ГУ «Медицинский центр Управления Делами Президента Республики Казахстан»

УТВЕРЖДАЮ

Руководитель
ГУ «Медицинский центр
Управления Делами Президента
Республики Казахстан»

_____ **Албаев Р.К.**
“____” _____ 2025 года

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

Системы лабораторного ассистента

на ____ листах

СОГЛАСОВАНО

Заместитель руководителя
ГУ «Медицинский центр
Управления Делами Президента
Республики Казахстан»

_____ **Бюрабекова Л.В.**
“____” _____ 2025 года

РАЗРАБОТАНО

Директор РГП на ПХВ
«Центр санитарно-эпидемиологической
экспертизы
Управления делами Президента
Республики Казахстан»

_____ **Шарипова С.Ф.**
“____” _____ 2025 года

г. Астана

ГУ «Медицинский центр
Управления Делами Президента Республики Казахстан»

Лист согласования

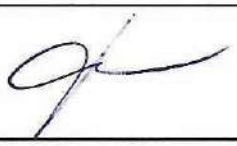
к техническому заданию

№	Ф.И.О.	Подпись
1.	Темиргалиева Айгуль Курганбековна	
2.	Вастьянов Александр Владимирович	
3.	Мукаев Нурлан Кайдарович	

РГП «Центр санитарно-эпидемиологической экспертизы Медицинского центра Управления делами Президента Республики Казахстан»

Лист согласования

к техническому заданию

№	Ф.И.О., должность	Подпись
1.	Биримкулова Жазира Бакытовна	
2.	Нусупбева Гаухар Есболатқызы	
3.	Муканов Данияр Еркинович	
4.	Бекмаганбетова Ляззат Сериковна	
5.	Раймбекова Асия Амангалиевна	
6.	Мұратханов Сұңғат Мұратханұлы	
7.	Омаргалиева Назым Кабдулмухитовна	
8.	Бейсембаева Алия Жетрубаевна	
9.	Ахметов Манат Зарбатович	
10.	Жусупова Баян Кабдигалимовна	
11.	Читимова Зарина Калижановна	
12.	Ибраева Алия Темирбековна	
13.	Сарсенгалиева Сания Жаксылыковна	

Содержание		
	Термины и определения	2
1	Общие сведения и границы	7
1.1.	Стандарты и нормативно правовые акты	7
1.2.	Шифр темы или шифр (номер) договора	7
1.3	Наименование предприятий разработчика и заказчика “Система лабораторного ассистента” и их реквизиты	7
1.4	Перечень документов, на основании которых внедряется система лабораторного ассистента	8
1.5	Плановые сроки начала и окончания работы по системе лабораторного ассистента	9
1.6	Порядок внесения изменений в Техническое задание и их характер	9
2.	Назначение и цели системы	9
2.1.	Цель создания система лабораторного ассистента	9
2.2.	Технические требования	10
2.3.	Требования к установке системы лабораторного ассистента на компьютеры лабораторий	11
2.4.	Объем требуемых работ и услуг	11
2.5.	Резервное копирование и восстановление	11
2.6.	Технологический процесс работы система лабораторного ассистента состоит из трех ключевых этапов внедрения	11
3.	Общие требования к архитектуре системы лабораторного ассистента	12
3.1.	Требования к инфраструктурному уровню	12
3.2.	Аутентификация и авторизация	12
3.3	Требования к эксплуатационному уровню	13
3.4	Дополнительные требования хранений данных	13
3.5	Требования к системе лабораторного ассистента	14
3.6	Требования к надежности и безопасности	14
4	Перечень элементов инфраструктуры РГП на ПХВ «ЦСЭЭ» УДП РК, который рекомендуется использовать для внедрения система лабораторного ассистента	15
4.1	SCNI платформа как стандартный набор компонентов, который применим для ИИ ассистента лабораторна	15
4.2	Технические рекомендации к инфраструктуре для платформы	16
5	Необходимый поддерживаемый функционал MVP для работы	18
5.1	Способы сбора обратной связи от пользователей	18
5.2	Рекомендации к разворачиванию обновление приложения	19
5.3	Описание интерфейса	19
5.4	Подготовка инфраструктуры	19
5.5	Рекомендуемые компоненты к программному обеспечению системы лабораторного ассистента	20
5.6	Подготовка базы знаний	23
5.7	Логирование и мониторинг	23
5.8	Обновление и синхронизация	23
5.9	Управление пользователями	24
5.10	Масштабирование и отказоустойчивость	24

Термины и определения

Используемые в настоящем документе термины и основные понятия определены в СП РК 3.02-145-2023. Также в тексте настоящего документа введены специальные термины:

Термины и определения	Описание
Автоматическая система	Совокупность управляемого объекта и устройств автоматического управления, функционирующих самостоятельно, без участия человека.
Автоматическое управление	Совокупность действий, направленных на поддержание или улучшение функционирования управляемого объекта без непосредственного участия человека в соответствии с заданной целью управления.
Информационная безопасность	Комплекс мер, направленных на защиту данных, систем и сетей от несанкционированного доступа, утечки, кибератак и других угроз, обеспечивая конфиденциальность, целостность и доступность информации.
Исполнительное устройство	Периферийное устройство с приводом (электрическим, пневматическим, гидравлическим и т.д.) для передачи управляющего воздействия на технологические объекты управления с целью изменения потока энергии или материала.
Технология API	Программный интерфейс с набором способов и правил, по которым различные программы взаимодействуют между собой и обмениваются данными, с помощью функций, классов, методов, структур, а иногда констант одной программы, к которой обращаются другие.
Протокол TCP/IP	Наборы правил, решающих задачу по передаче данных, сетевая модель, описывающая процесс передачи цифровых данных от источника информации к получателю через четыре уровня, каждый из которых описывается правилом (протоколом передачи).
ПО	Программное обеспечение
ППР	Планово-профилактические работы
ПТД	Проектно-техническая документация
API	Описание способов взаимодействия одной компьютерной программы с другими
АРМ	Автоматизированное рабочее место. Система, в которой все необходимые для работы инструменты и программы объединены в одну среду, часто под управлением специализированного программного обеспечения
ИИ	Искусственный интеллект
БД	База данных

