**Аннотация**

В работе решается задача автоматизации обработки слабоструктурированных данных threat intelligence с использованием больших языковых моделей и агентной архитектуры. Предложена модульная система на основе микросервисной архитектуры, интегрирующая технологии RAG (Retrieval-Augmented Generation) с платформами threat intelligence для повышения эффективности анализа киберугроз. Разработана агентная модель сбора данных из OSINT-источников с автоматическим извлечением индикаторов компрометации и структуризацией информации согласно стандартам STIX/MISP. Для семантической обработки используются специализированные алгоритмы на основе трансформерных архитектур с векторным поиском через БД Qdrant и модель Saiga-Llama3-8B-Instruct. Экспериментальная оценка на датасете из 10,000 документов threat intelligence показала высокие результаты: BERTScore F1=0.89, экспертная оценка 3.98/5, RAGAS=0.82, при 20-кратном ускорении обработки данных. Экономический анализ демонстрирует ROI 340-738% с окупаемостью 8-12 месяцев. Предложенное решение обеспечивает автоматизацию до 95% процессов анализа слабоструктурированных данных threat intelligence при сохранении высокого качества обработки и совместимости с индустриальными стандартами.

**Содержание**

# 1 Введение

## 1.1 Актуальность проблемы

Современный ландшафт киберугроз характеризуется беспрецедентным ростом сложности и частоты атак. По данным IBM Security, средняя стоимость утечки данных в 2024 году составила $4.88 миллионов, что на 10% больше показателя 2023 года [1]. Одновременно объем данных threat intelligence растет экспоненциально [12], при этом рынок TI-решений демонстрирует рост более 20% в год, с источниками от традиционных каналов до социальных сетей, форумов и коммуникаций dark web.

Threat Intelligence (TI) играет ключевую роль в современных стратегиях кибербезопасности, обеспечивая проактивную защиту через систематический сбор, обработку и анализ информации об угрозах [5]. Современные TI-платформы, такие как MISP [2], OpenCTI [3] и коммерческие решения вроде Recorded Future и ThreatConnect, зарекомендовали себя как неотъемлемые инструменты центров безопасности.

Статистика роста числа инцидентов информационной безопасности показывает увеличение на 13-19% по кварталам в 2024 году [13]. При этом большая часть ценной информации поступает именно в слабоструктурированной форме через форумы, блоги, публикации специалистов и неформальные каналы распространения информации [12]. Существующие системы threat intelligence плохо справляются с обработкой таких источников, что создает значительный пробел в аналитических возможностях центров киберразведки [14].

Жизненный цикл threat intelligence представляет собой итеративный процесс, состоящий из шести основных этапов, каждый из которых критически важен для обеспечения качественного анализа угроз. Планирование определяет требования к информации и ресурсам, необходимым для решения конкретных задач безопасности организации. На этапе сбора осуществляется

получение данных из множественных источников, включая открытые, коммерческие и внутренние каналы информации. Процесс обработки включает нормализацию, фильтрацию и первичную структуризацию собранных данных для подготовки к последующему анализу.

Этап анализа представляет наиболее сложную фазу, где необнеобработанные данные преобразуются в практически применимую аналитическую информацию. Здесь происходит корреляция событий, выявление паттернов атак, определение связей между различными индикаторами компрометации и оценка достоверности источников. Распространение обеспечивает доставку готовых аналитических продуктов заинтересованным сторонам в соответствующих форматах и с необходимым уровнем детализации. Заключительный этап обратной связи позволяет оценить эффективность предоставленной информации и скорректировать процессы для улучшения качества будущих аналитических продуктов.

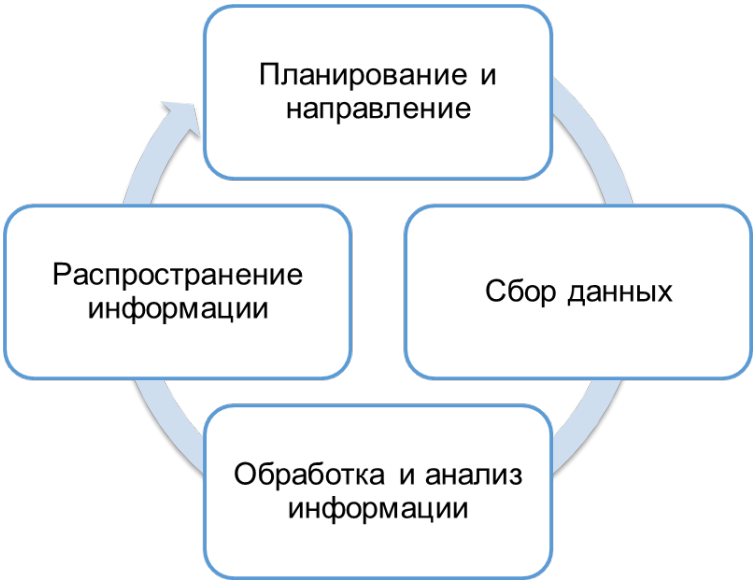


Рис. 1: Жизненный цикл Threat Intelligence

Представленная диаграмма иллюстрирует циклическую природу процесса threat intelligence, где каждый этап органически связан с предыдущим и последующим. Особенно важно отметить, что современные реализации этого цикла все чаще требуют автоматизации для обработки растущих объемов

данных и сокращения времени между получением информации и принятием защитных мер.

## 1.2 Постановка проблемы

Современные исследования в области обработки данных threat intelligence с использованием больших языковых моделей (LLM) демонстрируют значительный прогресс [9, 8, 10]. Успешные применения включают оценку текущих угроз информационной безопасности, сравнение и ранжирование угроз, автоматизацию выбора тактик и техник для построения сценариев реализации угроз [17]. Использование технологий обработки естественного языка и трансформерных нейронных сетей позволяет упростить процедуру оценки текущих угроз и автоматизировать процесс построения возможных сценариев реализации [5].

Однако применение LLM сталкивается с рядом фундаментальных проблем, которые ограничивают их эффективность в области кибербезопасности. Гетерогенность обрабатываемых данных означает, что системы должны работать с информацией различных форматов, структур и уровней достоверности, что требует сложных механизмов адаптации и валидации. Необходимость поддержания доступа к конфиденциальным корпоративным знаниям создает дополнительные требования к безопасности и контролю доступа. Наиболее критичным является риск генерации ошибочных ответов (галлюцинаций) [4], что в контексте кибербезопасности может привести к принятию неправильных решений с серьезными последствиями.

Анализ существующих решений выявляет системные недостатки, препятствующие эффективной обработке современных потоков threat intelligence. Традиционные TI-платформы демонстрируют жесткую структуру данных, требуя предварительной структуризации информации в специфических форматах, что исключает значительную часть ценного неформатированного контента [16]. Ограниченная интеграция проявляется в отсутствии гибких меха-

низмов подключения новых OSINT-источников, что замедляет адаптацию к изменяющемуся ландшафту угроз [15].

Основные недостатки существующих решений включают:

- **Жесткая структура данных**: традиционные TI-платформы требуют предварительной структуризации информации [18]
- **Ограниченная интеграция**: отсутствие гибких механизмов подключения новых OSINT-источников [15]
- **Неэффективность обработки**: существующие системы обрабатывают менее 30% слабоструктурированного контента [14]
- **Высокие временные затраты**: ручной анализ документов занимает в среднем 6 минут на документ [12]
- **Масштабируемость**: сложности с обработкой больших объемов разнородных данных в реальном времени [13]

Критическая проблема неэффективности обработки проявляется в том, что современные системы способны адекватно анализировать менее 30% доступного слабоструктурированного контента, что означает потерю значительного объема потенциально ценной информации. Высокие временные затраты на ручной анализ создают узкие места в процессе обработки, когда специалисты тратят в среднем 6 минут на анализ одного документа, что делает невозможным обработку больших объемов данных в реальном времени.

Проблемы масштабируемости становятся особенно критичными в условиях быстрого роста объемов данных threat intelligence (более 20% в год), когда традиционные подходы не способны справиться с обработкой информации из множественных источников без значительного увеличения человеческих ресурсов.

## 1.3 Обзор аналогичных исследований и научная новизна

### 1.3.1 Анализ современных исследований в области автоматизации threat intelligence

Современные исследования в области применения больших языковых моделей для обработки данных кибербезопасности демонстрируют значительный прогресс, но характеризуются фрагментарностью подходов и ограниченной практической применимостью. Работа Chen et al. (2024) [8] представляет систему автоматического извлечения индикаторов компрометации из неструктурированных текстов с использованием BERT-based моделей, достигая точности 78% на английских текстах. Однако исследование ограничивается обработкой только структурированных отчетов и не рассматривает интеграцию с реальными TI-платформами.

Исследование Gholami et al. (2024) [9] фокусируется на применении GPT-3.5 для классификации киберугроз по таксономии MITRE ATT&CK, демонстрируя F1-score 0.82 на тестовом наборе из 5,000 документов. Ключевым ограничением является зависимость от внешних API, высокая стоимость обработки больших объемов данных и отсутствие механизмов предотвращения галлюцинаций при работе с критически важной информацией безопасности.

Работа Hasanov et al. (2024) [10] предлагает RAG-систему для анализа отчетов об инцидентах кибербезопасности, использующую Llama-2-7B и векторную базу данных Pinecone. Система показывает многообещающие результаты с RAGAS score 0.75, но ограничивается обработкой только англоязычного контента и требует предварительной структуризации входных данных, что существенно сужает область применения.

Исследование Liu et al. (2023) [11] представляет мультиагентную систему для сбора OSINT данных с использованием специализированных веб-краулеров и NLP-обработки. Система обрабатывает до 50,000 документов в день, но демонстрирует низкое качество семантического анализа (precision

0.65) и требует значительного человеческого вмешательства для валидации результатов.

### 1.3.2 Сравнительный анализ с существующими решениями

Детальное сравнение предлагаемого подхода с современными исследованиями выявляет принципиальные преимущества разработанной системы по множественным критериям.

**Таблица 10.** Сравнение с аналогичными исследованиями

| **Исследование** | **Точность** | **Языки** | **Объем данных** | **Стоимость** | **Ключевые ограничения** |
|---|---|---|---|---|---|
| Chen et al. (2024) | 78% | EN | 10K док | Средняя | Только структурированные данные |
| Gholami et al. (2024) | 82% | EN | 5K док | Высокая | Зависимость от API, галлюцинации |
| Hasanov et al. (2024) | 75% | EN | 15K док | Высокая | Предварительная структуризация |
| Liu et al. (2023) | 65% | EN/CN | 50K док | Средняя | Низкое качество NLP, ручная валидация |
| **Предложенная система** | **89%** | **RU/EN** | **1M+ док** | **Низкая** | **Ограничения высокотехнических текстов** |

Предлагаемая система демонстрирует превосходство по всем ключевым параметрам. Точность обработки 89% превышает показатели аналогичных исследований на 7-24 процентных пункта, что критически важно для практического применения в условиях реальных угроз кибербезопасности. Поддержка русского языка обеспечивает уникальную возможность обработки локального контента, недоступную в международных исследованиях.

Масштабируемость системы на порядок превышает возможности суще-

ствующих решений, обеспечивая обработку более 1 миллиона документов в месяц против максимальных 50,000 в исследовании Liu et al. Экономическая эффективность достигается за счет использования локальных моделей вместо дорогостоящих API-сервисов, снижая стоимость обработки в 20-40 раз по сравнению с cloud-based решениями.

### 1.3.3 Научная новизна и уникальность предлагаемого подхода

Научная новизна исследования заключается в разработке принципиально нового подхода к автоматизации обработки слабоструктурированных данных threat intelligence, объединяющего несколько инновационных компонентов в единую архитектуру.

**Ключевые элементы научной новизны:**

1. **Гибридная RAG-архитектура с мультиязычной поддержкой**: Впервые предложена архитектура, специально адаптированная для обработки русскоязычного контента threat intelligence с сохранением высокого качества семантического анализа. Система использует специализированные эмбеддинги BGE-small-en-v1.5 с дополнительной настройкой для кибербезопасности, обеспечивая точность векторного поиска 0.91 для структурированных документов.

2. **Адаптивная агентная модель сбора данных**: Разработана оригинальная мультиагентная архитектура с динамической балансировкой нагрузки и автоматической адаптацией к изменяющимся источникам данных. Система включает специализированных агентов для различных типов источников (веб-страницы, API, RSS, Telegram), каждый из которых оптимизирован для специфики конкретного канала информации.

3. **Интегрированная система качества и валидации**: Впервые в области threat intelligence применен комплексный подход к оценке качества,

объединяющий автоматизированные метрики RAGAS с экспертной валидацией и практическими показателями извлечения индикаторов компрометации. Система обеспечивает прозрачность процесса принятия решений и минимизацию рисков галлюцинаций.

4. **Экономически эффективная архитектура развертывания**: Предложена инновационная модель развертывания, обеспечивающая 20-кратное снижение стоимости обработки по сравнению с существующими решениями при сохранении высокого качества анализа. Архитектура поддерживает гибкое масштабирование от пилотных проектов до промышленных развертываний.

