УДК 004.912
DOI: 10.26102/2310-6018/2026.60.9.006

Метод автоматического реферирования научных текстов на основе диагностически управляемой итерационной генерации

В.Н. Гокарев✉

Всероссийский институт научной и технической информации РАН, Москва, Российская Федерация

Резюме. В работе рассматривается задача автоматического абстрактивного реферирования русскоязычных научных текстов с использованием локальных больших языковых моделей. Показано, что однопроходная генерация реферата языковой моделью без учета характеристик исходного документа приводит к систематическим типам ошибок: семантической неполноте, механическому копированию, информационной избыточности или недостаточности. Для выявления и устранения указанных недостатков предложена гибридная программная система, реализующая замкнутый контур управления процессом генерации с обратной связью от модуля трехфакторной диагностической оценки качества. Контур осуществляет итерационную коррекцию параметров запроса к языковой модели, таких как температура, количество подаваемых ключевых слов, формат инструкции системного запроса, минимальное и максимальное количество генерируемых токенов, на основе классификации диагностического профиля машинного реферата. В составе диагностической методики предложено асимметричное штрафное преобразование *z*-нормализованных отклонений, отражающее содержательное различие между избыточностью и недостаточностью лексической, семантической и компрессионной близости реферата к референсному распределению. Описана модульная архитектура системы, включающая модули итерационной генерации рефератов, диагностической оценки и блок автоматического останова итерационного процесса. Приведены результаты экспериментальной апробации на корпусе русскоязычных научных статей, продемонстрирован статистически значимый и устойчивый прирост качества по комбинированной диагностической метрике относительно однопроходной базовой линии на отложенной тестовой выборке.

Ключевые слова: автоматическое реферирование, научные тексты, большие языковые модели, контур управления, диагностическая оценка качества, асимметричные штрафные функции.

Благодарности: Статья выполнена в рамках государственного задания: FFFU-2025-0010.

Для цитирования: Гокарев В.Н. Метод автоматического реферирования научных текстов на основе диагностически управляемой итерационной генерации. *Моделирование, оптимизация и информационные технологии.* 2026;14(9). URL: https://moitvivt.ru/ru/journal/article?id=2447 DOI: 10.26102/2310-6018/2026.60.9.006

A method for automatic summarization of scientific texts based on diagnostically driven iterative generation

V.N. Gokarev✉

All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences, Moscow, the Russian Federation

Abstract. This paper examines the problem of automatic abstract summarization of Russian-language scientific texts using local large-scale language models. It is shown that single-pass abstract generation by a language model without taking into account the characteristics of the source document leads to

systematic errors: semantic incompleteness, mechanical copying, and information redundancy or insufficiency. To identify and eliminate these deficiencies, a hybrid software system is proposed that implements a closed-loop control system for the abstract generation process with feedback from a three-factor diagnostic quality assessment module. The system iteratively adjusts query parameters to the language model, such as temperature, the number of submitted keywords, the format of the system query instruction, and the minimum and maximum number of generated tokens, based on the classification of the generated abstract diagnostic profile. The diagnostic methodology proposed includes an asymmetric penalty transformation of *z*-normalized deviations, reflecting the substantive difference between redundancies and insufficiencies in the lexical, semantic, and compressional proximity of an abstract to the reference distribution. The system's modular architecture is described, including modules for iterative abstract generation, diagnostic evaluation, and an automatic iteration process stop block. The results of experimental testing on a corpus of Russian-language scientific articles are presented, demonstrating a statistically significant and sustainable increase in quality for the combined diagnostic metric relative to a single-pass baseline on a delayed test sample.

Keywords: automatic summarization, scientific texts, large language models, control loop, diagnostic quality evaluation, asymmetric penalty functions.

Acknowledgements: The article was prepared within the framework of the state assignment: FFFU-2025-0010.

For citation: Gokarev V.N. A method for automatic summarization of scientific texts based on diagnostically driven iterative generation. *Modeling, Optimization and Information Technology.* 2026;14(9). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2447 DOI: 10.26102/2310-6018/2026.60.9.006

Введение

В Российской Федерации функционирует развитая инфраструктура информационного обеспечения научных исследований. Российские институты ВИНИТИ и ИНИОН РАН обрабатывают свыше 1 млн документов в год каждый, а также открытые электронные библиотеки НЭБ, eLibrary, КиберЛенинка и другие сервисы собирают, хранят, обрабатывают и распространяют миллионы полнотекстовых материалов научных публикаций. При ежегодном потоке, превышающем 5 млн статей, ручное реферирование становится экономически нецелесообразным из-за дефицита кадров и высокой стоимости ручного труда. Вследствие этого возникает критическая потребность в автоматизации процессов данной обработки научных текстов, однако существующие программные решения не дают требуемого качества для русскоязычных текстов.

Существующие подходы к использованию LLM в задаче реферирования сводятся к однопроходным схемам: формированию запроса по фиксированному шаблону и однократной генерации реферата. При этом отсутствует обратная связь между качеством сгенерированного реферата и параметрами запроса, что не позволяет автоматически устранять систематические недостатки конкретной модели на конкретном документе. Стандартные метрики оценки качества – ROUGE, BERTScore, BLEURT – оперируют поверхностными лексико-статистическими совпадениями и неприменимы как сигнал управления, поскольку не диагностируют тип ошибки.

В настоящей работе предложена гибридная программная система автоматического реферирования научных текстов, реализующая замкнутый контур управления процессом генерации. Система объединяет три компонента: алгоритм извлечения ключевых слов из русскоязычных научных текстов, локальную языковую модель в роли генератора и модуля трехфакторной диагностической оценки качества рефератов. Диагностическая модель выполняет роль обратной связи: на каждой итерации она оценивает машинный реферат, классифицирует его диагностический

профиль и направляет управляющее воздействие на параметры запроса к языковой модели.

