ВЕСТНИК НАЦИОНАЛЬНОГО ИССЛЕДОВАТЕЛЬСКОГО ЯДЕРНОГО УНИВЕРСИТЕТА «МИФИ», 2024, т. 13, № 5, с. 350–357

ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА

УДК 004.658.2

МЕТОДИКА ФОРМИРОВАНИЯ БАЗЫ ДАННЫХ ХАРАКТЕРИСТИК СЛОЖНОГО ТЕХНОЛОГИЧЕСКОГО ОБЪЕКТА С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

А.Р. Матвеева*, Е.В. Антонов
Национальный исследовательский ядерный университет «МИФИ», Москва, 115409, Россия
**e-mail: armatveeeva@gmail.com*

Поступила в редакцию: 05.07.2024
После доработки: 21.08.2024
Принята к публикации: 10.09. 2024

Атомная энергетика играет важную роль в обеспечении безопасности многих стран мира. При проектировании и эксплуатации таких сложных технологических объектов (СТО), как атомные электростанции (АЭС), критически важно учитывать их характеристики для обеспечения безопасной работы. Актуальность темы исследования заключается в необходимости разработки методики, позволяющей ускорить процесс выявления целевой информации, содержащейся в научных публикациях, для предприятий атомной отрасли. Отсутствие научных работ, описывающих применение языковых моделей для анализа и выделения характеристик из сложных технологических объектов, подчеркивает необходимость проведения исследования. В работе в качестве примера такого объекта выбрана АЭС. Для проведения ряда экспериментов по выделению технических характеристик СТО составлен перечень параметров профиля атомной электростанции (35 параметров) и сформирован набор данных по атомным электростанциям (60 научных публикаций, содержащих сведения об АЭС Линьао). Разработана программа, которая позволяет обрабатывать содержащиеся в научных публикациях данные путем загрузки статей в языковую модель, написания запросов и получения ответов для последующего составления профиля сложного технологического объекта. Результаты работы показали, что предложенная методика позволяет программно обрабатывать научные публикации для составления профиля АЭС.

Ключевые слова: большие языковые модели, атомная электростанция, научные публикации, автоматизированная обработка данных.
DOI: 10.26583/vestnik.2024.5.7
EDN PJFXNC

ВВЕДЕНИЕ

В настоящее время большой объем данных о сложных технологических объектах содержится в научных публикациях, но поиск и извлечение необходимых параметров из текстов статей представляет собой трудоемкий процесс [1]. С развитием технологий искусственного интеллекта появилась возможность использования больших языковых моделей для автоматизированного сбора и обработки больших массивов данных.

В современном контексте научно-технического прогресса активно развиваются большие языковые модели, представляющие собой нейросетевые архитектуры для анализа большого объема информации, основная масса которого представляет собой неструктурированные данные [2, 3]. С увеличением объема тестовых данных наблюдаются перманентное усовершенствование и расширение существующих алгоритмов. С течением времени становится возможной оценка их внедрения в актуальне научные задачи. Учитывая рост объема информации в интернете в геометрической прогрессии, первостепенной задачей многих исследователей становится качество обработки данных, включая их структурирование [4].

В рамках представленного исследования рассматриваются несколько больших языковых моделей таких, как модели GPT (Generative Pretrained Transformers), являющихся мощными и продвинутыми нейросетевыми архитектурами, обученными на огромных массивах текстовых источников [5–9]. GPT демонстрируют возможность для повышения производительности за счет автоматизации таких трудоемких задач,

как создание контента (написание сценариев или художественных произведений), составление документов (исковые заявления в юридических компаниях или сопроводительные письма для отправки резюме в компаниях), составление плана обучения, оказание технической помощи в чатах поддержки, оказание помощи в написании кода для программистов [10, 11]. Ряд исследований демонстрирует применимость GPT для извлечения данных из научных публикаций, например, в работе «Pre-trained Language Models in Biomedical Domain: A Systematic Survey» авторы представляют обзор применения языковых моделей для извлечения информации из научных публикаций в области биомедицины [12]. В статье обсуждается применимость моделей машинного обучения для решения комплекса задач в области биомедицины, например, информационный поиск, классификация текстов или извлечение сущностей.

