УДК 004.056
doi:10.31799/1684-8853-2025-2-50-67
EDN: QFQTPU

Научные статьи
Articles

Автоматизированная система анализа слабоструктурированных данных кибербразведки с использованием больших языковых моделей

А. И. Луцкович[а], *ассистент, orcid.org/0009-0000-7257-6947*
В. И. Васильев[а], *доктор техн. наук, профессор, orcid.org/0000-0002-6105-5481*
А. М. Вульфин[а], *доктор техн. наук, профессор, orcid.org/0000-0001-5857-2413, vulfin.am@ugatu.su*
А. Д. Кириллова[а], *канд. техн. наук, доцент, orcid.org/0009-0000-4164-2526*
А. Е. Сулавко[б], *доктор техн. наук, профессор, orcid.org/0000-0002-9029-8028*
[а]*Уфимский университет науки и технологий, К. Маркса ул., 12, Уфа, 450008, РФ*
[б]*Омский государственный технический университет, Мира пр., 11, Омск, 644050, РФ*

Введение: *актуальной проблемой является недостаточная эффективность и высокая трудоемкость анализа и управления слабоструктурированными и структурированными данными кибербразведки, собираемыми из различных источников.* ***Цель:*** *повышение эффективности (сокращение трудозатрат экспертов, обеспечение полноты и оперативности обновления базы знаний) анализа данных кибербразведки с помощью методов искусственного интеллекта.* ***Результаты:*** *проведенный анализ подходов к построению систем управления данными кибербразведки показал, что перспективным направлением является применение больших языковых моделей в составе вопросно-ответной системы поддержки принятия решений с механизмом расширенной дополненной генерации. Разработаны структурная схема и функциональная модель системы интеллектуального анализа данных кибербразведки на основе больших языковых моделей с применением конвейера расширенной дополненной генерации, алгоритм семантической разметки и извлечения информации из слабоструктурированных данных, исследовательский прототип (компонентная модель, микросервисная архитектура) программного обеспечения. Отличительной особенностью системы является комплексирование сведений из нескольких источников с помощью конвейера расширенной дополненной генерации. Вычислительный эксперимент на подготовленном наборе данных показал согласованность экспертных ответов с ответами, предложенными системой, на уровне 3,98 балла из пяти. Значение метрики BERTScore F1 составило 0,89, метрики «корректность ответов» (Answer Correctness) фреймворка RAGAS — 0,822.* ***Практическая значимость:*** *применение больших языковых моделей обеспечивает возможность аккумулировать знания об актуальных сценариях атак в рамках единой базы знаний. Это позволяет повысить эффективность обработки данных с целью оказать значимую поддержку специалистам по защите информации в ходе анализа актуальных угроз, уязвимостей и сценариев реализации компьютерных атак за счет снижения трудозатрат (времени сбора и обработки) и тем самым повысить защищенность корпоративных информационных систем.* ***Обсуждение:*** *дальнейшее повышение эффективности анализа данных кибербразведки возможно на основе построения гетерогенных «комитетов экспертов» из больших языковых моделей и мультиагентных систем.*

Ключевые слова *— информационная безопасность, кибербразведка, анализ уязвимостей, индикаторы компрометации, большие языковые модели, дополнительный расширенный поиск информации.*

Для цитирования: Луцкович А. И., Васильев В. И., Вульфин А. М., Кириллова А. Д., Сулавко А. Е. Автоматизированная система анализа слабоструктурированных данных кибербразведки с использованием больших языковых моделей. *Информационно-управляющие системы*, 2025, № 2, с. 50–67. doi:10.31799/1684-8853-2025-2-50-67, EDN: QFQTPU
For citation: Lutskovich A. I., Vasilyev V. I., Vulfin A. M., Kirillova A. D., Sulavko A. E. Automated system for analyzing weakly structured threat intelligence data using large language models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 2, pp. 50–67 (In Russian). doi:10.31799/1684-8853-2025-2-50-67, EDN: QFQTPU

Введение

Одним из ключевых трендов в обеспечении информационной безопасности (ИБ) корпоративных информационных систем (ИС) является все более широкое использование автоматизированных систем мониторинга событий и инцидентов ИБ, осуществляющих своевременное обнаружение аномалий в работе ИС, оповещение и предупреждение специалистов о возможных компьютерных атаках или инцидентах ИБ. Как правило, эти системы функционируют в составе центров мониторинга ИБ (Security Operation Center, SOC), собирающих и анализирующих информацию об угрозах, уязвимостях и сценариях реализации атак в ИС с использованием методов кибербразведки (Threat Intelligence, TI). Для автоматизации этих процессов создаются специальные системы управления данными кибербразведки (СУДК), при построении и функционировании которых используются перспективные технологии интеллектуального анализа данных и комплексирование информации, поступающей в SOC из различных источников слабоструктурированных данных [1].

Под кибербразведкой при этом понимается TI-сервис, владелец которого получает возмож-

ность следить за преступной деятельностью злоумышленников (хакерских группировок), анализировать используемые ими методы предварительного анализа информации, внедрения и закрепления в ИС, а также перемещения внутри взломанной инфраструктуры и совершения вредоносных действий, связанных с хищением данных, нарушением работы ИС жертвы, запуском программы-вымогателя и т. п.

Основное достоинство кибербразведки в том, что она позволяет компаниям действовать на опережение, предотвращая возникающие угрозы, оперативно реагируя на выявленные атаки, используя при этом накопленную аналитиками компании информацию о применяемых методах и характере действий киберпреступников.

Процесс кибербразведки включает в себя следующие основные этапы [2, 3]:

1) планирование (устанавливаются цели, требования к получаемой информации и расставляются приоритеты в дальнейшем анализе);

2) сбор информации;

3) обработку собранных данных (их интерпретация, трансляция, унификация и интеграция);

4) подготовку данных (уточнение и слияние обработанной информации);

5) распространение информации конечным потребителям (включая как внешних потребителей, так и собственные службы ИБ компании).

Для автоматизации этих процессов используются специализированные коммерческие платформы (Threat Intelligence Platform, TIP), предлагаемые на рынке рядом российских и международных компаний: Group-IB, Лабораторией Касперского, Positive Technologies, R-Vision, Anomali, ElectricIQ, Threat Connect, Threat Quotinet и др. Достаточно широкую известность также получили TIP-решения с открытым исходным кодом: MISP (Malware Information Sharing Platform) и Yeti (Your Everyday Threat Intelligence).

Источниками, или каналами («фидами»), получения данных кибербразведки могут быть [1] средства массовой информации, социальные сети, бюллетени рассылок, мессенджеры (платные и бесплатные, открытые), хакерские форумы, DarkNet (нелегальные), частные и государственные организации, предоставляющие информацию о киберугрозах. К наиболее крупным источникам данных TI относятся государственные структуры НКЦКИ (Национальный координационный центр по компьютерным инцидентам) и ФинЦЕРТ (Центр взаимодействия и реагирования на компьютерные атаки в кредитно-финансовой сфере) Банка России.

Основными типами данных, собираемых сервисами TIP, являются [1, 4]:

— индикаторы компрометации (Indicators of Compromise, IoC) — наборы данных о вредоносных объектах или подозрительных действиях (анализируются на оперативном уровне);

— тактики, техники и процедуры (Tactics, Techniques, Procedures, TTP) — описания способов действий злоумышленников, содержащие различную детализацию (анализируются на тактическом уровне);

— индикаторы атак (Indicators of Attacks, IoAs) — правила, содержащие описание подозрительного поведения в системе, которое может быть признаком целевой атаки (анализируются на стратегическом уровне).

Наибольшую ценность для кибербразведки представляют прежде всего IoC. Именно на выявлении и анализе подобных индикаторов сегодня сфокусированы исследования в области проактивного поиска и предупреждения угроз [5]. IoC [5] — это также признаки (технические данные), такие как IP- и URL-адреса, доменные имена, адреса электронной почты, темы писем, ссылки и вложения, ключи реестра, имена и хеш-суммы файлов и т. п., которые можно использовать для идентификации действий или инструментов атакующих. Помимо этих данных, IoC могут включать в себя также информацию о контексте, т. е. содержать дополнительное описание рассматриваемой угрозы (например, название вредоносного ПО, названия группировок атакующих, ранее пострадавших компаний и т. д.).

В то же время, как отмечается в работе [5], использование IoC на практике нередко встречается с определенными трудностями. Содержащиеся в них данные являются разнородными; как правило, рассредоточены в различных источниках (фидах), где представлены в виде фрагментов слабоструктурированного текста; могут содержать неполную или, наоборот, избыточную информацию, местами могут противоречить друг другу и т. п., что является предметом заботы и необходимостью ручной обработки данных с привлечением специалистов высокой квалификации.

Выходом из данной ситуации является применение современных методов и технологий обработки естественного языка (Natural Language Processing, NLP), интеллектуального анализа текстов (Text Mining) и больших языковых моделей (Large Language Models, LLMs) [6]. В настоящей статье рассмотрены подходы к автоматизации сбора и обработки слабоструктурированных текстовых данных TI с использованием LLMs и технологии расширенного поиска релевантной информации (Retrieval Augmented Generation, RAG). Представлены результаты разработки прототипа автоматизированной системы анализа данных кибербразведки на основе технологии RAG, а также результаты экспериментов по

оценке эффективности разработанной системы с использованием натуральных данных TI из источников ФинЦЕРТ и НКЦКИ.