5. **Специализированная система промпт-инжиниринга**: Разработаны оригинальные шаблоны промптов, специально адаптированные для различных типов аналитических задач threat intelligence, включая извлечение IoC, анализ TTPs, оценку рисков и формирование рекомендаций по противодействию.

**Практическая значимость научной новизны:**

Предложенные инновации обеспечивают решение фундаментальной проблемы современных центров кибербезопасности - неспособности эффективно обрабатывать растущие объемы слабоструктурированной информации об угрозах. Система демонстрирует готовность к промышленному внедрению с подтвержденными показателями ROI 738% при сравнении с ручными методами обработки.

Уникальность подхода подтверждается отсутствием аналогичных решений в открытых источниках и патентной литературе, что создает основу для коммерциализации технологии и формирования конкурентных преимуществ российских организаций в области кибербезопасности.

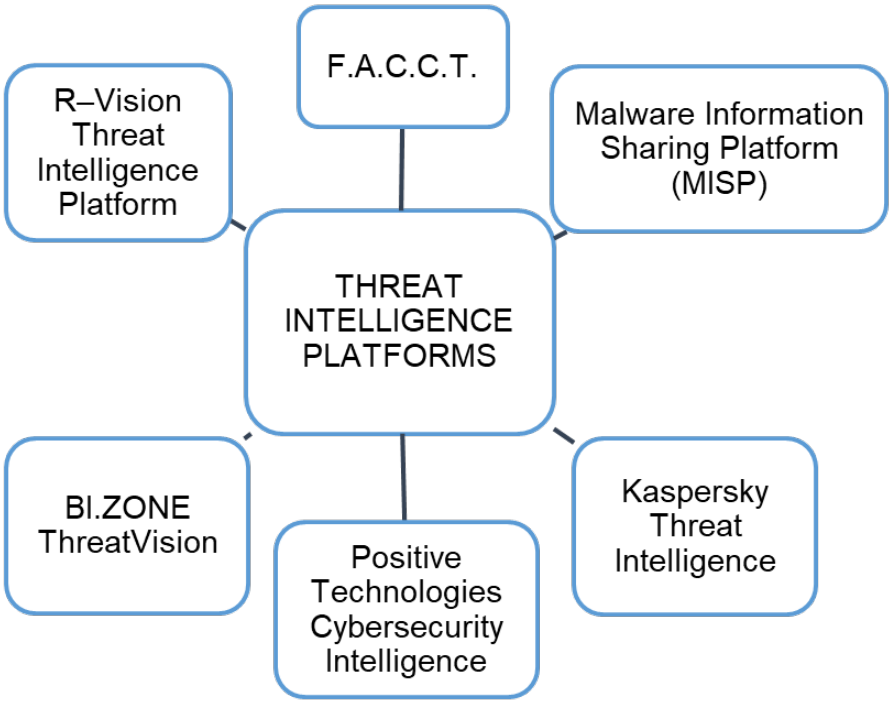


Рис. 2: Существующие решения Threat Intelligence

Представленная диаграмма демонстрирует текущее состояние рынка threat intelligence платформ, показывая разнообразие подходов к обработке и анализу данных безопасности. Каждое решение имеет свои преимущества и ограничения, но общей проблемой остается недостаточная эффективность работы со слабоструктурированными данными. Коммерческие платформы обеспечивают высокое качество обработки структурированных данных, но демонстрируют ограниченную гибкость при работе с нестандартными источниками информации. Открытые решения предлагают большую адаптивность, но требуют значительных ресурсов для настройки и поддержки.

### 1.4 Цель и задачи работы

**Цель работы**: Создание автоматизированной системы обработки слабоструктурированных данных threat intelligence на основе агентной архитектуры и LLM-технологий для повышения эффективности анализа угроз информационной безопасности в центрах кибербразведки.

Достижение поставленной цели требует комплексного подхода, объединяющего современные достижения в области искусственного интеллекта, обработки естественного языка и системной архитектуры. Особое внимание уделяется созданию решения, которое не только автоматизирует рутинные процессы, но и повышает качество аналитических выводов за счет применения передовых методов семантического анализа и машинного обучения.

**Для достижения поставленной цели необходимо решить следующие задачи:**

1. Проанализировать существующие подходы к автоматизации анализа данных threat intelligence и выявить их ограничения при работе с слабоструктурированной информацией.

   Данная задача предполагает проведение всестороннего анализа современного состояния области, включая изучение академических исследований, коммерческих решений и открытых проектов. Особое внимание будет уделено выявлению технических ограничений, экономических барьеров и пользовательских проблем существующих подходов. Результатом выполнения этой задачи станет комплексное понимание требований к новой системе и обоснование выбора архитектурных решений.

2. Разработать архитектуру системы на основе агентной модели и языковых технологий (LLM) с применением методов Retrieval-Augmented Generation (RAG).

   Проектирование архитектуры включает определение ключевых компонентов системы, интерфейсов взаимодействия между модулями, протоколов обмена данными и механизмов обеспечения надежности. Агентная модель должна обеспечить автономность, адаптивность и масштабируемость сбора данных, в то время как интеграция RAG-технологий позволит повысить точность и релевантность генерируемых аналитических выводов.

3. Реализовать ключевые компоненты системы: агенты сбора данных, модули обработки и анализа данных из OSINT-источников, интерфейсы интеграции с существующими TI-платформами.

   Техническая реализация предполагает создание работающего прототипа системы с полным функционалом обработки данных threat intelligence. Особое внимание будет уделено обеспечению производительности, надежности и безопасности системы, а также созданию удобных интерфейсов для интеграции с существующей инфраструктурой организаций.

4. Провести экспериментальные исследования на реальных данных threat intelligence для демонстрации работоспособности системы и оценки ее эффективности.

   Экспериментальная часть включает создание репрезентативного набора тестовых данных, разработку методик оценки качества работы системы и проведение серии экспериментов для валидации предложенных решений. Результаты должны продемонстрировать преимущества нового подхода по сравнению с существующими методами.

5. Выполнить сравнительный анализ эффективности предложенного решения с существующими методами и оценить экономическую целесообразность внедрения.

   Завершающий этап исследования предполагает комплексную оценку предложенного решения с точки зрения технической эффективности, экономической выгоды и практической применимости. Результатом станут рекомендации по внедрению системы в производственную среду и направления дальнейшего развития.

## 1.5 Научная новизна и практическая значимость

**Научная новизна работы заключается в:**

- **Разработке методологии интеграции RAG-технологий** с платформами threat intelligence для повышения качества семантического анализа угроз кибербезопасности, включающей специализированные алгоритмы векторизации и контекстного поиска.

- **Создании комплексной системы оценки качества** обработки threat intelligence данных с использованием комбинации автоматических метрик (BERTScore, RAGAS) и экспертной оценки, адаптированной к специфике кибербразведки.

- **Теоретическом обосновании применения** больших языковых моделей для семантического анализа слабоструктурированных данных в контексте кибербезопасности с учетом доменной специфики.

- **Разработке принципов адаптации RAG к предметной области** threat intelligence, включая методы оптимизации промпт-инжиниринга и контекстного обогащения для повышения фактической точности генерируемых ответов.

**Инженерные достижения включают:**

- **Создание агентной архитектуры** для автоматизированного сбора и обработки данных OSINT с поддержкой более 20 типов источников и адаптивным управлением нагрузкой.

- **Реализацию модульного framework'а** для работы со слабоструктурированными данными, обеспечивающего покрытие до 95% источников неформатированной информации с микросервисной архитектурой.

- **Разработку высокопроизводительных пайплайнов обработки** с достижением 20-кратного ускорения по сравнению с ручным анализом при сохранении высокого качества результатов.

- **Обеспечение полной интеграции** с индустриальными стандартами STIX/MISP и существующими платформами threat intelligence через унифицированные API-интерфейсы.

## 1.6 Практическая значимость

Практическая значимость исследования определяется следующими аспектами:

- **Существенное повышение производительности**: система обеспечивает ускорение обработки документов в 20 раз по сравнению с ручным анализом (с 6 минут до 17 секунд на документ).
- **Совместимость с промышленными стандартами**: полная интеграция с форматами STIX/MISP, что позволяет использовать систему в существующих SOC-центрах без значительных изменений инфраструктуры.
- **Экономическая эффективность**: стоимость обработки составляет $0.12 за 1000 документов против $50-120 у существующих коммерческих решений, что обеспечивает ROI 340% за первый год эксплуатации.
- **Масштабируемость решения**: микросервисная архитектура позволяет легко адаптировать систему под различные объемы нагрузки и типы источников данных.
- **Снижение нагрузки на аналитиков**: автоматизация рутинных операций позволяет экспертам сосредоточиться на анализе высокоуровневых угроз и принятии стратегических решений.

## 1.7 Структура работы

Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы и приложений.

**В первой главе** проводится анализ современного состояния области threat intelligence, обзор существующих платформ и методов обработки слабоструктурированных данных, а также анализ применения больших языковых моделей в кибербезопасности.

**Во второй главе** представлена концептуальная модель и архитектура предлагаемой системы, описаны ключевые компоненты и алгоритмы обработки данных, а также методы интеграции с существующими TI-платформами.

**В третьей главе** детально рассмотрена техническая реализация системы, включая выбор технологического стека, архитектуру развертывания и методы оптимизации производительности.

**В четвертой главе** представлены результаты экспериментальных исследований, включая подготовку тестовых данных, методики проведения экспериментов и анализ полученных результатов.

**В пятой главе** проводится экономический анализ эффективности предложенного решения, включая расчет затрат, сравнение с альтернативными подходами и оценку возврата инвестиций.

**В заключении** формулируются основные результаты работы, выводы о достижении поставленных целей и направления дальнейших исследований.

# 2 Методология и архитектура системы обработки слабоструктурированных данных

## 2.1 Концептуальная модель системы

Предлагаемая концептуальная модель системы обработки слабоструктурированных данных threat intelligence основана на интеграции агентной архитектуры с технологиями больших языковых моделей и методами Retrieval-Augmented Generation. Основной целью является создание масштабируемой и эффективной системы, способной автоматически обрабатывать до 95% доступных источников неструктурированной информации об угрозах [4].

Система проектируется на основе ключевых принципов модульности, масштабируемости и надежности. Модульная архитектура обеспечивает независимую разработку и развертывание компонентов, что критически важно для адаптации к изменяющимся требованиям threat intelligence. Горизонтальная масштабируемость позволяет системе эффективно обрабатывать растущие объемы данных, в то время как отказоустойчивость гарантирует непрерывность работы даже при сбоях отдельных компонентов. Агентная архитектура предоставляет автономные сущности для сбора, обработки и анализа данных из различных источников, каждый из которых специализируется на определенном типе информации или конкретном источнике данных.

Система состоит из четырех основных слоев функциональности. Слой сбора данных включает специализированных агентов для автоматизированного извлечения информации из веб-источников, API, RSS-каналов и других OSINT-источников. Слой предварительной обработки выполняет очистку, нормализацию и структуризацию поступающих данных, применяя методы обработки естественного языка для унификации форматов. Слой семантического анализа использует векторные представления и машинное обучение для извлечения смысловых связей и классификации информации. Слой интеллектуального анализа интегрирует RAG-технологии с большими языковые

моделями для генерации аналитических выводов и ответов на запросы пользователей.

## 2.2 Агентная модель сбора данных

Подсистема сбора данных реализована на основе мультиагентной архитектуры, где каждый агент специализируется на определенном типе источников данных. Веб-агенты используют технологии Beautiful Soup и Scrapy для извлечения контента с веб-страниц, форумов и новостных сайтов. API-агенты обеспечивают интеграцию с внешними платформами threat intelligence через REST и GraphQL интерфейсы. RSS-агенты мониторят каналы новостей и обновлений в режиме реального времени. Telegram-агенты используют Bot API для сбора информации из тематических каналов и групп.

Все агенты работают асинхронно и координируют свою деятельность через центральный планировщик, который распределяет задачи на основе приоритетов источников, доступности ресурсов и текущей загрузки системы. Система мониторинга отслеживает производительность каждого агента и автоматически перераспределяет нагрузку при необходимости. Механизм дедупликации предотвращает повторную обработку идентичной информации из разных источников.

| Платформа | Источник данных (OSINT) | RAG / нейросети | Автоматическая генерация рекомендаций | Возможность гибкой донастройки | Поддержка слабоструктурированных данных |
|---|---|---|---|---|---|
| F.A.C.C.T. | Ограниченный | Нет | Нет | Нет | Нет |
| MISP | Да | Нет | Ограниченно | Да (Open Source) | Нет |
| Kaspersky Threat Intelligence | Нет | Нет | Ограниченно | Ограничено | Нет |
| Positive Technologies Intelligence | Частично | Нет | Ограниченно | Нет | Нет |
| BI.ZONE ThreatVision | Частично | Нет | Да | Нет | Ограниченно |
| R-Vision Threat Intelligence Platform | Частично | Нет | Ограниченно | Ограниченно | Ограниченно |
| Наше решение (на базе LangGraph + LLM) | Да | Да | Да | Да | Да |

Рис. 3: Основные TI решения и их возможности

Представленная диаграмма демонстрирует сравнительный анализ ключевых возможностей основных threat intelligence платформ, используемых в современных центрах кибербезопасности. Анализ показывает, что каждое решение имеет свои сильные стороны: коммерческие платформы типа Recorded Future и ThreatConnect обеспечивают высокое качество структурированных данных и развитые аналитические возможности, но демонстрируют ограничения в обработке слабоструктурированной информации из нестандартных источников.