Современные исследования в области автоматического реферирования преимущественно сосредоточены на дообучении предобученных языковых моделей – трансформеров на специализированных корпусах [1, 2]. Для русского языка наиболее распространены модели семейств ruT5 и mBART, обученные на новостных данных Gazeta [3]. Применение этих моделей к научным текстам характеризуется падением качества при работе со специальной терминологией [4]. Альтернативные подходы предполагают «Instruction tuning» или адаптацию через PEFT-методы (LoRA, QLoRA), однако требуют значительных вычислительных ресурсов и размеченных корпусов.

Группа работ [5, 6] исследовала методы влияния на содержание порождаемого текста через модификацию запроса: подачу ключевых концепций, ограничений по стилю, желаемой длины. Большинство таких подходов реализуют однопроходные схемы. Работы по Retrieval-Augmented Generation (RAG) [7] вводят внешний контекст, однако ориентированы преимущественно на задачу ответов на вопросы, а не на реферирование длинных документов.

Стандартные метрики оценки качества рефератов делятся на лексические (ROUGE [8], BLEU) и семантические (BERTScore [9], BLEURT [10]). Их общими ограничениями являются скалярный характер и отсутствие гибкости в использовании. Большинство работ было создано для оценки качества машинного перевода, после этого они стали также применяться и в задаче машинного реферирования. Работы по фактологической корректности [11] показывают, что современные абстрактивные модели генерируют до 30 % рефератов с фактическими ошибками. Подходы к референс-независимой оценке [12] исследуются, однако предлагаемые метрики, как правило, не дифференцированы по типам недостатков.

Методы извлечения ключевых слов, такие как: TF-IDF, RAKE [13], YAKE, TextRank [14] разрабатывались преимущественно для английского языка и не учитывают морфосинтаксической специфики русского. Подходы, основанные на синтаксических деревьях зависимостей[1] [15], позволяют извлекать многословные терминологические конструкции, однако в существующих работах не интегрированы с модулем генерации.

Данная работа сочетает все перечисленные направления и предлагает систему автоматического реферирования, направляемого выделенными ключевыми словами и использующего итерационный контур генерации на основе обратной связи от модуля трехфакторной диагностической оценки. В качестве модуля выделения ключевых слов реализован предложенный алгоритм на основе обхода дерева синтаксических связей в тексте.

Материалы и методы

Предлагаемая система имеет конвейерно-циклическую архитектуру, состоящую из трех независимых модулей: модуль извлечения ключевых слов, модуль итеративной генерации рефератов и модуль диагностической оценки качества. Модуль диагностической оценки выполняет двойную роль: служит инструментом валидации результата и источником обратной связи для итеративного контура.

Общая схема обработки документа представлена на Рисунке 1. Входной документ d поступает в модуль извлечения ключевых слов, который формирует ранжированное множество кандидатов $K(d)$. Модуль генерации формирует параметризованный запрос

[1] Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. *Автоматическая обработка текстов на естественном языке и анализ данных.* Москва: Изд-во НИУ ВШЭ; 2017. 269 с.

$p_t = T(d, K, \theta_t)$ к локальной языковой модели и получает реферат r_t. Модуль диагностической оценки вычисляет диагностический вектор $s_t \in \mathbb{R}^3$ и классифицирует профиль реферата. На основе классификации управляющая политика π корректирует вектор параметров $\theta_{t+1} = \psi(\theta_t, \pi(s_t))$, и цикл повторяется. Завершение цикла происходит при попадании реферата в область приемлемости либо при исчерпании бюджета итераций; финальный реферат r^* выбирается по стратегии best-so-far относительно референс-независимой метрики качества.