Обработка слабоструктурированных данных каналов распространения IoC с помощью LLMs и технологии RAG

Большие языковые модели — это большие нейронные сети на основе архитектуры трансформеров с миллиардами параметров, обученные на очень больших корпусах текстовых данных, взятых из самых различных источников (веб-сайтов, энциклопедий, книг, статей, материалов конференций и др.). Появление LLMs вызвало настоящую революцию в сфере обработки слабоструктурированной текстовой информации, представленной на естественном языке. Использование методов глубокого обучения позволяет им понимать сложные грамматические конструкции, смысл слов и предложений в зависимости от контекста, генерировать тексты (отчеты, рефераты, переводы и др.) в различных предметных областях. Поэтому понятен тот интерес, который проявляется к этим моделям со стороны специалистов по ИБ, и в том числе аналитиков в области киберразведки. Будучи дополнительно дообученными (этап fine-tuning) на специально отобранных профильных текстах, относящихся к сфере TI, что требует значительно меньше времени по сравнению с предварительным обучением моделей (этап pre-traing), LLMs оказываются способны извлекать ключевые факты (такие как IoC) из «сырых» слабоструктурированных данных, выявляя закономерности и аномалии, указывающие на вредоносную активность, предоставляя информацию о потенциальных угрозах и методах реализации атак, предлагая стратегии смягчения последствий от возможных атак [6–8].

В литературе приводится много примеров успешного применения LLMs в задачах сбора и обработки данных TI (киберразведки). Так, в работе [9] представлены результаты разработки прототипа ПО для оценки актуальных угроз безопасности ИС, позволяющего сопоставить и ранжировать угрозы нарушения безопасности информации для выявленного специалистом перечня уязвимостей ИС из банка данных угроз (БДУ) безопасности информации ФСТЭК России, автоматизировать подбор тактик и техник для построения возможных сценариев реализации угроз. В основе предложенного подхода используется алгоритм сопоставления множества выявленных уязвимостей, соответствующих им тактик, техник и релевантных угроз безопасности информации путем оценки метрик семантической близости из текстовых описаний с использованием технологии обработки естественного языка (Word Embedding) и нейросетевой модели трансформера BERT3 компании Google. Применение предложенного подхода позволяет упростить процедуру оценки актуальных угроз безопасности информации, а также автоматизировать процесс построения возможных сценариев их реализации, сокращая временные затраты на проведение анализа.

В работе [4] выполнен подробный анализ использования больших языковых моделей для поиска угроз кибербезопасности в целях повышения защищенности ИС. Раскрыты особенности и выполнена классификация направлений применения LLM в задачах обеспечения кибербезопасности, определены трудности внедрения LLM в процессы поиска угроз. В работе [10] сформулированы признаки, характеризующие относительно низкую эффективность использования LLM для извлечения именованных сущностей в области кибербезопасности.

В работе [11] предложен подход к построению и обучению LLM для решения задач обнаружения и классификации компьютерных атак, а также извлечения из текста именованных сущностей (IoC). В качестве LLM-системы используется нейросетевая модель трансформера BERT. Предварительно обученная нейронная сеть дообучалась в ходе проведенных экспериментов на большом корпусе текстовых описаний атак из MITRE ATT & CK, CAPEC, учебников, статей из Википедии, материалов конференций по проблематике ИБ, описаний уязвимостей из CVE и CWE. Отмечается высокая точность обработки и классификации слабоструктурированных данных TI, обеспечиваемая с помощью дообученной LLM.

В [12] приводятся результаты разработки агента искусственного интеллекта, позволяющего автоматизировать извлечение важной информации (и в первую очередь IoC) из больших объемов текстовых данных (например, отчетов TI) с использованием LLMs. Помимо этого, разработанный агент позволяет на основе полученной информации устанавливать зависимости между компонентами IoC, генерировать регулярные выражения (шаблоны для поиска фрагментов в тексте), фильтровать выходные реакции LLMs в ответ на запрос пользователя, чтобы добиться низкого уровня ошибочных решений. Подчеркивается, что использование данного агента позволяет аналитикам SOC лучше понять модели атак и сократить время реагирования на атаки, высвобождая в конечном итоге время специалиста на решение более творческих задач.

В [13] рассмотрены возможности применения LLMs, обученных на большом объеме неразмеченных данных, для построения плана поведения интеллектуальных агентов в многоагентных системах.

В работе [14] авторы оценивают перспективы применения больших языковых моделей и подчеркивают необходимость особого подхода со стороны регулирующих органов, так как LLMs являются глобальными моделями, и для них может потребоваться новая нормативная категория (согласно национальной стратегии развития искусственного интеллекта, при внедрении любых систем искусственного интеллекта должен соблюдаться принцип технологического суверенитета, включая преимущественное использование отечественных разработок).

В работах рассматриваются вопросы обеспечения безопасности интеграции LLMs в корпоративные информационные системы [15, 16].

В исследовании [17] авторы предлагают методическое, алгоритмическое и программное обеспечение для создания при ограниченных вычислительных ресурсах на основе больших языковых моделей автоматических систем обработки данных, в которых отсутствуют катастрофическое забывание, риск переобучения, галлюцинации.

Вместе с тем, несмотря на указанные преимущества LLMs, их применение на практике встречается с рядом вопросов. В работе [18] показано, что при решении практических задач в различных проблемных областях (доменах знаний) возникает множество трудностей, вызванных неоднородностью и сложностью обрабатываемых данных, вариабельностью целей анализа и синтеза решений с помощью LLM и разнообразными ограничениями. Кроме того, необходимо поддерживать доступ к конфиденциальным корпоративным знаниям и реализовывать механизм обновления базы знаний на основе LLM [19]. В ряде случаев LLM может генерировать ошибочные ответы и неверные утверждения, выдавая их как вполне обоснованные, — то, что называется галлюцинациями («связным бредом»). Это обычно происходит, когда запрос пользователя требует наличия знаний, которые выходят за рамки дополнительного обучения LLM (т. е. модель «этому не учили»).

На сегодня предложено [18, 20] несколько способов устранения указанных проблем (табл. 1).

В работе [21] представлен подход к созданию и дообучению русскоязычных LLMs (XGLM, LLaMA, ruGPT-3.5) для применения в различных доменах знаний. Приведено сравнение двух основных методик дообучения: дообучение всех параметров модели и дообучение слоев с использованием LoRA. Авторами показано, что качество предоставляемых для обучения данных существенно влияет на возможности модели, оцениваемые с помощью автоматических метрик качества MT-BENCH и MMLU.

Одним из наиболее гибких и эффективных способов борьбы с галлюцинациями LLM стало

■ ***Таблица 1.*** Способы адаптации LLM для решения прикладных задач специфической предметной области
■ ***Table 1.*** Methods of adapting LLM to solve applied problems of a specific subject area

Метод адаптации LLM	Характеристика	Особенности
Передача дополнительного контекста через запрос к LLM (prompt)	Разработка наиболее эффективных запросов для генерации ответов требуемой структуры и содержания	Не изменяет параметры LLM
Тонкая настройка модели (fine tuning, FT)	Адаптация параметров LLM в ходе дообучения на небольшом (по сравнению с исходным обучающим набором) объеме специально подготовленных данных: – адаптация нескольких (выходных) слоев модели; – адаптация всех параметров модели	Обновление весовых коэффициентов всех слоев LLM может существенно ухудшить способность к генерации ответов Необходимы существенные вычислительные ресурсы Возможно существенное возрастание качества генерируемых ответов
Генерация с расширенным поиском (RAG)	Сочетание достоинств традиционных информационно-поисковых систем с возможностями LLM	Не изменяет параметры LLM
Использование LLM для построения графа знаний и применения графа знаний как источника данных для LLM	Сочетание возможностей построения структурированной интерпретируемой модели внешней среды (графа знаний) и извлечения закономерностей из имеющихся данных с помощью LLM	Перспективное направление, позволяющее повысить эффективность анализа данных кибербразведки

появление новой технологии RAG — генерации ответов с расширенным поиском, впервые предложенной в 2020 г. в работе [22] (подробнее см. обзоры [23–25]).

Суть данной технологии заключается в формировании такой среды искусственного интеллекта, которая объединяет силу традиционных информационно-поисковых систем с возможностями современных LLMs.

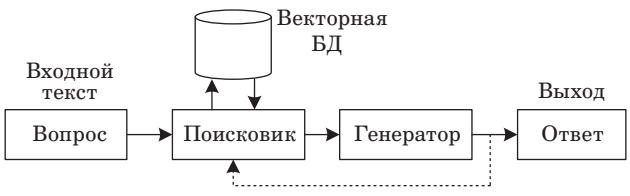
Архитектура RAG-системы в общем виде (рис. 1) состоит из двух основных частей — поисковика и генератора. Процесс ее работы начинается с ввода вопроса (или запроса) пользователя в поисковую систему. Поисковый модуль (поисковик) преобразует текст в последовательность числовых векторов (операция векторного вложения слов, Word Embedding), после чего производится семантический поиск релевантных документов в векторной базе данных, где хранятся документы компании, касающиеся в данном случае вопросов обеспечения ее ИБ. Из каждого документа извлекается текст, который используется для построения векторного вложения. Текст и вложение хранятся в базе данных. Поисковик на основе оценки семантической близости выбирает из базы данных несколько (Top-k) наиболее близких к сути вопроса (запроса) пользователя документов, содержащих максимально полезную контекстную информацию, необходимую для последующей обработки вопроса (запроса) модулем генератора.

Модуль генератора обычно реализуется на базе LLM трансформера типа GPT или BERT. Информация, полученная с поисковика (вопрос + найденный контекст), при этом также подвергается определенным NLP-операциям (токенизации, нормализации, стеммизации, лемматизации и др.) в целях приведения входных данных LLM к формату, более удобному для их обработки. Генератор генерирует ответ (выход) или инициализирует продолжение ответа, используя для этого предварительно обученную модель трансформера. Привлечение дополнительной релевантной информации (контекста) позволяет LLM дать более точный и обоснованный ответ на поступивший в систему вопрос (запрос).

Известны отдельные публикации, относящиеся к применению RAG-систем в области TI. Так, в работе [26] рассматриваются вопросы интеграции RAG с платформами TI в целях автоматизации процедур анализа и интерпретации киберугроз. Обсуждаются пути использования RAG для синтеза и контекстного обогащения данных TI из различных источников (фидов), включая логи, фиды угроз, отчеты об инцидентах ИБ, что позволяет более глубоко понимать причины и ландшафт угроз безопасности информации, повышать оперативность и эффективность принятия решений по предотвращению потенциальных угроз и смягчению последствий от их возникновения. Подчеркивается синергетический эффект от взаимодействия специалистов-профессионалов и применения технологии RAG, что является предпосылкой к достижению безопасной цифровой среды.