GPU	графический процессор, ускоряет параллельные вычисления,
CPU	центральный процессор, выполняет последовательные вычисления и управляет системой
Bare metal	физический сервер без виртуализации
Docker	система контейнеризации приложений для изоляции, упрощенного развертывания
LLM	большая языковая модель, способная генерировать текстовые ответы.
FAISS	библиотека для быстрого поиска по векторным embedding
Embedding	числовое векторное представление текста
PVC	запрос на постоянное хранилище для контейнеров
ETCD	распределенное хранилище конфигураций Kubernetes
Fallback	резервный режим работы при ошибках или недоступности сервиса.
IndexedDB	локальная база данных в браузере
Zero-downtime	обновления без остановки работы системы

Ollama	локальный inference-движок для LLM
DICOMweb	стандарт веб-доступа к медицинским изображениям
JWT	формат токена для аутентификации пользователей
RAG	метод, комбинирующий генерацию текста и поиск по базе знаний
REST API	протокол взаимодействия сервисов через HTTP
HTTP	протокол передачи данных между клиентом (браузером) и сервером
Кеш	временное хранилище данных, позволяющее ускорить доступ к часто используемой информации
Кеширование	процесс сохранения данных в кэше для сокращения времени обработки запросов и снижения нагрузки на систему
Чат-бот	программный интерфейс, позволяющий пользователю взаимодействовать с системой с помощью текстовых запросов, получать ответы, инструкции, справочную информацию и выполнять базовые операции через диалог
HL7 v2	стандарт обмена медицинскими сообщениями, используемый для передачи лабораторных результатов
ASTM	стандарт обмена данными для лабораторного оборудования

2FA	двухфакторная аутентификация
Коннекторы/шины	программные компоненты, которые обеспечивают обмен данными между медицинскими системами
Mirth Connect	интеграционная шина для приема, преобразования и маршрутизации
KVM over IP	удаленная консоль сервера для аварийного управления
Security & Backup	комплекс мер по защите, резервному копированию и восстановлению
Логи пользователей	мониторинг, графики состояния системы и аудит действий
MVP	минимально жизнеспособная версия с простым сетевым хранилищем и офлайн-резервами
Шифрование AES-256	защита конфиденциальных данных на уровне шифрования
PostgreSQL/MySQL	реляционные логи данных
Оптические патч-корды	кабели для соединения сетевых узлов по оптике
Разделение сетей (VLAN/VRF)	изоляция трафика и управление входящими запросами
Veeam, SAN snapshots	протокол резервного копирования данных
Схема A/B	Разделение всей инфраструктуры на два независимых контура для высокой доступности

Data Layer	слой хранения данных, куда система складывает структурированную и неструктурированную информацию
PDU	блок распределения питания в серверной стойке
Резервная нода	дублирующий сервер, работающий как необходимый при отказе основной ноды
REST API (/ask, /files, /upload, /doc, /health)	точки входа для запросов таких как: задавать вопросы, загружать файлы, получать документы, проверять состояние системы
AD/LDAP	система доменной аутентификации сотрудников (аккаунты, роли, доступы)
Fallback	резервный сценарий работы, когда основная модель или сервис недоступны
Embedding	векторное представление текста для поиска знаний и RAG
Межкоммутатор	канал между сетевыми коммутаторами, обеспечивающий транспорт данных
Compute Layer	вычислительный слой, на котором исполняются сервисы, API и ИИ-модели
Вычислительный сервер 1U	компактный сервер высотой 1U для развёртывания контейнерной инфраструктуры Kubernetes
RAG	генерационная модель модуля ИИ

SAN/NVMe SSD	быстрые дисковые массивы для высокоскоростного хранения локальных накопителей
HAPI FHIR	сервер/библиотека для работы с протоколом FHIR
PDF, DOCX	форматы документов для генерации отчётов, протоколов, нормативных файлов

1. Общие сведения и границы

Документ разработан в соответствии с требованиями стандарта «СТ РК 34.015-2002. Информационная технология. Техническое задание на создание "Система лабораторного ассистента". Документ предназначен для специалистов, ответственных за разработку и тестирование программного обеспечения, а также для специалистов по сопровождению и эксплуатации инфраструктуры и модуля система лабораторного ассистента. Разработка и внедрение системы лабораторного ассистента, обеспечивающей централизованный доступ к нормативной, методической и справочной информации для повышения эффективности и информационной точности лабораторных исследований. Реализация проекта нацелена на сокращение времени на поиск информации, снизить человеческий фактор во время обучения, ускорить обучение персонала и повысить качество результатов лабораторных исследований.

