Р. И. Кувшанов, Д. В. Семихин
Тюменский государственный университет, г. Тюмень
УДК 004.891.2

СРАВНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ГЕНЕРАЦИИ АННОТАЦИЙ К НАУЧНЫМ СТАТЬЯМ

Аннотация. В статье рассмотрено сравнение открытых и проприетарных больших языковых моделей для генерации аннотаций к научным статьям. Сравнение было произведено по метрикам ROUGE-1, ROUGE-2, ROUGE-L и BERTScore. Была составлена база промптов. Затем была отобрана комбинация лучшего промпта и модели. Полученные результаты могут быть использованы для обогащения баз данных научных статей.

Ключевые слова: большие языковые модели, аннотация, промпт, научные статьи, суммаризация, генерация текста.

Введение. В современном мире существует проблема точной аккумуляции научного знания в единых базах данных. Существующие проекты агрегируют записи о научных статьях, однако существует проблема отсутствия аннотаций к некоторым научным статьям по различным причинам, что обедняет эти базы данных.

Так, например, в наиболее полной открытой базе научных статей openAlex [1] содержится 2 340 000 научных статей на русском языке. Из них только 724 100 имеют аннотации. Или же 30,94% от всего числа статей на русском языке.

Наличие аннотации является зачастую обязательным условием научной публикации. Традиционно максимальная длина аннотации должна быть не более 500 знаков. Аннотация должна придерживаться следующей структуры: наличие характеристики основной темы статьи, затем краткая формулировка рассматриваемой проблемы, указание на новизну в сравнении с другими материалами, также указание на полученные результаты в заключении. Эти требования были учтены при формировании базы промптов.

Таким образом, задачей исследования было определение сочетания модели и промпта для лучшей генерации аннотации к научной статье.

Задача аннотирования — задача не новая, но имеет свои особенности. К примеру, в работах может присутствовать текст или картинки, значимые для понимания статьи, и это важно учитывать при оценке результатов.

Обзор публикаций, посвященных теме исследования, позволяет выделить несколько статей, и первая из них – «Guiding Large Language Models via External Attention Prompting for Scientific Extreme Summarization» (Chang Y., Li Z., Le X.) [2], посвящена решению задачи суммаризации научного текста. В ней авторы сосредоточились на экстремальной суммаризации аннотации до одного абзаца с помощью промптинга больших языковых моделей. Особенностью их промптов является использование языка разметки «Markdown», которым они подчеркивают важность некоторых частей промпта, например – заголовка или ключевых слов. Кроме того, авторы решают задачу для английского языка.

В статье «Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке» (Мельничук Д. В., Носкина А. В.) [3], авторы описывают процесс создания

аннотации по тексту научной статьи, а затем сравнение ее с авторской с помощью метрик BLEU, ROUGE-1, ROUGE-2, ROUGE-L, Perplexity. Однако, они предлагают подход, при котором используются специализированные модели для суммаризации: mBART, T5, GPT-3. Наш же подход использует подбор промпта для генерации аннотации с использованием возможностей генеративных больших языковых моделей.

Среди существующих сервисов, решающих похожую задачу, можно выделить Scholarcy [4], который позволяет суммаризировать научные статьи для подготовки обзора. Сервис также основан на работе с большими языковыми моделями, по типу GPT-4. Однако, он сфокусирован на иной целевой аудитории, не решает задачу генерации аннотации к научным статьям, а также не делает акцент на русском языке. Другой сервис — SciSummary [5] выглядит более простым, чем Scholarcy. Сервис позволяет только суммаризировать научные статьи. Он не решает задачу генерации аннотации к научным статьям и не делает акцент на русском языке.

В ходе нашего исследования для измерения качества генерации использовались метрики ROUGE-1, ROUGE-2, ROUGE-L и BERTScore. Метрики ROUGE основаны на статистике. Так метрики ROUGE-1 и ROUGE-2 оценивают количества совпадений по соответствующим N-граммам. Метрика ROUGE-L оценивает наибольшую общую последовательность [6]. Метрика BERTScore основана на оценке схожести эмбеддингов двух текстов. По оригинальному и сгенерированному тексту вычисляются эмбеддинги с помощью модели BERT. Затем для двух эмбеддингов вычисляется косинусное сходство [7]. Все метрики выдают оценки от нуля до единицы.