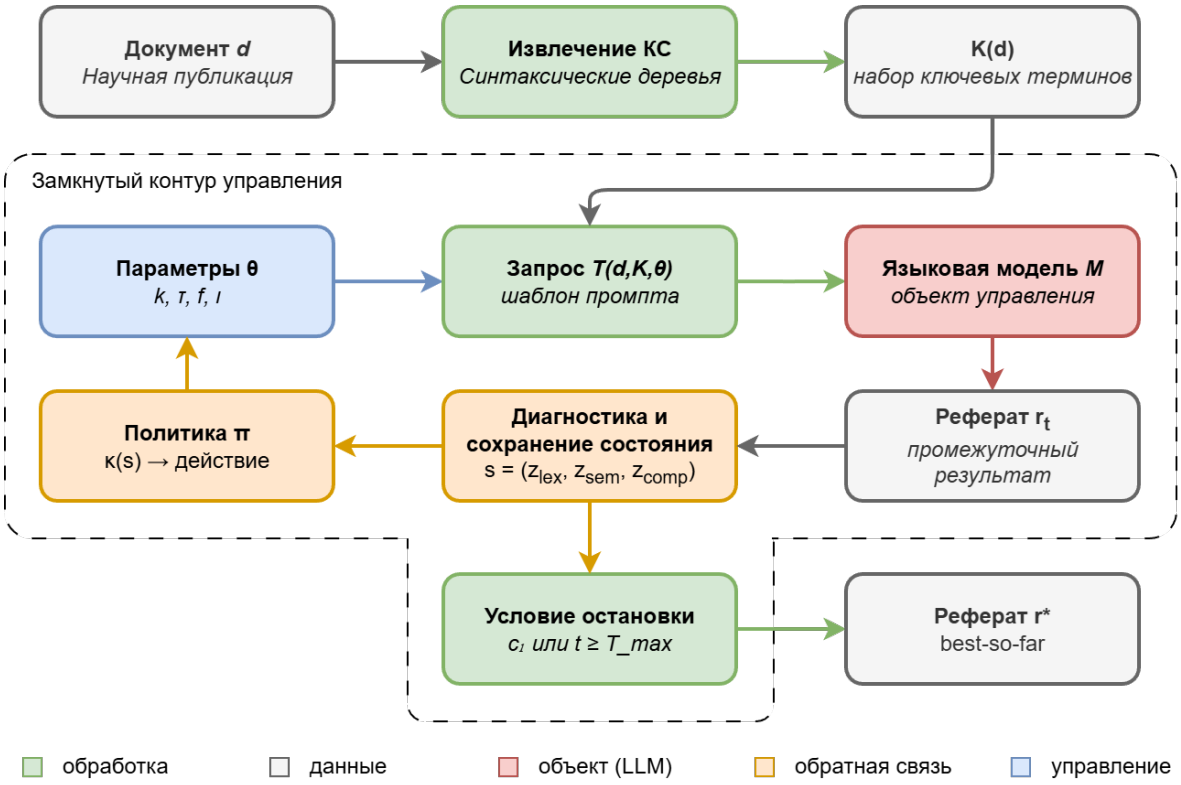

Рисунок 1 – Архитектура программной системы автоматического реферирования с итеративным контуром управления

Figure 1 – Architecture of a software system for automatic abstracting with an iterative control loop

Программная реализация выполнена на языке Python 3.10. Модуль извлечения ключевых слов использует библиотеки Stanza (морфосинтаксический анализ), pymorphy3 (лемматизация), NumPy, scikit-learn (вычисление признаков). Модуль генерации построена на основе библиотеки Transformers с поддержкой бэкендов vLLM и llama.cpp для эффективного локального инференса. Модуль диагностической оценки реализует вычисление лексических метрик (ROUGE-L Precision), семантических метрик (BERTScore F1, BLEURT), z-нормализацию, асимметричное штрафное преобразование и классификацию диагностических состояний. Калибровочные параметры референсного распределения $(\mu_x^{OA}, \sigma_x^{OA})$ вычисляются однократно на этапе подготовки системы и сохраняются как ее параметры; в режиме эксплуатации авторский реферат не требуется. Модули взаимодействуют через типизированные структуры данных, что обеспечивает их независимое тестирование и возможность замены отдельных компонентов.

В основу диагностической оценки положено сравнение машинного реферата с референсным распределением, сформированным на парах «оригинальный документ – авторский реферат». Параметры распределения $(\mu_x^{OA}, \sigma_x^{OA})$ для каждой из трех метрик $x \in [\text{lex}, \text{sem}, \text{comp}]$ – лексической (ROUGE-L), семантической (BERTScore) и

коэффициента сжатия $SR = |r|/|d|$ – вычисляются однократно на калибровочной выборке с предварительной фильтрацией аномалий по правилу 3σ.

Для произвольной пары «оригинальный документ d – машинный реферат r» нормализованные отклонения определяются как:

$$z_{\text{lex}} = \frac{s_{\text{lex}}^{OS} - \mu_{\text{lex}}^{OA}}{\sigma_{\text{lex}}^{OA}}, \quad z_{\text{sem}} = \frac{s_{\text{sem}}^{OS} - \mu_{\text{sem}}^{OA}}{\sigma_{\text{sem}}^{OA}}, \quad z_{\text{comp}} = \frac{SR^{OS} - \mu_{\text{comp}}^{OA}}{\sigma_{\text{comp}}^{OA}}, \tag{1}$$

где надстрочный индекс OS обозначает пары «оригинал-машинный реферат».

Тройка $(z_{\text{lex}}, z_{\text{sem}}, z_{\text{comp}}) \in \mathbb{R}^3$ играет роль диагностического вектора, описывающего положение машинного реферата относительно целевой зоны.

Естественным выбором штрафной функции для z-отклонения служит ядро распределения Гаусса – $\exp(-z^2/2)$. Данная функция является симметричной, однако симметричный штраф некорректно отражает специфику ряда метрик реферирования. К таким метрикам относится семантическая близость. Большое положительное значение z_{sem} означает, что машинный реферат семантически ближе к оригиналу, чем типичный авторский. Если такая близость не сопровождается дословным копированием (т. е. z_{lex} остается в норме), это указывает на полноту покрытия идей оригинала и не должно интерпретироваться как недостаток. Однако большое отрицательное значение z_{sem} соответствует семантической неполноте – наиболее критическому недостатку реферата.

Кроме семантического отклонения асимметрией обладает также коэффициент сжатия. Положительное значение z_{comp} соответствует превышенному объему реферата, относительно типичного авторского. Превышенный объем текста – менее серьезный недостаток, чем чрезмерная краткость, при которой теряется существенная информация. Для научно-информационной деятельности избыточный реферат сохраняет ценность как источник, тогда как краткий может оказаться непригодным.

Лексическое отклонение, наоборот, должно быть симметричным с обоих сторон. Большое положительное значение z_{lex} соответствует механическому копированию фрагментов оригинала, а отрицательное – потере специальной терминологии при чрезмерной перефразировке. Оба отклонения нежелательны и требуют симметричного штрафа.

Симметричное ядро $\exp(-z^2/2)$ для семантики и сжатия приводит к необоснованному снижению оценки качественных рефератов: реферат с $z_{\text{sem}} = +1{,}2$ получает штраф $\exp(-0{,}72) \approx 0{,}49$, эквивалентный штрафу за столь же сильную семантическую неполноту, хотя содержательно эти ситуации существенно различны.