1.1. Стандарты и нормативно правовые акты

Полное наименование модуля и ее условное обозначение

Полное наименование модуля – Система лабораторного ассистента

1.2 Шифр темы или шифр (номер) договора

Шифр темы: ——---

1.3 Наименование предприятий разработчика и заказчика "Система лабораторного ассистента" и их реквизиты

Наименование и реквизиты заказчика:

РГП «Центр санитарно-эпидемиологической экспертизы Медицинского центра Управления делами Президента Республики Казахстан», г.Астана, АЛЕКСАНДР СЫЗГАНОВ, 3/2, БИН 780140000023, РНН 600700013091, АО «Народный Банк Казахстана».

Наименование и реквизиты разработчика:

ТОО «DODA Digital» Адрес: Казахстан, Астана, ул.Ж.Тархана, д.4, оф. БИН: 250640024138 Банк: АО «Kaspi Bank» КБЕ: 17 БИК: CASPKZKA Номер счета: KZ04722S000046836511

1.4 Перечень документов, на основании которых внедряется система лабораторного ассистента

Основанием для развития система лабораторного ассистента является программа цифровой трансформации и развитие существующей медицинской информационной инфраструктуры. Работы выполняются в рамках её масштабирования и модернизации, с последующей интеграцией в уже функционирующие информационные системы. Согласно следующим документациям:

- Кодекс Республики Казахстан от 7 июля 2020 года № 360-VI «О здоровье народа и системе здравоохранения»;
- Закон Республики Казахстан от 21 мая 2013 года N 94-V «О персональных данных и их защите»;
- Приказ и.о. Министра здравоохранения Республики Казахстан от 30 октября 2020 года № ҚР ДСМ-175/2020 «Об утверждении форм учетной документации в области здравоохранения»;
- Постановление Правительства Республики Казахстан от 26 декабря 2019 года № 982 «Об утверждении Государственной программы развития здравоохранения Республики Казахстан на 2020 – 2025 годы»;

Все работы по проектированию, внедрению, настройке и интеграции системы лабораторного ассистента должны соответствовать требованиям следующих нормативных актов и стандартов:

- ISO/IEC/IEEE 42020 «Программное обеспечение, системы и корпоративная среда. Процессы, связанные с разработкой архитектуры»;
- СТ РК ISO 15704-2021 «Корпоративное моделирование и архитектура Требования к стандартным архитектурам и методологиям предприятия»;
- Постановления Правительства Республики Казахстан от 20 декабря 2016 года № 832 «Об утверждении Единых требований в области информационно-коммуникационных технологий и обеспечения информационной безопасности»;
- СТ РК IEC 62443-3-3-2021 Сети коммуникационные промышленные Безопасность сети и системы. Часть 3-3. Требования к системной безопасности и уровни безопасности.
- Закон Республики Казахстан «О связи» от 8 января 2003 года № 379.

1.5 Плановые сроки начала и окончания работы по системе лабораторного ассистента

Сроки оказания услуг (работ) по разработке модуля “Система лабораторного ассистента” октябрь 2025 г. - октябрь 2026 г.

1.6 Порядок внесения изменений в Техническое задание и их характер

Изменения, вносимые в ТЗ, должны быть технически обоснованными, содержать ссылки на актуальные нормативно-технические документы и иметь обозначенное авторство.

Изменения касаются усовершенствования действующих компонентов и развитию новых функций на базе текущей архитектуры предназначенной для системы лабораторного ассистента.

Изменения и дополнения к настоящему ТЗ должны оформляться Дополнениями или Протоколами к ТЗ в порядке, указанном в пп. 5.1.7 СТ РК 34.015-2002, согласованными и утвержденными организациями, участвующими в разработке, внедрении, сопровождении и эксплуатации системы лабораторного ассистента.

После подписания вышеназванные документы становятся неотъемлемой частью настоящего ТЗ.

2. Назначение и цели системы

Система лабораторного ассистента лаборанта включает функциональные модули для поиска и анализа нормативных документов, контроля условий проведения исследований, поддержки пользователей в ходе пробоподготовки и аналитических процедур, а также для ведения обучающих и справочных материалов.

- Система должна автоматизировать инструктаж и информированность сотрудника лаборатории на базе модуля искусственного интеллекта.
- Система должна быть легко масштабируемой для возможности дальнейшего расширения функциональности.
- Обеспечивать быстрый и точный контекстный поиск нормативной, методической и справочной информации;

2.1. Цель создания система лабораторного ассистента

Система лабораторного ассистента предназначен для поддержки лаборантов в работе по предоставлению понятных и актуальных инструкций по выполнению лабораторных процедур, ознакомлению с последними изменениями нормативных документов и стандартов, ознакомлению о правилах и требованиях техники безопасности, помощь в подготовке рабочего места в соответствии со стандартами техники безопасности, контроль соблюдения последовательности рабочих процессов в лаборатории, информирование по корректному обращению с лабораторным оборудованием, информирование о возможных ошибках работы с лабораторным оборудованием и реагентами и способах их предотвращения, сопровождение вступивших на должность сотрудников во время ознакомления с работой,

предоставлению справочной информации по реагентам, расходным материалам и оборудованию; помощи в ведении лабораторных журналов и заполнении форм, а также поддержке в оперативном пояснении регламентов, статусов анализов и организационных процедур. Развитие цифрового контура учреждения за счет внедрения локальной серверной инфраструктуры, оптимизированной под работу LLM-моделей, AI модулей, использования bare metal подхода в пределах закрытой сети учреждения.