В статье «Challenges and Advances in Information Extraction from Scientific Literature: a Review» рассматриваются проблемы и перспективы автоматизированного извлечения информации из научных публикаций в области материаловедения [13]. Отмечается, что именно в научных статьях, справочниках и лабораторных журналах содержится большой объем экспериментальных данных, однако извлечение такого рода информации вручную из массива публикаций трудоемко. Авторы подробно рассматривают основные препятствия, лежащие в основе практического применения технологий информационного поиска к научным публикациям: формат файлов, дефицит или отсутствие структурированности данных, нарушение целостности представления полезной информации, а также сложность адаптации языковых моделей к научному стилю речи. Извлечение информации представляет собой важный шаг к использованию моделей машинного обучения для содействия научным и инженерным исследованиям. Решение проблем, которые в настоящее время препятствуют возможности извлекать информацию из документов, свидетельствует о перспективах дальнейшей разработки и совершенствования методик автоматизированного анализа данных искусственным интеллектом в материаловедении и инженерии.

В работе «Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing» подчеркивается, что социальные детерминанты здоровья редко доступны в структурированных электронных медицинских картах, чаще упоминаются в неструктурированных клинических записях. Однако извлечение подобной информации ручным способом является трудоемким и затратным по времени процессом [14]. Авторы статьи предлагают использовать для этой цели современные методы обработки естественного языка. Статья демонстрирует применимость языковых моделей на основе глубокого обучения для анализа научных текстов и извлечения необходимых характеристик, в данном случае – факторов, влияющих на здоровье людей.

В статье «Natural Language Processing for Swedish Nuclear Power Plants» описывается исследование, в котором рассмотрены возможности использования методов обработки естественного языка для анализа текстовых данных и извлечения из них пользы в шведской атомной промышленности [15]. Исследование сосредоточено на интервью с представителями шведских и международных компаний в сфере ядерной энергетики для определения текущих потребностей, проблем и возможностей использования обработки естественного языка для обслуживания АЭС.

В исследовании с целью разработки методики формирования базы данных характеристик сложного технологического объекта с использованием программных средств в качестве примера СТО выбрана атомная электростанция. Языковые модели, в частности, адаптированные к предметной области, могли бы внести важный вклад в понимание и оптимизацию рабочих процессов в ядерной энергетике за счет автоматизации анализа больших массивов данных.

Обзор литературы позволяет сделать вывод, что в представленных источниках не обнаружено удовлетворительного ответа на поставленный вопрос, касающийся возможности применения языковой модели для выделения характеристик атомной электростанции. Была показана применимость моделей в таких сферах, как атомная энергетика, биомедицина и материаловедение. Отсутствие научных работ, описывающих применение языковых моделей для анализа и выделения характеристик атомных электростанций, подчеркивает необходимость проведения исследования.

МЕТОДИКА

В рамках исследования была разработана методика формирования базы данных характе-

ристик атомных электростанций, основанная на применении анализа текстовых данных с использованием языковых моделей (Рис. 1).

Этапы предложенной методики направлены на анализ текстовых данных и формирование информационных профилей сложного технологического объекта, обогащая тем самым процесс принятия решений с учетом более полной и точной информации.

Рассмотрим каждый этап предложенной методики формирования базы данных характеристик сложного технологического объекта.

На первом этапе осуществляется подбор исходных материалов. В качестве входных данных рассматриваются документы различных форматов, например, PDF, HTML, TXT и др., которые охватывают спектр информации о научных исследованиях, технической документации, отчетах и других источниках, содержащих значимую информацию о функционировании исследуемого объекта. Рассматриваемый этап направлен на обеспечение многостороннего охвата научно-технического контекста, необходимого для последующего детального анализа. В результате отбора формируется набор входных данных, предоставляющих базу для детального исследования характеристик.