В [27] предложен метод автоматизированного построения графов атак (названный авторами Crystal Ball — «хрустальный мяч»), основанный также на применении технологии RAG. Поисковик RAG-системы при этом извлекает релевантную текстовую информацию из базы данных уязвимостей CVE, содержащей описания уязвимостей в виде цепочки основных этапов их реализации (предусловия, постусловия). В качестве источников данных TI могут также использоваться отчеты об угрозах безопасности информации. Обогащенный полученным контекстом, запрос пользователя поступает на вход LLM, которая строит граф компьютерной атаки, реализующей ту или иную уязвимость ИС, в форме графического представления путей атаки (действий злоумышленника), которые могут привести его к достижению поставленных целей. Приводятся примеры построения графов атак для конкретных уязвимостей с использованием различных LLMs (GPT-3.5, GPT-4, BARD), отмечается высокая эффективность, точность и релевантность обработки информации TI, в том числе возможность обработки данных в режиме реального времени, с использованием предложенного подхода.

В [28] предложен оригинальный подход к построению динамичных и адаптируемых сценариев учений по кибербезопасности для специалистов и сотрудников организаций. В основе разработанной автоматизированной системы генерации сценариев используется ансамбль из нескольких LLMs, одна из которых играет роль «эксперта по кибербезопасности», в функцию которой входит оценка текущего ландшафта киберугроз, тогда как другие LLMs выступают в качестве «исполнительных руководителей» организации, отвечающих за финансы, управление персоналом, инфраструктуру информационных технологий и т. д. Обогащение контекстом первоначального запроса на создание сценария



■ ***Рис. 1.*** Архитектура RAG-системы
■ ***Fig. 1.*** RAG system architecture

учений реализуется с использованием технологии RAG. Вместе с тем, как отмечают авторы, если одним из предназначений RAG обычно считается борьба с галлюцинациями LLMs, то в данном случае, наоборот, это вредное качество LLMs признается полезным свойством, позволяющим LLM генерировать более разнообразные (возможно, на первый взгляд, даже необычные) сценарии реализации угроз и выбора соответствующих стратегий и мер защиты. Однако окончательное решение в пользу выбора того или иного сгенерированного варианта сценария учений остается за человеком — профессионалом в области кибербезопасности. Авторами рассмотрено несколько примеров построения с использованием RAG детализированных сценариев проведения учений с учетом таких факторов, как выбор множества киберугроз и атак, особенностей информационной инфраструктуры организации, имеющихся технических средств защиты информации, распределения ролей участников в обеспечении кибербезопасности и т. д., что делает предложенные системой сценарии вполне обоснованными и реалистичными.

Повышение эффективности анализа данных кибербразведки и управления информацией о киберугрозах возможно [4] на основе перспективных подходов построения «комитетов экспертов» LLM, а также мультиагентных систем [29].

Разработка структурно-функциональной организации системы интеллектуального анализа данных кибербразведки

Существующие [30] подходы к обмену данными кибербразведки в машиночитаемых форматах характеризуются разнообразием стандартов, среди которых наиболее часто используются MISP и STIX. В то же время существенное количество данных, размещаемых на открытых платформах, представлено в структурированных текстовых форматах (plain text, CSV) и в виде слабоструктурированных отчетов на естественном языке. Как отмечалось ранее, ценным источником данных кибербразведки (но уже не в машиночитаемом формате) являются тематические блоги, каналы в мессенджерах и специализированные онлайн-ресурсы. Многообразие источников и форматов представления данных осложняет их преобразование в машиночитаемый формат конкретной платформы TI, так как необходимо сохранить семантический смысл исходных данных и контекст в рамках схемы хранения, определяемой выбранным форматом.

Развитие платформ TI тесно связано с обеспечением интерпретируемости данных — повышением эффективности преобразования форматов представления данных «человек-машина» и «машина-машина» — т. е. с простотой подключения новых источников данных, извлечением и сохранением максимального объема полезной информации при преобразовании форматов, организацией семантического поиска по анализируемым данным и диалогового интерфейса для специалистов.

Следовательно, для актуализации возможностей существующих СУДК значимо внедрение подсистем и модулей, позволяющих:

— предоставить интерфейс человеко-машинного взаимодействия в режиме «вопрос-ответ» для оперативного анализа собираемых данных;

— автоматизировать структурирование собираемых данных в конкретный формат платформы TI.

Возможным решением является разработка моделей извлечения знаний и автоматического построения онтологии и графа знаний проблемно-ориентированного корпуса для конкретной предметной области. Однако этот процесс требует трудоемкой разметки корпуса текстовых документов. Например [31], для неразмеченного корпуса текстов предложенное решение демонстрирует удовлетворительную работоспособность ввиду выделения атомарных фрагментов, включаемых в автоматически формируемую онтологию.

Другим рассматриваемым подходом для актуализации возможностей СУДК является применение RAG, позволяющей использовать открытые предобученные большие языковые модели общего назначения без необходимости дообучения для конкретной предметной области, а также без предварительного создания онтологии.

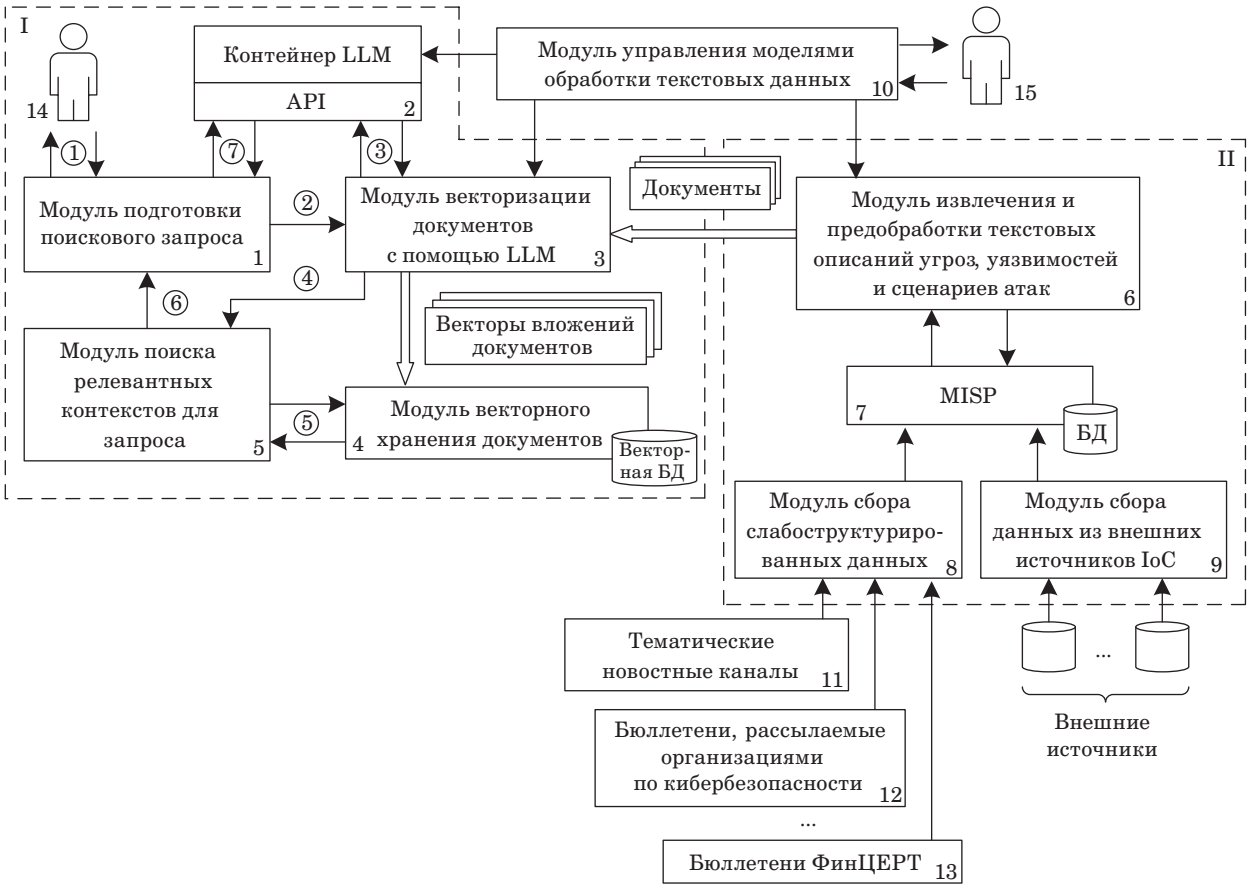
Таким образом, целью создания системы является повышение эффективности анализа слабоструктурированных и структурированных данных из различных источников с помощью развернутой платформы MISP [27].

Структурная схема представляемой системы анализа данных кибербразведки, включающей LLM с механизмом RAG в составе СУДК, представлена на рис. 2.

Обсуждаемый подход позволяет построить вопросно-ответную систему поддержки принятия решений для ранжированного набора собранных фрагментов документов, направленную на повышение оперативности анализа тактических и стратегических данных кибербразведки.

Предлагается выделить две основные подсистемы:

— подсистема I предназначена для обеспечения поддержки принятия решений на основе интеллектуального анализа данных с помощью конвейера RAG (цепочки предобработки и подготовки данных, поиска контекста и генерации ре-



■ ***Рис. 2.*** Структурная схема системы анализа данных киберразведки в составе SOC
■ ***Fig. 2.*** Structural diagram of the TI data analysis system as part of the SOC

шения) путем реализации человеко-машинного взаимодействия в режиме «вопрос-ответ»;

— подсистема II реализует текущий вариант сбора и обработки данных на основе платформы MISP в составе действующей СУДК и предназначена для автоматизации структурирования собираемых данных по определенным шаблонам в целях последующего использования.