Функциональные требования

- Разработка архитектуры и алгоритмов ИИ для поиска, анализа и выдачи нормативной и методической информации.
- Пилотное внедрение системы с последующей корректировкой и оптимизацией.
- Создание модуля интеллектуальных подсказок и контроля параметров исследования;
- Разработка единого хранилища нормативных документов, инструкций и методик исследований;
- Импорт нормативных документов должен поддерживаться из утвержденных источников.
- Реализация интеллектуальный поиск по текстам ГОСТ, СТРК и внутренним регламентам;
- Обновление нормативной информации. **указать обновление информации**
- Разработка обучающего модуля с краткими инструкциями и справочными материалами; **обучающая программа согласовать образовательную программу инструктаж. по технике безопасности**
- Система выдачи рекомендаций с помощью чат-бота при выполнении лабораторных процедур;
- Защита данных и ведение журнала действий пользователей;

2.2. Технические требования

- Предоставлять автоматическую проверку условий проведения исследований и выдачу рекомендаций при отклонениях от нормативов работы в ходе пользования системой лабораторного ассистента
- Уведомлять пользователя о загруженности системы если запрос идет более 10 секунд
- Поддерживать работу не менее 100 параллельных пользователей.
- Обеспечивать кеширование данных при сбое и обработать запрос после сбоя
- Система должна обеспечивать шифрование данных, разграничение прав доступа и ведение журнала действий пользователей;
- В системе должны поддерживаться резервная копия данных и механизм восстановления после сбоев;
- Система должна быть спроектирована с учетом возможного расширения, включая:

Возможность модернизации ПО для интеграции с новыми технологиями.

- Создание единого диспетчерского пункта для: мониторинга и управления учетными записями.

- Реализация заранее заданных сценариев: возможность настройки пользовательских сценариев, интерфейс должен соответствовать ролям пользователей.
- Адаптироваться под контекст работы лаборатории и сохранять его в ходе выдачи рекомендаций пользователю.

2.3. Требования к установке системы лабораторного ассистента на компьютеры лабораторий

- Установка системы должны быть выполнены квалифицированными специалистами с опытом работы с аналогичными системами
- По окончании установки необходимо предоставить в бумажном и электронном виде инструкции по эксплуатации и обслуживанию на казахском и русском языках.
- Обязательный мониторинг уязвимостей после завершения установки и настройки. Проведение тестирования перед запуском системы.
- Предоставление акта тестирования после завершения всех проверок системы.

2.4 Объем требуемых работ и услуг

В целях разработки системы лабораторного ассистента Заказчик предоставляет Поставщику для изучения существующую техническую, исполнительную документацию, без выноса за территорию объекта. Поставщик после изучения документации совместно с Заказчиком разрабатывают и утверждают график производства работ, с точной спецификацией оборудования и детализацией работ в течении 10 дней после заключения договора.

2.5 Резервное копирование и восстановление

- Система должна поддерживать автоматическое резервное копирование конфигурации, журналов событий и критически важных данных.
- Бэкап должен выполняться по расписанию и перед критическими обновлениями.
- Данные резервных копий должны храниться на отдельном защищенном сервере или носителе.
- Восстановление системы из резервной копии должно быть оперативным и не требовать полной остановки работы.
- Логирование всех операций по резервному копированию и восстановлению.
- Возможность ручного запуска резервного копирования по запросу администратора.

2.6 Технологический процесс работы система лабораторного ассистента состоит из трех ключевых этапов внедрения:

Первый этап внедрения системы лабораторного ассистента будет охватывать следующие подразделы:

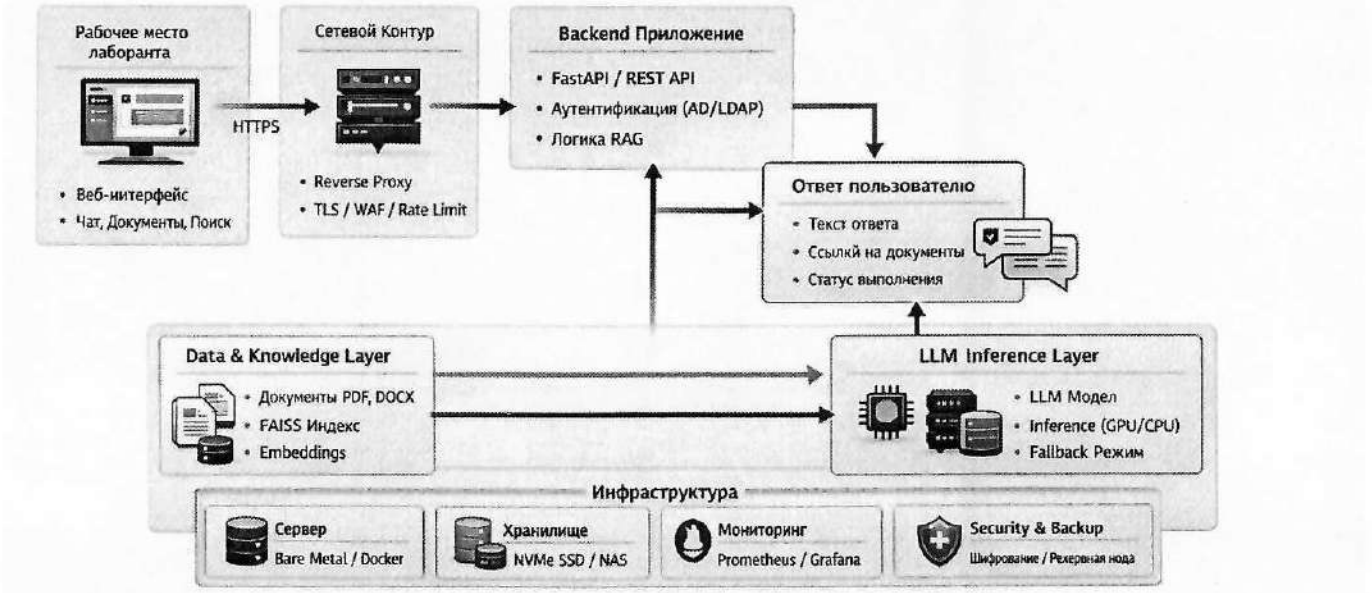
- Оценку существующей рабочих процессов ознакомления с методической информацией сотрудников и цифровых систем: набор подключенного оборудования, качество и полнота методических документов, доступность инструкций и СОП для персонала, текущее соблюдение техники безопасности,

уровень автоматизации операций, а также особенности взаимодействия лаборантов с системой и внутренними регламентами.
- Анализ выявляет узкие места, дублирующие операции, риски ошибок и информационные разрывы.