Проблема исследования. Основная проблема исследования заключается в необходимости определить наилучшее сочетание модели и промпта для задачи генерации аннотации к научной статье по ее тексту. Основной задачей было сравнение авторской аннотации и сгенерированной с помощью метрик: ROUGE-1, ROUGE-2, ROUGE-L и BERTScore и обоснование выбора лучшей модели и промпта (пример использованных промптов представлен на рис. 1).

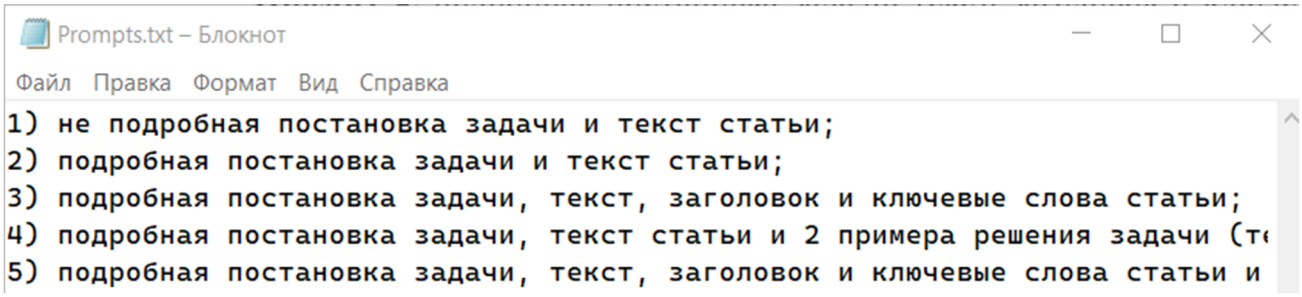

Рис. 1. Варианты промптов для решения задачи

Материалы и методы. Для выполнения исследования была использована облачная платформа Google Collab и инструмент с открытым исходным кодом Ollama, позволяющий запускать LLM локально для моделей «Gemma3 4b» и «T-lite-1.0». Также мы использовали API для обращения к сервисам «Google AI Studio» и «OVH Cloud» для получения генераций для моделей «Gemini 2.0 flash» и «Llama 3.3 70b» соответственно. В качестве датасета был

выбран датасет «Russian scientific articles» (Русские научные статьи), размещенный на платформе Kaggle [8]. Все модели, кроме «Gemini 2.0 flash», имеют открытые веса.

Тестовая платформа для проведения наших экспериментов была построена на основе нескольких Jupyter-ноутбуков. Каждый отвечал за свою часть работы (рис. 2), что позволило избежать выполнения уже сделанной работы дважды и лишних временных и вычислительных затрат.