На рис. 2 введены следующие обозначения: *1* — подготовка поискового запроса на естественном языке; *2* — предобработанный поисковой запрос эксперта; *3* — векторное представление текстового запроса (embedding) эксперта; *4*, *5* — поиск релевантных запросу фрагментов хранимых документов на основе меры сходства в пространстве векторных представлений; *6* — передача релевантных фрагментов документов для формирования контекста ответа; *7* — передача контекста и запроса в LLM для формирования ответа.

Работа модуля *1* начинается с подготовки поискового запроса. На этом этапе запрос проходит предобработку, включающую нормализацию и очистку текстовых данных для обеспечения качества и корректности дальнейшей обработки. После предобработки запрос преобразуется *3* в векторное представление с использованием модели LLM *2*. Этот этап критически важен, так как векторное представление запроса используется для поиска релевантных фрагментов документов в векторной базе предобработанного массива собранных данных *4* и последующего формирования контекста.

После создания векторного представления запроса начинается поиск *5* релевантных фрагментов документов. Поиск осуществляется на основе меры сходства в пространстве векторных представлений, что позволяет найти фрагменты, максимально соответствующие запросу пользователя. Найденные фрагменты извлекаются из векторной базы данных и передаются на следующий этап для формирования контекста ответа.

Формирование контекста *5* включает сбор и подготовку релевантных фрагментов, которые будут использованы для ответа на запрос пользователя. Контекст и исходный запрос передаются в контейнер LLM *2*. На этом этапе LLM использует предоставленный контекст для генерации окончательного ответа, который затем возвращается пользователю.

Модуль *10* управления моделями обработки текстовых данных координирует работу различных моделей, обеспечивая их интеграцию в общий процесс. Модуль *3* векторизации документов с помощью LLM преобразует текстовые документы в векторные представления. Модуль *6* извлечения и предобработки текстовых описаний угроз, уязвимостей и сценариев атак обрабатывает и подготавливает данные для анализа. Модуль *5* поиска релевантных контекстов проводит поиск на основе векторных представлений запросов. Модуль *4* векторного хранения документов обеспечивает хранение документов в векторной форме.

Платформа MISP *7* используется для сбора и обмена информацией о вредоносных программах, обеспечивая интеграцию данных из различных источников. Модули сбора слабоструктурированных данных *8* и сбора данных из внешних источников *9* собирают информацию из тематических новостных каналов и бюллетеней, обеспечивая актуальность и полноту базы данных. IoC используются для идентификации угроз, включая данные о вредоносных программах и уязвимостях.

Функциональная модель процесса анализа данных киберразведки приведена на рис. 3.

Далее при построении прототипа системы применяется схема Advanced RAG [32], использующая увеличенное контекстное окно анализа и методы суммаризации и сжатия контекста (рис. 4). Схема поясняет блок *4* «Поиск релевантного контекста и его обработка» функциональной модели и в свою очередь включает следующие блоки.

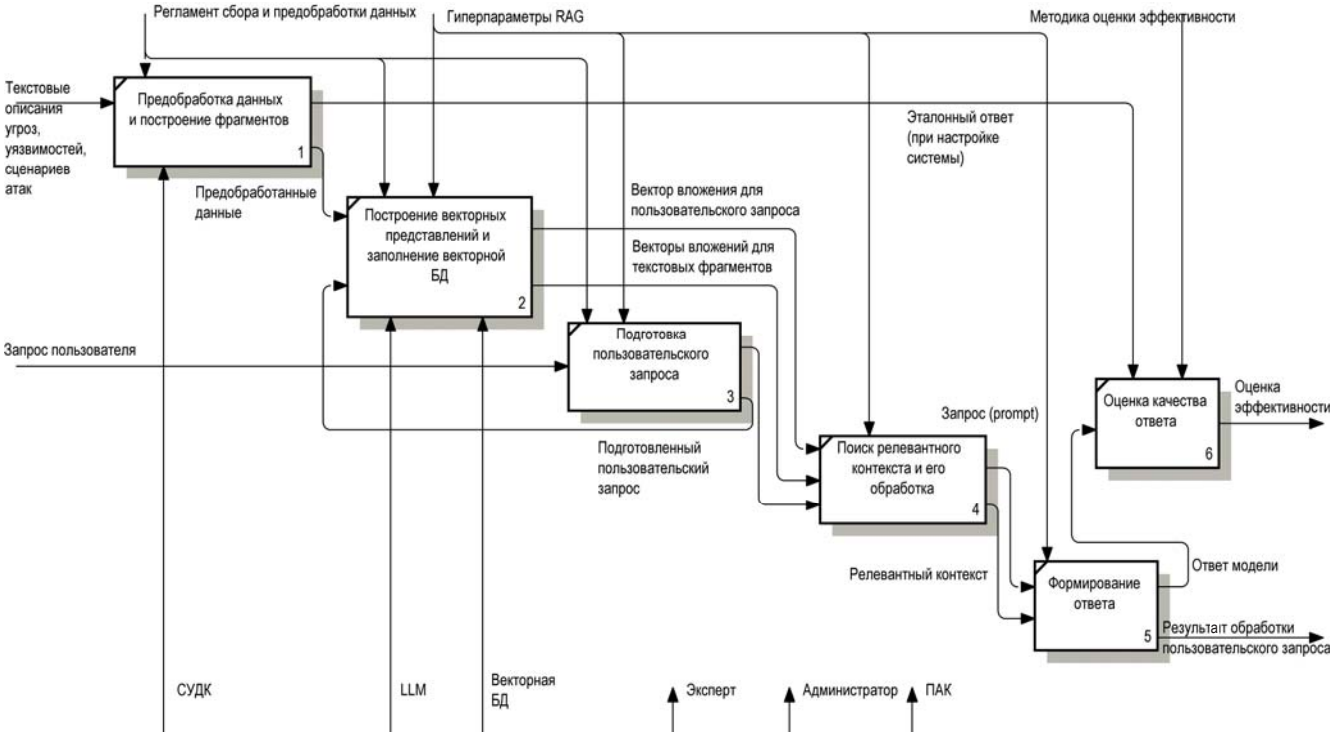
VectorStore (хранилище векторных представлений) — база данных, хранящая векторные представления документов. Используется для быстрого поиска релевантной информации на основе векторных запросов.

Search Sparse + Dense Vector (поиск с разреженными и плотными векторами) — выполняет комбинированный поиск на основе разреженного вектора признаков текстового описания (Sparse) (например, с помощью TF-IDF) и семантический поиск с помощью векторных (Dense) вложений (например, полученных с помощью LLM).

Eliminate Redundant Embeddings (удаление избыточных векторов вложений) — исключение избыточных векторов вложений, уменьшает количество нерелевантных или дублирующихся данных.

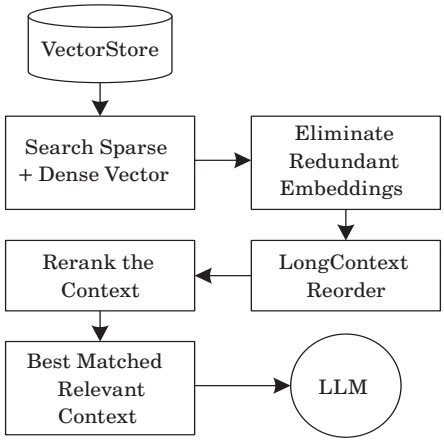
Rerank the Context (переранжирование контекста) — найденные текстовые фрагменты упорядочиваются по степени релевантности с помощью моделей ранжирования на основе LLM.

LongContext Reorder (перераспределение длинного контекста) — найденные фрагменты



■ ***Рис. 3.*** Функциональная модель процесса анализа данных киберразведки: *ПАК* — программно-аппаратный комплекс

■ ***Fig. 3.*** Functional model of the TI data analysis process



■ ***Рис. 4.*** Конвейер обработки данных Advanced RAG с суммаризацией контекста

■ ***Fig. 4.*** Advanced RAG pipeline with context summarization

переструктурируются так, чтобы их подача была наиболее логичной для генерации ответа.

Best Matched Relevant Context (наиболее релевантный контекст) — после ранжирования и перераспределения извлекается финальный контекст, который будет передан в LLM.

Алгоритм на рис. 5 иллюстрирует ключевые шаги, необходимые для реализации эффективной системы анализа семантической разметки и извлечения информации из слабоструктурированных данных киберразведки с точки зрения последующей программной реализации.

Схема описывает процесс извлечения и генерации текстовой информации с использованием больших языковых моделей и технологии RAG. LLM_1 применяется для векторизации документов, LLM_2 — для подготовки ответа на пользовательский запрос.

■ ***Рис. 5.*** Алгоритм семантической разметки и извлечения информации из слабоструктурированных данных

■ ***Fig. 5.*** The algorithm for semantic tagging and information extraction from semi-structured data

Проектирование исследовательского прототипа программного обеспечения анализа данных киберразведки

Компонентная модель исследовательского прототипа программного обеспечения для анализа данных киберразведки представлена на рис. 6. Экспертная оценка возможных кандидатов для использования в качестве ядра системы приведена в табл. 2. С целью обеспечить конфиденциальность обрабатываемых данных принято решение использовать развернутую локально LLM.

Переобученная LLM с параметрами 7.3B построена на основе Mistral, где ключевыми компонентами являются saiga_mistral_7b_gguf. Это обновление версии Mistral-7B-v0.2, которая занимала 52-е место в рейтинге Chatbot Arena, что выше, чем у GPT-3.5-Turbo-1106.

Алгоритмы предобработки текстовых данных детально проанализированы в работе авторов [33].

Механизм RAG реализован с использованием компонентных фреймворков LangChain [34] и OpenAI [35].

