста статьи. Второй подход (далее подход 2) является основным для цитатно-осведомленной суммаризации, в нем входными данными служит набор цитатных фрагментов, извлеченных из цитирующих публикаций. Третий подход (далее подход 3) отличается этапом предварительного отбора цитатных фрагментов и соответствует тому, который описан в предлагаемой модели генерации обзорных текстов.

На Рисунке 7 приведена визуализация результатов второго этапа эксперимента. В качестве модели использована Claude-Sonnet-4-5-20250929, показавшая лучший результат на первом этапе эксперимента.

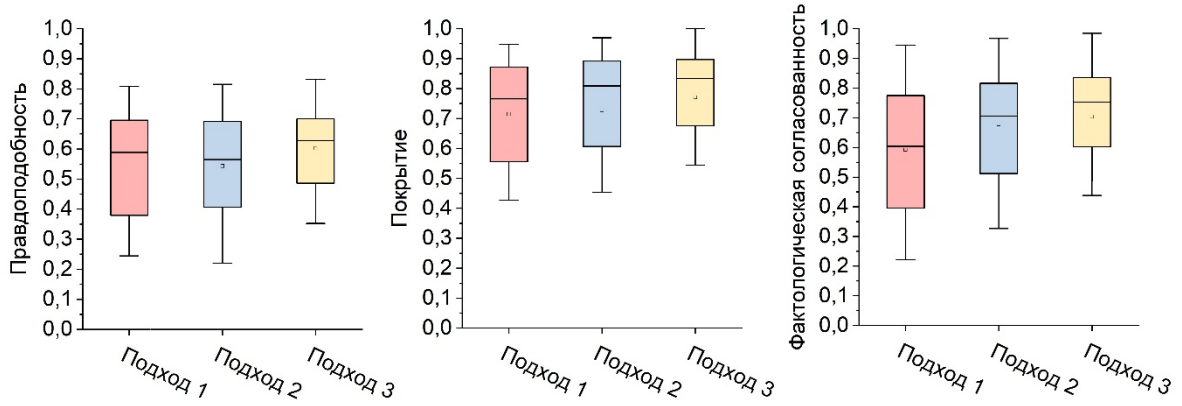

Рисунок 7 – Значения критериев качества генерации саммари для сравниваемых подходов к генерации саммари
Figure 7 – Values of summary generation quality criteria for compared summary generation approaches

Подход 2 демонстрирует более высокие значения фактологической согласованности и покрытия по сравнению с подходом 1, но уступает ему по правдоподобности. Это позволяет предположить, что наличие фоновых и обобщенных цитатных фрагментов может снижать логическую согласованность. Введение этапа предварительного отбора цитатных фрагментов в подходе 3 обеспечивает наилучшие значения по всем критериям (на 11,17 % для правдоподобности, 3,08 % для покрытия и 6,69 % для фактологической достоверности в сравнении с подходом 2), что свидетельствует о снижении доли некорректных и искаженных утверждений, повышении логической согласованности и подтверждает целесообразность использования данного подхода при построении научных суммаризирующих и/или обзорных текстов.

В Таблице 3 приведены показатели медианных значений критериев по результатам второго этапа эксперимента.

Таблица 3 – Медианные значения критериев качества для сравниваемых подходов к генерации
Table 3 – Median values of quality criteria for the compared approaches to generation

Подход	Критерий правдоподобности	Критерий покрытия	Критерий фактологической достоверности
Подход 1	0,5893	0,7664	0,6045
Подход 2	0,5657	0,8092	0,706
Подход 3	0,6289	0,8341	0,7533

На третьем этапе проводилась генерация обзорного текста на основании сгенерированных цитатно-осведомленных саммари. В качестве модели использовалась Claude-Sonnet-4-5-20250929. Для сравнения программного решения и отдельной LLM проводилась генерация обзоров для тех же тематик, но с использованием полных текстов статей вместо цитатно-осведомленных саммари. После генерации для всех обзоров была проведена оценка по критериям правдоподобности, покрытия и фактологической достоверности, а также LLM-оценка, путем сравнения с эталонным обзором из датасета. Для LLM-оценки использовалась модель GPT-5-Mini, поскольку рекомендуется использовать модель, отличную от использованной при генерации, для избежания искусственного завышения оценки [46]. Для повышения стабильности LLM-оценки для каждого обзора она производилась 20 раз, в качестве итогового показателя бралось среднее значение. Результаты третьего этапа эксперимента приведены в Таблице 4.

