

# ОЦЕНКА ОБЩИХ И СПЕЦИАЛЬНЫХ ЗНАНИЙ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ ДЛЯ РУССКОГО ЯЗЫКА ПОСРЕДСТВОМ ВОСПРОИЗВЕДЕНИЯ ЭНЦИКЛОПЕДИЧЕСКИХ СТАТЕЙ

© 2025 г. Д. А. Григорьев<sup>1,\*</sup>, Д. И. Чернышев<sup>1,\*\*</sup>

Представлено кем-то

Поступило 16.08.2025

После доработки 20.08.2025

Принято к публикации 31.08.2025

Работа исследует методы сжатия художественных текстов с помощью языковых моделей и предлагает улучшенные подходы для точного реферирования в условиях ограниченного контекста.

*Ключевые слова и фразы:* LLM, реферирование, литература, книги, краткий пересказ

## ВВЕДЕНИЕ

**Реферирование художественной литературы** Автоматическое реферирование текста — одна из ключевых задач в области обработки естественного языка. Суть этой задачи заключается в создании информативной аннотации исходного текста с сохранением основного смысла содержания. В последние годы, с появлением больших языковых моделей, резко возрос интерес к автоматизации реферирования в самых разных жанрах текстов, включая художественные произведения. В отличие от научных, новостных или технических текстов, художественные произведения характеризуются высокой степенью стилистической и семантической сложности. Нелинейность повествования, образность, метафоричность и стилистические приёмы делают задачу написания краткого содержания особенно трудоёмкой. Ограниченное контекстное окно современных моделей дополнительно осложняет работу с длинными произведениями.

Теоретически автоматическое реферирование может выполняться двумя основными способами: извлекающим (выбор ключевых фрагментов текста) и абстрактным (генерация нового текста на основе содержания оригинала). Для художественной литературы более уместен второй подход, поскольку он позволяет передать смысл и стиль произведения, не нарушая его целостности.

## НАБОР ДАННЫХ

На момент начала исследования не существовало открытых и репрезентативных корпусов, предназначенных специально для задачи реферирования художественных текстов на русском языке. Для экспериментов был сформирован собственный набор данных, включающий:

- более 5 000 пользовательских пересказов с ресурса «Народный Брифли»;
- исходные произведения из электронной библиотеки LibRuSec (публичное достояние или тексты с разрешения правообладателей);
- очищенные версии аннотаций, полученные с помощью LLM Meta-Llama 3–70B–Instruct.

Тексты аннотаций проходили автоматическую очистку от HTML-тегов, комментариев и служебных пометок. Затем к каждой аннотации добавлялся оригинальный текст произведения. Такая пара «текст – аннотация» использовалась для тренировки и тестирования методов реферирования.

## ОЦЕНИВАНИЕ МЕТОДОВ

На момент начала исследования в русскоязычном сегменте отсутствовали открытые и репрезентативные наборы данных, специально предназначенные для задачи автоматического реферирования

<sup>1</sup> Московский государственный университет им. М. В. Ломоносова, Москва, Россия

\* E-mail: dagrig14@yandex.ru

\*\* E-mail: chdanorbis@yandex.ru

художественных текстов. В отличие от аналогичных англоязычных проектов, которые уже включали крупные корпуса литературных произведений с высококачественными аннотациями, для русскоязычных текстов подобные ресурсы не были представлены. Это существенно затрудняло объективное тестирование и сравнение эффективности методов автоматического реферирования в условиях художественных произведений на русском языке.

С целью проведения экспериментов и оценки различных подходов к генерации аннотаций был создан собственный корпус, состоящий из художественных текстов и соответствующих кратких пересказов. В качестве источника для аннотаций был выбран ресурс «Народный Брифли» [1] — платформа, где пользователи публикуют краткие пересказы литературных произведений. Несмотря на вариативность качества и стиля пользовательских аннотаций и наличие нерелевантной информации, такой как учебные вопросы или редакторские замечания, после тщательной предварительной обработки удалось получить достаточно надёжный и чистый набор данных.

Художественные тексты были отобраны из электронной библиотеки LibRuSec — одного из крупнейших русскоязычных ресурсов художественной литературы, содержащего свыше 400 тысяч текстов. Отбор произведений осуществлялся на основании наличия аннотаций на выбранном ресурсе. Каждый текст проходил автоматическую предварительную обработку: удалялась метаинформация (например, заголовки, описания глав и технические вставки), после чего текст форматировался в единый стандартизированный вид, подходящий для дальнейшего использования в моделях. Важно отметить, что при создании корпуса использовались только тексты, находящиеся в общественном достоянии или распространяемые свободно с разрешения правообладателей, что обеспечивает соблюдение требований авторского права.

Для иллюстрации структуры сформированного корпуса ниже представлен пример пары «фрагмент художественного текста — соответствующая аннотация».

## ПРИМЕНЕНИЕ МЕТОДОВ

Экспериментальная часть включает проверку базовых стратегий (иерархическая, итеративная, «чертёжная») и двух усовершенствований.

**Влияние предварительной очистки** Очистка исходного текста от технических артефактов положительно сказалась на ROUGE-L и BERTScore для моделей MetaLlama 370B и RuadaptQwen2.5-32BProBeta, повысив также полноту покрытия ключевых вопросов.

**Сравнение базовых методов** Иерархический и итеративный подходы продемонстрировали сопоставимое качество (ROUGE-L  $\approx 0.48$ , BERTScore  $\approx 0.70$ ), заметно превосходя псевдо-генерацию без доступа к исходному тексту.

**Иерархический метод с фильтрацией узлов** Добавление фильтра по семантической близости (порог 0.85) позволило убрать избыточные фрагменты и ускорить генерацию на длинных текстах в среднем на 15 %.

**«Чертёжный» метод с кластеризацией вопросов** Кластеризация эмбедингов вопросов алгоритмом  $k$ -means и последующее сэмплирование 10–30 вопросов из каждого кластера уменьшили число обращений к модели и ускорили работу метода в 1.22 раза без потери качества.

## СПИСОК ЛИТЕРАТУРЫ

- [1] *Народный Брифли*. Электронная библиотека кратких пересказов литературных произведений. <https://wiki.briefly.ru/> (дата обращения: 16.07.2025).

## REFERENCES

- [1] *Народный Брифли*. Электронная библиотека кратких пересказов литературных произведений. <https://wiki.briefly.ru/> (дата обращения: 16.07.2025).