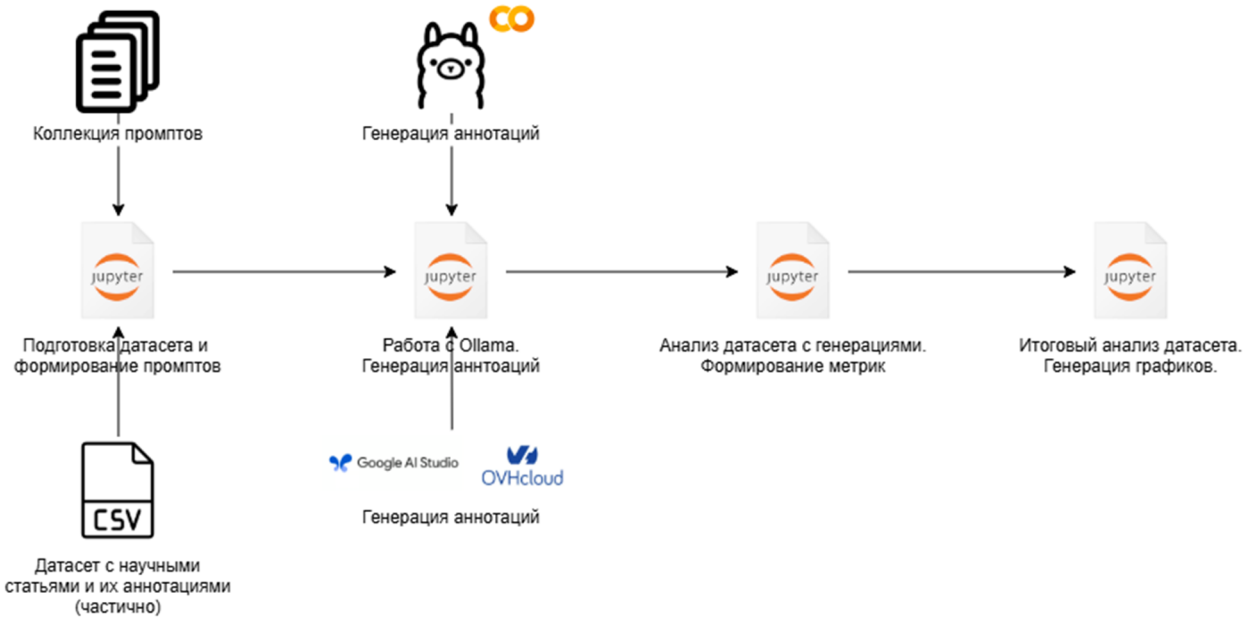

Рис. 2. Схема работы

В первом ноутбуке – «Подготовка датасета и формирование промптов», загружается датасет, выполняется предварительная обработка данных и создание готовых для последующей загрузки в нейросеть текстов, в которых объединяются промпт и текст аннотируемой статьи. Результат выгружается в формате csv.

Во втором ноутбуке – «Работа с Ollama и API. Генерация аннотаций», загружается результат предыдущего ноутбука и для каждой модели генерируется аннотация по соответствующему тексту на основе промпта. Результат также выгружается в формате csv.

В третьем ноутбуке – «Анализ датасета с генерациями. Формирование метрик», для каждой аннотации и по каждой модели получаем метрики суммаризации ROUGE и BERTScore, сравнивая каждую генерацию с авторской из датасета.

В последнем четвертом ноутбуке – «Итоговый анализ датасета» результат предыдущего ноутбука используется для итогового анализа. Суммируя все метрики, группируя их по модели и промпту, находится среднее и создаются графики для наглядного сравнения, что и позволяет выбрать наилучшее сочетание модели и промпта.

Результаты. В результате работы была выявлена комбинация лучшего промпта и лучшей модели как для открытых, так и для закрытых моделей. Лучшим промптом оказался промпт номер четыре, а лучшей моделью «Gemini 2.0 flash». Среди открытых моделей, лучшей моделью оказалась модель «Llama 3.3 70B» с комбинацией промпта номер два. Модель

«Gemma 3 4b» демонстрировала результаты хуже. Важно обратить внимание, что результаты по метрикам были незначительно хуже результатов модели «Llama 3.3 70B» несмотря на разницу в количестве параметров в 17,5 раз. 70 миллиардов параметров против 4 миллиардов параметров. Вероятно, это связано с наличием большего количества русскоязычных текстов в обучающем корпусе «Gemma 3 4b», по сравнению с «Llama 3.3 70B». Кроме того, стоит обратить внимание, что открытые модели в целом справились с задачей хуже по сравнению с закрытой моделью «Gemini 2.0 flash». Модель «T-lite-1.0» показала средний результат. Все результаты по всем комбинациям промптов и моделей представлены на рис. 3.