Таблица 4 – Медианные значения критериев качества для генерации обзорных текстов
Table 4 – Median values of quality criteria for the compared approaches to generation

Подход к генерации	Критерий правдоподобности	Критерий покрытия	Критерий фактологической достоверности	LLM-оценка
Генерация на основе цитатно-осведомленных саммари	0,8744	0,9356	0,9713	0,9232
Генерация на основе полного текста цитируемых статей	0,8121	0,8942	0,8893	0,8841

Генерация на основе цитатно-осведомленных саммари показала лучшие результаты в сравнении с генерацией на основе полного текста: на 7,67 % для правдоподобности, 4,63 % для покрытия, 9,22 % для фактологической согласованности и 4,42 % для LLM-оценки. Это доказывает целесообразность применения такого подхода, при этом наибольшая разница показателей наблюдается для правдоподобности и фактологической согласованности, что позволяет предполагать, что описываемые в модели этапы действительно повышают точность передачи ключевых фактов из цитируемых статей, а также снижают галлюцинации и искажения.

Обсуждение

Задача сопоставления результатов с другими работами является сложной, поскольку на текущий момент в области автоматизированной генерации научных текстов отсутствует единый стандарт оценки результатов. Различные работы используют отличающиеся наборы критериев и метрик, включая как автоматические показатели (семантическое сходство, логическая согласованность, полнота охвата), так и экспертные оценки (с человеком или LLM в роли эксперта). Близкие по смыслу критерии могут вычисляться различными способами, в различных диапазонах и опираться на разные исходные данные.

В связи с этим представленные сравнительные оценки носят ограниченно сопоставимый характер и не предполагают строгого эквивалентного сравнения. Тем не менее, анализ показателей позволяет получить представление о текущем развитии методов и определить положение предлагаемого подхода. В Таблице 5 даны показатели критериев из ряда работ для этой области, для которых имеется близкое смысловое соответствие. В ряде случаев представленные в исходной работе показатели нормированы для возможности сравнения с показателями текущей работы. Малое число пересечений критериев объясняется разнообразием способов оценки, различной направленностью оценки в работах, и тем, что в ряде работ результаты описаны без конкретных численных показателей. Для каждой работы приведены наилучшие описанные показатели.

Таблица 5 – Медианные значения критериев качества для генерации обзорных текстов
Table 5 – Median values of quality criteria for the compared approaches to generation

Источник	Критерий правдоподобности	Критерий покрытия	Критерий фактологической достоверности	LLM-оценка
Текущая работа	0,8744	0,9356	0,9713	0,9232
[9]	0,4636	0,9763	–	0,9612
[24]	–	0,8310	–	–
[10]	–	–	–	0,84
[6]	0,9337	0,9237	–	0,62
[26]	–	0,864	0,692	0,884
[7]	0,7758	–	–	0,8037
[5]	–	–	0,993	–
[31]	–	0,9163	–	0,9358
[34]	0,8012	0,685	–	0,6875

По результатам сравнения можно утверждать, что показатели, полученные для программной реализации модели в рамках текущей работы не хуже, а в ряде случаев превосходят схожие показатели из других работ. Отдельную важность имеет высокий показатель фактологической достоверности, поскольку именно искажение фактов и

галлюцинации зачастую являются слабым местом современных систем генерации обзоров. Таким образом, разработанная система на основе предложенной модели обеспечивает достаточно высокие значения по всем используемым критериям, что подтверждает ее применимость для задач цитатно-осведомленной генерации обзоров на основе научных текстов. Кроме того, система предполагает возможность контроля результатов на всех основных этапах работы, что позволяет повышать итоговое качество работы системы и генерации обзоров за счет повышения качества этих этапов.

Заключение

В рамках работы выполнена формализация процесса автоматизированного построения обзорных текстов на основе научных публикаций с использованием цитатно-осведомленной суммаризации в виде обобщенной математической модели. Предложенная модель задает единое формальное описание всех ключевых этапов обработки научной литературы от отбора публикаций и извлечения цитирующих фрагментов до генерации частных саммари и итогового обзорного текста. В отличие от существующих подходов, модель обеспечивает согласованную интеграцию этапов извлечения, оценки и агрегации информации, а также вводит систему критериев качества, позволяющую формализовать процесс управления качеством на всех уровнях построения обзора.