Рис. 1. Методика формирования профиля сложного технологического объекта

В рамках предложенной методики на втором этапе осуществляется выделение текстовых данных или загрузка полного текста научной публикации. Рассматриваемый процесс охватывает разделение отобранных текстов на такие разделы, как введение, эксперимент, заключение и приложения. Выделение частей научных публикаций позволяет извлечь технические параметры сложных технологических объектов. Второй этап важен, поскольку не только обеспечивает структурирование данных для дальнейшего анализа, но и выделяет ключевые компоненты текста, на которые будет сфокусирован последующий интеллектуальный поиск, способствуя детальному исследованию характеристик и особенностей функционирования атомных электростанций. Также на данном этапе происходит разделение текста на блоки в соответствии с параметрами языковой модели. Рассматриваемый процесс необходим, поскольку языковые модели имеют ограничения на объем текста, который может быть принят для реализации последующей работы. Разделение текста на смысловые блоки позволяет управлять ограничениями по объему текста, предоставляя возможность детального и точного анализа каждого сегмента информации. Разбираемый подход обеспечивает не только соответствие требованиям моделей, но и сохранение семантической целостности каждого выделенного блока, что

является ключевым моментом при последующем интеллектуальном анализе.

На третьем этапе методики происходит разработка запросов (промтов), направленных на извлечение характеристик СТО. В рамках предварительно составленного перечня интересующих характеристик создаются запросы, направленные на извлечение соответствующей информации из текстовых блоков научных публикаций. Этап требует анализа и определения технических параметров, которые важны для формирования информационного профиля сложного технологического объекта. Созданные запросы, которые основаны на понимании предметной области, становятся инструментом для выделения релевантных характеристик, предоставляя возможность целенаправленного анализа научных статей.

На следующем этапе методики осуществляется внедрение созданных запросов в языковую модель. Рассматриваемый этап предполагает фактическую реализацию разработанных запросов, направленных на выделение интересующих характеристик атомной электростанции с использованием выбранной языковой модели. Этап является ключевым в анализе, так как именно здесь происходит взаимодействие с технологическими возможностями модели, способной выделить и интерпретировать содержание текста в соответствии с предложенными запросами.

На пятом этапе методики происходит обработка полученных результатов, представленных языковой моделью. Основная задача этого этапа – провести анализ выделенных данных с целью извлечения информации о характеристиках атомной электростанции.

Важным аспектом для рассмотрения становится вопрос о том, как отличить действительные данные от возможных галлюцинаций, которые могут возникнуть в результате особенностей работы модели. Для решения возникшей проблемы предлагаются определенные пути решения:

- использование статистики, ориентированной на характеристики сложного технологического объекта, с установлением пороговых значений для количества рассматриваемых научных публикаций;
- использование экспертного мнения, позволяющего провести проверку данных и выявить возможные галлюцинации большой языковой модели. Эксперт, опираясь на свой профессиональный опыт, может дать оценку достоверности информации, что повышает уровень доверия к результатам анализа;
- использование таких реферированных исходных данных, как рецензируемые научные публикации, официальные отчеты, подготовленные экспертами в соответствующей предметной области, поскольку эти источники подвергаются тщательной экспертной оценке и строгому рецензированию, что гарантирует высокий уровень валидности представленной информации;
- помимо вышеописанных способов решения проблемы, связанной с галлюцинациями, необходимо упомянуть о надстройке (корректировка параметров большой языковой модели для оптимизации ее работы) нейросетевой архитектуры согласно относящейся к ней документации.

Таким образом, обработка результатов запросов представляет этап, который предоставляет механизмы для выявления и фильтрации галлюцинаций в отношении характеристик сложного технологического объекта.