Для устранения указанного недостатка вводится асимметричное преобразование z-отклонения:

$$g(z, h, \beta) = \frac{z}{1 + (h-1), \sigma(\beta, z)}, \qquad \sigma(y) = \frac{1}{1 + e^{-\beta y}}, \tag{2}$$

где $h \geq 1$ – коэффициент подавления штрафа в положительной полуоси, $\beta > 0$ – крутизна перехода между режимами. Преобразование (2) обладает следующими свойствами:

1. $g(z, 1, \beta) = z$ – при $h = 1$ восстанавливается симметричный случай;
2. $g(0, h, \beta) = 0$ – нулевое отклонение не штрафуется при любых параметрах;
3. $\sigma\big(g(z, h, \beta)\big) = \sigma(z)$ – преобразование сохраняет знак отклонения;
4. $\lim_{z \to -\infty} g(z, h, \beta)/z = 1$ – отрицательные отклонения штрафуются как при симметричном ядре;
5. $\lim_{z \to +\infty} g(z, h, \beta)/z = 1/h$ – положительные отклонения штрафуются ослабленно с асимптотическим коэффициентом $1/h$;

6. При $\beta \in [1,4]$ и $h \in [1,3]$ преобразование g монотонно возрастает по z, что обеспечивает интерпретируемость штрафной функции.

Замена стандартного z на $g(z, h_x, \beta)$ в гауссовом ядре дает асимметричную штрафную функцию:

$$\Phi(z, h, \beta) = \exp\left(-\frac{g(z,h,\beta)^2}{2}\right), \quad (3)$$

принимающую максимальное значение $\Phi(z, h, \beta) = 1$ при $z = 0$, убывающую как $\exp(-z^2/2)$ при $z \to -\infty$ и как $\exp\left(-z^2/(2h^2)\right)$ при $z \to +\infty$. Параметр β определяет ширину переходной зоны вокруг $z = 0$. Применение описанных преобразований для метрик семантики и сжатия показано на Рисунке 2.

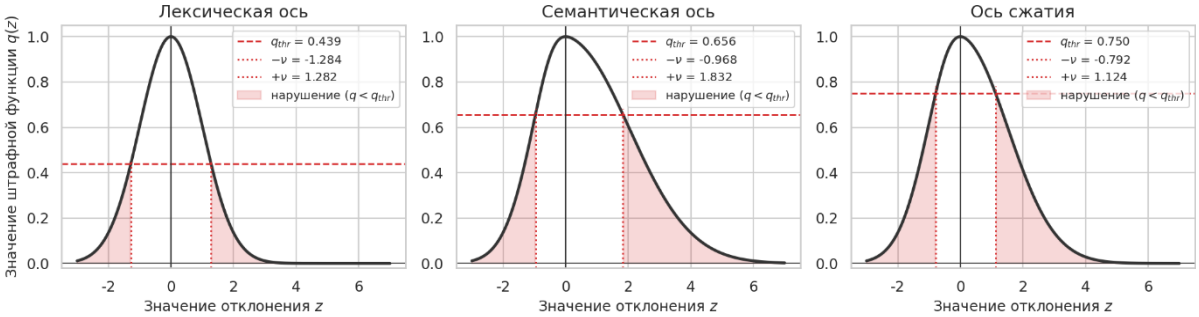

Рисунок 2 – Формы ядер $q_i(z)$ с пороговыми q_{thr} и эффективными $\pm\nu$
Figure 2 – Shapes of kernels $q_i(z)$ with threshold q_{thr} and effective $\pm\nu$

Коэффициенты $(h_{\text{lex}}, h_{\text{sem}}, h_{\text{comp}}, \beta)$ являются гиперпараметрами, подбираемыми на валидационной выборке исходя из содержательной интерпретации каждой метрики. Разумные диапазоны коэффициентов имеют вид:

$$h_{\text{lex}} = 1, \qquad h_{\text{sem}} > 1, \qquad h_{\text{comp}} > 1, \qquad h_{\text{sem}} \geq h_{\text{comp}}. \quad (4)$$

Условие $h_{\text{lex}} = 1$ соответствует симметричной штрафной функции для лексической метрики. Неравенство $h_{\text{sem}} \geq h_{\text{comp}}$ отражает большую содержательную допустимость семантической избыточности по сравнению с компрессионной. Параметр β выбирается из интервала $[-4,4]$, что гарантирует одновременно монотонность g и достаточную ширину переходной зоны вокруг типичных значений $|z| \leq 2$.

Границы целевой области для классификации диагностических состояний определяются не непосредственно по перцентилям распределения z-отклонений, а через пороговые значения штрафной функции Φ, что позволяет автоматически учесть введенную асимметрию.

Процедура определения порогов для каждой оси $x \in [\text{lex}, \text{sem}, \text{comp}]$:

1. на калибровочной выборке из M пар «оригинал – авторский реферат» вычислить z-отклонения z_x по формуле (1);
2. для каждого z_x вычислить значение штрафной функции $q_x = \Phi(z_x, h_x, \beta)$;
3. определить пороговое значение как перцентиль α полученного распределения:

$$q_{\text{thr},x} = P_\alpha(q_x); \quad (5)$$

4. решить уравнение $\Phi(z, h_x, \beta) = q_{\text{thr},x}$ относительно z, получив два корня $-\nu_x <$ и $+\nu_x >$, определяющих границы целевой зоны по оси x.