На основе разработанной модели реализована программная система в виде модульного конвейера, обеспечивающего воспроизводимость всех этапов обработки и возможность настройки параметров в зависимости от решаемой задачи. Проведенные эксперименты на наборе данных, основанном на датасете SurGE (114 тематик, 7103 целевых публикации и более 73 тыс. цитирующих работ) показали высокую эффективность предложенного подхода. Для задачи цитатно-осведомленной генерации саммари наилучшие результаты достигнуты при использовании модели Claude-Sonnet-4-5-20250929. При этом применение цитатно-осведомленного подхода с предварительным отбором фрагментов показало более высокие результаты по сравнению с альтернативными схемами генерации (на 11,17 % для правдоподобности, 3,08 % для покрытия и 6,69 % для фактологической достоверности в сравнении с цитатно-осведомленным подходом без предварительного отбора фрагментов).

Для генерации обзорных текстов на основе цитатно-осведомленных саммари получены следующие результаты для критериев качества: правдоподобность – 0,8744, покрытие – 0,9356, фактологическая достоверность – 0,9713, LLM-оценка качества – 0,9232, что превосходит результаты для генерации обзоров на основе полного текста статей (на 7,67 % для правдоподобности, 4,63 % для покрытия, 9,22 % для фактологической согласованности и 4,42 % для LLM-оценки). Полученные результаты свидетельствуют о том, что разработанная модель и ее программная реализация обеспечивают эффективное формирование содержательных, достоверных и структурированных обзорных текстов, а также создают теоретическую основу для дальнейшего развития системы автоматизированной генерации научных обзоров с контролем качества на всех этапах обработки. Разработанная модель и программная система могут использоваться для подготовки обзорных разделов научных статей, аналитических отчетов, а также в информационно-аналитических системах научных организаций. Возможность контроля качества на всех этапах обработки и формализованная оценка достоверности и полноты результатов позволяют применять систему в задачах, требующих высокой степени обоснованности выводов и достоверности генерируемых текстов.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Syed A.A., Gaol F.L., Boediman A., et al. A survey of abstractive text summarization utilising pretrained language models. In: *Intelligent Information and Database Systems: 14th Asian Conference (ACIIDS 2022): Proceedings: Part I, 28–30 November 2022, Ho Chi Minh City, Vietnam.* Cham: Springer; 2022. P. 532–544. https://doi.org/10.1007/978-3-031-21743-2_42
2. Afsharizadeh M., Ebrahimpour-Komleh H., Bagheri A., et al. A survey on multi-document summarization and domain-oriented approaches. *Journal of Information Systems and Telecommunication.* 2022;10(37):68–78. https://doi.org/10.52547/jist.16245.10.37.68
3. Keya F., Jaradeh M.Y., Auer S. Leveraging LLMs for scientific abstract summarization: unearthing the essence of research in a single sentence. In: *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries, 16–20 December 2024, Hong Kong, China.* New York: ACM; 2024. https://doi.org/10.1145/3677389.3702588
4. Scherbakov D., Hubig N., Jansari V., et al. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association.* 2025;32(6):1071–1086. https://doi.org/10.1093/jamia/ocaf063
5. Wu Sh., Ma X., Luo D., et al. Automated literature research and review-generation method based on large language models. *National Science Review.* 2025;12(6):nwaf169. https://doi.org/10.1093/nsr/nwaf169
6. Guo B., Wen Zh., Yang Y., et al. SGSimEval: a comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems. In: *Advanced Data Mining and Applications: 21st International Conference (ADMA 2025): Proceedings: Part II, 22–24 October 2025, Kyoto, Japan*. Singapore: Springer; 2025. P. 393–407. https://doi.org/10.1007/978-981-95-3456-2_27
7. Wu S., Liang Ch., Bi Z., et al. *AutoSurvey2: empowering researchers with next level automated literature surveys.* arXiv. URL: https://doi.org/10.48550/arXiv.2510.26012 [Accessed 7th April 2026].
8. Li Y., Datta S., Rastegar-Mojarad M., et al. Enhancing systematic literature reviews with generative artificial intelligence: development, applications, and performance evaluation. *Journal of the American Medical Informatics Association.* 2025;32(4):616–625. https://doi.org/10.1093/jamia/ocaf030
9. Su W., Xie A., Ai Q., et al. *SurGE: a benchmark and evaluation framework for scientific survey generation.* arXiv. URL: https://doi.org/10.48550/arXiv.2508.15658 [Accessed 19th April 2026].
10. Zhao J., Zhang Sh., Xu N., et al. *SurveyEval: towards comprehensive evaluation of LLM-generated academic surveys.* arXiv. URL: https://doi.org/10.48550/arXiv.2512.02763 [Accessed 14th April 2026].
11. Ji Z., Lee N., Frieske R., et al. Survey of hallucination in natural language generation. *ACM Computing Surveys.* 2023;55(12):248. https://doi.org/10.1145/3571730
12. Maynez J., Narayan Sh., Bohnet B., et al. On faithfulness and factuality in abstractive summarization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 05–10 July 2020, Virtual Event.* Association for Computational Linguistics; 2020. P. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173
13. Lamba N., Tiwari S., Gaur M. Hallucinations in scholarly LLMs: a conceptual overview and practical implications. In: *The Second Bridge on Artificial Intelligence for Scholarly*