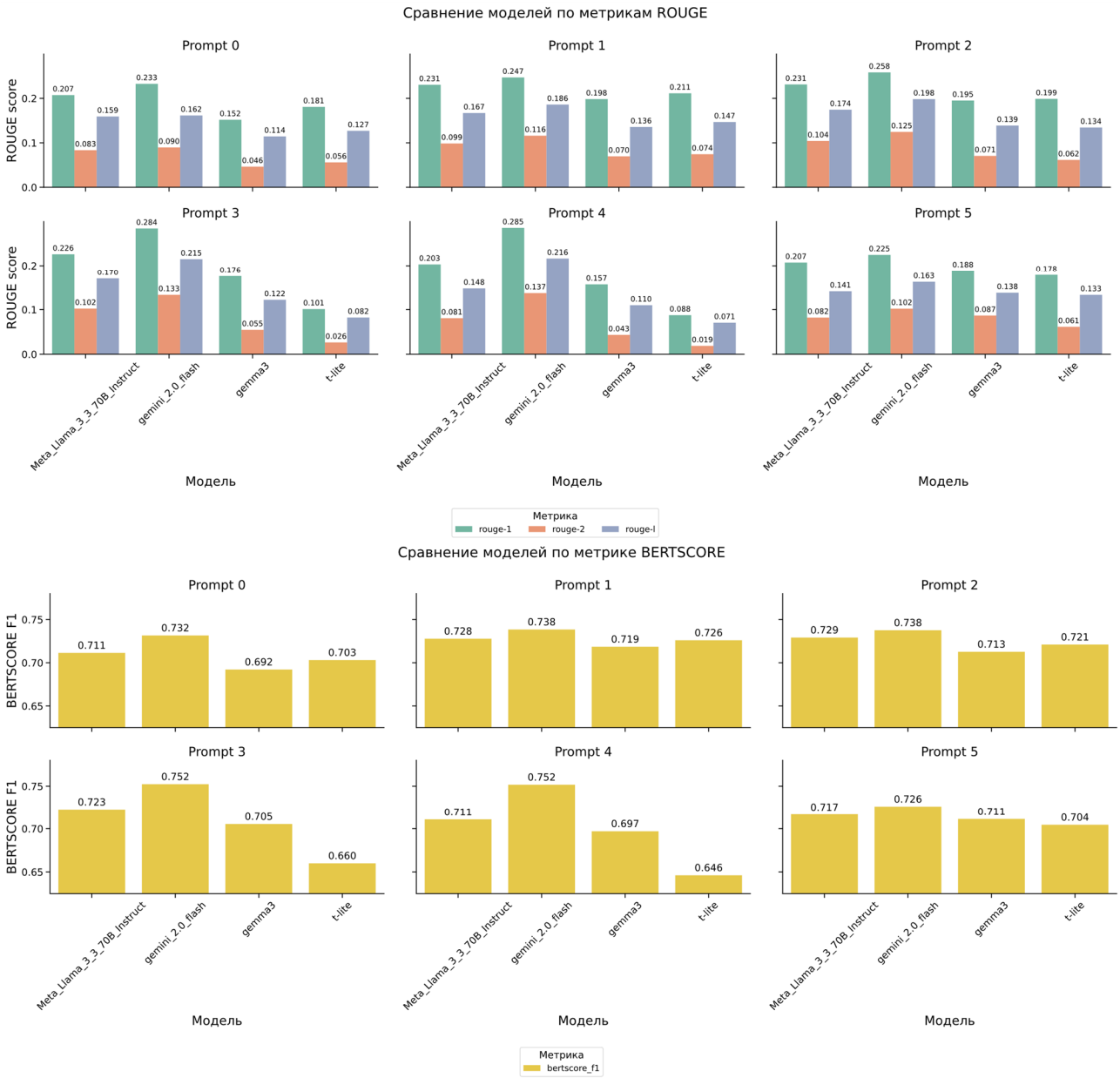

Рис. 2. Сравнение различных моделей по метрикам ROUGE и BERTSCORE

Рассмотрев результаты по промптам, сделаем следующие предположения по влиянию промпта на результат: для рассмотренной задачи необходимо использовать промпт с подробным описанием задачи. Это улучшает метрики. Также, было обнаружено, что увеличение количества параметров модели не всегда значимо улучшает значения метрик.

В целях оценивания практической значимости найденной комбинации модели и промпта, был создан сервис на StreamLit, позволяющий получить аннотацию для статьи, загруженной из PDF или введенной в поле ввода. Обработка происходит на StreamLit, затем запрос отправляется платформе, выполняющей вычисления. Далее ответ показывается пользователю (рис. 4).