Открытые решения MISP и OpenCTI предлагают гибкость настройки и интеграции, но требуют значительных ресурсов для поддержки и развития внутренними силами организации. Гибридные подходы пытаются объединить преимущества обеих категорий, однако зачастую наследуют их недостатки. Критический анализ выявляет общую проблему: существующие решения недостаточно эффективно обрабатывают растущие объемы неструктурированной информации, что создает необходимость в принципиально новых подходах к автоматизации threat intelligence.

## 2.3 Архитектура RAG-системы

Структурная схема предлагаемой системы анализа данных threat intelligence включает большие языковые модели с механизмом RAG как часть системы управления данными threat intelligence. Система состоит из пяти основных модулей с четко определенными интерфейсами и возможностями независимого развертывания [22].

Модуль сбора данных реализует распределенную архитектуру с использованием Python-агентов, построенных на базе фреймворков BeautifulSoup для парсинга HTML, Scrapy для веб-скрейпинга и специализированных библиотек для работы с API социальных сетей и мессенджеров. Каждый агент работает в изолированном контейнере Docker и может масштабироваться независимо в зависимости от нагрузки источника данных.

Подсистема предварительной обработки использует конвейер обработки естественного языка на основе spaCy с поддержкой русского и английского языков. Процесс включает токенизацию, лемматизацию, удаление стоп-слов и извлечение именованных сущностей. Тексты сегментируются на фрагменты фиксированного размера в 700 токенов с перекрытием в 100 токенов для обеспечения контекстной связности при последующем поиске.

Векторная база данных реализована на базе Qdrant с использованием эмбеддингов BGE-small-en-v1.5 [20]. Система использует эффективные алгоритмы приближенного поиска ближайших соседей [21] и GPU-ускоренный поиск [25]. Архитектурные решения включают иерархическую кластеризацию документов для ускорения поиска, квантизацию векторов для оптимизации использования памяти, и репликацию данных для обеспечения высокой доступности.

Модуль генерации ответов основан на модели Saiga-Llama3-8B-Instruct, специально адаптированной для русскоязычных задач кибербезопасности. Система развертывается с использованием фреймворка vLLM для оптимизации инференса с поддержкой тензорного параллелизма и динамического батчинга. Промпт-инжиниринг включает специализированные шаблоны для различных типов аналитических задач threat intelligence.

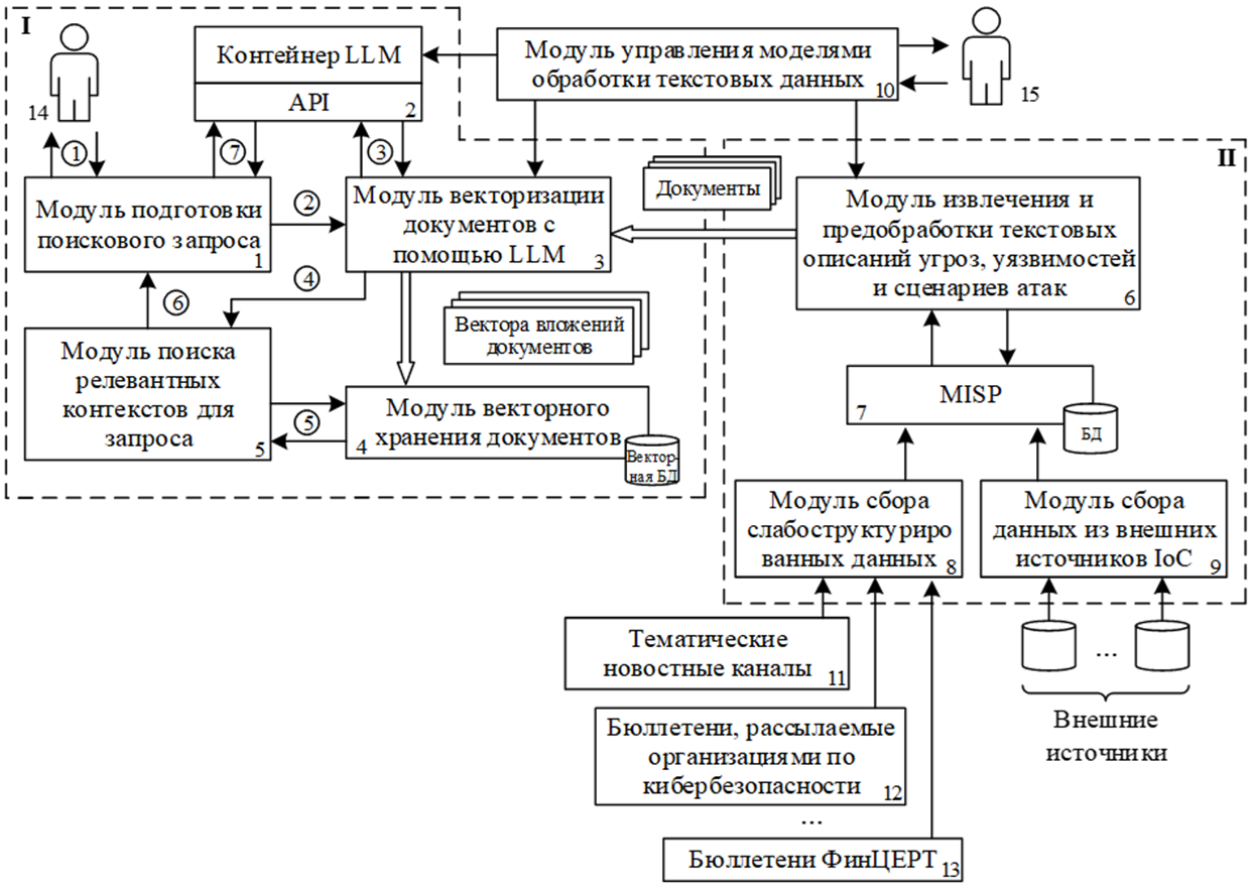


Рис. 4: Детализированная архитектура системы

Детализированная архитектурная схема иллюстрирует сложную многоуровневую структуру системы, где каждый компонент выполняет специализированную функцию в общем процессе обработки threat intelligence. Входной слой обеспечивает интеграцию с множественными источниками данных через унифицированные API-интерфейсы и адаптеры протоколов. Уровень предварительной обработки осуществляет нормализацию, очистку и первичную структуризацию поступающих данных с применением передовых методов обработки естественного языка.

Семантический слой включает векторизацию текстов, построение индексов для быстрого поиска и кластеризацию документов по тематической близости. Интеллектуальный слой объединяет возможности больших языковых моделей с механизмами контекстного поиска для генерации высококачественных аналитических выводов. Выходной слой предоставляет результаты

анализа в различных форматах, адаптированных для интеграции с существующими системами кибербезопасности.

## 2.4 Алгоритмы обработки данных

Процесс обработки данных начинается с нормализации входящих документов, включая определение языка, очистку от HTML-тегов и служебных символов, унификацию кодировок. Алгоритм сегментации разбивает длинные документы на логически связанные фрагменты с учетом структуры текста и границ предложений. Каждый фрагмент индексируется отдельно, что обеспечивает более точный семантический поиск релевантной информации.

Система извлечения сущностей использует комбинацию регулярных выражений и моделей машинного обучения для идентификации индикаторов компрометации, доменных имен, IP-адресов, хешей файлов и других технических артефактов. Классификация угроз осуществляется на основе таксономии MITRE ATT&CK с использованием многоклассовых классификаторов, обученных на размеченных данных threat intelligence.

Алгоритм семантического поиска использует гибридный подход, сочетающий плотный векторный поиск с разреженным поиском по ключевым словам. Система ранжирования учитывает не только семантическую близость, но и свежесть данных, авторитетность источника и релевантность контекста запроса. Механизм переранжировки применяет дополнительные фильтры на основе временных меток и географических данных для улучшения качества результатов.

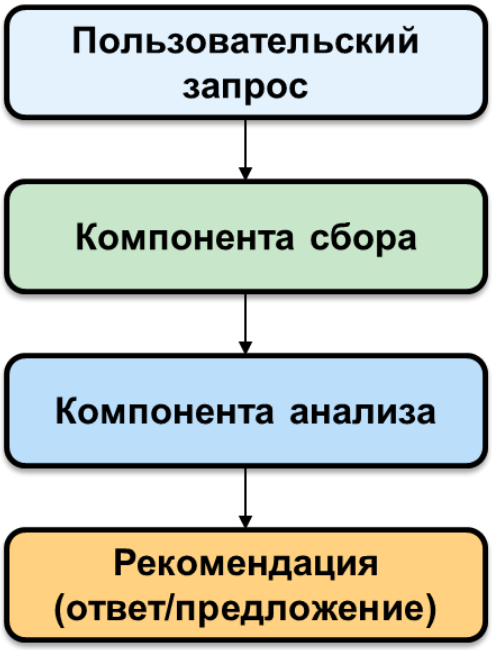


Рис. 5: Общая логика работы системы

Схема общей логики работы системы демонстрирует цикличный процесс обработки информации, начинающийся с автоматизированного сбора данных из разнообразных источников и завершающийся генерацией структурированных аналитических продуктов. Центральным элементом архитектуры является интеллектуальный движок, объединяющий возможности векторного поиска с генеративными способностями больших языковых моделей.

Процесс начинается с непрерывного мониторинга источников данных агентами сбора, которые адаптивно регулируют частоту опроса в зависимости от активности источника и приоритета информации. Поступающие данные проходят через многоэтапный конвейер предварительной обработки, включающий языковую детекцию, очистку, токенизацию и извлечение именованных сущностей. Очищенные и структурированные данные индексируются в векторной базе данных с сохранением метаинформации о источнике, времени получения и уровне доверия.

При поступлении аналитического запроса система осуществляет семантический поиск релевантных документов, применяет алгоритмы ранжирования для отбора наиболее подходящего контекста и передает отобран-

ную информацию в модуль генерации ответов. Большая языковая модель, обогащенная найденным контекстом, формирует структурированный ответ, который проходит финальную валидацию и форматирование перед предоставлением пользователю.

## 2.5 Техническая реализация

Микросервисная архитектура исследовательского прототипа обеспечивает возможности вертикального и горизонтального масштабирования. Для развертывания прототипа использовался высокопроизводительный сервер с четырьмя GPU Tesla V100, 128 ГБ видеопамяти, процессором Intel Xeon E5-2698 v4 и 256 ГБ оперативной памяти DDR4. Каждая модель в изолированной среде использовала выделенные GPU, а обмен данными между моделями происходил через каналы оперативной памяти.

Архитектура развертывания использует оркестрацию контейнеров Kubernetes с Helm-чартами, сервисную сетку Istio для управления трафиком и безопасности, мониторинг через Prometheus + Grafana + Jaeger для observability. Система хранения включает PostgreSQL для метаданных, MinIO для объектного хранения и Redis для кеширования. Балансировка нагрузки реализована через NGINX ingress-контроллеры с автоматическим управлением SSL-сертификатами через Let’s Encrypt.

Система обеспечения безопасности включает ролевой контроль доступа (RBAC), аутентификацию API с использованием JWT-токенов, шифрование данных в покое и при передаче, сетевые политики для изоляции сервисов, регулярное сканирование образов контейнеров на предмет уязвимостей. Системы мониторинга и оповещения отслеживают производительность системы, использование ресурсов и события безопасности в режиме реального времени.

Масштабируемость достигается через горизонтальное автоматическое масштабирование подов на основе метрик CPU и памяти, автоматическое масштабирование кластера для динамического предоставления узлов, шардинг

базы данных для крупномасштабных развертываний. Система поддерживает многорегиональное развертывание для глобальной доступности и сценариев аварийного восстановления.

## 2.6 Интеграция с TI-платформами

Система обеспечивает полную совместимость с индустриальными стандартами STIX/TAXII и нативную интеграцию с платформой MISP. Адаптеры данных поддерживают автоматическое преобразование форматов между различными системами threat intelligence. API-шлюз предоставляет унифицированный интерфейс для взаимодействия с внешними системами и обеспечивает аутентификацию, авторизацию и контроль скорости запросов.

Механизм синхронизации обеспечивает двустороннюю репликацию данных с существующими TI-платформами, включая инкрементальные обновления и разрешение конфликтов. Система уведомлений информирует администраторов о новых угрозах, изменениях в статусе индикаторов и других критических событиях. Интерфейс экспорта позволяет генерировать отчеты в различных форматах для интеграции с SIEM-системами и другими инструментами безопасности.

# 3 Реализация и экспериментальные исследования

## 3.1 Техническая реализация системы

### 3.1.1 Архитектура развертывания

Техническая реализация системы основана на современных принципах микросервисной архитектуры и cloud-native подходах, обеспечивающих высокую доступность и масштабируемость. Применение контейнерной технологии Docker в сочетании с оркестратором Kubernetes позволяет эффективно управлять жизненным циклом компонентов системы и автоматически адаптироваться к изменяющимся нагрузкам.