Communication (AAAI-26), 20 January 2026, Singapore. 2026. https://doi.org/10.52825/ocp.v8i.3175

14. Lin S., Hilton J., Evans O. TruthfulQA: measuring how models mimic human falsehoods. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 22–27 May 2022, Dublin, Ireland.* Association for Computational Linguistics; 2022. P. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229
15. Alkaissi H., McFarlane S.I. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. 2023;15(2):e35179. https://doi.org/10.7759/cureus.35179
16. Izacard G., Lewis P., Lomeli M., et al. Atlas: few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research.* 2023;24:251.
17. Kung T.H., Cheatham M., Medenilla A., et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digital Health.* 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
18. Anh-Hoang D., Tran V., Nguyen L.-M., et al. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence.* 2025;8:1622292. https://doi.org/10.3389/frai.2025.1622292
19. Guu K., Lee K., Tung Z., et al. Retrieval augmented language model pre-training. In: *Proceedings of the 37th International Conference on Machine Learning, 13–18 July 2020, Virtual Event.* PMLR; 2020. P. 3929–3938.
20. Han B., Susnjak T., Mathrani A. Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview. *Applied Sciences.* 2024;14(19):9103. https://doi.org/10.3390/app14199103
21. Lewis P., Perez E., Piktus A., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, 06–12 December 2020, Virtual Event.* 2020.
22. Li Y., Du X., Wang Y., et al. AI-assisted literature screening: a hybrid approach using large language models and retrieval-augmented generation. *International Journal of Medical Informatics.* 2025;207:106205. https://doi.org/10.1016/j.ijmedinf.2025.106205
23. Asai A., He J., Shao R., et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature.* 2026;650(8103):857–863. https://doi.org/10.1038/s41586-025-10072-4
24. Bao T., Nayeem M.T., Rafiei D., et al. SurveyGen: quality-aware scientific survey generation with large language models. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 04–09 November 2025, Suzhou, China.* Association for Computational Linguistics; 2025. P. 2712–2736. https://doi.org/10.18653/v1/2025.emnlp-main.136
25. Shi X., Kou Q., Li Y., et al. *SciSage: a multi-agent framework for high-quality scientific survey generation.* arXiv. URL: https://doi.org/10.48550/arXiv.2506.12689 [Accessed 12th April 2026].
26. Sun Zh., Zhu X., Zhou X., et al. *SurveyBench: can LLM(-agents) write academic surveys that align with reader needs?* arXiv. URL: https://doi.org/10.48550/arXiv.2510.03120 [Accessed 11th April 2026].
27. Syed Sh., Hakimi A.D., Khatib Kh.A., et al. Citance-contextualized summarization of scientific papers. In: *Findings of the Association for Computational Linguistics (EMNLP 2023), 06–10 December 2023, Singapore.* Association for Computational Linguistics; 2023. P. 8551–8568. https://doi.org/10.18653/v1/2023.findings-emnlp.573
28. Saini N., Kumar S., Saha S., et al. Scientific document summarization using citation context and multi-objective optimization. In: *2020 25th International Conference on*

Pattern Recognition, 10–15 January 2021, Milan, Italy. IEEE; 2021. P. 4290–4295. https://doi.org/10.1109/icpr48806.2021.9412201