На заключительном этапе методики выполняется процесс загрузки обработанных данных в базу данных. Основываясь на результатах анализа и обработки запросов языковой моделью, формируется итоговый профиль атомной электростанции. Рассматриваемый профиль включает в себя разнообразные элементы, среди которых название сложного технологического объекта, его технические характеристики, результаты выделения ключевых характеристик атомных электростанций, а также научные публикации, связанные с выбранным сложным технологическим объектом.

База данных служит центральным хранилищем для практичного хранения полученных данных, обеспечивая доступ к информации для последующего использования. Профиль сложного технологического объекта представляет собой набор данных, обогащенный аналитической информацией, что позволяет представить характеристики и особенности рассматриваемой электростанции.

В рамках апробации предложенной методики проведены:

1) сравнительный анализ доступных больших языковых моделей;

2) автоматическое выделение характеристик атомной электростанции из научных публикаций с использованием языковых моделей для выбора нейросетевой архитектуры, которая будет развернута локально;

3) автоматическое выделение характеристик атомной электростанции из полного текста научных публикаций с поочередной загрузкой и последующей обработкой каждого файла по очереди с использованием языковой модели (ход эксперимента: загрузка одной научной публикации, написание запросов для получения 35 характеристик, получение ответов на запросы, удаление рассматриваемой статьи, загрузка следующего файла и т.д.);

4) автоматическое выделение характеристик атомной электростанции из полного текста научных публикаций с одновременной загрузкой и последующей одновременной обработкой всех файлов с использованием языковой модели (ход эксперимента: загрузка 60 научных публикаций, написание запросов для получения 35 характеристик, получение ответов на запрос, удаление всех файлов);

5) автоматическое выделение характеристик атомной электростанции из полного текста научных публикаций с поочередной загрузкой файлов и обработкой каждого из них пять раз подряд с использованием языковой модели (ход эксперимента: загрузка 60 научных публикаций, написание запросов пять раз подряд для получения 35 характеристик, получение ответов на запрос, удаление всех файлов).

Проведенные эксперименты с целью определения наилучшего подхода для автоматического выделения характеристик сложного технологического объекта из научных публикаций позволили оценить различные методы обработки данных. Выделенные характеристики применены для тестового наполнения базы данных, что предоставило возможность оценить предложенную методику формирования базы данных характеристик сложного технологического объекта.

РЕЗУЛЬТАТЫ

Для выделения характеристик атомной электростанции из научных публикаций на основе критериев (доступность, языки обучения, работа с текстовыми файлами, объем принимаемого текста, количество запросов в диалоге, наличие API, opensource) выбраны такие языковые модели, как Mistral 7B (PrivateGPT), ChatGPT и Claude.

Последующий эксперимент заключался в автоматическом выделении характеристик атомной электростанции из научных публикаций с использованием языковых моделей, последние продемонстрировали способность извлекать фактическую информацию из статей, но также выявили некоторые области для улучшения [2]. В итоге для создания и апробации методики выбрана Mistral 7B (PrivateGPT), которая для проведения исследования была развернута локально, на кафедре № 65 «Анализ конкурентных систем» ФГАОУ ВО «Национальный исследовательский ядерный университет «МИФИ». Для применения искусственного интеллекта необходимы такие технические характеристики компьютера, как:

- операционная система Microsoft Windows 10 Pro;
- процессор Intel(R) Core(TM) i7-6850K CPU @ 3.60GHz, 3601 МГц, ядер – 6, логических процессоров – 12;
- установленная оперативная память (RAM) 64,0 ГБ.

Рассматриваемые характеристики демонстрируют, что компьютер оснащен достаточной вычислительной мощностью и объемом оперативной памяти для работы с языковыми моделями.