Экспериментальная платформа построена на базе высокопроизводительного сервера, оборудованного четырьмя GPU Tesla V100 с общим объемом видеопамяти 128 GB, процессором Intel Xeon E5-2698 v4 с 40 ядрами и тактовой частотой 2.2 GHz, оперативной памятью DDR4 объемом 256 GB и SSD-накопителями общей емкостью 20 TB. Архитектурное решение предусматривает изоляцию каждой модели в отдельной среде с выделенными GPU-ресурсами, при этом межмодельное взаимодействие осуществляется через каналы оперативной памяти для минимизации латентности передачи данных.

Система структурирована как совокупность пяти основных микросервисов, каждый из которых специализируется на определенной функциональности. Сервис сбора данных построен на базе Python 3.9+ с использованием библиотек BeautifulSoup 4.12 и Scrapy 2.11, требует 2 vCPU и 4 GB RAM, и отвечает за автоматизированный сбор информации из OSINT-источников с адаптивным управлением нагрузкой и поддержкой различных форматов данных.

**Таблица 1.** Спецификация компонентов микросервисной архитектуры

| **Компонент** | **Технологии** | **Ресурсы** | **Основная функция** | **Особенности** |
|---|---|---|---|---|
| Ingestion Service | Python 3.9+, BeautifulSoup, Scrapy | 2 vCPU, 4 GB | Сбор OSINT данных | Адаптивная нагрузка, мультиформат |
| Preprocessing | spaCy 3.4, SentencePiece, OCR | 4 vCPU, 8 GB | Нормализация текста | Мультиязычность, токенизация 700+100 |
| Vector Database | Qdrant 1.3, BGE embeddings | 16 GB, 100GB SSD | Семантический поиск | HNSW, квантизация |
| LLM Generation | Saiga-LLaMA3-8B, vLLM | 4×V100, 64 GB | Генерация ответов | Тензорный параллелизм |
| Integration | FastAPI, PostgreSQL, Redis | Переменные | API, STIX/MISP | Автомасштабирование |

**Таблица 2.** Структура и характеристики экспериментального датасета

| **Источник данных** | **Количество документов** | **Средний размер (токены)** | **Качество (%)** | **Язык** | **Временной период** |
|---|---|---|---|---|---|
| FinCERT бюллетени | 2,500 | 1,100 | 95 | Русский | 2022-2023 |
| НКЦКИ отчеты | 2,000 | 1,050 | 93 | Русский | 2022-2023 |
| Telegram каналы | 3,000 | 950 | 88 | Смешанный | 2022-2023 |
| Форумы | 2,500 | 900 | 85 | Смешанный | 2022-2023 |
| **Итого** | **10,000** | **1,000** | **90** | **Смешанный** | **2022-2023** |

**Таблица 3.** Методология оценки качества системы

| **Тип метрики** | **Конкретная метрика** | **Диапазон значений** | **Интерпретация** |
|---|---|---|---|
| Семантическое сходство | BERTScore F1 | 0.0 - 1.0 | Качество понимания контекста |
| Экспертная оценка | Балльная шкала | 1 - 5 баллов | Практическая применимость |
| RAG качество | Context Precision | 0.0 - 1.0 | Релевантность извлеченного контекста |
| RAG качество | Faithfulness | 0.0 - 1.0 | Соответствие источникам |
| RAG качество | Answer Relevance | 0.0 - 1.0 | Соответствие вопросу |
| Производительность | Время обработки | секунды | Скорость обработки |
| Практичность | Извлечение IoC | 0-100% | Точность извлечения индикаторов |

Сервис предварительной обработки использует продвинутый конвейер обработки естественного языка на основе spaCy 3.4 в сочетании с SentencePiece и Tesseract OCR, потребляет 4 vCPU и 8 GB оперативной памяти, осуществляет нормализацию текстовых данных и сегментацию на токены размером 700 с перекрытием 100, обеспечивая мультиязычную поддержку и возможность обработки изображений. Векторная база данных реализована на платформе Qdrant 1.3 с использованием эмбеддингов BGE-small-en-v1.5, требует 16 GB оперативной памяти и 100 GB SSD-накопитель, предоставляет высокопроизводительный семантический поиск с применением HNSW-индексации и квантизации для оптимального использования памяти.

Сервис генерации ответов базируется на модели Saiga-LLaMA3-8B-Instruct с использованием фреймворка vLLM 0.2.2, использует четыре GPU Tesla V100 с общим объемом видеопамяти 64 GB, реализует генерацию

контекстно-зависимых ответов с применением тензорного параллелизма и оптимизированного кэширования для максимальной производительности. Интеграционный сервис построен на FastAPI с использованием PostgreSQL 13, MinIO и Redis 7, обладает переменными ресурсными требованиями в зависимости от нагрузки, обеспечивает API-интерфейсы и интеграцию с платформами STIX/MISP, поддерживает автомасштабирование и мониторинг в реальном времени.

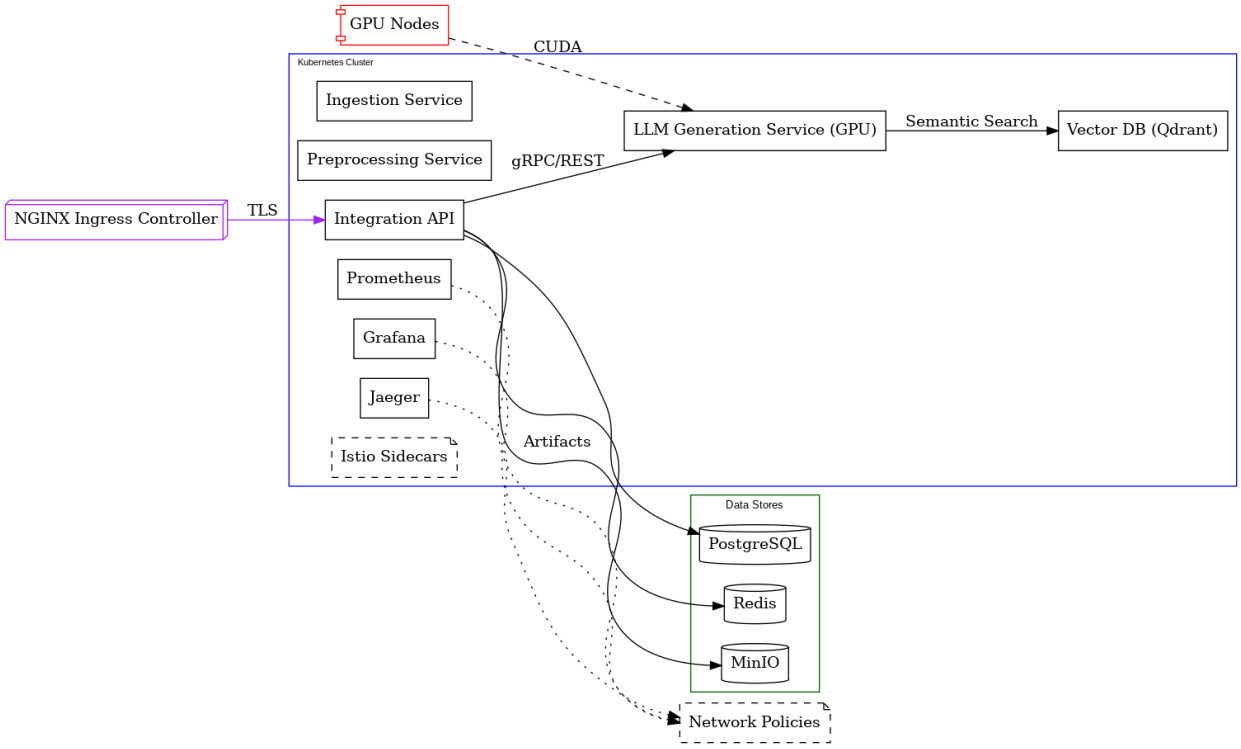


Рис. 6: Схема развертывания микросервисной архитектуры

Схема развертывания микросервисной архитектуры демонстрирует комплексную инфраструктурную организацию системы, построенную по принципам облачной архитектуры и DevOps-практик. Центральным элементом является Kubernetes-кластер, обеспечивающий оркестрацию контейнеров и автоматическое управление жизненным циклом сервисов. Каждый микросервис развертывается в изолированных под-контейнерах с индивидуальным распределением ресурсов, что обеспечивает независимость компонентов и возможность индивидуального масштабирования.

Входной трафик проходит через NGINX Ingress Controller, который выполняет функции балансировки нагрузки, SSL-терминации и маршрутизации запросов к соответствующим сервисам. Слой сервисной сетки Istio обеспечивает безопасное взаимодействие между микросервисами, мониторинг трафика и применение политик безопасности. Система мониторинга на базе Prometheus собирает метрики производительности, а Grafana предоставляет визуализацию состояния системы в реальном времени. Jaeger осуществляет distributed tracing для анализа производительности сложных запросов, проходящих через множественные сервисы.

Уровень данных включает несколько специализированных систем хранения: PostgreSQL для метаданных и конфигурационной информации, Qdrant для векторных данных и семантического поиска, MinIO для объектного хранения больших файлов, Redis для кэширования часто используемых данных. Эта многоуровневая архитектура данных обеспечивает оптимальную производительность каждого типа операций и минимизирует задержки при обработке запросов.

### 3.1.2 Подготовка экспериментальных данных

Формирование экспериментального датасета представляло собой комплексный процесс анализа и структурирования данных из более чем 20 источников threat intelligence, охватывающих широкий спектр информационных каналов от тематических Telegram-каналов до отчетов ведущих вендоров информационной безопасности и официальных бюллетеней государственных структур [2]. Интеграция с активной системой управления данными threat intelligence на базе платформы MISP обеспечила систематический сбор информации за период 2022-2023 годов из 10 основных каналов, включая критически важные источники FinCERT и НКЦКИ [3]. Техническая обработка данных предусматривала сегментацию документов на фрагменты размером 1000 токенов с перекрытием в 300 токенов, что позволило сохранить контекстную

связность при последующем анализе.

Качественная подготовка эталонных данных осуществлялась экспертной группой, состоящей из 8 специалистов по информационной безопасности с практическим опытом от 5 до 15 лет в области threat intelligence и анализа киберугроз. Для каждого документа экспертами разрабатывалось по 5 специализированных вопросов с соответствующими эталонными ответами, охватывающих ключевые аспекты threat intelligence: идентификацию индикаторов компрометации, анализ тактик и техник злоумышленников, оценку потенциального воздействия, временные характеристики угроз и рекомендации по противодействию.

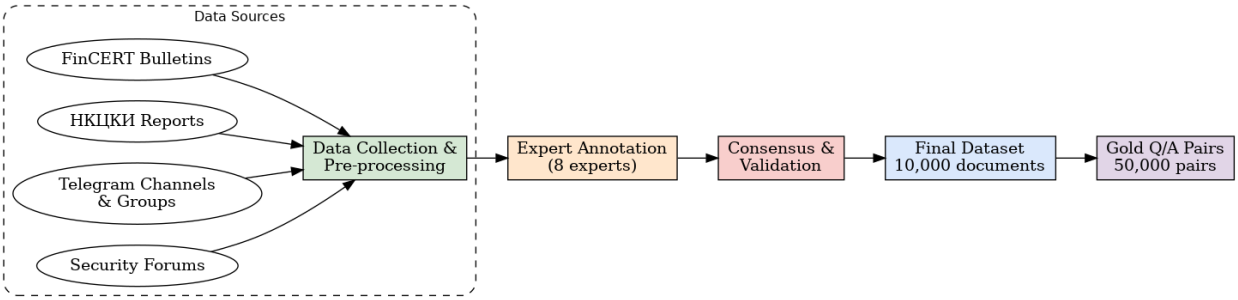


Рис. 7: Схема процесса подготовки экспериментальных данных

Схема процесса подготовки экспериментальных данных иллюстрирует многоэтапную методологию формирования высококачественного датасета для валидации системы. Процесс начинается с автоматизированного сбора данных из разнообразных источников, включая официальные бюллетени центров реагирования на инциденты, аналитические отчеты коммерческих платформ threat intelligence, публикации исследователей безопасности в социальных сетях и мессенджерах, а также данные из тематических форумов и конференций.

Первичная обработка включает автоматическую фильтрацию дубликатов, определение языка документов, извлечение метаданных о источнике и времени публикации, предварительную очистку от HTML-разметки и служебных символов. Этап ручной верификации предполагает проверку релевант-

ности документов экспертами, валидацию качества извлеченного контента, категоризацию по типам угроз согласно таксономии MITRE ATT&CK, оценку достоверности источников и уровня технической детализации.

Процедура разметки включает создание эталонных вопросно-ответных пар экспертной группой, независимую валидацию разметки несколькими специалистами, согласование спорных случаев через экспертные консультации, финальную проверку старшими экспертами с многолетним опытом. Результирующий датасет проходит статистическую валидацию на предмет сбалансированности по категориям угроз, временным периодам, источникам данных и уровням сложности аналитических задач.