Работа программного прототипа начинается с получения данных из различных источников, таких как *MISP* и локальные файлы (например, текстовые файлы и PDF-документы, собранные из внешних источников). Данные из PDF-файлов извлекаются с помощью компонента *MISP PDF OCR Parser* (извлечение текстового слоя или оптическое распознавание символов). Извлеченные данные передаются в *LangChain Text Splitter*, который разделяет текст на фрагменты для дальнейшей обработки.

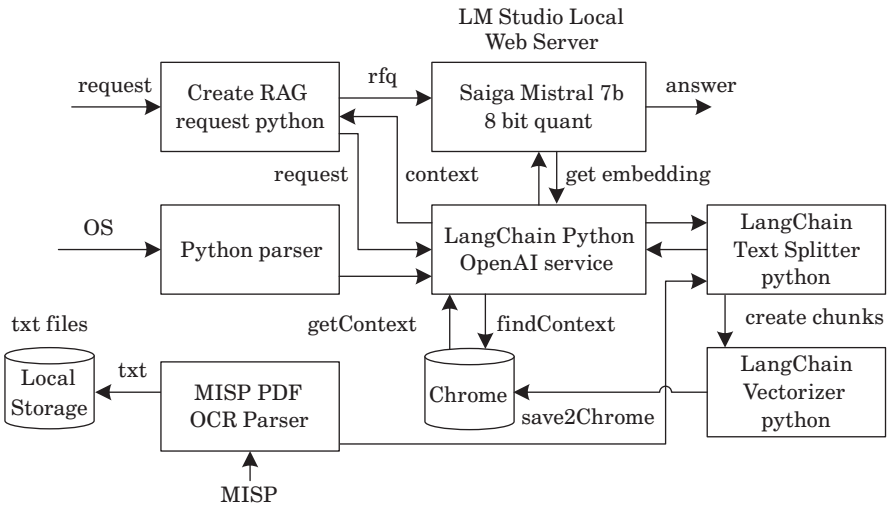
Затем данные проходят этап векторизации в *LangChain Vectorizer*, где текстовые фрагменты преобразуются в векторное представление. Для мультиязычных данных были использованы LLMs:

– sentence-paraphrase-multilingual-mpnet-base-v2;

■ ***Таблица 2.*** Оценка возможностей использования LLM в задаче обработки данных киберразведки

■ ***Table 2.*** Evaluation of the possibilities of using LLM in the task of processing TI data

Модель	Возможность локального запуска	Условия использования	Уровень адаптации к русскому языку	Быстродействие модели
ChatGPT	Нет	Платный API, нет доступа в РФ	Высокий	Высокое
Llama 13B	Да	Бесплатная	Низкий	Среднее
YandexGPT	Нет	Бесплатный API	Высокий	Высокое
Mistral 7B	Да	Бесплатная	Низкий	Среднее
Gemini	Нет	Платный API, нет доступа в РФ	Низкий	Высокое
Llama-3-Saiga	Да	Бесплатная	Высокий	Высокое



■ ***Рис. 6.*** Компонентная модель исследовательского прототипа программного обеспечения анализа данных киберразведки

■ ***Fig. 6.*** Component model of a research prototype of cyber intelligence data analysis software

— DeepPavlov/rubert-base-cased;
— multilingual-e5-large.

Векторные представления текстовых фрагментов сохраняются в векторной базе данных *Chroma*, обеспечивая быстрый доступ и поиск данных.

Для генерации запросов используется компонент *Create RAG request*, который отправляет запросы на обработку в *Saiga Mistral 7b 8 bit quant* — предобученную LLM.

Запросы пользователя и контексты обрабатываются с использованием *Python parser* и фреймворков *LangChain* и *OpenAI*, которые обеспечивают предобработку данных и взаимодействие с LLM. Полученные данные сохраняются и обрабатываются локально с помощью *Local Storage* и *LM Studio Local Web Server*, обеспечивая интеграцию всех компонентов системы и доступ к функционалу модуля через веб-интерфейс.

Микросервисная архитектура исследовательского прототипа (рис. 7) обеспечивает возможности вертикальной и горизонтальной масштабируемости.

Развертывание прототипа осуществлялось высокопроизводительным сервером со следующими параметрами:

— 20-ядерный процессор Intel Xeon E5-2698 v4 2,2 ГГц;
— 256 Гб ОЗУ RDIMM DDR4;
— четыре видеоускорителя Tesla V100 с суммарным объемом видеопамяти 128 Гб.

Каждая из моделей в изолированном окружении использовала выделенные графические процессоры, обмен данными между моделями осуществлялся через каналы в ОЗУ.

Проведение вычислительного эксперимента на собранном наборе данных

Цель вычислительного эксперимента — оценка работоспособности программного прототипа при реализации:

— вопросно-ответного диалогового взаимодействия эксперта с базой подготовленных документов об угрозах, уязвимостях, тактиках и техниках их реализации;

— структурирования текстового документа в формат MISP для последующего использования в СУДК.

В ходе разработки системы был создан набор данных на основе анализа более 20 источников, включающих тематические TI-каналы Telegram, отчеты о ландшафте угроз ведущих вендоров в области защиты информации (F.A.C.C.T., Лаборатория Касперского, Cisco и т. п.). С помощью действующей СУДК на основе платформы MISP были собраны данные за период 2022–2023 гг. из 10 основных каналов, включая, бюллетени ФинЦЕРТ и НКЦКИ. Данные представ-

■ ***Рис. 7.*** Микросервисная архитектура исследовательского прототипа программного обеспечения
■ ***Fig. 7.*** Microservice architecture of a research software prototype

лены в открытых источниках, собраны в текстовом формате и далее предобработаны.

Размер окна анализа текстового документа после предобработки составил 1000 токенов с перекрытием 300 токенов (1 токен = 2–4 символа текста). Общая база данных текстовых описаний и векторов вложений включала более 10 тыс. записей.

Вопросы к модели задавались в соответствии с эталонным списком подготовленных запросов — для каждого документа группой экспертов были подготовлены пять вопросов и эталонных ответов. Результаты генерации модели добавлялись в базу для дальнейшего сравнения с эталонными ответами эксперта. Итоговый набор текстов включал 100 предобработанных документов, для каждого из которых было подготовлено пять эталонных вопросов и ответов.

Модель способна находить правильные фрагменты текста и составлять по ним ответ на поставленный вопрос. В лучших итерациях ответ модели может практически полностью совпадать с эталонным ответом.

Анализ полученных результатов и оценка эффективности предложенного решения

Для оценки результатов работы системы были использованы следующие метрики.

1. Экспертная оценка: оценка по пятибалльной шкале (от единицы — худшее совпадение до пяти — полное совпадение), предоставленная экспертом на основе субъективного восприятия качества сгенерированных ответов. Эта оценка важна для понимания того, насколько результаты модели соответствуют ожиданиям специалистов предметной области.

2. BertScore: оценка (безразмерная величина в диапазоне [0, 1]) семантического сходства сгенерированного ответа и эталонного ответа. Использует модели BERT для вычисления векторных представлений текста и вычисления косинусного сходства между ними. Включает оценки точности (Precision), полноты (Recall) и гармонического среднего (F1-score):

— BertScore_pre: точность (Precision);

— BertScore_rec: полнота (Recall).

Высокие значения этих метрик показывают, что текст модели семантически близок к эталонному.

Результаты расчета среднего значения каждой из метрик приведены в табл. 3.

Метрики, рассчитанные индивидуально для каждого ответа модели, представлены в табл. 4.

Необходимо оценивать как эффективность работы модулей поиска и генерации по отдельности, так и работу конвейера RAG в целом. Помимо экспертной оценки результатов работы, существуют и иные метрики: ROUGE [36], BLEU [37] и RAGAS [38].

Применение фреймворка RAGAS (Retrieval Augmented Generation Automated Scoring) [38] позволяет автоматизировать оценку качества ответов модели в виде набора метрик (табл. 5). Состав предобработанного набора данных показан на рис. 8.

Использование как можно меньшего количества данных, аннотированных экспертом вручную, позволит снизить стоимость разработки систем и повысить оперативность оценки эффективности системы. Реализация фреймворка RAGAS позволяет проводить оценку эффективности системы в каждом из этих сценариев (с использованием аннотированных экспертами данных и без, когда в качестве эксперта выступает отдельная LLM-«критик»). Все значения метрик находятся в диапазоне [0, 1], при этом более высокие значения указывают на лучшую производительность. Для собранного набора данных ре-

■ ***Таблица 3.*** Результаты подсчета метрик и их анализ

■ ***Table 3.*** Results of metrics calculation and their analysis

Обозначение метрики	Среднее значение	Значение метрики	Анализ результата
Экспертная оценка	3,98	Высокие значения экспертной оценки говорят о том, что, с точки зрения профессионалов в данной области, результаты считаются качественными и удовлетворительными	Оценка показывает, что ответы модели находятся на высоком уровне и близки к эталонным. Это свидетельствует о том, что модель успешно генерирует релевантные и полезные ответы на вопросы
BERTScore_pre	0,9	Высокие баллы BERTScore означают, что ответы языковой модели семантически и синтаксически схожи с рефератами	Высокие значения BERTScore показывают, что модель хорошо справляется с задачами, где важна семантическая точность, способна генерировать тексты, близкие по смыслу к эталонным
BERTScore_rec	0,88		
BERTScore F1	0,89		

■ ***Таблица 4.*** Рассчитанные для ответов метрики
■ ***Table 4.*** Calculated metrics for responses

Экспертная оценка	BERTScore_pre	BERTScore_rec	BERTScore F1
5	0,99	0,94	0,96
5	0,76	0,45	0,57
5	0,93	0,88	0,9
5	0,94	0,94	0,94
5	0,92	0,88	0,9
5	0,94	0,94	0,94
5	0,92	0,92	0,92