Проведенные эксперименты, направленные на выделение характеристик сложных технологических объектов из текстовых массивов данных, для последующего тестового наполнения базы данных продемонстрировали [3–5]:

- в результате первого эксперимента составлен профиль атомной электростанции Линьао, состоящий из 25 извлеченных языковой моделей характеристик, время выполнения эксперимента составило 22 часа;
- в результате второго эксперимента получен профиль атомной электростанции Линьао, состоящий из 18 характеристик, время выполнения эксперимента составило два часа;
- в результате третьего эксперимента получен профиль атомной электростанции Линьао, состоящий из 10 характеристик, время выполнения эксперимента составило три часа.

Таким образом, автоматическое выделение характеристик атомной электростанции из полного текста научных публикаций с загрузкой и последующей обработкой файлов по очереди с использованием языковой модели Mistral 7B (PrivateGPT) показало наилучшие результаты в сравнении с другими проведенными экспериментами (извлечены 25 из 35 запрашиваемых характеристик) без галлюцинаций благодаря раннее проведенной надстройке большой языковой модели. Для повышения качества работы модели и получения полноценного информационного профиля СТО необходимо расширение

ее знаний в соответствующей области, проведение настройки алгоритмов, внедрение дополнительных механизмов проверки извлеченных данных, а также увеличение объема рассматриваемых данных.

Для проверки корректности вышеописанных действий выполнено тестовое наполнение базы данных, полученной в результате проведения второго эксперимента, информацией (60 научных публикаций, 35 характеристик, 36 запросов, 2160 ответов).

Результаты апробации подтвердили корректность работы разработанной методики и ее готовность к использованию с большим массивом информации. Применение искусственного интеллекта позволяет извлекать необходимые данные из большого количества научных статей, в частности, характеристики атомных электростанций. Для практического использования полученных результатов можно рекомендовать применение больших языковых моделей в информационных системах предприятий атомной отрасли для ускорения обработки и анализа больших массивов информации.

ЗАКЛЮЧЕНИЕ

В представленном исследовании разработана и апробирована методика формирования базы данных характеристик сложного технологического объекта на примере атомной электростанции. Методика основана на применении больших языковых моделей для автоматизированного извлечения необходимой информации из научных публикаций.

Проведенные эксперименты продемонстрировали способность языковых моделей выделять характеристики сложных технологических объектов из текстовых массивов данных. Наилучшие результаты достигнуты при обработке публикаций по очереди с использованием языковой модели Mistral 7B (PrivateGPT), что позволило извлечь 25 из 35 запрашиваемых характеристик из 60 научных публикаций.

Разработанная методика включает этапы сбора входных данных, выделения и разделения текста на блоки, составления запросов к языковой модели, обработки результатов и загрузки данных в базу. В рамках апробации подтверждена работоспособность предложенного подхода.

Применение искусственного интеллекта и больших языковых моделей открывает возможности для ускорения процесса выявления целевой информации, содержащейся в научных публикациях. Предложенная методика может быть использована в информационных системах предприятий для обработки больших массивов данных.

БЛАГОДАРНОСТИ

Автор выражает искреннюю благодарность и признательность за неоценимую поддержку и профессиональное руководство коллективу кафедры № 65 «Анализ конкурентных систем» ФГАОУ ВО «Национальный исследовательский ядерный университет «МИФИ».

СПИСОК ЛИТЕРАТУРЫ

1. *Polak M.P., Morgan D.* Extracting accurate materials data from research papers with conversational language models and prompt engineering // Nature Communications, 2024.V. 15(1), 1569.

2. *Yao Y., Duan J., Xu K., Cai Y., Sun Z., Zhang Y.* A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly // High-Confidence Computing, 2024. 100211.

3. *Sui Y., Zhou M., Zhou M., Han S., Zhang D.* Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study // Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024. P. 645–654.

4. *Федорова А.А.* Необходимые правила потребления информации для снижения негативного влияния информационного общества // Скиф. Вопросы студенческой науки, 2020. № 5–1. С. 157–162.