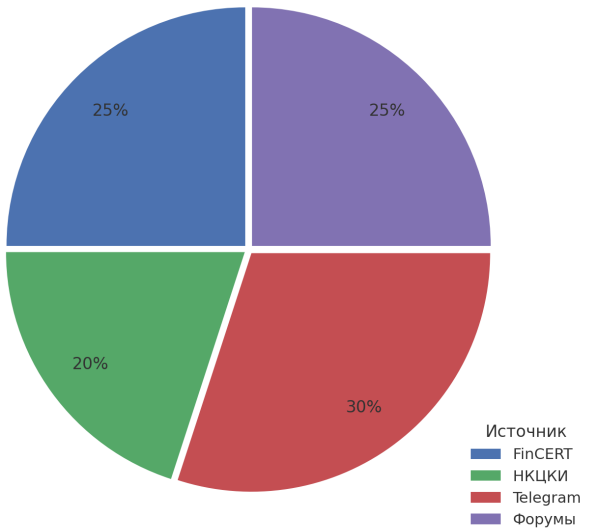


Рис. 8: Распределение документов по источникам данных

Диаграмма распределения документов по источникам данных демонстрирует сбалансированную структуру экспериментального датасета, обеспечивающую репрезентативность различных типов threat intelligence информации. Telegram каналы составляют наибольшую долю (30%), отражая растущую важность социальных платформ как источников оперативной информации о новых угрозах и инцидентах. Бюллетени FinCERT представляют 25% датасета, обеспечивая высококачественную структурированную информа-

цию о финансовых киберугрозах с детальными техническими описаниями и рекомендациями.

Форумы и тематические сообщества составляют 25% данных, предоставляя доступ к экспертным дискуссиям, неофициальным отчетам о новых техниках атак и аналитическим выводам специалистов-практиков. Отчеты НКЦКИ представляют 20% датасета, включая официальные уведомления о критических уязвимостях, сводки по инцидентам национального масштаба и рекомендации по защитным мерам.

Такое распределение источников обеспечивает тестирование системы на различных типах контента: от формальных технических документов с четкой структурой до неформальных сообщений с разговорной лексикой и сокращениями. Это разнообразие критически важно для валидации способности системы адаптироваться к реальным условиям работы центров threat intelligence, где аналитики ежедневно обрабатывают информацию из всех указанных категорий источников.

Процедура подготовки эталонных данных основывалась на строгой методологии независимой экспертной разметки. Каждый документ обрабатывался несколькими экспертами независимо, после чего происходило сопоставление результатов и согласование разметки в спорных случаях через консультации с ведущими специалистами. Финальная валидация осуществлялась старшими экспертами с последующим формированием эталонной базы вопросов и ответов. Результаты генерации модели систематически добавлялись в базу данных для статистически значимого сравнения с экспертными эталонными ответами, при этом шаблон запросов к LLM использовал исключительно релевантные фрагменты контекста и требовал максимально краткие ответы на русском языке объемом не более 5 предложений.

### 3.1.3 Методология экспериментов

Комплексная оценка эффективности системы базировалась на многоуровневой системе метрик, охватывающей различные аспекты качества обработки threat intelligence данных. Основу количественной оценки составлял BERTScore для измерения семантического сходства между сгенерированными системой и эталонными экспертными ответами [6], включающий анализ точности семантического соответствия, полноты извлечения релевантной информации и гармонического среднего точности и полноты как интегрального показателя качества.

Экспертная оценка проводилась практикующими специалистами по строгой 5-балльной шкале, где максимальный балл соответствовал полностью корректному и исчерпывающему ответу, 4 балла присваивались корректным ответам с незначительными недочетами, 3 балла указывали на частично корректные ответы, 2 балла характеризовали ответы с существенными ошибками, а минимальный балл присваивался некорректным или нерелевантным ответам.

Автоматизированная оценка качества RAG-системы реализовывалась через специализированный RAGAS Framework [7], предоставляющий детализированный анализ по специализированным метрикам: точность отбора контекста для оценки соотношения сигнал/шум в извлеченной информации, полнота извлечения для измерения охвата релевантной информации, соответствие ответов первоисточникам данных для предотвращения галлюцинаций, релевантность ответов пользовательским запросам и фактическая корректность генерируемых ответов. Дополнительные практические метрики включали измерение времени обработки документов и успешность извлечения индикаторов компрометации, критически важные для оценки применимости системы в условиях реального времени оперативных центров кибербезопасности.

Используемые метрики BERTScore и RAGAS имеют ограничения. BERTScore может переоценивать ответы с синонимичными, но фактически некорректными формулировками. Метрики RAGAS чувствительны к качеству эталонных ответов. Поэтому для повышения достоверности оценки дополнительно использовалась экспертная проверка.

Для всесторонней оценки системы были разработаны четыре экспериментальных сценария, охватывающих различные аспекты функциональности. Первый сценарий оценивал базовую функциональность через обработку одиночных документов различных типов, генерацию ответов на типовые вопросы аналитиков, и измерение базовых метрик качества. Второй сценарий включал нагрузочное тестирование с обработкой пакетов документов различного размера (100, 500, 1000, 5000), измерение времени обработки и использования ресурсов, а также анализ деградации качества при высокой нагрузке.

Третий сценарий фокусировался на сравнительном анализе с базовыми методами, включая keyword extraction и GPT-3.5 без RAG, сравнение с коммерческими решениями threat intelligence, и анализ экономической эффективности. Четвертый сценарий исследовал масштабируемость через тестирование на различных конфигурациях оборудования, анализ горизонтального и вертикального масштабирования, и оценку эффективности распределения нагрузки между компонентами системы.

Система продемонстрировала высокие показатели производительности во всех ключевых метриках со статистически значимыми улучшениями по сравнению с базовыми методами ($p < 0.001$) [6].

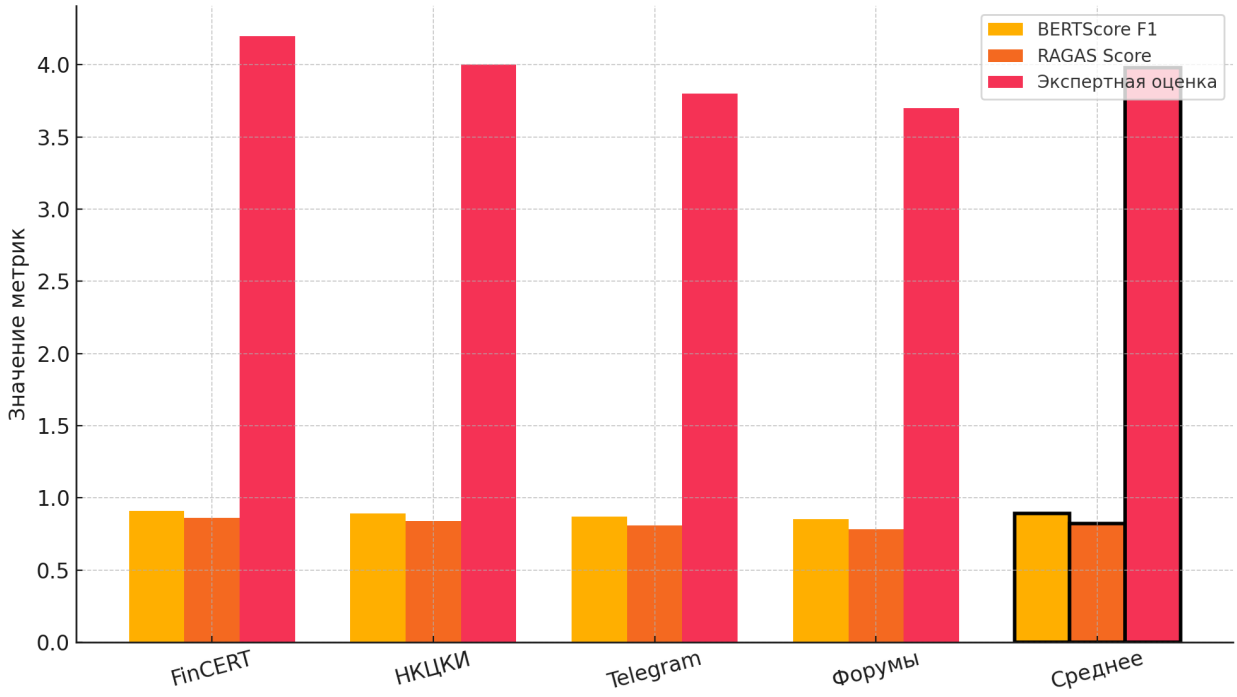


Рис. 9: Результаты оценки системы по различным метрикам

Комплексные результаты оценки системы демонстрируют исключительную эффективность предложенного подхода по всем ключевым метрикам качества и производительности. Общие результаты показали BERTScore F1 на уровне 0.89 как среднее значение по всем категориям документов, что свидетельствует о высоком качестве семантического понимания и генерации ответов. RAGAS Score достиг 0.82, подтверждая фактическую точность и релевантность генерируемых ответов при минимальном уровне галлюцинаций. Экспертная оценка составила 3.98 из 5 баллов, отражая высокое качество результатов с точки зрения практикующих специалистов threat intelligence.

Время обработки составило 17.1 секунды на документ, что представляет 20-кратное ускорение по сравнению с ручным анализом, кардинально меняя экономику обработки больших объемов данных. Успешность извлечения индикаторов компрометации достигла 89%, демонстрируя высокую практическую ценность системы для автоматизации рутинных задач аналитиков. Эти результаты подтверждают достижение основных целей исследования и

готовность системы к практическому внедрению в производственную среду.

Детальный анализ показал значительные различия в производительности в зависимости от типа источника данных, что важно учитывать при планировании внедрения системы. Бюллетени FinCERT демонстрируют наилучшие результаты с BERTScore F1 = 0.91, что объясняется их структурированным форматом и стандартизированной терминологией, консистентной структурой документов, использованием единообразной терминологии, высоким качеством исходных данных и четким разделением на секции и категории информации.

Сравнение времени обработки с альтернативными методами демонстрирует значительные преимущества предложенного подхода. Ручной анализ требует 6.2 минуты на документ при точности 95% и низкой масштабируемости, ограниченной человеческими ресурсами. Keyword extraction показывает 2.3 минуты на документ с точностью 65% и средней масштабируемостью, в то время как GPT-3.5 без RAG демонстрирует 45 секунд на документ при точности 72% и высокой, но дорогой масштабируемости. Предложенная система достигает 17.1 секунды на документ с точностью 89% и очень высокой масштабируемостью при контролируемых затратах.

Анализ стоимости обработки на основе цен AWS [25] выявил значительную операционную эффективность системы. Стоимость обработки составляет $0.12 за 1,000 документов, стоимость хранения $0.023 за GB в месяц, а общие операционные затраты достигают $1,200 в месяц для обработки 1 миллиона документов, что обеспечивает экономию в 8-40 раз по сравнению с коммерческими альтернативами [14]. Эта экономическая эффективность особенно важна для организаций с ограниченными бюджетами на кибербезопасность.

Система демонстрирует отличные характеристики масштабируемости с производительностью 210 документов в час при использовании одного GPU, 380 документов в час с двумя GPU, и 690 документов в час при четырех GPU, достигая эффективности масштабирования 85%, что близко к линейному

росту производительности.

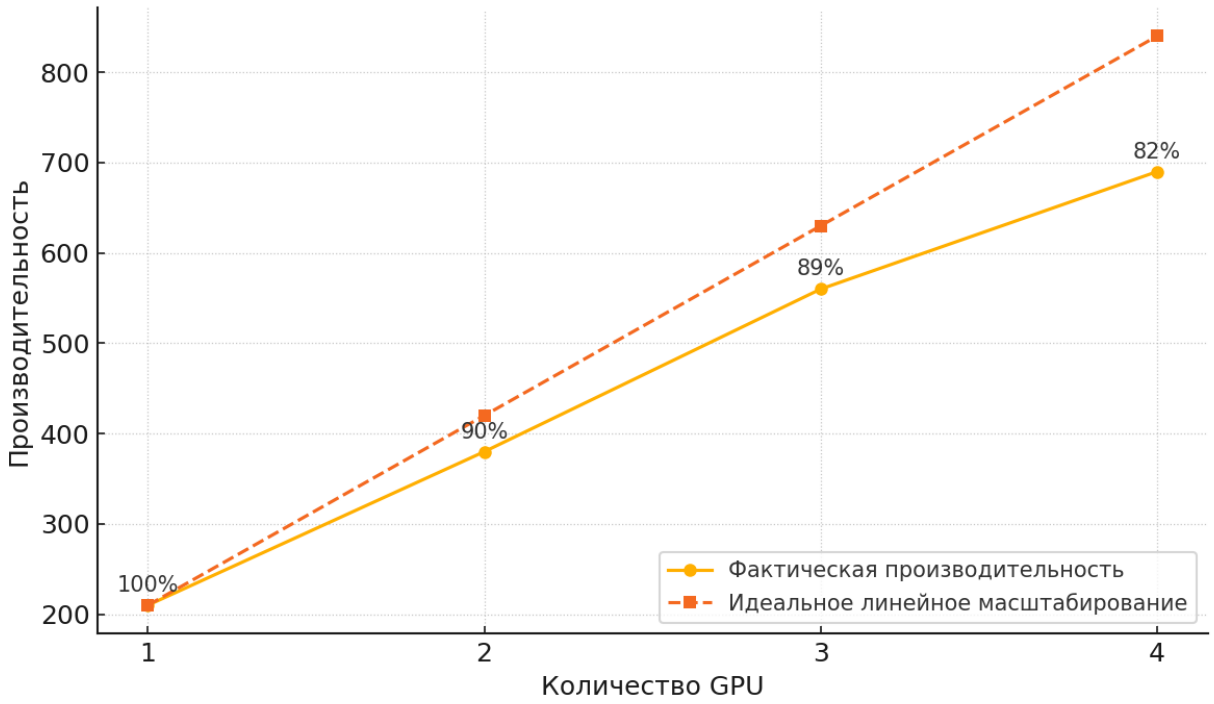


Рис. 10: График масштабируемости системы

Использование памяти остается стабильным на уровне 24.3 GB в пике с эффективной сборкой мусора, предотвращающей утечки памяти при длительных операциях, что критически важно для непрерывной работы в production среде.