Второй этап внедрения целевой архитектуры:
- Определение будущей архитектуры цифрового контура лаборатории, включая локальную платформу для LLM, стандартизированные интерфейсы обмена (HL7 v2, ASTM), единую базу методик и инструкций.
- Формирование целевой модели взаимодействия: автоматизированные подсказки, контроль последовательности действий, оперативное предоставление актуальных нормативов, поддержка техники безопасности и интеллектуальная помощь при работе с оборудованием. Этот этап должен описать, алгоритм функционирования система лабораторного ассистента в работе лаборанта.

Третий этап:
- Составление технических рекомендаций.
- Подготовка перечня технических требований и шагов по переходу от текущего состояния к целевому: требования к серверным мощностям для LLM и ИИ, архитектурным компонентам, безопасности данных, интеграциям с информационной системой лаборатории и оборудованием;
- Рекомендации по цифровизации методик, стандартизации документирования, формированию единого справочника лабораторных данных; требования к эксплуатационной поддержке, обновлению моделей и мониторингу качества работы система лабораторного ассистента

Описательная бизнес-схема взаимодействия лаборанта с LLM:



3. Общие требования к архитектуре системы лабораторного ассистента

3.1 Требования к инфраструктурному уровню

- В основе архитектуры должны использоваться cloud-native подходы, применимые внутри закрытого контура: контейнеризация, оркестрация, сервисная сегментация.
- Должен быть реализован единый слой данных с интеграцией в информационную систему лаборатории через стандартизированные протоколы (HL7 v2, ASTM или API).
- Доступ к данным должен быть строго ролевым: ассистент может использовать только те сведения, которые разрешены для контекстных ответов
- Система должна поддерживать версионность документов, контроль обновлений и синхронизацию с внутренними справочниками лаборатории
- Логика работы должна предусматривать строгое ограничение функционала: ответы только в рамках инструкций, нормативов, оборудования, процедур и техники безопасность
- Обновление модели и базы знаний должно осуществляться через централизованный механизм, без влияния на текущие операции

3.2 Аутентификация и авторизация: реализовать механизмы строгой аутентификации пользователей и разграничения их прав доступа в соответствии с их ролями и обязанностями.

- Журналирование действий: вести подробные журналы (логи) действий пользователей и системных событий для последующего анализа и аудита (Prometheus, Grafana, файлы логов пользователей).
- Вести логи всех действий пользователей и системных событий с хранением не менее 6 месяцев.
- Мониторинг уязвимостей: регулярно проводить сканирование системы на наличие уязвимостей и своевременно устранять выявленные проблемы.
- Обновление программного обеспечения: обеспечить своевременное обновление всех компонентов системы для защиты от известных угроз.
- Интеграции должны быть построены по принципам безопасного обмена, с журналированием событий
- Должно быть предусмотрено управление соединениями и механизм контроля ошибок интеграции
- Безопасность и резервирование (Security & Backup)
- Резервное копирование конфигураций, моделей и базы знаний (Veeam, SAN snapshots)
- Разделение сетей: локальный VLAN/VRF, reverse proxy (Nginx/Traefik) для контроля доступа.
- Rate-limiting, мониторинг состояния LLM, fallback при зависании модели.

3.3 Требования к эксплуатационному уровню

- Архитектура должна быть прозрачной для последующей эксплуатации: обновление модулей, мониторинг производительности, контроль логов, отладка.
- Должен быть внедрен механизм мониторинга работы ассистента: состояние сервисов, задержки, ошибки, загрузка сервера.
- Должна обеспечиваться возможность расширения: добавление новых модулей ассистента (оборудование, СОП, контроль качества и т.д.).

- Все обновления должны проходить на тестовом окружении перед вводом в промышленную эксплуатацию
- Пользователями системы лабораторного ассистента являются: лабораторный персонал обслуживаемого объекта и иные лица по согласованию с заведующим участка.

3.4 Дополнительные требования хранений данных:

- Все конфиденциальные данные зашифрованы алгоритмом ГОСТ 28147-2009 / AES-256.
- Резервные копии шифруются и хранятся на физически защищенных носителях (NAS, offline HDD).
- Реализовать хеширование файлов (SHA-256, ГОСТ Р 34.11-2012) и контроль изменений.
- Автоматическая проверка целостности ПО и данных при каждом запуске системы.
- Вход в систему только по 2FA (пароль, токен/смарт-карта).
- Пароли пользователей должны соответствовать СТ РК ISO/IEC 27001-2021 (не менее 12 символов, регулярная смена).
- Вход в систему только с доверенных IP-адресов (белый список).

Таблица 1. Функциональные роли пользователей

Роль	Функции
Администратор	Управление доступом пользователей к системе, формирование шаблонов; установка плановых показателей.
Пользователь	Авторизированный по постоянной учетной записи пользователь системы.