Поскольку функция Φ при $h_x > 1$ является асимметричной функцией, корни $-\nu_x$ и $+\nu_x$ в общем случае не равны по модулю – $|\nu_x^-| \neq |\nu_x^+|$. В частности, при $h_x > 1$

положительная граница $+\nu_x$ превышает по модулю отрицательную $|-\nu_x|$, что отражает большую допустимость положительных отклонений. При $h_x = 1$ оба корня симметричны: $\nu_x^- \approx -\nu_x^+$.

С учетом введенного асимметричного преобразования компонентные оценки в составе комбинированной метрики качества принимают вид:

$$q_x = \Phi(z_x, h_x, \beta),\ \ q_{\text{align}} = \exp\left(-\frac{\left(g(z_{\text{lex}}, h_{\text{lex}}, \beta) - g(z_{\text{sem}}, h_{\text{sem}}, \beta)\right)^2}{2}\right). \quad (6)$$

где $x \in [\text{lex}, \text{sem}, \text{comp}]$.

Компонент q_{align} играет роль предохранителя против эксплуатации асимметрии: при высоком z_{sem} и одновременно высоком z_{lex} (характерно для механического копирования) рассогласование $g(z_{\text{lex}}) - g(z_{\text{sem}})$ остается малым, но при «истинной» семантической избыточности без копирования компонент q_{align} приближается к единице. Итоговая референсная метрика выглядит следующим образом:

$$Q = 0{,}25 \cdot \left(q_{\text{sem}} + q_{\text{lex}} + q_{\text{align}} + q_{\text{comp}}\right). \quad (7)$$

Метрика $Q \in (0,1]$ принимает максимальное значение тогда и только тогда, когда $z_{\text{lex}} = z_{\text{sem}} = z_{\text{comp}} = 0$, и используется как критерий в итеративном контуре управления.

Каждому диагностическому вектору $s = \left(z_{\text{lex}}, z_{\text{sem}}, z_{\text{comp}}\right) \in \mathbb{R}^3$ ставится в соответствие класс из множества $\mathcal{C} = \{c_1, \dots, c_7\}$ посредством функции к: $\mathbb{R}^3 \to \mathcal{C}$. Классификация основана на эффективных границах $\pm\nu_x$, определенных через пороговые значения штрафной функции $q_{\text{thr},x}$. Реферат диагностируется как нарушающий норму по оси x, если $\Phi(z_x,\ h_x, \beta) < q_{\text{thr},x}$, что эквивалентно условию $z_x \notin [-\nu_x, +\nu_x]$.

Целевая область приемлемости определяется как

$$\mathcal{R}_* = \{\, s \in \mathbb{R}^3 \mid z_x \in [-\nu_x, +\nu_x],\ x \in [\text{lex}, \text{sem}, \text{comp}] \,\}. \quad (8)$$

Границы $\pm\nu_x$ при $h_x > 1$ асимметричны: положительная граница шире отрицательной, отражая ослабленный штраф за положительные отклонения. При $h_x = 1$ (лексическая ось) границы симметричны.

Диагностические классы определены следующим образом:

1. c_1 – целевая зона ($s \in \mathcal{R}_*$);
2. c_2 – избыточное копирование ($z_{\text{lex}} > +\nu_{\text{lex}}$);
3. c_3 – семантическая неполнота ($z_{\text{sem}} < -\nu_{\text{sem}}$);
4. c_4 – лексическая неполнота ($z_{\text{lex}} < -\nu_{\text{lex}}$);
5. c_5 – недостаточное сжатие ($z_{\text{comp}} > +\nu_{\text{comp}}$);
6. c_6 – избыточное сжатие ($z_{\text{comp}} < -\nu_{\text{comp}}$);
7. c_7 – семантический выброс ($z_{\text{sem}} > +\nu_{\text{sem}}$).

Схематично зоны показаны на Рисунке 3. Зоны «Недостаточного сжатия» и «Избыточного сжатия» лежат в трехмерном пространстве на оси аппликата.

Классификация выполняется по приоритетной логике с порядком $c_7 \succ c_2 \succ c_3 \succ c_4 \succ c_5 \succ c_6 \succ c_1$.

Запрос к языковой модели формируется параметризованным шаблоном $T: \mathcal{D} \times \mathcal{K} \times \Theta \to \mathcal{P}$, где вектор управления:

$$\theta = (k, \tau, \iota, f, s) \in \Theta, \quad (9)$$

который включает: k – количество подаваемых ключевых терминов (КТ) из $K(d)$; τ – температура декодирования; ι – текстовая инструкция, задающая стилистический и

содержательный акцент машинного реферата; f – формат подачи; s – целевой объем текста генерации.