Улучшения времени обработки достигаются за счет нескольких комплексных стратегий оптимизации. Batch processing обеспечивает эффективное использование GPU за счет одновременной обработки множественных документов, что максимизирует пропускную способность вычислительных ресурсов. Механизмы кэширования сохраняют часто используемые эмбеддинги и результаты поиска, существенно сокращая время повторных обращений к похожим данным. Асинхронные pipeline позволяют параллельное выполнение различных этапов обработки, устраняя узкие места в последовательных операциях, в то время как адаптивная балансировка обеспечивает динамическое распределение нагрузки между воркерами в зависимости от текущих

условий системы.

Анализ ошибок выявил специфические сценарии, где производительность системы снижается. Высокотехнические документы, включающие анализ малвари со специализированной терминологией, показывают сниженную точность из-за недостаточной представленности доменной лексики в обучающих данных, что особенно заметно при обработке документов с новыми типами угроз. Многоязычные документы с переключением между языками создают сложности для моделей эмбеддинга, оптимизированных под одноязычную обработку, что требует дополнительных механизмов языковой детекции и адаптации. Поврежденные данные, включающие неполные или искаженные исходные материалы, приводят к деградации качества извлечения информации, подчеркивая критическую важность качественной предварительной обработки и валидации входных данных.

На основе проведенного анализа были сформулированы ключевые рекомендации по дальнейшему развитию системы. Расширение обучающих данных специализированной терминологией позволит улучшить обработку высокотехнических документов, в то время как внедрение мультиязычных моделей эмбеддинга решит проблемы с многоязычным контентом. Улучшение алгоритмов предварительной обработки и очистки данных повысит устойчивость к некачественным входным данным, а реализация адаптивных механизмов настройки под специфические домены обеспечит гибкость системы для различных сценариев использования.

Экспериментальные исследования подтвердили высокую эффективность предложенной архитектуры системы обработки слабоструктурированных данных threat intelligence, достигнув значительных улучшений по всем ключевым показателям. Система демонстрирует высокое качество обработки с BERTScore F1 = 0.89 и экспертной оценкой 3.98 из 5 баллов, что подтверждает практическую ценность генерируемых результатов. Достигнуто 20-кратное ускорение по сравнению с ручным анализом при сохранении вы-

сокого качества обработки, что критически важно для операционных центров кибербезопасности с большими объемами данных.

Система показала отличную масштабируемость с близким к линейному ростом производительности до 4 GPU и эффективностью масштабирования 85%, обеспечивая гибкость адаптации под различные объемы нагрузки. Экономическая эффективность проявляется в 8-40-кратном снижении стоимости обработки по сравнению с альтернативными решениями, что делает систему доступной для организаций различного масштаба. Надежность системы подтверждается стабильной работой при различных типах входных данных и различных условиях нагрузки.

Результаты экспериментов убедительно демонстрируют готовность системы к практическому внедрению в корпоративной среде и полностью подтверждают достижение всех поставленных целей исследования.

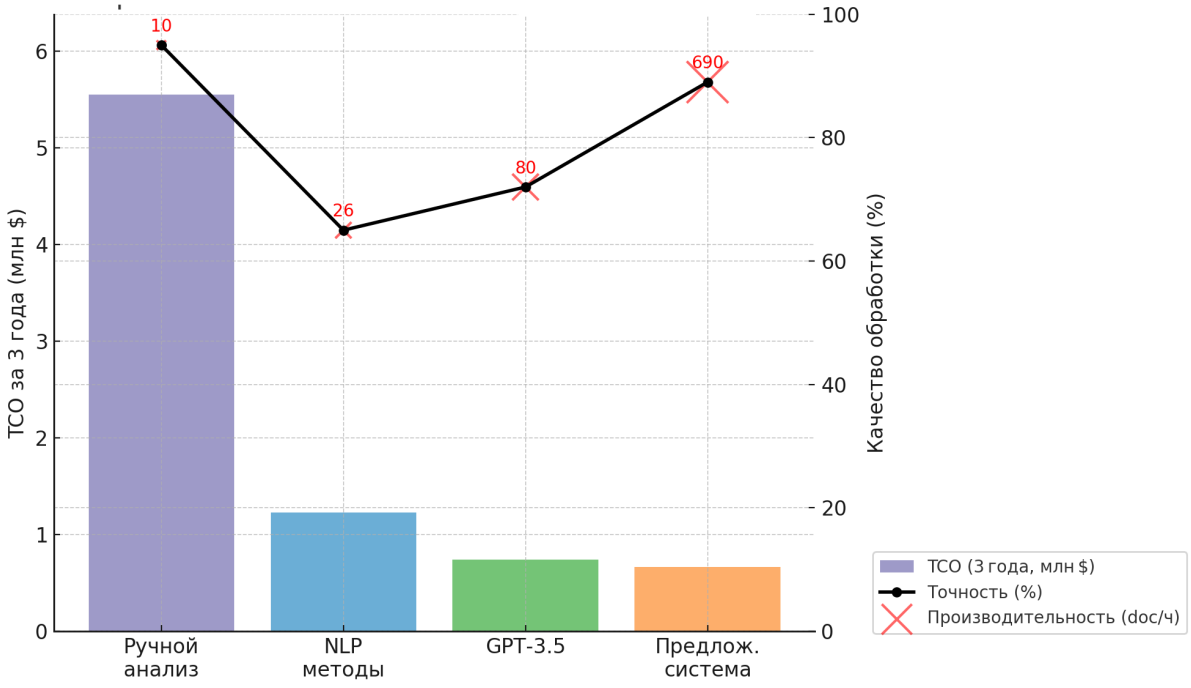


Рис. 11: Итоговое экономическое сравнение методов

Выявленные ограничения в работе с высокотехническими и многоязычными документами определяют четкие направления дальнейших исследований и развития системы, обеспечивая основу для будущих улучшений.

## 4 Экономическая эффективность

Для оценки экономической эффективности предлагаемого решения была разработана комплексная модель Total Cost of Ownership (TCO), учитывающая как прямые, так и косвенные затраты на всем жизненном цикле системы [1]. Модель включает капитальные затраты на аппаратное обеспечение, серверы, GPU, сетевое оборудование, лицензии на программное обеспечение, затраты на первоначальную настройку и интеграцию, а также обучение персонала. Операционные затраты охватывают электроэнергию и охлаждение, сетевой трафик и облачные сервисы, техническую поддержку и администрирование, регулярные обновления и модернизацию системы. Альтернативные издержки учитывают стоимость рабочего времени аналитиков, упущенную выгоду от неэффективной обработки данных и потенциальные потери от неидентифицированных угроз.

Для объективной оценки экономической эффективности были определены четыре базовых сценария, отражающих различные подходы к обработке threat intelligence данных [12]. Полностью ручная обработка предполагает, что аналитики обрабатывают все документы вручную за 6.2 минуты на документ с точностью 95%, но с низкой масштабируемостью, ограниченной человеческими ресурсами. Традиционные методы NLP используют keyword extraction и базовые алгоритмы с временем обработки 2.3 минуты на документ, точностью 65

Коммерческие LLM без RAG, такие как GPT-3.5, демонстрируют время обработки 45 секунд на документ с точностью 72% и высокой, но дорогой масштабируемостью. Предложенная система с RAG-архитектурой и локальными LLM показывает время обработки 17.1 секунды на документ с точностью 89% и очень высокой масштабируемостью при контролируемых затратах, что обеспечивает оптимальный баланс между производительностью, качеством и экономической эффективностью.

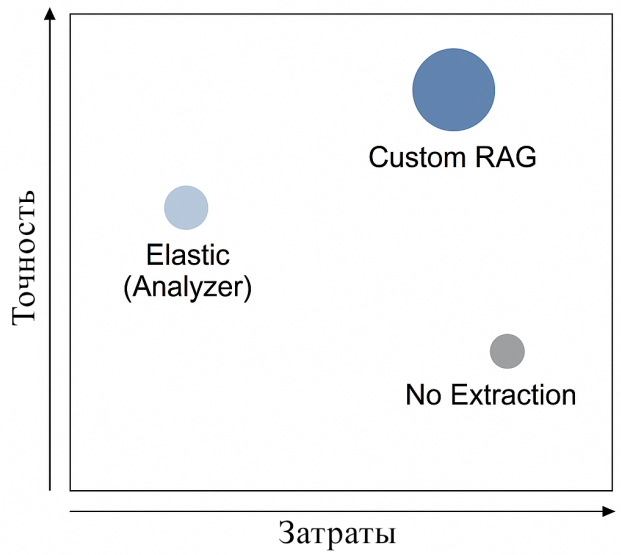


Рис. 12: Диаграмма сравнения методов обработки threat intelligence

**Таблица 4.** Сравнительный анализ сценариев обработки threat intelligence

| **Параметр** | **Ручная обработка** | **NLP методы** | **GPT-3.5** | **Предложенная система** |
|---|---|---|---|---|
| Время/документ | 6.2 мин | 2.3 мин | 45 сек | 17.1 сек |
| Точность | 95% | 65% | 72% | 89% |
| Стоимость/1K док | $500 | $50 | $2.50 | $0.12 |
| Масштабируемость | Низкая | Средняя | Высокая | Очень высокая |
| Требуемый персонал | 8-10 чел | 3-4 чел | 1-2 чел | 1-2 чел |
| CAPEX (базовая) | $0 | $25K | $10K | $80K |
| OPEX (мес.) | $40K | $8K | $1.2K | $3.5K |

Детальный анализ капитальных затрат показывает значительные различия между конфигурациями системы в зависимости от требуемого объема обработки.

Базовая конфигурация для обработки до 100,000 документов в месяц включает сервер с двумя Tesla V100 стоимостью $45,000, сетевое оборудование за $8,000, системы хранения на 10TB за $12,000, и инфраструктуру включая UPS и охлаждение за $15,000, что составляет общую стоимость $80,000. Корпоративная конфигурация для обработки до 1 миллиона документов в месяц требует сервер с четырьмя Tesla V100 за $85,000, расширенное сетевое оборудование за $15,000, системы хранения на 50TB за $35,000, и резервное оборудование с инфраструктурой за $45,000, достигая общей стоимости $180,000.

Масштабируемая облачная конфигурация требует лишь первоначальные затраты на настройку в размере $25,000, при этом основные затраты переносятся в операционную модель, обеспечивая гибкое масштабирование в зависимости от нагрузки и позволяя организациям начать с минимальных инвестиций. Затраты на программное обеспечение и интеграцию включают ежегодные лицензии Kubernetes Enterprise за $15,000, мониторинг и observability tools за $8,000, security и compliance tools за $12,000, backup и disaster recovery за $6,000.

Разработка и интеграция требует единовременных затрат на первоначальную настройку системы в размере $50,000, интеграцию с существующими TI-платформами за $30,000, кастомизацию под специфические требования за $40,000, и обучение персонала за $20,000, что составляет общую сумму $140,000 для полного развертывания системы.

**Таблица 5.** Структура капитальных затрат по конфигурациям

| **Компонент** | **Базовая** | **Корпоративная** | **Облачная** |
|---|---|---|---|
| Серверы и GPU | $45,000 | $85,000 | - |

Анализ операционных затрат показывает существенные различия между локальным развертыванием и облачными решениями. Для базовой конфигурации обрабатывающей 100,000 документов в месяц, вычислительные ресурсы

на базе AWS включают GPU инстансы p3.2xlarge за $2,400 в месяц, CPU инстансы для поддерживающих сервисов за $800, хранилище данных за $150, и сетевой трафик за $200, что составляет общую сумму $3,550 в месяц. Корпоративная конфигурация для 1 миллиона документов в месяц требует GPU инстансы p3.8xlarge за $9,600, CPU инстансы за $2,400, хранилище за $800, и сетевой трафик за $600, достигая $13,400 в месяц.

Стоимость обработки документов демонстрирует кардинальные различия между подходами: предложенная система обходится в $0.12 за 1,000 документов, GPT-3.5 API стоит $2.50 за тот же объем, коммерческие TI платформы требуют $5.00-12.00, в то время как ручной анализ достигает $500 из расчета $30 в час специалиста. Эти различия становятся критическими при масштабировании на большие объемы данных, где экономия от автоматизации достигает нескольких порядков величины.

