УДК 004.85

АНАЛИЗ ПРИМЕНЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ РАБОТЫ С КОМПЛЕКСНЫМИ ДОКУМЕНТАМИ

Дедкова А.В. (ИТМО), **Маракулин А.А.** (ИТМО), **Федорова М.В.** (ИТМО)
Научный руководитель – преподаватель, Терещенко В.В.
(ИТМО)

Введение. Эффективная работа со сложными документами (техническая документация, научные отчеты, научные статьи) критически важна для ускорения и повышения качества научных исследований. Одним из ключевых вызовов современной науки является растущий объем научных публикаций. Исследователям приходится анализировать огромные массивы информации, что требует больших затрат человеческих усилий.

Эксперименты, проведенные в этой области [1], показывают, что большие языковые модели (Large language model, LLM), как инструмент автоматизации рутинных задач, способны существенно упростить такую работу и оказать значительное влияние на научные исследования и образовательные процессы. Например, проводились исследования применения LLM в задачах резюмирования [2] и извлечения данных из опубликованных исследований [3].

Цель работы – проанализировать применение больших языковых моделей с открытым исходным кодом (open-source LLM) в рамках решения задачи рецензирования результатов образовательной деятельности, представленных в виде комплексных документов.

Основная часть. В последние годы наблюдается активное развитие больших языковых моделей среднего размера (3–14 миллиардов параметров) с открытым исходным кодом, которые представляют собой компромисс между вычислительной сложностью и качеством генерации текста, что делает их доступными для исследований и практического применения в науке и образовании.

К преимуществам таких LLM можно отнести:

1. Доступность и воспроизводимость – открытый код моделей позволяет исследователям разрабатывать собственные решения, адаптируя их под конкретные задачи;
2. Сбалансированное качество – модели такого размера показывают сравнимую с более крупными моделями точность, но требуют меньше вычислительных ресурсов;
3. Гибкость дообучения – благодаря открытому доступу можно адаптировать модель под специализированные домены и задачи;
4. Эффективность на локальном оборудовании – возможность работы без облачных сервисов снижает затраты и повышает конфиденциальность данных.

Исследование включает сравнительный анализ возможностей предобученных LLM среднего размера в извлечении информации из сложных документов, поиске ошибок и структурировании данных, а также оценку влияния на качество работы моделей формата взаимодействия [4] и специализации предметной области, к которой относится анализируемый документ.

Выводы. Результаты проведенного исследования показывают, что open-source LLM среднего размера при работе со сложными документами на русском языке позволяют автоматизировать рутинные процессы и повысить эффективность обработки информации. Использование LLM позволяет значительно сократить время первоначального анализа документов. Например, модель Vikhr-Nemo-12B-Instruct-R-21-09-24 обрабатывает документ объемом 20 000 символов в среднем за 2,5 минуты, что делает применение подобных моделей эффективным инструментом для ускорения работы с текстами большой сложности.

Предобученные LLM, адаптированные для работы с русскоязычными текстами, могут выполнять начальную проверку качества представленных сложных документов, включая анализ структуры текста и выявление базовых несоответствий. В частности, LLM демонстрируют способность к автоматической верификации формальных критериев, таких как наличие ключевых разделов, логическая последовательность и соответствие формату.

Однако, на данный момент нет достаточных доказательств эффективности LLM в качестве инструмента для полноценного научного рецензирования. Тем не менее, дальнейшее тестирование программных решений в этой области может быть полезным, особенно для автоматического выявления явно не соответствующих стандартам работ, а также для предоставления авторам ранней обратной связи для улучшения их работ до отправки на рецензирование.

Дальнейшие исследования будут направлены на разработку специализированных решений на основе open-source LLM, адаптированных к работе со сложными документами и задачам обработки информации.

Список использованных источников:

1. Liu R., Shah N. B. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. // arXiv preprint arXiv:2306.00622. – 2023.
2. Wolfee C. et al. Prompting for directed content in literature summarization: Fine-tuning to steer large language models in academic text analysis //Authorea Preprints. – 2024.
3. Gartlehner G. et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. - DOI 10.1002/jrsm.1710 // Research Synthesis Methods Volume 15, Issue 4, Pages 523-699.
4. Santu S. K. K., Feng D. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks. // arXiv preprint arXiv:2305.11430. – 2023.