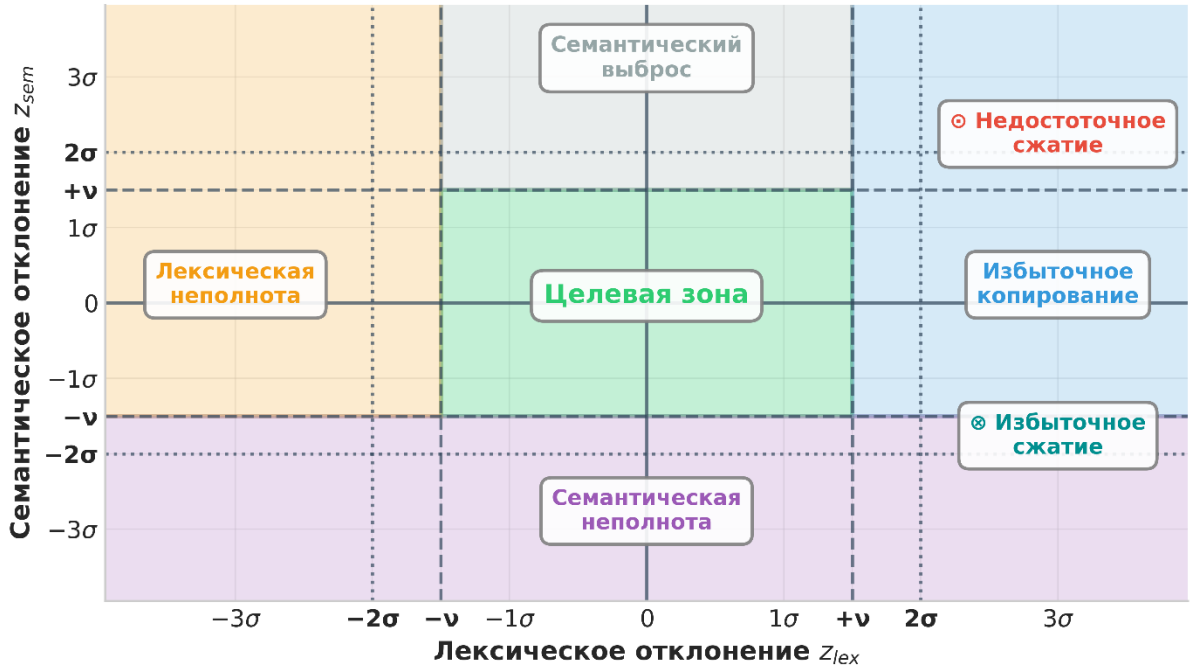

Рисунок 3 – Распределение диагностических зон в z-нормализованном пространстве
Figure 3 – Distribution of diagnostic zones in z-normalized space

Управляющая политика $\pi: \mathcal{C} \to \mathcal{A}$ задает отображение диагностического класса в действие коррекции вектора θ. В данной работе π реализована как детерминированная экспертно-построенная таблица соответствий (Таблица 1).

Таблица 1 – Управляющая политика: отображение диагностического класса в действие коррекции вектора параметров запроса
Table 1 – Control policy: mapping a diagnostic class to a query parameter vector correction action

Класс c	**Действие $\pi(c)$**
1. Недостаточное сжатие	Сжимающая инструкция: $\iota \to \iota_{\text{concine}}$; $s \leftarrow s - \Delta s$
2. Избыточное копирование	Больше перефразирования: $\iota \to \iota_{\text{paraphrase}}$; $k \leftarrow k - \Delta k$; $\tau \leftarrow + \Delta\tau$
3. Семантическая неполнота	Больше внимания на КС: $\iota \to \iota_{\text{terms}}$; $f \to f_{\text{glossary}}$; $k \leftarrow k + \Delta k$
4. Избыточное сжатие	Расширяющая инструкция: $\iota \to \iota_{\text{expand}}$; $s \leftarrow s + \Delta s$; $k \leftarrow k + \Delta k$
5. Лексическая неполнота	Меньше перефразирования: $\iota \to \iota_{\text{preserve}}$; $f \to f_{\text{verbatim}}$; $\tau \leftarrow -\Delta\tau$
6. Целевая зона	Цель достигнута, действия отсутствуют – остановка контура
7. Семантический выброс	Сброс в начальное состояние

Предложенный алгоритм представляет собой дискретный контур управления с обратной связью, в котором диагностический вектор s играет роль наблюдаемого

выхода, θ – управляющего воздействия, а композиция (κ, π, ψ) – регулятора. В отличие от классических задач управления, динамика объекта определяется не известной моделью, а стохастическим чёрным ящиком $\mathcal{M}$, что относит метод к классу управления чёрным ящиком на основе диагностических признаков.

Алгоритм завершает работу не более чем за T_{max} итераций для любого входа. Стратегия «best-so-far» обеспечивает соотношение $Q(d, r^*) \geq Q(d, r_0)$. Поскольку языковая модель стохастична, последовательность Q_t в общем случае не монотонна, однако возвращаемое значение заведомо не хуже результата однопроходной генерации с начальными параметрами θ_0.

Условие стагнации с окном n и порогом ε предотвращает бесполезные итерации в случаях, когда политика π переключается между двумя несбалансированными состояниями (например, $c_2 \leftrightarrow c_3$). Вместе с условием $t \geq T_{max}$ оно гарантирует завершение алгоритма за конечное число шагов.

Доминирующий вклад вносит инференс языковой модели: $\mathcal{O}\big(T \cdot C_{\mathcal{M}}(L_d, L_r)\big)$, где T – фактическое число итераций, L_d, L_r – длины документа и реферата в токенах. Стоимость диагностической оценки на каждой итерации пренебрежимо мала по сравнению со стоимостью инференса для моделей рассматриваемого класса (7–14 млрд параметров).

Результаты

Эксперимент проведен на выборке из 478 статей с использованием локальных языковых моделей Qwen2.5-7B-Instruct, GigaChat3-10B-A1.8B и YandexGPT-5-Lite-8B-instruct. Сравнивались три режима генерации: E1 – однопроходная генерация без интеграции ключевых терминов (базовая линия); E2 – однопроходная генерация с интеграцией ключевых терминов в запрос; E3 – итерационная генерация по алгоритму без интеграции ключевых терминов; E4 – итерационная генерация с интеграцией КТ. Оценка проводилась по референс-независимой метрике Q, доле рефератов, попадающих в целевую диагностическую зону (класс c_1), и средним z-отклонениям по трём осям. Результаты приведены на Рисунке 4 и в Таблице 2.