29. Mao Y., Zhong M., Han J. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 07–11 December 2022, Abu Dhabi, UAE*. Association for Computational Linguistics; 2022. P. 10922–10935. https://doi.org/10.18653/v1/2022.emnlp-main.750
30. Gu N., Hahnloser R.H.R. SciLit: A platform for joint scientific literature discovery, summarization and citation generation. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), 09–14 July 2023, Toronto, Canada.* Association for Computational Linguistics; 2023. P. 235–246. https://doi.org/10.18653/v1/2023.acl-demo.22
31. Zhu K., Feng X., Feng X., et al. Hierarchical catalogue generation for literature review: a benchmark. In: *Findings of the Association for Computational Linguistics (EMNLP 2023), 06–10 December 2023, Singapore.* Association for Computational Linguistics; 2023. P. 6790–6804. https://doi.org/10.18653/v1/2023.findings-emnlp.453
32. Groppe S., Hartung L. ReViz: a tool for automatically generating citation graphs and variants. In: *Digital Libraries at Times of Massive Societal Transition: 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020), 30 November – 01 December 2020, Kyoto, Japan.* Cham: Springer; 2020. P. 107–121. https://doi.org/10.1007/978-3-030-64452-9_10
33. Hu Y., Lei Zh., Dai Zh., et al. CG-RAG: research question answering by citation graph retrieval-augmented LLMs. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 13–18 July 2025, Padua, Italy.* New York: ACM; 2025. P. 678–687. https://doi.org/10.1145/3726302.3729920
34. Ding H., Zhao Y., Hu T., et al. SciRAG: adaptive, citation-aware, and outline-guided retrieval and synthesis for scientific literature. In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, 24–29 March 2026, Rabat, Morocco.* Association for Computational Linguistics; 2026. P. 6440–6460. https://doi.org/10.18653/v1/2026.eacl-long.303
35. Menick J., Trebacz M., Mikulik V., et al. *Teaching language models to support answers with verified quotes.* arXiv. URL: https://doi.org/10.48550/arXiv.2203.11147 [Accessed 7th May 2026].
36. Li W., Xiao X., Liu J., et al. Leveraging graph to improve abstractive multi-document summarization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 05–10 July 2020, Virtual Event.* Association for Computational Linguistics; 2020. P. 6232–6243. https://doi.org/10.18653/v1/2020.acl-main.555
37. An C., Zhong M., Chen Y., et al. Enhancing scientific papers summarization with citation graph. *Proceedings of the AAAI Conference on Artificial Intelligence.* 2021;35(14):12498–12506. https://doi.org/10.1609/aaai.v35i14.17482
38. Chen J., Cai Ch., Jiang X., et al. Comparative graph-based summarization of scientific papers guided by comparative citations. In: *Proceedings of the 29th International Conference on Computational Linguistics, 12–17 October 2022, Gyeongju, Republic of Korea.* International Committee on Computational Linguistics; 2022. P. 5978–5988.
39. Wadden D., Lo K., Kuehl B., et al. SciFact-open: Towards open-domain scientific claim verification. In: *Findings of the Association for Computational Linguistics (EMNLP 2022), 07–11 December 2022, Abu Dhabi, UAE.* Association for Computational Linguistics; 2022. P. 4719–4734. https://doi.org/10.18653/v1/2022.findings-emnlp.347

40. Zheng T., Deng Zh., Tsang H.T., et al. From automation to autonomy: a survey on large language models in scientific discovery. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 04–09 November 2025, Suzhou, China.* Association for Computational Linguistics; 2025. P. 17733–17750. https://doi.org/10.18653/v1/2025.emnlp-main.895
41. Beltagy I., Lo K., Cohan A. SciBERT: a pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 03–07 November 2019, Hong Kong, China.* Association for Computational Linguistics; 2019. P. 3615–3620. https://doi.org/10.18653/v1/d19-1371
42. Lin Ch.-Y. ROUGE: a package for automatic evaluation of summaries. In: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 25–26 July 2004, Barcelona, Spain.* Association for Computational Linguistics; 2004. P. 74–81.
43. Zhang T., Kishore V., Wu F., et al. BERTScore: evaluating text generation with BERT. In: *8th International Conference on Learning Representations, 26–30 April 2020, Addis Ababa, Ethiopia.* 2020. URL: https://openreview.net/pdf?id=SkeHuCVFDr
44. Deutsch D., Bedrax-Weiss T., Roth D. Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics.* 2021;9:774–789. https://doi.org/10.1162/tacl_a_00397
45. Chen Y., Eger S. MENLI: robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics.* 2023;11:804–825. https://doi.org/10.1162/tacl_a_00576
46. Zhang Q., Wang Y., Jiang Y., et al. Crowd comparative reasoning: unlocking comprehensive evaluations for LLM-as-a-judge. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 27 July – 01 August 2025, Vienna, Austria.* Association for Computational Linguistics; 2025. P. 5059–5074. https://doi.org/10.18653/v1/2025.acl-long.252

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Кузнецов Илия Игоревич, аспирант, Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство), Москва, Российская Федерация.
e-mail: iliya-kuznetsov@mail.ru
ORCID: 0009-0001-6287-8295

Iliya I. Kuznetsov, Postgraduate, The Kosygin State University of Russia, Moscow, the Russian Federation.

Статья поступила в редакцию 06.05.2026; одобрена после рецензирования 15.06.2026; принята к публикации 24.06.2025.

The article was submitted 06.05.2026; approved after reviewing 15.06.2026; accepted for publication 24.06.2025.