УДК 004.91

# АНАЛИЗ БОЛЬШИХ ДОКУМЕНТОВ ПРИ ПОМОЩИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

**Богданов М. А.** (ИТМО), **Никифоров М. А.** (ИТМО), **Аминов Н. С.** (ИТМО), **Федоров Д.А.** (ИТМО)

**Научный руководитель – кандидат технических наук, доцент Федоров Д.А.** (ИТМО)

**Введение.** Среди многочисленных документов образовательного процесса особое место занимают результаты научно-исследовательских работ, к которым относятся: курсовые работы; выпускные квалификационные работы; диссертации, а также отчеты научно-исследовательских работ.

Для обработки таких документов, в т. ч. создания кратких описаний, резюме, поиска в тексте нужной информации, извлечения таблиц и других аналогичных задач эффективно применение больших языковых моделей (Large Language Models, LLM). Однако различные LLM имеют определенные ограничения, а также потери качества и точности выходных ответов при увеличении объема документа. Например, максимальная длина текста, которая может быть обработана за один запрос, составляет всего лишь 3000 слов, в различии от модели. Кроме того, при анализе объемных документов, языковая модель «теряет внимание», что приводит к потере релевантности и слабой связанности ответов и к более частому появлению «галлюцинаций». Для больших документов языковые модели затрачивают огромные вычислительные мощности, а значит увеличивается время и стоимость выполнения задачи. Также, возможность обработки зависит от используемого языка, например слова на русском языке занимают больше токенов, чем слова на английском.

Несмотря на существующие методы и подходы решения этой проблемы, на практике возникает много сложностей. В связи с этим, актуальной задачей внедрения LLM является разработка и совершенствование инструментов анализа больших текстов.

**Основная часть**. Для преодоления указанных ограничений без ущерба для целостности информации прежде всего могут использоваться различные методы уменьшения объема текста, в т. ч. предварительная обработка текста (лемматизация, стеминг, удаление N-грамм [1], шумовых слов и предложений), разбиение на сегменты, фильтрация дубликатов или терминов или не имеющих значения и т. п. Предлагаемой оптимальной стратегией решения проблемы является комплексный подход, то есть разработка приложений, сочетающих несколько взаимодополняющих методов.

При проведении анализа нескольких текстов через языковые модели, которым задавали различного рода вопросы по приведенному тексту на русском языке. Результат показал, что языковые модели имеют несколько типов ответов:

1. Верный ответ - языковая модель дала полный ответ, который содержит в себе точную верную информацию по заданному тексту. Такие ответы делились на отрицательные и положительные ответы.
2. Ложный ответ - языковая модель может утверждать, что указанной информации в тексте нет, что является ложью.
3. Неполный ответ - искусственный интеллект может предоставить верный ответ, при этом не предоставив часть информации из текста
4. Искаженный ответ - языковая модель преподносит информацию, которой нет в источнике. Также в данный раздел попадали выдуманные ответы и отсутствие ответа на поставленный вопрос.

Данное исследование показало, что при анализе документов языковая модель может дать неточный ответ или же вовсе искаженный.

Для решения вышеуказанных проблем есть несколько различных методов:

производительность и минимизировать ресурсные затраты. Выполнение семантического поиска [3] в документе с передачей в OpenAI наиболее релевантных вложений. Для улучшения эффективности языковых моделей можно использовать параллельные вычисления [4], чтобы увеличить эффективность работы с большими промптами. Требуется генерировать точные ответы. Ответ, который получает пользователь, должен указывать на источник информации, например, создавать ссылку на предложение из текста. Это увеличивает доверие к предоставленным данным и облегчает поиск нужной информации. На данный момент у языковых моделей увеличивается количество возможных токенов (GPT-3.5 - 4096 токенов, GPT-4 - 32768 токенов [5]), что позволит анализировать тексты большего размера.

**Выводы.** Применение языковых моделей типа GPT для анализа объемных тематических текстовых документов в образовательном процессе имеет свои ограничения и трудности. Основные проблемы связаны с ограничением объема текста, которое может быть обработано за один запрос, потерей релевантности и связанности результатов. Используя приведенные методы для усовершенствования анализа текста, появилась возможность улучшить процесс распознавания текста и его анализа

В будущем, усовершенствование алгоритмов обработки и адаптация моделей под конкретные типы документов и языки могут значительно повысить эффективность анализа больших текстов в образовательной сфере.

**Список использованных источников**:

1. Almgerbi M. et al. Improving Topic Modeling Performance through N-gram Removal //IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. – 2021. – С. 162-169.
2. Liu S., Healey C. G. Abstractive Summarization of Large Document Collections Using GPT //arXiv preprint arXiv:2310.05690. – 2023.
3. Bast H. et al. Semantic search on text and knowledge bases //Foundations and Trends in Information Retrieval. – 2016. – Т. 10. – №. 2-3. – С. 119-271.
4. Tekiner F. et al. Parallel text mining for large text processing //Proceedings of IEEE CSNDSP2010. – 2010. – С. 348-353.
5. Kalyan K. S. A survey of GPT-3 family large language models including ChatGPT and GPT-4 //Natural Language Processing Journal. – 2023. – С. 100048.