Система лабораторного ассистента на всех ее уровнях должна отвечать следующим принципам:

- выдача релевантной запросу пользователя информации и ускорять существующие рабочие процессы поиска информации и выполнения СОП;
- непрерывный сбор информации, обработка данных;
- возможность масштабирования и модернизации системы лабораторного ассистента.

3.5 Требования к системе лабораторного ассистента

- Семантический поиск должен интерпретировать смысловые запросы и выдавать точные ответы на основе предзагруженной базы данных.
- Каждый ответ ассистента должен сопровождаться ссылкой на конкретный нормативный документ, пункт или фрагмент текста.
- В системе должны быть реализованы шаблоны ответов и раздел часто задаваемых вопросов, позволяющие ускорить обработку типовых запросов.

- Система должна обеспечивать возможность обновления нормативной базы с автоматической индексацией новых документов.
- Резервное копирование должно охватывать базу знаний, конфигурации и данные модели для безопасного восстановления в случае сбоя.
- Диалоговый чат должен обеспечивать интерактивное взаимодействие с LLM с сохранением контекста запросов.
- Экспорт ответов и выдержек из документов должен позволять сохранять информацию в формате PDF, Word или аналогов и содержать контекст запроса связанный с историей чата
- Поддержка автономной работы внутри периметра лаборатории должна исключать зависимость от внешних сервисов и Интернета.
- Отказоустойчивость системы должна обеспечивать продолжение работы при сбоях отдельных компонентов и предотвращать потерю данных

3.6 Требования к надежности и безопасности

- Обеспечение бесперебойной и отказоустойчивой работы с учетом ГОСТ 27.002-2015;
- Резервирование критичных компонентов системы в соответствии с СНиП РК 4.01-02-2001;
- Полная автономность системы без постоянного подключения к интернету;
- Запрет на передачу данных через облачные сервисы и сторонние серверы;
- Локальное хранение данных на защищенных серверах;
- Использование многоуровневой аутентификации и контроля доступа;
- Круглосуточная техническая поддержка 24/7 в течении гарантийного срока.
- Время реакции на заявки технической поддержки по уровню критичности:

- критический инцидент (полный отказ системы/авария)- реакция в течении часа, устранение - не более 4 часов;

- **средний уровень** (частичное нарушение функционала): реакция в течении 4 часов, устранение – в течении рабочего дня;

- низкий уровень (консультации, некритичные замечания): реакция в течении 8 часов, устранение - до 3 рабочих дней.

- Предоставление отчета по выполненным заявкам и времени реагирования;
- Предоставление актуальных контактных данных службы технической поддержки (телефон, email, ФИО ответственного лица) для оперативной связи.

4. Перечень элементов инфраструктуры РГП на ПХВ «ЦСЭЭ» УДП РК, который рекомендуется использовать для внедрения система лабораторного ассистента:

Интеграционный фундамент: HL7 v2 и web-сервис позволит быстро поставить шину/коннектор (Mirth Connect) и развернуть FHIR-витрину (HAPI FHIR).
SAN/FC и 10 GbE: достаточно для MVP;
Unity/3PAR имеют готовые CSI-драйверы для Kubernetes.
Backup: Veeam — база для резервирования конфигураций платформы (etcd/PVC/БД).

4.1 SCNI платформа как стандартный набор компонентов, который применим для ИИ ассистента лабораторна

- Интеграционный слой (Integration Layer)
- Подключение к информационной системе лаборатории через HL7 v2, FHIR, web-сервисы.
- Коннекторы/шины сообщений (Mirth Connect, HAPI FHIR) для синхронизации данных
- Обеспечивает поток данных для LLM-ассистента: нормативные документы.
- Хранилище данных (Data Layer)
- Реляционные базы данных (PostgreSQL/MySQL) для метаданных документов и сессий пользователей
- Файловое хранилище для документов (PDF, DOCX) и embedding-файлов FAISS Опционально SAN/NVMe SSD для быстрого доступа и масштабирования.
- Вычислительный слой (Compute Layer) GPU-серверы для inference LLM (NVIDIA A40/L40S или аналог). CPU, RAM для FastAPI, RAG, кэширования, фоновых задач.
- Контейнеризация с помощью Docker развертывание для упрощённого обновления и масштабирования
- LLM и AI-модули (AI Layer) LLM-инференс через Ollama (Qwen2 7B Instruct) или аналог
- Модули обработки документов: извлечение текста, разбиение на фрагменты, генерация embedding, FAISS-индексация.
- RAG-подход для повышения точности ответов на вопросы лаборантов
- Поддержка локальных моделей без выхода в интернет
- Сервисный слой (Service Layer) Backend (FastAPI) с REST API (/ask, /files, /upload, /doc, /health).
- Frontend: веб-интерфейс с вкладками «Чат», «Документ»,

4.2 Технические рекомендации к инфраструктуре для платформы

Рекомендуемое оборудование, которое должно учитываться в ресурсах

№	Наименование	Кол-во	Характеристика
1	Вычислительный сервер 1U для K8s (control/worker)	3 шт	2×CPU (уровня Xeon Silver/Gold/экв. AMD), 256–512 ГБ RAM, 2×NVMe (boot) + 6–10×NVMe/SAS (данные), 2×10 GbE SFP+, iDRAC/iLO/IPMI; роль — Виртуализация для кластеров Kubernetes/микросервисы