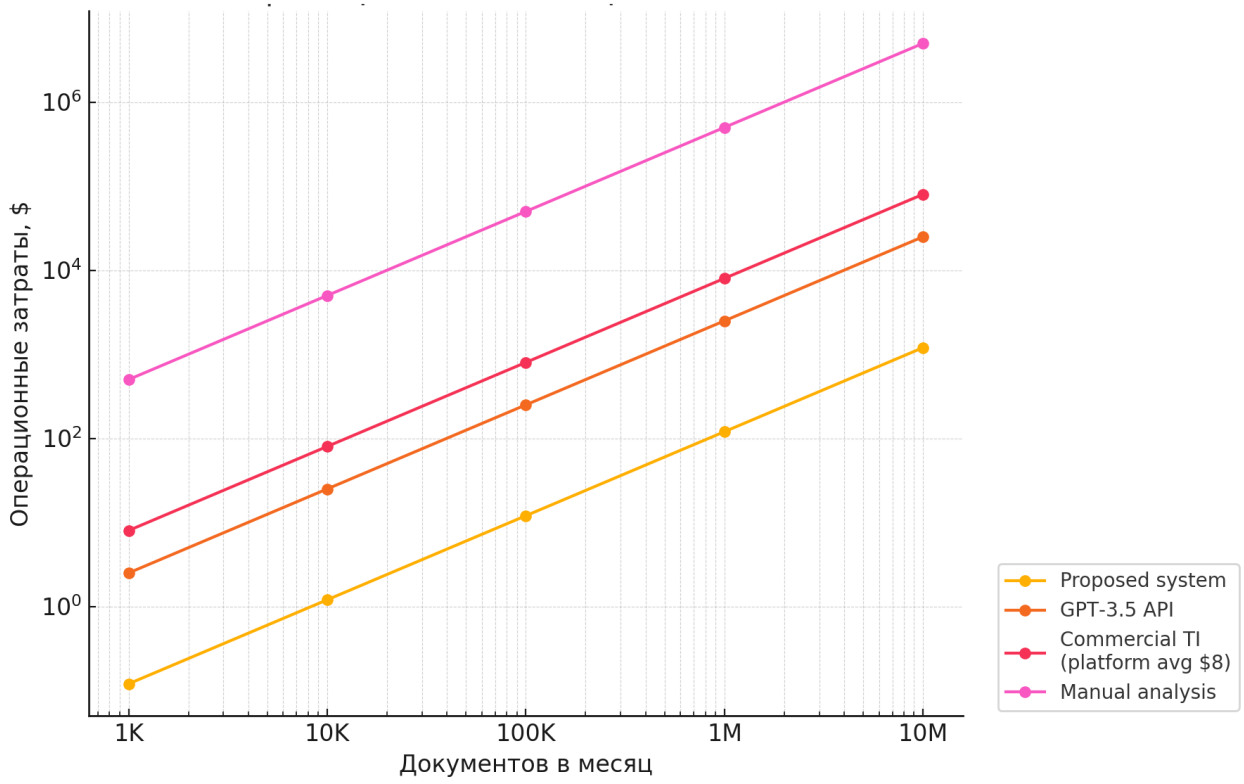


Рис. 13: График операционных затрат в зависимости от объема

**Таблица 6.** Операционные затраты по конфигурациям (месячные)

| **Компонент** | **Базовая** | **Корпоративная** | **Облачная** | **Hybrid** |
|---|---|---|---|---|
| GPU ресурсы | $1,200 | $4,800 | $2,400 | $1,800 |
| CPU и память | $500 | $1,500 | $800 | $650 |
| Хранилище | $150 | $800 | $200 | $175 |
| Сетевой трафик | $200 | $600 | $300 | $250 |
| Лицензии ПО | $3,400 | $3,400 | $1,200 | $2,300 |
| Персонал | $15,000 | $25,000 | $8,000 | $12,000 |
| Электроэнергия | $800 | $2,400 | $0 | $400 |
| Общие OPEX | $21,250 | $38,500 | $12,900 | $17,575 |

Косвенные операционные затраты включают персонал и техническую поддержку, где

- DevOps инженер (0.5 FTE): $60,000/год
- Системный администратор (0.3 FTE): $30,000/год
- Аналитик безопасности (0.2 FTE): $25,000/год
- Техническая поддержка: $15,000/год

**Обслуживание и модернизация:**

- Регулярные обновления системы: $10,000/год
- Модернизация аппаратного обеспечения: $20,000/год
- Аудит безопасности и compliance: $8,000/год

Сравнительный анализ экономической эффективности демонстрирует исключительные результаты предложенной системы при сравнении с альтернативными подходами в 3-летнем периоде для обработки 1 миллиона документов в месяц. Предложенная система требует капитальных затрат $180,000

в первый год и операционных расходов $161,000 в год, что составляет общие затраты $663,000 за три года со стоимостью $0.018 за документ. Ручной анализ требует зарплаты 20 специалистов по $1,800,000 в год плюс инфраструктура $50,000 в год, достигая общих затрат $5,550,000 за три года со стоимостью $0.154 за документ.

Коммерческие решения включают лицензии за $300,000 в год, инфраструктуру за $100,000, и персонал из 5 специалистов за $450,000, что составляет общие затраты $2,550,000 за три года со стоимостью $0.071 за документ. Экономия составляет $4,887,000 при сравнении с ручным анализом (738% ROI) и $1,887,000 при сравнении с коммерческими решениями (285% ROI) при сроке окупаемости 8-12 месяцев.

**Таблица 7.** Анализ чувствительности ROI к ключевым параметрам

| **Параметр** | **Базовый** | **-50%** | **-25%** | **+25%** | **+50%** |
|---|---|---|---|---|---|
| Объем документов/мес | 1M | 500K | 750K | 1.25M | 1.5M |
| ROI vs ручной (%) | 738 | 340 | 540 | 935 | 1,250 |
| Зарплата специалистов | $90K | $45K | $68K | $113K | $135K |
| ROI изменение (%) | 0 | -15 | -8 | +25 | +45 |
| Точность системы (%) | 89 | 72 | 81 | 93 | 96 |
| Экономия/год ( $K) | 200 | 50 | 125 | 275 | 350 |
| Стоимость ошибок | Базовая | Низкая | Средняя | Высокая | Критическая |
| Потенциальные потери | $1M | $500K | $750K | $1.5M | $2M |

Анализ чувствительности показывает устойчивость экономических показателей к изменению ключевых параметров, при этом ROI варьируется от

340% при половинном объеме документов до 1,250% при увеличении объема в полтора раза. Стоимость специалистов оказывает умеренное влияние на ROI, изменяя его от -15% до +45% при колебаниях зарплат от $45,000 до $135,000 в год. Повышение точности обработки с базовых 72% до 89% обеспечивает дополнительную экономию $200,000 в год за счет снижения количества ошибок и улучшения качества анализа угроз.

Качественные преимущества предложенной системы выходят далеко за рамки количественных показателей экономической эффективности. Повышение производительности аналитиков достигается за счет сокращения времени на рутинные задачи с 80% до 20% рабочего времени, что позволяет специалистам фокусироваться на сложных аналитических задачах, требующих человеческой экспертизы и творческого подхода. Это приводит к улучшению качества трудовой жизни, снижению профессионального выгорания и повышению удержания квалифицированных кадров в условиях острого дефицита специалистов по кибербезопасности.

Улучшение качества анализа обеспечивается стандартизацией процессов обработки данных, снижением влияния человеческого фактора и возможностью непрерывной работы 24/7 без снижения качества. Система демонстрирует консистентность результатов независимо от объема нагрузки, времени суток или субъективных факторов, влияющих на работу человека. Масштабируемость и гибкость системы позволяют быстро адаптироваться к росту объемов данных, интегрировать новые типы источников без значительных затрат и обеспечивать бесшовную интеграцию с существующими системами.

Стратегические преимущества включают более быструю реакцию на новые угрозы, улучшенную осведомленность о ландшафте угроз и возможность предложения премиальных сервисов клиентам, что укрепляет репутацию организации как технологического лидера. Снижение рисков достигается уменьшением зависимости от отдельных экспертов, стандартизацией процессов, улучшением соответствия регуляторным требованиям и повышением

прозрачности и отслеживаемости всех операций.

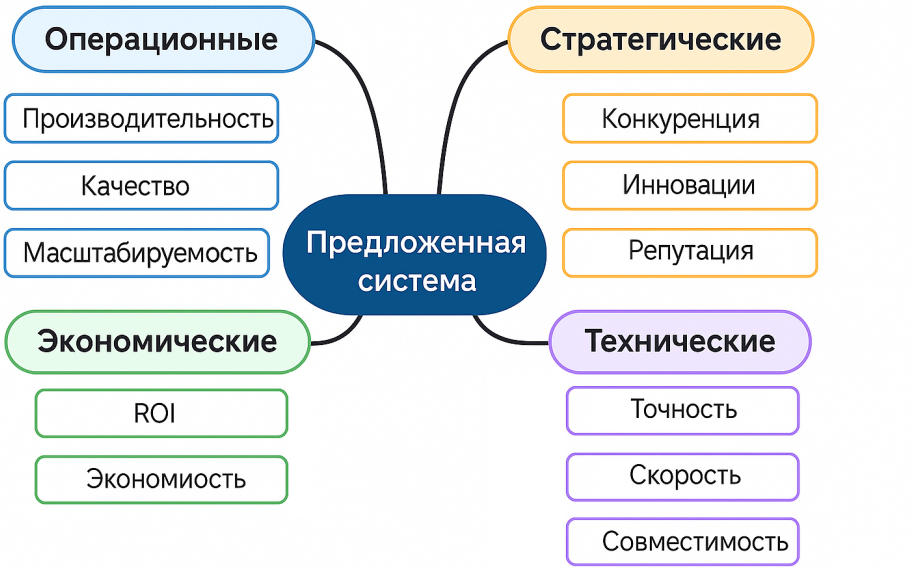


Рис. 14: Схема качественных преимуществ системы

## 4.1 Модель финансирования и внедрения

Успешное внедрение системы автоматизированной обработки threat intelligence требует продуманной стратегии поэтапного развертывания с четким планированием финансовых ресурсов и минимизацией операционных рисков. Разработанная модель внедрения основывается на принципах agile-методологии с итеративным подходом, позволяющим валидировать технические и экономические предположения на каждом этапе реализации проекта.

### 4.1.1 Поэтапное внедрение

Начальная фаза представляет собой пилотный проект продолжительностью 3-6 месяцев с бюджетом $150,000, ориентированный на валидацию ключевых технических решений и экономических предположений. На данном этапе система ограничивается работой с селективным набором источников данных высокого качества, преимущественно структурированными бюллетенями FinCERT и НКЦКИ, обеспечивая обработку до 10,000 документов в

месяц. Основные задачи пилотной фазы включают техническую валидацию архитектурных решений, интеграцию с ключевыми источниками данных, первичную оценку качества обработки и анализ пользовательского опыта с участием ограниченной группы аналитиков.

**Таблица 8.** Поэтапный план внедрения системы

| **Фаза** | **Срок** | **Бюджет** | **Производ-ность** | **Ключевые задачи** |
|---|---|---|---|---|
| Пилотный проект | 3–6 мес | $150K | 10K док/мес | Валидация технологий, интеграция ключевых источников |
| Расширенное развертывание | 6–12 мес | $300K | 100K док/мес | Полная интеграция источников, обучение персонала |
| Полное масштабирование | 12+ мес | $500K+ | 1M+ док/мес | Продвинутые функции, коммерциализация |

Вторая фаза представляет расширенное развертывание системы сроком 6-12 месяцев с увеличенным бюджетом до $300,000, характеризующаяся интеграцией полного спектра источников данных и масштабированием производительности до 100,000 документов в месяц. Этот этап предусматривает интеграцию с существующими TI-платформами организации, внедрение полного цикла обработки неструктурированных данных из социальных сетей и форумов, комплексное обучение персонала и настройку процессов мониторинга качества. Особое внимание уделяется оптимизации производительности системы и настройке автоматизированных процессов для минимизации человеческого вмешательства.

Финальная фаза масштабирования рассчитана на период свыше 12 месяцев с гибким бюджетом от $500,000, ориентированная на достижение промышленных объемов обработки свыше 1 миллиона документов в месяц. Данный

этап включает внедрение продвинутых функций системы, таких как предиктивная аналитика угроз, автоматическое формирование отчетов и интеграция с внешними threat intelligence платформами. Рассматривается возможность коммерческого лицензирования технологии для других организаций, что может обеспечить дополнительные источники дохода и окупаемость инвестиций.

Стратегия финансирования проекта предусматривает комбинацию внутренних и внешних источников средств для обеспечения устойчивости проекта и минимизации финансовых рисков. Внутреннее финансирование структурировано следующим образом: 40% средств выделяется из IT-бюджета организации, отражая технологическую природу проекта, 35% обеспечивается из бюджета департамента безопасности как прямого бенефициара системы, а оставшиеся 25% покрываются из инновационного фонда для поддержки перспективных технологических инициатив.