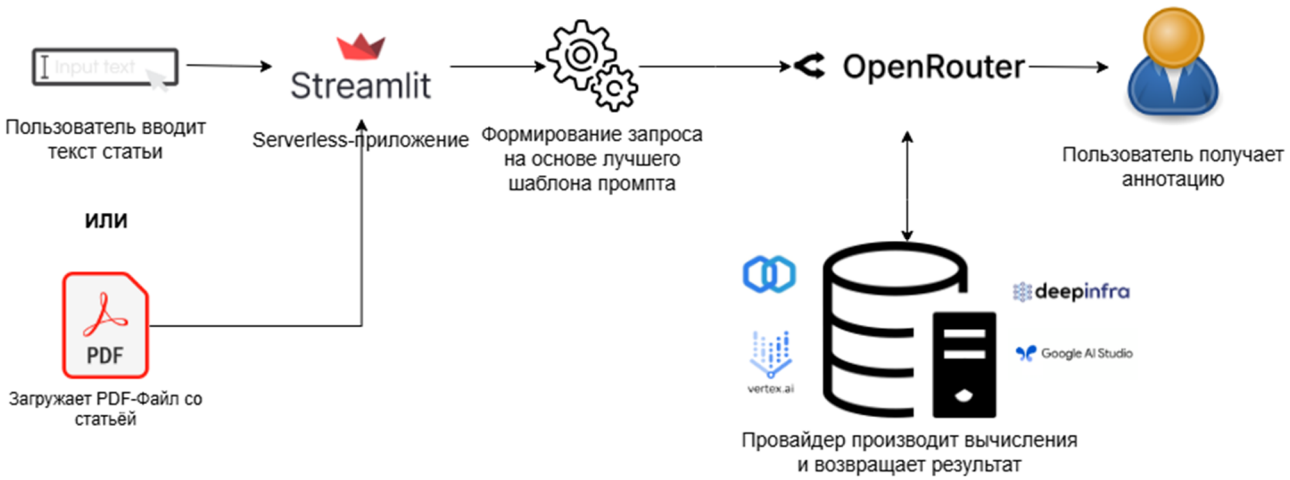

Рис. 4. Архитектура демо-приложения

Заключение. В результате работы была решена задача создания аннотации по тексту научной статьи, составлена база промптов и отобрана модель «Gemini 2.0 flash», как наиболее подходящая для аннотирования научных текстов. Далее для этой модели был определен оптимальный промпт, дающий лучшие результаты (Средние значения метрик: rouge-1: 0.2854; rouge-2: 0.1370; rouge-l: 0.2164; bertscore_f1: 0.7517). Были сделаны предположения по влиянию промпта на результат.

Также в качестве демонстрации возможности практического использования решения, был разработан сервис на платформе Streamlit, позволяющий получить аннотацию по тексту статьи.

Тем не менее, необходимо сказать и об ограничениях этой работы. Из-за ограничения на количество вычислительного времени мы полноценно проанализировали всего 100 статей из датасета в 2400 комбинациях модели и промпта. Таким образом, необходимо проведение дальнейшего исследования на большем количестве статей.

Исследование выполнено при поддержке Министерства науки и высшего образования Российской Федерации в рамках проекта "Искусственный интеллект" (FEWZ-2024-0052).

СПИСОК ЛИТЕРАТУРЫ

1. Priem J., Piwowar H., Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts // arXiv preprint arXiv:2205.01833. — 2022.

2. Chang Y., Li Z., Le X. Guiding Large Language Models via External Attention Prompting for Scientific Extreme Summarization // Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024). — 2024. — С. 226-242.
3. Мельничук Д. В., Носкина А. В. Сравнение NLP-моделей на задаче суммаризации академических текстов на русском языке // Компьютерная лингвистика и вычислительные онтологии. Выпуск 7 (Труды XXVI Международной объединенной научной конференции «Интернет и современное общество», IMS-2023, Санкт-Петербург, 26–28 июня 2023 г. Сборник научных статей). — Санкт-Петербург : Университет ИТМО, 2023. С. 54–59. DOI: 10.17586/2541-9781-2023-7-54–59
4. Scholarcy — Knowledge made simple // Scholarcy : сайт. — URL: https://www.scholarcy.com/ (дата обращения: 11.04.2025).
5. Use AI To Summarize Scientific Articles — SciSummary // SciSummary : сайт. — URL: https://scisummary.com/ (дата обращения: 11.04.2025).
6. Lin C. Y. Rouge: A package for automatic evaluation of summaries // Text summarization branches out. — 2004. — С. 74-81.
7. Zhang T. et al. Bertscore: Evaluating text generation with bert // arXiv preprint arXiv:1904.09675. — 2019.
8. Russian scientific articles // Kaggle : сайт. — URL: https://www.kaggle.com/datasets/megass/russian-scientific-articles (дата обращения: 11.04.2025).