5. *Jiang A.Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D.S., etc.* Mistral 7B. arXiv preprint arXiv:2310.06825.

6. *Ali A.H., Alajanbi M., Yaseen M.G., Abed S.A.* Chatgpt4, DALL· E, Bard, Claude, BERT: Open Possibilities // Babylonian Journal of Machine Learning, 2023. P. 17–18.

7. *Dubois Y., Li C.X., Taori R., Zhang T., Gulrajani I., Ba J., etc.* AlpacaFarm: A simulation framework for methods that learn from human feedback // Advances in Neural Information Processing Systems, .2024. V. 1. URL: https://doi.org/10.48550/arXiv.2305.14387

8. *Chaka C.* Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools // Journal of Applied Learning and Teaching, 2023. V. 6(2). P. 1–11.

9. *Кумратова А.М., Морозова Н.В., Василенко А.И., Когай И.Е.* Анализ возможностей нейронной сети на основе языковой модели GPT-3 и способы ее применения на производстве // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки, 2023. Вып. 1 (316). С. 80–85.

10. *Zhan T., Shi C., Shi Y., Li H., Lin Y.* Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3) // Applied and Computational Engineering, 2024. V.67(1). P.41-47.

11. *Yenduri G., Ramalingam M., Selvi G.C., Supriya Y., Srivastava G., etc.* GPT (generative pre-trained transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions // IEEE Access, 2023. DOI: 10.1109/ACCESS.2024.3389497

12. *Wang B., Xie Q., Pei J., Chen Z., Tiwari P., Li Z., Fu J.* Pre-trained language models in biomedical domain: A systematic survey // ACM Computing Surveys, 2023. V. 56(3), P. 1–52.

13. *Hong Z., Ward, L., Chard, K., Blaiszik, B., Foster I.* Challenges and advances in information extraction from scientific literature: a review // JOM, 2021. V. 73(11), P. 3383–3400.

14. *Han S., Zhang R.F., Shi L., Richie R., Li, H., Tseng A., etc*. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing // Journal of biomedical informatics, 2022. V. 127. 103984.

15. Kåhrström, F. Natural Language Processing for Swedish Nuclear Power Plants: A study of the challenges of applying Natural language processing in Operations and Maintenance and how BERT can be used in this industry. Visby: Uppsala Universitet, 2022. URL: http://uu.diva-portal.org/smash/get/diva2: 1678697/ FULLTEXT01.pdf

Vestnik Natsional'nogo Issledovatel'skogo Yadernogo Universiteta «MIFI», 2024, vol. 13, no. 5, pp. 350–357

METHODOLOGY FOR FORMING A DATABASE OF CHARACTERISTICS OF A COMPLEX TECHNOLOGICAL OBJECT USING LARGE LANGUAGE MODELS

A.R. Matveeva*, E.V. Antonov
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, 115409, Russia
**e-mail: armatveeeva@gmail.com*

Received July 05, 2024; revised August 21, 2024; accepted September 10, 2024

Nuclear energy plays an important role in ensuring the safety of many countries in the world. When designing and operating complex technological objects (CTO) such as nuclear power plants (NPP), it is critical to take into account their characteristics to ensure safe operation. The relevance of the research topic lies in the need to develop a methodology that can speed up the process of identifying target information contained in scientific publications for nuclear industry enterprises. The lack of scientific papers describing the use of language models for analyzing and extracting characteristics from complex technological objects emphasizes the need for research. In this paper, a NPP is chosen as an example of such an object. To conduct a series of experiments to identify the technical characteristics of the CTO, a list of parameters of the nuclear power plant profile (35 parameters) was compiled and a data set on nuclear power plants was formed (60 scientific publications containing information about the Ling Ao NPP). A program was developed that allows processing the data contained in scientific publications by loading articles into a language model, writing queries and receiving responses for subsequent compilation of a profile of a complex technological object. The results of the work showed that the proposed technique allows programmatic processing of scientific publications to compile a profile of a NPP.