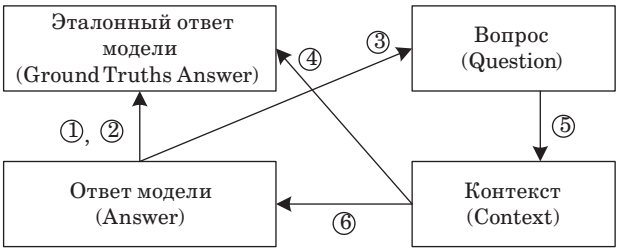
■ ***Таблица 5.*** Метрики RAGAS для полного набора данных
■ ***Table 5.*** RAGAS metrics for the full dataset

Метрика	Характеристика	Оцениваемая часть конвейера RAG	Значение
Точность контекста	Оценивает соотношение сигнала к шуму извлеченного контекста, вычисляется с использованием вопроса и контекстов	Отдельные этапы анализа RAG	0,803
Полнота контекста	Измеряет, была ли извлечена вся необходимая для ответа на вопрос информация. Эта метрика вычисляется на основе эталонных данных — ground_truth (это единственная метрика в фреймворке, которая опирается на аннотированные человеком метки истинности) и контекстов	То же	0,794
Релевантность контекста	Оценивает релевантность извлеченного контекста, определяемую на основе как вопроса, так и контекстов	–//–	–
Верность	Измеряет фактическую согласованность сгенерированного ответа по отношению к заданному контексту. Количество правильных утверждений из заданных контекстов делится на общее количество утверждений в подготовленном ответе (на основе вопросов, контекстов и ответов)	–//–	0,831
Релевантность ответа	Оценивает, насколько релевантен ответ, подготовленный моделью, соответствует вопросу. Более низкий балл присваивается ответам, которые являются неполными или содержат избыточную информацию, а более высокие баллы указывают на более высокую релевантность. Для оценки метрики используются вопросы и ответы	–//–	0,813
Семантическое сходство ответов	Оценивает семантическое сходство между сгенерированным ответом и истинным значением	Сквозная оценка конвейера RAG	–
Корректность ответов	Включает в себя измерение точности сгенерированного ответа по сравнению с истинным значением	То же	0,822

зультаты оценки с помощью RAGAS приведены в табл. 5.

В ходе эксперимента было установлено, что модель показывает высокие результаты по большинству метрик. Экспертная оценка 3,98 подтверждает, что генерируемые ответы воспринимаются как качественные и полезные. Метрики BERTScore показывают, что модель генерирует ответы, которые семантически схожи с эталонными, что также поддерживает высокую оценку качества модели. Совокупность метрик RAGAS также подтверждает способность системы корректно отвечать на поставленные вопросы, что делает ее полезным инструментом для анализа неструктурированных и структурированных данных кибербразведки.

Для оценки эффективности системы (снижение трудоемкости обработки данных и повышение оперативности анализа) эксперт в области кибербразведки анализировал открытые новостные сообщения тематического канала и структурировал информацию вручную, во втором сцена-



① корректность ответа (answer correctness)
② семантическое сходство ответов (answer similarity)
③ релевантность ответа (answer relevancy)
④ полнота контекста (context recall)
⑤ точность контекста (context precision)
⑥ верность (faithfulness)

■ ***Рис. 8.*** Основные сущности RAGAS
■ ***Fig. 8.*** Main entities of RAGAS

■ ***Таблица 6.*** Результаты эксперимента
■ ***Table 6.*** Results of the experiment

Сценарий и длительность	Ручная обработка одним экспертом	Автоматическая обработка
Сбор информации за один день	20 записей киберразведки из 10 источников за 30 мин	400 записей киберразведки из 10 источников за 30 мин
Анализ и структурирование информации за один день	20 записей из 10 источников за 150 мин	400 записей из 10 источников за 33 мин
Сбор информации за неделю	140 записей за 250 мин	2100 записей за 250 мин
Анализ и структурирование информации за неделю	140 записей за 1050 мин	2100 записей за 210 мин

рии та же задача выполнялась с использованием предлагаемой системы (табл. 6).

Таким образом, на сбор данных для одной записи в базу TI эксперт тратил в среднем 1,5 мин, при использовании предлагаемой системы — 0,075 мин. При структурировании одной записи в формат MISP эксперт тратил 7,5 мин, при использовании системы — 0,083 мин.

Заключение

Рассмотрены способы повышения эффективности (повышение оперативности и снижение трудоемкости) анализа данных киберразведки с помощью методов искусственного интеллекта:

— разработана структурно-функциональная организация (структурная схема и функциональная модель) системы интеллектуального анализа данных киберразведки на основе больших языковых моделей с применением конвейера RAG;

— разработан алгоритм семантической разметки и извлечения информации из слабоструктурированных данных киберразведки для последующего анализа с помощью LLM в составе вопросно-ответной системы поддержки принятия решений;

— разработан исследовательский прототип (компонентная модель, диаграммы вариантов использования, микросервисная архитектура) программного обеспечения для анализа данных киберразведки с помощью больших языковых моделей, оценена эффективность предложенного решения в задаче обработки слабоструктурированных данных из каналов TI.

Проведенный эксперимент показал согласованность экспертных ответов по набору собранных документов с ответами, предложенными моделью, на уровне 3,98 балла из пяти. Значение метрики BERTScore F1 составило 0,89, метрик RAGAS «корректность ответов» (Answer Correctness) — 0,822. Временные затраты на сбор и подготовку данных для СУДК снизились более чем в 20 раз.

Выявлены особенности реализации систем интеллектуального анализа данных TI:

— сложность проектирования и реализации с использованием современных фреймворков программного прототипа системы, использующей конвейер RAG для управления данными киберразведки, ниже, чем сложность настройки гиперпараметров отдельных этапов обработки и формирования набора метрик для непрерывной оценки эффективности анализа и адекватности ответов модели;

— оценка эффективности системы требует высококачественного набора проверочных данных, подготовленного экспертами, что является сложным, трудоемким и дорогостоящим процессом.

Новизна предлагаемых алгоритма и модели основана на алгоритмах поисковой дополненной генерации с помощью больших языковых моделей, отличительной особенностью является комплексирование сведений из нескольких источников распространения IoC и последующее извлечение информации из слабоструктурированных накапливаемых данных с помощью моделей машинного обучения в составе конвейера RAG. Это позволяет повысить эффективность анализа данных киберразведки для оказания значимой поддержки специалистам по защите информации в ходе анализа актуальных угроз и

уязвимостей за счет снижения трудозатрат (времени сбора и обработки данных).

Несмотря на наблюдающиеся тенденции увеличения контекстного окна LLM, механизм RAG позволяет реализовать разделение прав и управление доступом к модели, что невозможно реализовать только лишь на уровне LLM.

Перспективным направлением дальнейших исследований является построение «графов знаний» (Knowledge Graphs) с метаданными (Knowledge Maps), описывающими структуру хранения и связи между сущностями предметной области (Graph Retrieval Augmented Generation).

Финансовая поддержка

Работа выполнена в ОмГТУ в рамках государственного задания Минобрнауки России на 2023–2025 гг. № FSGF-2023-0004.

Литература

1. **Вульфин А. М.** Система управления данными кибербразведки. *Моделирование, оптимизация и информационные технологии*, 2021, т. 9, № 1 (32). doi:10.26102/2310-6018/2021.32.1.020, EDN: RDVDTE
2. **Ramsdale A., Shiaeles S., Kolokotronis N.** A comparative analysis of cyber-threat intelligence sources, formats and languages. *Electronics (MDPI)*, 2020, vol. 9, Article 824. doi:10.3390/electronics9050824. https://www.mdpi.com/journal/electronics (дата обращения: 10.12.2024).
3. **Ainslie S., Thompson D., Maynard S., Ahmad A.** Cyber-threat intelligence for security decision-making: A review and research agenda for practice. *Computers & Security*, 2023, vol. 132, pp. 103352. doi:10.1016/j.cose.2023.103352
4. **Котенко И. В., Абраменко Г. Т.** Использование больших языковых моделей для поиска угроз кибербезопасности на основе методов глубокого обучения: анализ современных исследований. *Правовая информатика*, 2024, № 3, с. 32–42. doi:10.24682/1994-1404-2024-3-32-42
5. **Мещеряков Р. В., Исхаков С. Ю.** Исследование индикаторов компрометации для средств защиты информационных и киберфизических систем. *Вопросы кибербезопасности*, 2022, № 5, с. 82–99. doi:10.21681/2311-3456-2022-5-82-99
6. **Gholami Y.** Large Language Models (LLMs) for cybersecurity: A systematic review. *World Journal of Advanced Engineering Technology and Sciences*, 2024, vol. 13(01), pp. 057–069. doi:10.30574/wjaets.2024.13.1.0395
7. **Chen Y., Cui M., Wang D., Cao Y., Yang P., Jiang B., Lu Zh., Liu B.** A survey of large language models for cyber threat detection. *Computers & Security*, 2024, vol. 145, iss. C, pp. 104016. doi:10.1016/j.cose.2024.104016, EDN: NVEXHQ
8. **Hasanov I., Virtanen S., Hakkala A., Isoaho J.** Application of Large Language Models in cybersecurity: A systematic literature review. *IEEE Access*, 2024, vol. 12, pp. 176751–176778. doi:10.1109/ACCESS.2024.3505983
9. **Васильев В. И., Вульфин А. М., Кучкарова Н. В.** Оценка актуальных угроз безопасности информации с помощью технологии трансформеров. *Вопросы кибербезопасности*, 2022, № 2 (48), с. 27–38. doi:10.21681/2311-3456-2022-2-27-38
10. **Bolla A., Talentino F.** *Threat Hunting Driven by Cyber Threat Intelligence*. Diss. Polytechnic University of Turin, 2022. 162 p.
11. **Park Y., You W.** A pretrained language model for cyber threat intelligence. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, December 6–10, 2023, pp. 113–122. doi:10.18653/v1/2023.emnlp-industry.12
12. **Kucsván Z. L., Caselli M., Peter A., Continella A.** Inferring recovery steps from cyber threat intelligence reports. *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessmen*, LNCS, 2024, vol. 14828, pp. 330–349.
13. **Ковалев А. К., Панов А. И.** Применение предобученных больших языковых моделей в задачах воплощенного искусственного интеллекта. *Доклады Российской академии наук. Математика, информатика, процессы управления*, 2022, т. 508, с. 94–99. doi:10.31857/S268695432207013X
14. **Андрейченко А. Е., Гусев А. В.** Перспективы применения больших языковых моделей в здравоохранении. *Национальное здравоохранение*, 2023, т. 4, № 4, с. 48–55. doi:10.47093/2713-069X.2023.4.4.48-55, EDN: MPDTJU
15. **Евглевская Н. В., Казанцев А. А.** Обеспечение безопасности сложных систем с интеграцией больших языковых моделей: анализ угроз и методов защиты. *Экономика и качество систем связи*, 2024, № 4 (34), с. 129–144. EDN: CJEAAZ. https://journal-ekss.ru/wp-content/uploads/2024/12/129-144.pdf (дата обращения: 10.12.2024).
16. **Мударова Р. М., Намиот Д. Е.** Противодействие атакам типа инъекция подсказок на большие языковые модели. *International Journal of Open Information Technologies*, 2024, т. 12, № 5, с. 39–48. EDN: LVERMB. http://injoit.org/index.php/j1/article/view/1845/1682 (дата обращения: 10.12.2024).
17. **Самонов А. В., Бурова И. О.** Методика разработки автоматизированных средств генерации программного кода посредством настройки больших языковых моделей. *Вопросы кибербезопасности*, 2024, № 3 (61), с. 68–75. doi:10.21681/2311-3456-2024-3-68-75, EDN: NPSBQW
18. **Зеленков Ю. А.** Управление знаниями организации и большие языковые модели. *Российский жур-*