2	Сервер с GPU	1–2 шт	1–2×GPU (уровня NVIDIA A2/A10/экв.) для NLP/CV, 128–256 ГБ RAM, 2×10 GbE SFP+; роль — инференс ИИ
3	Система хранения (СХД/NAS)	1 шт	12–24 диска, суммарно от 100 ТБ raw; кэш NVMe, зеркальный boot; 2×10 GbE SFP+; поддержка снапшотов/репликации
4	Коммутатор 10 GbE SFP+ (L2/L3)	2 шт	24×SFP+ 10 GbE, 2–4×uplink, поддержка LACP/VLAN/ACL; установка парой для отказоустойчивости
5	Коммутатор 1 GbE для mgmt/OOB	1 шт	24×RJ-45 1 GbE, VLAN для изоляции mgmt/IPMI/консолей
6	Оптические модули SFP+ 10G SR	16 шт	LC, 850 нм, для серверов/СХД/коммутаторов (12 — серверы, 2 — СХД, 2 — запас)
7	Кабели DAC 10G SFP+ (короткие)	4 шт	Для межкоммутаторных связей и коротких подключений в стойке
8	Оптические патч-корды OM3/OM4 LC-LC	10–12 шт	1–3 м (по трассе), для подключений серверов/СХД к коммутаторам
9	Медные патч-корды Cat6a	24 шт	1–3 м, для mgmt/OOB и служебных подключений
10	Патч-панель LC (оптика)	1 шт	На 24 порта LC, с кросс-кассетами

11	Патч-панель RJ-45	1 шт	На 24 порта, Cat6a
12	Вертикальная PDU (электропитание)	2 шт	A/B питательные линии, защита, счётчики нагрузки
13	Источник бесперебойного питания (UPS) (опц.)	2 шт	По 3–6 кВА, с байпасом/сетевым управлением; ёмкость под время автономии по проекту
14	Органайзеры кабеля, лотки, стяжки/Velcro	Комплект	Для коммутации и разгрузки кабелей в стойке
15	Комплект крепежа/салазки	По месту	Для каждого сервера/СХД/коммутатора (обычно в комплекте поставки)
16	Консольный доступ/KVM over IP (опц.)	1 шт	Для аварийного локального доступа; альтернатива — только IPMI/iDRAC/iLO
17	Датчики температуры/влажности (SNMP)	2 шт	Мониторинг микроклимата стойки/зала
18	Набор ЗИП	1 комплект	2×SFP+ в запасе, вентиляторы/диски (1–2 шт критичных типоразмеров), кабели питания C13–C14

5. Необходимый поддерживаемый функционал MVP для работы:

- Импорт нормативных и лабораторных документов (DOCX, PDF) в локальную базу знаний
- Полный автоматический конвейер обработки данных:

извлечение текста → разбиение на фрагменты → генерация embedding → индексирование во FAISS или аналоги.

- Семантический поиск по документам с использованием Sentence Transformers.
- Подготовка контекста и использование RAG-подхода для повышения точности ответов.
- Генерация ответа моделью LLM (через Ollama или аналоги с совместимым интерфейсом).
- Возврат пользователю ответа с указанием источников, включая нормализованное читаемое название документов.
- Веб-интерфейс с двумя режимами работы: «Чат» и «Документ».
- Предпросмотр документов с поддержкой больших файлов, корректным отображением текста и скроллированием.
- Отображение состояния LLM-сервера и диагностики конфигурации.
- Поддержка доиндексации при загрузке новых документов без остановки сервиса.

5.1 Способы сбора обратной связи от пользователей

- Серверные логи запросов и ошибок.
- Возможность добавления модулей анализа качества ответов.
- Возможность интеграции кнопок оценки ответа («полезно / не полезно») и отправки фидбека.
- Анализ вопросов, по которым LLM возвращает ошибки или низкую уверенность.

5.2 Рекомендации к разворачиванию обновление приложения

Backend (FastAPI):

- Обновление через перезапуск Docker-контейнера.
- zero-downtime: новый контейнер поднимается параллельно старому.

Frontend:

- Простая замена статических файлов (index.html, JS).

LLM:

- Обновление модели через Ollama (ollama pull) без остановки FastAPI.

Индекс документации:

- Новые файлы доиндексируются без остановки системы.
- Процесс поддерживает разворачивание в закрытой сети без выхода в интернет.

5.3 Описание интерфейса

- Интерфейс состоит из двух основных зон:

Левая панель:

- Список всех загруженных документов.
- Поле поиска для фильтрации.

- Отображение размера файла.
- Открытие документа в режиме предпросмотра.

Правая панель:

- Вкладка «Чат»: история диалога, входное поле, кнопка отправки.
- Вкладка «Документ»: вертикальный просмотр полного текста, скроллирование без ограничений.
- Отображение источников ответов рядом с сообщением.
- Статус сервера LLM и текущей модели.
- Интерфейс работает полностью в браузере без внешних зависимостей.

5.4 Подготовка инфраструктуры

- Выделяется сервер для LLM и FastAPI с минимальной конфигурацией рассчитанную для 100 параллельных пользователей: GPU NVIDIA A40/L40S (48 GB VRAM) для ускоренного inference, CPU 16–24 ядер серверного класса, RAM 64–128 GB DDR4/DDR5 ECC, хранилище 1–2 TB NVMe SSD (PCIe 4.0), сетевой интерфейс 1–10 Gbit/s.
- Для масштабирования рекомендуется резервный сервер с предзагруженной моделью и балансировщик нагрузки.
- Вся инфраструктура разворачивается в закрытой локальной сети с VLAN/VRF и firewall.
- Reverse proxy (Nginx/Traefik) обеспечивает единую точку доступа, TLS-терминацию и rate-limiting.
- Аутентификация пользователей через AD/LDAP с выдачей JWT-токенов для API.