Итерационный контур с ключевыми терминами (E4) обеспечивает прирост $\bar{Q}$ на 3,7 % относительно базовой линии E1 и увеличение доли рефератов в целевой зоне на 16 % (в среднем с 67 % до 83 %). Однопроходная интеграция ключевых терминов (E2) самостоятельного прироста относительно E1 не обеспечила: подача 15 ключевых терминов в запрос увеличивает среднее семантическое отклонение $\bar{z}_{\text{sem}}$ с +0,07 до +0,15, смещая реферат в сторону избыточной семантической близости к оригиналу. Итерационный контур без ключевых слов (E3) показал улучшение в среднем с 67 % до 78 %, тем самым подтверждая работу корректирующей политики. Этот результат не обесценивает роль ключевых терминов в составе итерационного контура: в режиме E4 ключевые термины выступают не статической добавкой к запросу, а управляемым параметром, количество и формат подачи которого корректируются обратной связью от диагностической оценки. Таким образом, источником прироста качества в E3 и E4 является именно диагностическая обратная связь, а не сам факт интеграции ключевых терминов в запрос.

Рисунок 4 – Пилотные результаты итерационного алгоритма
Figure 4 – Pilot results of the iterative algorithm

Таблица 2 – Сравнение режимов генерации
Table 2 – Comparison of generation modes

Режим	Целевая зона	$\bar{Q}$	$\bar{z}_{lex}$	$\bar{z}_{sem}$	$\bar{z}_{comp}$
Qwen2.5					
E1 (без КТ)	76 %	0,880	−0,07	+0,19	−0,10
E2 (с КТ)	74 %	0,872	−0,07	+0,10	−0,13
E3 (итерационный, без КТ)	**88 %**	**0,916**	−0,26	+0,20	−0,08
E4 (итерационный, с КТ)	88 %	0,913	−0,24	+0,16	−0,09
GigaChat					
E1 (без КТ)	52 %	0,816	−0,81	+0,16	+0,19
E2 (с КТ)	57 %	0,834	−0,70	+0,11	+0,13
E3 (итерационный, без КТ)	67 %	0,853	−0,83	+0,08	+0,05
E4 (итерационный, с КТ)	**72 %**	**0,867**	−0,73	+0,05	+0,01

Таблица 2 (продолжение)
Table 2 (continued)

YandexGPT-5					
E1 (без КТ)	73 %	0,878	−0,08	+0,07	−0,05
E2 (с КТ)	73 %	0,875	−0,02	+0,15	−0,03
E3 (итерационный, без КТ)	79 %	0,896	−0,16	+0,06	−0,07
E4 (итерационный, с КТ)	**89 %**	**0,910**	−0,29	+0,28	−0,02

Экспериментальные результаты подтвердили эффективность всех трех разработанных компонентов системы. Алгоритм извлечения ключевых терминов превосходит существующие аналоги по основным метрикам качества. Методика трехфакторной диагностической оценки с асимметричными штрафными функциями обеспечивает диагностику типов недостатков машинных рефератов, невозможную при использовании стандартных метрик. Метод диагностически управляемого реферирования обеспечивает прирост качества генерируемых рефератов относительно однопроходной базовой линии за счет замкнутой обратной связи от диагностической оценки к параметрам запроса.

Обсуждение

По вычисленным результатам отсутствует явный прирост качества в режиме E2 относительно начального E1. Для разных моделей характерно различное поведение при интеграции 15 ключевых терминов в запрос. Модели Qwen 2.5 и GigaChat 3 значительно снизили абсолютное семантическое отклонение в то время, как модель YandexGPT-5 увеличило отклонение $\overline{z_{\text{sem}}}$ с +0,07 до +0,15. При этом также заметно падение показателей лексики $\overline{z_{\text{lex}}}$ в среднем на 0,05 пунктов. Получив список терминов модели, стремится включить их все в итоговый реферат, что порождает перегруженный специальной лексикой текст более близкий к пересказу, чем к реферату. Данное явление связано с эффектом позиционного смещения внимания в больших языковых моделях [16, 17]. Структурированные вставки в запросе (списки, перечни) непропорционально влияют на распределение внимания модели, сдвигая генерацию в сторону буквального воспроизведения подаваемого контекста. В работе [18] показано, что количество подаваемых ключевых фраз контролирует баланс полноты и точности реферирования, однако чрезмерное число фраз может приводить к деградации качества – результат, согласующийся с наблюдаемым нами эффектом.

Существенно, что этот отрицательный результат не обесценивает роль ключевых терминов в составе итеративного контура. В режиме E4 ключевые слова выполняют иную функцию: они являются не статической добавкой к запросу, а управляемым параметром, количество и формат подачи которого корректируются диагностической обратной связью. Политика π при обнаружении семантической избыточности (c_2) снижает число подаваемых терминов, а при обнаружении семантической неполноты (c_3) – увеличивает. Таким образом, в итеративном контуре ключевые термины играют роль «регулируемого сигнала», а не фиксированной подсказки.

Заключение

В работе представлена гибридная программная система автоматического реферирования русскоязычных научных текстов, основанная на замкнутом контуре управления генерацией с обратной связью от трехфакторной диагностической оценки.

Основные результаты включают:

1) Архитектуру итеративного контура, осуществляющего коррекцию параметров запроса к локальной языковой модели на основе классификации диагностического профиля машинного реферата. Контур не требует дообучения языковой модели и доступа к ее внутренним представлениям, что обеспечивает применимость к произвольным локально развёрнутым моделям.

2) Асимметричные штрафные функции для z-нормализованных диагностических признаков, корректно отражающие содержательное различие между избыточностью и недостаточностью семантической близости и компрессионной близости реферата к референсному распределению.

3) Референс-независимую форму комбинированной метрики качества Q, применимую в режиме инференса при отсутствии авторского реферата и используемую в качестве критерия отбора в итеративном контуре.