нал менеджмента, 2024, т. 22, № 3, с. 573–601. doi:10.21638/spbu18.2024.309, EDN: KKHPPZ

19. **O'Leary D. E.** The rise and design of enterprise large language models. *IEEE Intelligent Systems*, 2024, vol. 39, no. 1, pp. 60–63. doi:10.1109/MIS.2023.3345591
20. **Yu W., Zhu C., Li Z., Hu Z., Wang Q., Ji H., Jiang M.** A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 2022, vol. 54, no. 11s, pp. 1–38. doi:10.1145/3512467
21. **Косенко Д. П., Куратов Ю. М., Жарикова Д. Р.** Большие языковые модели для следования инструкциям на русском языке: модели и датасеты с открытой лицензией для коммерческого использования. *Доклады Российской академии наук. Математика, информатика, процессы управления*, 2023, т. 514, № 2, с. 262–269. doi:10.31857/S2686954323602063, EDN: GECIST
22. **Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D.** Retrieval-Augmented Generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 9459–9474.
23. **Cai D., Wang Y., Liu L., Shi S.** Recent advances in retrieval-augmented text generation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11–15, 2022. New York, 2022, pp. 3417–3419. doi:10.1145/3477495.3532682
24. **Fan W., Ding Y., Ning L., Wang S., Li H., Yin D., Chua T.-S., Li Q.** A survey on rag meeting llms: Towards retrieval-augmented large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, Spain, August 25–29, 2024. New York, 2024, pp. 6491–6501. doi:10.1145/3637528.3671470
25. **Li Z., Li C., Zhang M., Mei Q., Bendersky M.** Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Miami, Florida, November 12–16, 2024, pp. 881–893. doi:10.18653/v1/2024.emnlp-industry.66
26. **Singh J. P., Agrawal Sh.** Automating threat intelligence analysis with Retrieval Augmented Generation (RAG) for enhanced cybersecurity posture. *International Journal of Science and Research (IJSR)*, 2024, vol. 13, iss. 5, pp. 251–255. doi:10.21275/SR24502103758
27. **Zhang Y., Du T., Ma Y., Wang X., Xie Y., Yang G., Lu Y., Chang E. C.** AttacKG+: Boosting attack graph construction with Large Language Models. *Computers & Security*, 2024, vol. 150, pp. 104220. doi:10.1016/j.cose.2024.104220
28. **Yamin M. M., Hashmi E., Ullah M., Katt B.** Application of LLMs for generating cyber security exercise scenarios. *Research Square*, Feb. 2024. doi:10.21203/rs.3.rs-3970015/v1. https://www.researchsquare.com/article/rs-3970015/v1 (дата обращения: 10.12.2024).
29. **Wu Q., Bansal G., Zhang J., Wu Y., Li B., Zhu E., Jiang L., Zhang X., Zhang S., Liu J., Awadallah A. H., White R. W., Burger D., Wang C.** AutoGen: Enabling next-gen LLM applications via multi-agent conversations. *First Conference on Language Modeling*, Pennsylvania, PA, October 7–9, 2024, pp. 1–46. https://openreview.net/pdf?id=BAakY1hNKS (дата обращения: 10.12.2024).
30. **Mavroeidis V., Bromander S.** Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. *2017 European Intelligence and Security Informatics Conference (EISIC)*, Greece, September 11–13, 2017. IEEE, 2017, pp. 91–98. doi:10.1109/EISIC.2017.20
31. **Зулкарнеев Р. Х., Юсупова Н. И., Сметанина О. Н., Гаянова М. М., Вульфин А. М.** Методы и модели извлечения знаний из медицинских документов. *Информатика и автоматизация*, 2022, т. 21, № 6, с. 1169–1210. doi:10.15622/ia.21.6.4, EDN: ASOOVS
32. **Rajapaksha S., Rani R., Karafili E.** A RAG-based question-answering solution for cyber-attack investigation and attribution. *Workshop on Security and Artificial Intelligence 2024, 29th European Symposium on Research in Computer Security*, 2024. http://eprints.soton.ac.uk/id/eprint/493520 (дата обращения: 10.12.2024).
33. **Васильев В. И., Вульфин А. М., Кучкарова Н. В.** Автоматизация анализа уязвимостей программного обеспечения на основе технологии Text Mining. *Вопросы кибербезопасности*, 2020, № 4 (38), с. 22–31. doi:10.21681/2311-3456-2020-04-22-31, EDN: TMRYIT
34. **Topsakal O., Akinci T. C.** Creating Large Language Model applications utilizing LangChain: A primer on developing LLM apps fast. *International Conference on Applied Engineering and Natural Sciences*, 2023, vol. 1, no. 1, pp. 1050–1056. doi:https://doi.org/10.59287/icaens.1127
35. **Auger T., Saroyan E.** *Overview of the OpenAI APIs.* Generative AI for Web Development: Building Web Applications Powered by OpenAI APIs and Next.js. Berkeley, CA, Apress, 2024, pp. 87–116.
36. **Lin C. Y.** ROUGE: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26, 2004, pp. 74–81.
37. **Papineni K., Roukos S., Ward T., Zhu W. J.** BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135
38. **Es S., James J., Anke L. E., Schockaert S.** RAGAs: Automated evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158. doi:10.48550/arXiv.2309.15217

UDC 004.056
doi:10.31799/1684-8853-2025-2-50-67
EDN: QFQTPU

Automated system for analyzing weakly structured threat intelligence data using large language models

A. I. Lutskovich[a], Assistant Professor, orcid.org/0009-0000-7257-6947
V. I. Vasilyev[a], Dr. Sc., Tech., Professor, orcid.org/0000-0002-6105-5481
A. M. Vulfin[a], Dr. Sc., Tech., Professor, orcid.org/0000-0001-5857-2413, vulfin.am@ugatu.su
A. D. Kirillova[a], PhD, Tech., Associate Professor, orcid.org/0009-0000-4164-2526
A. E. Sulavko[b], Dr. Sc., Tech., Professor, orcid.org/0000-0002-9029-8028
[a]Ufa University of Science and Technology, 12, K. Marx St., 450008, Ufa, Russian Federation
[b]Omsk State Technical University, 11, Mira Pr., 644050, Omsk, Russian Federation

Introduction: The lack of efficiency and the complexity of the analysis and management of semi-structured and structured Threat Intelligence data collected from various sources is a pressing issue. **Purpose:** To increase the efficiency (which implies reducing the labor costs of experts and ensuring the completeness and promptness of the knowledge base updates) of Threat Intelligence data analysis using artificial intelligence methods. **Results:** The conducted analysis of approaches to the construction of Threat Intelligence data management systems has shown as a promising direction the use of large language models as part of a "question-answer" decision support system with an extended augmented generation mechanism. We develop a structural diagram and functional model of a data mining system for the Threat Intelligence data. It is based on large language models using a retrieval augmented generation pipeline. In addition, we develop an algorithm for semantic tagging and extracting information from weakly structured data. We present a research prototype (component model, microservice architecture) of the software. A distinctive feature of the system is the integration of information from multiple sources using a retrieval augmented generation pipeline. A computational experiment on the prepared data set has shown the consistency of expert responses for the set of collected documents with the responses proposed by the system at the level of 3.98 points out of 5. The BERTScore F1 metric value is 0.89, and the RAGAS Answer Correctness metric value is 0.822. **Practical relevance:** The use of large language models provides the accumulation of knowledge about current attack scenarios within a single knowledge base, which makes it possible to enhance the efficiency of data processing to enable a significant support to information security specialists during the analysis of current threats, vulnerabilities and cyber attack scenarios by reducing labor costs (time for collecting and processing), and thereby to increase the security of corporate information systems. **Discussion:** A further improvement of the efficiency of the Threat Intelligence data analysis is possible on the basis of constructing heterogeneous "expert committees" out of large language models and multi-agent systems.

Keywords — information security, Threat Intelligence, vulnerability analysis, Indicators of Compromise, Large Language Models, Retrieval-Augmented Generation.