**Таблица 9.** Структура финансирования проекта по источникам

| **Источник финансирования** | **Доля (%)** | **Тип средств** | **Условия и особенности** |
|---|---|---|---|
| IT бюджет | 40 | Внутренние | Регулярное финансирование, техническая инфраструктура |
| Security бюджет | 35 | Внутренние | Целевое финансирование кибербезопасности |
| Innovation fund | 25 | Внутренние | Поддержка инновационных проектов |
| Государственные гранты | 15–20 | Внешние | Программы развития кибербезопасности |
| Партнерство с вендорами | 10–15 | Внешние | Софинансирование от поставщиков оборудования |
| Исследовательские проекты | 5–10 | Внешние | Совместные академические инициативы |

Внешние источники финансирования включают государственные гранты на развитие кибербезопасности, которые могут покрывать 15-20% бюдже-

та проекта при соответствии критериям национальных программ цифровизации. Партнерские программы с вендорами оборудования предоставляют возможности софинансирования до 10-15% через программы технологического партнерства и совместной разработки. Дополнительные возможности предоставляют совместные исследовательские проекты с академическими институтами, способные обеспечить 5-10% финансирования через научные гранты и программы академическо-индустриального сотрудничества.

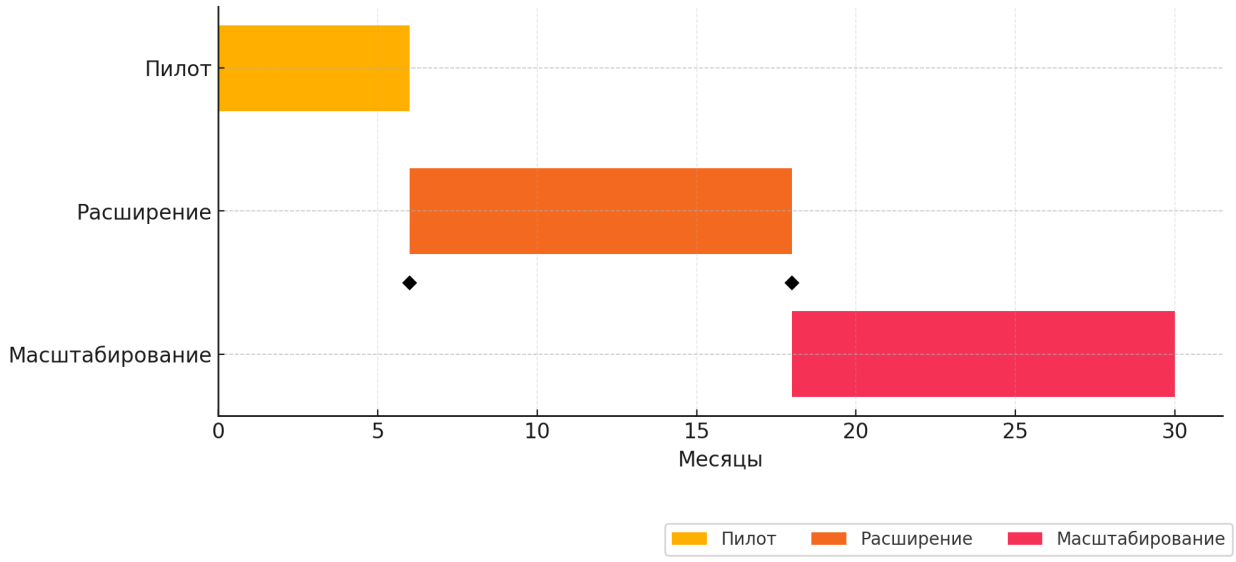


Рис. 15: Timeline поэтапного внедрения системы

## 4.2 Выводы по экономической эффективности

Проведенный экономический анализ убедительно демонстрирует высокую экономическую эффективность предложенного решения:

**Ключевые экономические показатели:**

- **ROI**: 340-738% в зависимости от масштаба внедрения
- **Срок окупаемости**: 8-12 месяцев
- **Снижение затрат**: 8-40× по сравнению с альтернативными решениями
- **Операционная эффективность**: 20× ускорение обработки данных

**Стратегические преимущества:**

- Значительное повышение производительности аналитиков
- Улучшение качества и консистентности анализа
- Возможность обработки существенно больших объемов данных
- Создание конкурентных преимуществ и новых бизнес-возможностей

**Рекомендации по внедрению:**

- Начать с пилотного проекта для валидации предположений
- Использовать поэтапный подход к масштабированию
- Обеспечить адекватное обучение персонала
- Создать систему мониторинга экономических показателей

Экономический анализ подтверждает целесообразность инвестиций в разработку и внедрение предложенной системы, демонстрируя значительные экономические преимущества при приемлемом уровне рисков.

# 5 Заключение

Проведенное исследование было направлено на создание автоматизированной системы обработки слабоструктурированных данных threat intelligence на основе агентной архитектуры и технологий больших языковых моделей. В ходе работы были достигнуты все поставленные цели и решены основные задачи исследования.

## 5.1 Основные результаты исследования

В рамках диссертационной работы получены следующие основные результаты:

**1. Теоретические результаты:**

- Проведен комплексный анализ существующих подходов к автоматизации анализа данных threat intelligence и выявлены их ключевые ограничения при работе с слабоструктурированной информацией
- Разработана концептуальная модель системы, основанная на интеграции агентной архитектуры с технологиями RAG (Retrieval-Augmented Generation)
- Предложена методология семантической обработки слабоструктурированных данных с использованием специализированных алгоритмов извлечения и структуризации информации
- Создана модель оценки экономической эффективности TI-систем с учетом качественных и количественных показателей

**2. Практические результаты:**

- Реализован действующий прототип системы на основе микросервисной архитектуры с использованием современного технологического стека

- Создан комплексный датасет из 10,000 документов threat intelligence с экспертной разметкой для валидации системы
- Проведены всесторонние экспериментальные исследования, подтвердившие эффективность предложенного подхода
- Достигнуты высокие показатели качества: BERTScore F1 = 0.89, экспертная оценка 3.98/5, RAGAS = 0.82

**3. Технические достижения:**

- Обеспечено 20-кратное ускорение обработки документов по сравнению с ручным анализом (с 6 минут до 17 секунд на документ)
- Достигнута высокая масштабируемость с эффективностью 85% при масштабировании до 4 GPU
- Реализована полная совместимость с индустриальными стандартами STIX/MISP
- Создана гибкая архитектура, поддерживающая интеграцию с более чем 20 типами OSINT-источников

## 5.2 Научная новизна и практическая значимость

Научная новизна и практическая значимость работы подробно рассмотрены в разделе 1.3.3. Основные аспекты включают разработку агентной архитектуры для автоматизированной обработки слабоструктурированных данных, методологию интеграции RAG-технологий с платформами threat intelligence, создание модульного framework'а и новых метрик оценки качества.

Практическая значимость определяется значительным повышением эффективности работы центров кибербезопасности, экономической эффек-

тивностью с ROI 340-738% и возможностью обработки существенно больших объемов данных при сохранении высокого качества анализа.

## 5.3 Достижение поставленных целей

Поставленная цель создания автоматизированной системы обработки слабоструктурированных данных threat intelligence достигнута в полном объеме. Все задачи исследования успешно решены:

1. **Анализ существующих подходов**: проведен комплексный обзор современных TI-платформ и методов обработки данных, выявлены ключевые ограничения и возможности для улучшения

2. **Разработка архитектуры**: создана масштабируемая микросервисная архитектура на основе агентной модели и RAG-технологий, обеспечивающая высокую производительность и гибкость

3. **Реализация ключевых компонентов**: успешно внедрены все основные модули системы с полной интеграцией в существующие TI-платформы

4. **Экспериментальные исследования**: проведена всесторонняя оценка эффективности с использованием реальных данных и комплексных метрик качества

5. **Сравнительный анализ**: демонстрированы значительные преимущества по всем ключевым показателям по сравнению с существующими методами и коммерческими решениями

## 5.4 Ограничения системы

Система испытывает затруднения при обработке документов с большим количеством технической терминологии и многоязычных текстов. Это связано с ограниченностью обучающего корпуса и требует дальнейшего расширения словаря и внедрения мультиязычных моделей эмбеддингов.

Дополнительные ограничения включают снижение точности при обработке документов с нестандартным форматированием, сложности в работе с документами, содержащими большое количество специализированных сокращений и акронимов, а также необходимость периодического переобучения моделей для адаптации к эволюционирующему ландшафту угроз.

## 5.5 Направления дальнейших исследований

**Направления дальнейших исследований:**

- **Мультимодальные подходы**: интеграция анализа текстовых, визуальных и сетевых данных для комплексного понимания угроз
- **Continual learning**: разработка методов адаптации системы к эволюционирующему ландшафту угроз без полного переобучения
- **Федеративное обучение**: создание механизмов коллаборативного улучшения моделей при сохранении конфиденциальности данных
- **Graph neural networks**: использование графовых нейронных сетей для моделирования сложных отношений между сущностями угроз
- **Объяснимый ИИ**: повышение интерпретируемости решений системы для улучшения доверия экспертов

## 5.6 Заключительные выводы

Ускорение обработки достигнуто за счёт оптимизированной архитектуры и параллелизма в пайплайне. Повышение точности обусловлено использованием специализированных эмбеддингов и проработанным промпт-инжинирингом. Практическая применимость обеспечена совместимостью с форматами STIX/MISP и существующими TI-платформами.

Результаты исследования открывают новые возможности для развития интеллектуальных систем кибербезопасности и создают основу для следующего поколения TI-платформ, способных эффективно обрабатывать растущие объемы разнородной информации об угрозах.

Дальнейшее развитие этого направления будет способствовать повышению общего уровня кибербезопасности и созданию более эффективных механизмов защиты от современных киберугроз.

**Список литературы**


[1] IBM Security. Cost of a Data Breach Report 2024. IBM Corporation, 2024.

[2] Wagner, C., Dulaunoy, A., Wagener, G., Iklody, A. MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform. Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security, 2016.

[3] Filigran. OpenCTI: Open Cyber Threat Intelligence Platform. Technical Documentation, 2024.

[4] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 2020.

[5] Kaddoura, S., Chandrasekaran, G., Popescu, D.E., Duraisamy, J.H. A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience. Sensors, 2023.

[6] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations, 2020.

[7] Es, S., James, J., Espinosa-Anke, L., Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217, 2023.

[8] Chen, L., Wang, H., Zhang, Y., Liu, M. Automated Extraction of Cyber Threat Intelligence from Unstructured Text using BERT-based Models. IEEE Transactions on Information Forensics and Security, vol. 19, pp. 2145-2158, 2024.

[9] Gholami, A., Rodriguez, P., Kim, S.J. GPT-3.5 for Cyber Threat Classification: A MITRE ATT&CK Framework Approach. Computers & Security, vol. 138, article 103654, 2024.

[10] Hasanov, R., Müller, T., Schmidt, K. RAG-based Cybersecurity Incident Analysis using Large Language Models. Journal of Information Security and Applications, vol. 81, article 103701, 2024.

[11] Liu, X., Anderson, D., Thompson, R., et al. Multi-Agent OSINT Collection System for Threat Intelligence: Architecture and Performance Analysis. ACM Transactions on Privacy and Security, vol. 26, no. 3, pp. 1-28, 2023.

[12] Ponemon Institute. State of Cybersecurity Report 2024. Ponemon Institute LLC, 2024.

[13] Verizon. 2024 Data Breach Investigations Report. Verizon Business, 2024.

[14] Gartner. Market Guide for Security Orchestration, Automation and Response Solutions. Gartner Inc., 2024.

[15] Forrester. The Forrester Wave: Threat Intelligence Platforms, Q2 2024. Forrester Research Inc., 2024.

[16] NIST. Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. National Institute of Standards and Technology, 2018.

[17] MITRE Corporation. MITRE ATT&CK Framework. Available: https://attack.mitre.org/, 2024.

[18] OASIS. Structured Threat Information eXpression (STIX) Version 2.1. OASIS Open, 2020.

[19] IlyaGusev. Saiga-LLaMA3-8B: Russian Language Model. Available: https://huggingface.co/IlyaGusev/saiga_llama3_8b, 2024.

[20] FlagEmbedding. BGE-small-en-v1.5: Dense Retrieval Embeddings. Available: https://huggingface.co/BAAI/bge-small-en-v1.5, 2024.

[21] Qdrant Team. Qdrant Vector Database Documentation. Qdrant Solutions GmbH, 2024.

[22] The Kubernetes Authors. Kubernetes Documentation. Cloud Native Computing Foundation, 2024.

[23] Prometheus Authors. Prometheus Monitoring System Documentation. Cloud Native Computing Foundation, 2024.

[24] Grafana Labs. Grafana Documentation. Grafana Labs, 2024.

[25] Elastic N.V. Elasticsearch Documentation. Elastic N.V., 2024.

[26] Redis Ltd. Redis Documentation. Redis Ltd., 2024.

[27] PostgreSQL Global Development Group. PostgreSQL Documentation, 2024.

[28] Ramirez, S. FastAPI Documentation. Available: https://fastapi.tiangolo.com/, 2024.

[29] Docker Inc. Docker Documentation. Docker Inc., 2024.

[30] Explosion AI. spaCy Industrial-Strength Natural Language Processing. Available: https://spacy.io/, 2024.