5.5 Рекомендуемые компоненты к программному обеспечению системы лабораторного ассистента:

№	Наименование	Количество	Характеристика
1	Bare-metal сервер с операционной системой	1	Сервер с GPU (NVIDIA A40/L40S, 48 GB VRAM), CPU 16–24 ядра, RAM 128 GB, NVMe SSD 1–2 TB, Linux серверная сборка; управляет LLM и FastAPI контейнерами

2	Docker	1	Контейнеризация Ollama, FastAPI, вспомогательных сервисов; GPU поддержка через NVIDIA Container Toolkit
3	LLM inference (Ollama)	1	Qwen2 7B Instruct в формате GGUF, локальный REST API, управление моделью и очередностью запросов
4	FastAPI backend	1	REST API для обработки запросов, интеграции с FAISS-индексом и базой знаний, кэширование запросов, управление токенами JWT
5	FAISS / база embedding	1	Быстрый семантический поиск по документам, доиндексация без остановки сервиса, поддержка больших документов
6	Кэширование (in-memory / Redis)	1	Хранение embedding, результатов поиска и промптов для ускорения отклика LLM
7	Объектное хранилище	1	Хранение исходных документов лабораторий (DOCX, PDF), raw,

			cleansed, curated. S3-совместимость
8	CI/CD (GitLab CE совместно с Runner)	1	Build/test/deploy пайплайны для FastAPI и фронтенда, blue-green/canary обновления локально
9	Мониторинг (Prometheus + Grafana)	1	Метрики работы сервиса и GPU, алерты по SLO/SLI, дашборды состояния LLM и FastAPI
10	Логи (ELK/EFK), трейсинг (OTel/Jaeger)	1	Централизованные логи FastAPI, Ollama; трассировка запросов, аудит и диагностика отказов
11	API-шлюз / Reverse proxy (Nginx/Traefik)	1	TLS-терминация, rate-limit, балансировка запросов, единая точка доступа к сервису
12	Аутентификация (Keycloak / LDAP)	1	SSO/OIDC, MFA, RBAC для пользователей, интеграция с корпоративным каталогом
13	Secret-manager / KMS (Vault)	1	Хранение секретов и ключей, управление токенами, ротация секретов

14	Резервное копирование (Velero)	1	Бэкапы контейнеров, конфигураций FastAPI, базы embedding и MinIO (объектное хранилище), проверка восстановления
15	Песочница с синтетическими данными	1	Тестовые наборы документов без персональных данных для проверки LLM, возможность тестирования новых функций и RAG-конвейеров

Развертывание LLM и вычислительного слоя

FastAPI и LLM развертываются в отдельных Docker-контейнерах с GPU-поддержкой через NVIDIA Container Toolkit. Ollama используется как inference-движок для Qwen2 7B Instruct (GGUF), обеспечивая совместимый REST API с компонентом ИИ.

Ollama управляет загрузкой модели, очередностью запросов и генерацией ответов. FastAPI предоставляет REST API для запросов к модели, работы с документами и мониторинга сервера, включая эндпоинты /ask, /files, /upload, /doc, /health, /debug/llm.

Backend валидирует токены и управляет доступом пользователей.

5.6 Подготовка базы знаний

- Документы лабораторий импортируются в локальную базу знаний.
- Данные проходят pipeline обработки: извлечение текста
- Для семантического поиска используются Sentence Transformers.
- RAG-подход позволяет формировать контекст для запросов и повышать точность ответов.
- Новые документы можно доиндексировать.
- Кэширование embedding и результатов поиска осуществляется в памяти сервера

- Клиентский интерфейс Веб-интерфейс обеспечивает работу в браузере. Интерфейс разделен на вкладки «Чат» с историей диалогов, вводом запроса и отображением источников, и «Документ» для просмотра полного текста и скроллинга.
- Статус сервера и модели отображается в интерфейсе.
- Локальный кеш сохраняет историю диалогов и открытые документы в памяти браузера

5.7 Логирование и мониторинг

- Логи фиксируются в файловом логе на сервере.
- При использовании PostgreSQL возможна запись событий и расширенных логов для аудита.
- Мониторинг включает метрики времени ответа LLM, успешные и неуспешные запросы, ошибки сети и зависания.
- Настроен автоматический restart сервисов
- Fallback при недоступности LLM позволяет пользователю продолжать поиск по документам с уведомлением «Модель временно недоступна».

5.8 Обновление и синхронизация

- Обновление всех компонентов выполняется локально без выхода в интернет
- Выполняется синхронизация между клиентом и сервером осуществляется
- При временной недоступности сети используется локальный кеш для сохранения истории диалогов и открытых документов.

5.9 Управление пользователями

- Аутентификация осуществляется через AD/LDAP или локальную схему
- Осуществляется авторизация API
- Логируются все попытки входа, успешные и неуспешные.
- История чатов хранится с ретенцией 6–12 месяцев, периодическая очистка
- Закладывается механизм оценки ответа LLM пользователем («полезно / не полезно») для сбора обратной связи и улучшения модели.

5.10 Масштабирование и отказоустойчивость

- Резервная нода с предзагруженной моделью обеспечивает отказоустойчивость.
- Автоматический перезапуск ИИ модуля при зависании
- Возможность расширения RAM и GPU при росте количества пользователей.

6. Список приложений

- Приложение 1 - бизнес-схема системы
- Приложение 2 - функциональная-схема системы
- Приложение 3 - техническая-схема системы