For citation: Lutskovich A. I., Vasilyev V. I., Vulfin A. M., Kirillova A. D., Sulavko A. E. Automated system for analyzing weakly structured threat intelligence data using large language models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2025, no. 2, pp. 50–67 (In Russian). doi:10.31799/1684-8853-2025-2-50-67, EDN: QFQTPU

Financial support
The research was carried out in Omsk State Technical University within the State assignment of the Ministry of Science and Higher Education of Russian Federation (theme No. FSGF-2023-0004).

References

1. Vulfin A. M. Cyber threat intelligence data management system. *Modeling, Optimization and Information Technology*, 2021, vol. 9, no. 1 (32) (In Russian). doi:10.26102/2310-6018/2021.32.1.020, EDN: RDVDTE
2. Ramsdale A., Shiaeles S., Kolokotronis N. A comparative analysis of cyber-threat intelligence sources, formats and languages. *Electronics (MDPI)*, 2020, vol. 9, p. 824. doi:10.3390/electronics9050824. Available at: https://www.mdpi.com/journal/electronics (accessed 10 December 2024).
3. Ainslie S., Thompson D., Maynard S., Ahmad A. Cyber-threat intelligence for security decision-making: A review and research agenda for practice. *Computers & Security*, 2023, vol. 132, pp. 103352. doi:10.1016/j.cose.2023.103352
4. Kotenko I. V., Abramenko G. T. Using large language models for cyber threat hunting based on deep learning methods: Analysis of modern research. *Legal Informatics*, 2024, no. 3, pp. 32–42 (In Russian). doi:10.24682/1994-1404-2024-3-32-42
5. Meshcheryakov R. V., Iskhakov S. U. Study of compromet-ation indicators for improvement of information and cyber-physical systems protection facilities. *Cybersecurity Issues*, 2022, no. 5, pp. 82–99 (In Russian). doi:10.21681/2311-3456-2022-5-82-99
6. Gholami Y. Large Language Models (LLMs) for cybersecurity: A systematic review. *World Journal of Advanced Engineering Technology and Sciences*, 2024, vol. 13(01), pp. 057–069. doi:10.30574/wjaets.2024.13.1.0395
7. Chen Y., Cui M., Wang D., Cao Y., Yang P., Jiang B., Lu Zh., Liu B. A survey of large language models for cyber threat detection. *Computers & Security*, 2024, vol. 145, iss. C, pp. 104016. doi:10.1016/j.cose.2024.104016, EDN: NVEXHQ
8. Hasanov I., Virtanen S., Hakkala A., Isoaho J. Application of Large Language Models in cybersecurity: A systematic literature review. *IEEE Access*, 2024, vol. 12, pp. 176751–176778. doi:10.1109/ACCESS.2024.3505983
9. Vasilyev V. I., Vulfin A. M., Kuchkarova N. V. Assessment of current threats to information security using transformer technology. *Cybersecurity Issues*, 2022, no. 2 (48), pp. 27–38 (In Russian). doi:10.21681/2311-3456-2022-2-27-38
10. Bolla A., Talentino F. *Threat Hunting Driven by Cyber Threat Intelligence*. Diss. Polytechnic University of Turin, 2022. 162 p.
11. Park Y., You W. A pretrained language model for cyber threat intelligence. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2023, pp. 113–122. doi:10.18653/v1/2023.emnlp-industry.12
12. Kucsván Z. L., Caselli M., Peter A., Continella A. Inferring recovery steps from cyber threat intelligence reports. *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessmen*, LNCS, 2024, vol. 14828, pp. 330–349.
13. Kovalev A. K., Panov A. I. Application of pretrained large language models in embodied artificial intelligence. *Doklady Rossijskoj Akademii Nauk. Mathematika, Informatika, Processy Upravlenia*, 2022, vol. 508, pp. 94–99 (In Russian). doi:10.31857/S268695432207013X
14. Andreychenko A. E., Gusev A. V. Perspectives on the application of large language models in healthcare. *National Health Care*, 2023, vol. 4, no. 4, pp. 48–55 (In Russian). doi:10.47093/2713-069X.2023.4.4.48-55, EDN: MPDTJU
15. Evlevskaya N. V., Kazantsev A. A. Ensuring the security of complex systems with the integration of large language models: Threat analysis and protection methods. *Ekonomika i kachestvo sistem svyazi*, 2024, no. 4 (34), pp. 129–144.

EDN: CJEAAZ. Available at: https://journal-ekss.ru/wp-content/uploads/2024/12/129-144.pdf (accessed 10 December 2024) (In Russian).
16. Mudarova R. M., Namiot D. E. Countering prompt injection attacks on large language models. *International Journal of Open Information Technologies*, 2024, vol. 12, no. 5, pp. 39–48. EDN: LVERMB. Available at: http://injoit.org/index.php/j1/article/view/1845/1682 (accessed 10 December 2024) (In Russian).
17. Samonov A. V., Burova I. O. Methodology for the development of automated software code generation tools by fine-tuning large language models. *Cybersecurity Issues*, 2024, no. 3 (61), pp. 68–75 (In Russian). doi:10.21681/2311-3456-2024-3-68-75, EDN: NPSBQW
18. Zelenkov Yu. A. Knowledge management in organization and the large language models. *Russian Management Journal*, 2024, vol. 22, no. 3, pp. 573–601 (In Russian). doi:10.21638/spbu18.2024.309, EDN: KKHPPZ
19. O'Leary D. E. The rise and design of enterprise large language models. *IEEE Intelligent Systems*, 2024, vol. 39, no. 1, pp. 60–63. doi:10.1109/MIS.2023.3345591
20. Yu W., Zhu C., Li Z., Hu Z., Wang Q., Ji H., Jiang M. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 2022, vol. 54, no. 11s, pp. 1–38. doi:10.1145/3512467
21. Kosenko D. P., Kuratov Y. M., Zharikova D. R. Accessible Russian large language models: Open-source models and instructive datasets for commercial applications. *Doklady Rossijskoj Akademii Nauk. Mathematika, Informatika, Processy Upravlenia*, 2023, vol. 514, no. 2, pp. 262–269 (In Russian). doi:10.31857/S2686954323602063, EDN: GECIST
22. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 9459–9474.
23. Cai D., Wang Y., Liu L., Shi S. Recent advances in retrieval-augmented text generation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 3417–3419. doi:10.1145/3477495.3532682
24. Fan W., Ding Y., Ning L., Wang S., Li H., Yin D., Chua T.-S., Li Q. A survey on rag meeting LLMs: Towards retrieval-augmented large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6491–6501. doi:10.1145/3637528.3671470
25. Li Z., Li C., Zhang M., Mei Q., Bendersky M. Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 881–893. doi:10.18653/v1/2024.emnlp-industry.66
26. Singh J. P., Agrawal Sh. Automating threat intelligence analysis with Retrieval Augmented Generation (RAG) for enhanced cybersecurity posture. *International Journal of Science and Research (IJSR),* 2024, vol. 13, iss. 5, pp. 251–255. doi:10.21275/SR24502103758
27. Zhang Y., Du T., Ma Y., Wang X., Xie Y., Yang G., Lu Y., Chang E. C. AttacKG+: Boosting attack graph construction with Large Language Models. *Computers & Security*, 2024, vol. 150, pp. 104220. doi:10.1016/j.cose.2024.104220
28. Yamin M. M., Hashmi E., Ullah M., Katt B. Application of LLMs for generating cyber security exercise scenarios. *Research Square*, Feb. 2024. doi:10.21203/rs.3.rs-3970015/v1. Available at: https://www.researchsquare.com/article/rs-3970015/v1 (accessed 10 December 2024).
29. Wu Q., Bansal G., Zhang J., Wu Y., Li B., Zhu E., Jiang L., Zhang X., Zhang S., Liu J., Awadallah A. H., White R. W., Burger D., Wang C. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. *First Conference on Language Modeling*, Pennsylvania, PA, 2024, pp. 1–46. Available at: https://openreview.net/pdf?id=BAakY1hNKS (accessed 10 December 2024).
30. Mavroeidis V., Bromander S. Cyber threat intelligence model: An evaluation of taxonomies, sharing standards, and ontologies within cyber threat intelligence. *2017 European Intelligence and Security Informatics Conference (EISIC).* IEEE, 2017, pp. 91–98.
31. Zulkarneev R. Kh., Yusupova N. I., Smetanina O. N., Gayanova M. M., Vulfin A. M. Method and models of extraction of knowledge from medical documents. *Informatics and Automation*, 2022, vol. 21, no. 6, pp. 1169–1210 (In Russian). doi:10.15622/ia.21.6.4, EDN: ASOOVS
32. Rajapaksha S., Rani R., Karafili E. A RAG-based question-answering solution for cyber-attack investigation and attribution. *Workshop on Security and Artificial Intelligence 2024, 29th European Symposium on Research in Computer Security*, 2024. Available at: http://eprints.soton.ac.uk/id/eprint/493520 (accessed 10 December 2024).
33. Vasilyev V. I., Vulfin A. M., Kuchkarova N. V. Automation of software vulnerabilities analysis on the basis of text mining technology. *Cybersecurity Issues*, 2020, no. 4 (38), pp. 22–31 (In Russian). doi:10.21681/2311-3456-2020-04-22-31, EDN: TMRYIT
34. Topsakal O., Akinci T. C. Creating Large Language Model applications utilizing LangChain: A primer on developing LLM apps fast. *International Conference on Applied Engineering and Natural Sciences*, 2023, vol. 1, no. 1, pp. 1050–1056. doi:https://doi.org/10.59287/icaens.1127
35. Auger T., Saroyan E. *Overview of the OpenAI APIs*. In: *Generative AI for Web Development: Building Web Applications Powered by OpenAI APIs and Next.js*. Berkeley, CA, Apress, 2024, pp. 87–116.
36. Lin C. Y. ROUGE: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, 2004, pp. 74–81.
37. Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318. https://doi.org/10.3115/1073083.1073135
38. Es S., James J., Anke L. E., Schockaert S. RAGAs: Automated evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150–158. doi:10.48550/arXiv.2309.15217