4) Программную реализацию системы и результаты экспериментальной апробации на корпусе русскоязычных научных статей, подтверждающие эффективность предложенного подхода.

Перспективы дальнейших исследований включают: применение bootstrap-схем для оценки доверительных интервалов параметров калибровочного распределения и реализации схемы leave-one-out с целью устранения возможного смещения; разработку адаптивного шага коррекции Δk, пропорционального величине z-отклонения; экспериментальное сравнение с более широким спектром локальных языковых моделей; обучение управляющей политики π методами обучения с подкреплением на корпусе диагностических траекторий; экспертную валидацию метрики Q корреляцию с ранжированием профессиональных референтов.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Lewis M., Liu Y., Goyal N., et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 05–10 July 2020, Online.* ACL; 2020. P. 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703
2. Raffel C., Shazeer N., Roberts A., et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research.* 2020;21(140):1–67.
3. Gusev I. Dataset for Automatic Summarization of Russian News. In: *Artificial Intelligence and Natural Language: 9th Conference (AINL 2020), 07–09 October 2020, Helsinki, Finland*. Cham: Springer; 2020. P. 122–134. https://doi.org/10.1007/978-3-030-59082-6_9
4. Balde G., Roy S., Mondal M., Ganguly N. Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings. In: *Findings of the Association for Computational Linguistics, 27 July – 01 August 2025, Vienna, Austria*. ACL; 2025. P. 22989–23004. https://doi.org/10.18653/v1/2025.findings-acl.1179
5. Keskar N.Sh., McCann B., Varshney L.R., Xiong C., Socher R. *CTRL: A Conditional Transformer Language Model for Controllable Generation.* arXiv. URL: https://arxiv.org/abs/1909.05858 [Accessed 6th March 2026].
6. He J., Kryściński W., McCann B., Rajani N., Xiong C. *CTRLsum: Towards Generic Controllable Text Summarization.* arXiv. URL: https://arxiv.org/abs/2012.04281 [Accessed 6th March 2026].

7. 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 2020 (NeurIPS 2020), 06–12 December 2020, Online*. 2020. URL: https://arxiv.org/abs/2005.11401
8. Lin Ch.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedings of Workshop on Text Summarization Branches Out: Post-Conference Workshop of ACL 2004, Barcelona, Spain.* ACL; 2004. P. 74–81.
9. Zhang T., Kishore V., Wu F., Weinberger K.Q., Artzi Y. *BERTScore: Evaluating Text Generation with BERT.* arXiv. URL: https://arxiv.org/abs/1904.09675 [Accessed 6th March 2026].
10. Sellam Th., Das D., Parikh A.P. BLEURT: Learning Robust Metrics for Text Generation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 05–10 July 2020, Online*. ACL; 2020. P. 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704
11. Maynez J., Narayan Sh., Bohnet B., McDonald R.T. On Faithfulness and Factuality in Abstractive Summarization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 05–10 July 2020, Online*. ACL; 2020. P. 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173
12. Vasilyev O.V., Dharnidharka V., Bohannon J. Fill in the BLANC: Human-free Quality Estimation of Document Summaries. In: *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP 2020), 20 November 2020, Online*. ACL; 2020. P. 11–20. https://doi.org/10.18653/v1/2020.eval4nlp-1.2
13. Rose S., Engel D., Cramer N., Cowley W. Automatic Keyword Extraction from Individual Documents. In: *Text Mining: Applications and Theory*. Chichester: John Wiley & Sons; 2010. P. 1–20. https://doi.org/10.1002/9780470689646.ch1
14. Mihalcea R., Tarau P. TextRank: Bringing Order into Text. In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), 25–26 July 2004, Barcelona, Spain*. ACL; 2004. P. 404–411.
15. Лукашевич Н.В., Логачёв Ю.М. Комбинирование признаков для автоматического извлечения терминов. *Вычислительные методы и программирование.* 2010;11(4):108–116.
    Lukashevich N.V., Logachev Yu.M. Automatic term extraction based on feature combination. *Numerical Methods and Programming.* 2010;11(4):108–116. (In Russ.).
16. Liu N.F., Lin K., Hewitt J., et al. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics.* 2024;12:157–173. https://doi.org/10.1162/tacl_a_00638
17. Hsieh Ch.-Y., Chuang Y.-S., Li Ch.-L., et al. Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization. In: *Findings of the Association for Computational Linguistics (ACL 2024), 11–16 August 2024, Bangkok, Thailand, Online.* ACL; 2024. P. 14982–14995. https://doi.org/10.18653/v1/2024.findings-acl.890
18. Xu L., Karim M.A., Dingliwal S., Elangovan A. Salient Information Prompting to Steer Content in Prompt-based Abstractive Summarization. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024): Industry Track, 12–16 November 2024, Miami, FL, USA*. ACL; 2024. P. 35–49. https://doi.org/10.18653/v1/2024.emnlp-industry.4

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Гокарев Вадим Николаевич, аспирант Московского государственного технического университета имени Н.Э. Баумана; ведущий специалист Всероссийского института научной и технической информации РАН, Москва, Российская Федерация.
e-mail: vckain6387@mail.ru
ORCID: 0009-0002-4082-2925

Vadim N. Gokarev, Postgraduate at Bauman Moscow State Technical University; Leading Specialist at the All-Russian Institute of Scientific and Technical Information of the Russian Academy of Sciences, Moscow, the Russian Federation.

Статья поступила в редакцию 23.05.2026; одобрена после рецензирования 09.09.2026; принята к публикации 15.09.2026.

The article was submitted 23.05.2026; approved after reviewing 09.09.2026; accepted for publication 15.09.2026.