Keywords: large language models, nuclear power plant, scientific publications, automated data processing.

REFERENSES

1. *Polak M.P., Morgan D.* Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 2024. Vol. 15(1), 1569.

2. *Yao Y., Duan J., Xu K., Cai Y., Sun Z., Zhang Y.* A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 2024. 100211.

3. *Sui Y., Zhou M., Zhou M., Han S., Zhang D.* Table meets LLM: Can large language models understand structured table data? A benchmark and empirical study. Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024. Pp. 645–654.

4. *Fedorova A.A.* Neobhodimye pravila potrebleniya informacii dlya snizheniya negativnogo vliyaniya informacionnogo obshchestva [Necessary rules for information consumption to reduce the negative impact of the information society]. Skif. Voprosy studencheskoj nauki, 2020. No. 5–1. Pp. 157–162 (in Russian).

5. *Jiang A.Q., Sablayrolles A., Mensch A., Bamford C., Chaplot D.S., etc*. Mistral 7B. arXiv preprint arXiv:2310.06825.

6. *Ali A.H., Alajanbi M., Yaseen M.G., Abed S.A.* Chatgpt4, DALL· E, Bard, Claude, BERT: Open Possibilities. Babylonian Journal of Machine Learning, 2023. Pp. 17–18.

7. *Dubois Y., Li C.X., Taori R., Zhang T., Gulrajani I., Ba J., etc.* AlpacaFarm: A simulation framework for methods that learn from human feedback. Advances in Neural Information Processing Systems, 2024. V. 1. https://doi.org/10.48550/arXiv.2305.14387

8. *Chaka C.* Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. Journal of Applied Learning and Teaching, 2023. Vol. 6(2). Pp. 1–11.

9. *Kumratova A.M., Morozova N.V., Vasilenko A.I., Kogaj I.E.* Analiz vozmozhnostej nejronnoj seti na osnove yazykovoj modeli GPT-3 i sposoby ee primeneniya na proizvodstve [Analysis of the capabilities of a neural network based on the GPT-3 language model and methods of its application in production]. Vestnik Adygejskogo gosudarstvennogo universiteta. S.4: Estestvenno-matematicheskie i tekhnicheskie nauki, 2023. Iss. 1 (316). Pp. 80–85 (in Russian).

10. *Zhan T., Shi C., Shi Y., Li H., Lin Y.* Optimization Techniques for Sentiment Analysis Based on LLM (GPT-3). Applied and Computational Engineering, 2024. Vol. 67(1). Pp. 41–47.

11. *Yenduri G., Ramalingam M., Selvi G.C., Supriya Y., Srivastava G., etc.* GPT (generative pre-trained transformer) – a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, 2023. DOI:10.1109/ACCESS.2024.3389497

12. *Wang B., Xie Q., Pei J., Chen Z., Tiwari P., Li Z., Fu J.* Pre-trained language models in biomedical domain: A systematic survey. ACM Computing Surveys, 2023. Vol. 56(3). Pp. 1–52.

13. *Hong Z., Ward, L., Chard, K., Blaiszik, B., Foster I.* Challenges and advances in information extraction from scientific literature: a review. JOM, 2021. Vol. 73(11), Pp. 3383–3400.

14. *Han, S., Zhang R. F., Shi L., Richie R., Li, H., Tseng A., etc*. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. Journal of biomedical informatics, 2022. Vol. 127. 103984.

15. Kåhrström F. Natural Language Processing for Swedish Nuclear Power Plants: A study of the challenges of applying Natural language processing in Operations and Maintenance and how BERT can be used in this industry. Visby. Uppsala Universitet, 2022. URL: http://uu.diva-portal.org/smash/get/diva2:1678697/FULLTEXT01.pdf