ТЕХНИЧЕСКОЕ ЗАДАНИЕ (MVP)

Проект: Daryn Online
Интеллектуальный чат-ассистент на казахском и русском языках, отвечающий на вопросы пользователей по базе знаний платформы на основе искусственного интеллекта (RAG)

Наименование	ИИ-ассистент поддержки учащихся, родителей и преподавателей (MVP)
Компания-заявитель (заказчик)	ТОО «Daryn.online» — владелец и оператор образовательной платформы Daryn Online
Контактное лицо	Мәдениет Дидар — ответственное лицо заказчика
Пользователи MVP	Учащиеся, родители и преподаватели Daryn Online; оператор поддержки; администратор
Данные	Публичная база знаний платформы: раздел «Помощь/FAQ» (~150 вопросов-ответов, HTML), инструкции (~30 документов PDF/DOCX), правила и тарифы (~10 документов), перечень 100 типовых вопросов поддержки (XLSX). Итого до 400 страниц, до 100 МБ, на казахском и русском. Передаёт заказчик в течение 5 рабочих дней (ZIP-архив + список URL). Реальные обращения и персональные данные не передаются (раздел 3.1)
Срок разработки MVP	6 недель: разработка MVP — 4 недели, пилот — 2 недели (разделы 3.4 и 10)
Бюджет (диапазон)	от 800 000 до 1 300 000 тенге (договорная, в пределах диапазона), раздел 11
Форма вознаграждения	Договорная: от 800 000 до 1 300 000 тенге. Оплата поэтапно — 30% аванс, 40% после демонстрации промежуточного результата, 30% после приёмки MVP.
Примечание	Только резиденты РК. К заявке приложить описание решения и архитектуры, состав команды, портфолио проектов RAG/LLM, календарный план и смету. Источники и формат базы знаний, способ доступа и интеграции — в разделах 3.1–3.2; границы пилота — в разделе 3.4. Реальные обращения и персональные данные не передаются. Исключительные права на результат переходят заказчику. Оплата LLM-API или GPU-сервера и хостинга — за счёт заказчика.
Дата	23.11.2026 (редакция 1)

1. Проблема

Служба поддержки получает большой поток однотипных вопросов: как пройти тест, где посмотреть результаты, как восстановить доступ, как оплатить подписку, что делать при технической ошибке. Ответ приходит через несколько часов, в вечернее время и выходные вопросы остаются без ответа. Операторы перегружены повторяющимися обращениями, а пользователи уходят с платформы, не решив проблему.

2. Цель MVP

За минимальный срок создать рабочий чат-ассистент, который отвечает на вопросы пользователей на казахском и русском языках строго по базе знаний платформы, указывает источник ответа и передаёт вопрос оператору, если не уверен в ответе, — и проверить его в двухнедельном пилоте на тестовой странице сайта Daryn Online. MVP и пилот должны подтвердить, что ассистент самостоятельно и правильно закрывает не менее 70% типовых вопросов и не выдумывает факты.

Ассистент не изменяет данные пользователей, не принимает оплату и не меняет оценки — он только консультирует и при необходимости передаёт обращение человеку.

3. Границы MVP

Входит в MVP	Не входит (следующие версии)
Загрузка базы знаний (PDF, DOCX, HTML, FAQ)	Доступ к личному кабинету и данным пользователя
Ответы на казахском и русском языках	Голосовой ассистент, мессенджеры (Telegram, WhatsApp)
Ссылка на источник в каждом ответе	Выполнение действий (сброс пароля, оплата)
Передача оператору при низкой уверенности	Интеграция с CRM / Help Desk заказчика
Веб-виджет чата и панель администратора базы знаний	Мобильное приложение
Журнал диалогов и оценка ответа пользователем («полезно» / «не полезно»)	Расширенная аналитика и отчёты PDF/XLSX
REST API модуля	SSO и встраивание в интерфейс Daryn Online

3.1. Источники данных и база знаний

Ассистент работает только с публичной базой знаний платформы. Все источники передаёт заказчик в течение 5 рабочих дней после подписания договора.

Источник	Формат	Объём	Способ доступа в MVP	Кто предоставляет
Раздел «Помощь / FAQ» платформы Daryn Online	HTML	~150 вопросов-ответов (каз./рус.)	Список URL публичных страниц + выгрузка CSV/HTML	Заказчик
Инструкции для учащихся, родителей и преподавателей	PDF, DOCX	~30 документов, ~250 страниц	ZIP-архив в облачной папке или Git-репозитории заказчика	Заказчик
Правила пользования, публичная оферта, тарифы и способы оплаты	PDF, HTML	~10 документов, ~60 страниц	ZIP-архив + URL	Заказчик
Перечень типовых вопросов поддержки (топ-100) с эталонными ответами, без переписки и персональных данных	XLSX / CSV	100 вопросов	Файл в той же папке	Заказчик (служба поддержки)
Синтетические тестовые вопросы	CSV	≥ 600 вопросов (каз./рус. поровну)	Создаются исполнителем	Исполнитель

Общий объём базы знаний: около 50 документов, до 400 страниц, до 100 МБ; языки — казахский и русский. Контрольный набор — 300 вопросов из перечня заказчика и синтетических вопросов с эталонными ответами, согласованными с заказчиком.

Документация. Вместе с материалами заказчик передаёт «паспорт базы знаний» — таблицу с перечнем документов (название, раздел, язык, URL, дата актуализации) и описание структуры раздела FAQ.

Обновление базы. В MVP — повторная загрузка документов через панель администратора или переобход списка URL по кнопке. Автоматическая синхронизация через API платформы — в следующей версии; при наличии API заказчик предоставит его описание.

Не используются: реальные обращения пользователей, личные кабинеты, ФИО, ИИН, телефоны и иные персональные данные.

3.2. Информационные системы и интеграции

Система	Назначение	Интеграция в MVP	Кто обеспечивает
Сайт Daryn Online — тестовая страница	Размещение виджета чата на пилоте	JS-код виджета (одна строка <script>) на отдельной странице; без доступа к личным кабинетам	Заказчик размещает код; исполнитель передаёт инструкцию
Облачная папка или Git-репозиторий заказчика	Передача документов базы знаний	ZIP-архив, список URL, паспорт базы знаний	Заказчик
Модуль ассистента (REST API)	Вопросы и ответы, загрузка документов	FastAPI, документация OpenAPI, авторизация по ключу	Исполнитель
Языковая модель (LLM)	Формирование ответа	Вариант А: внешний LLM-API (ключ на аккаунт заказчика). Вариант Б: открытая модель (ISSAI KazLLM, Qwen) на сервере с GPU. Выбор — на неделе 1	Заказчик оплачивает; исполнитель настраивает
Поиск по базе знаний	Индекс и поиск фрагментов	Многоязычные эмбеддинги (multilingual-e5), PostgreSQL + pgvector	Исполнитель
Канал передачи оператору	Обращения, которые ассистент не закрыл	E-mail службы поддержки заказчика + список в панели администратора; Help Desk / CRM — следующая версия	Заказчик даёт e-mail; исполнитель настраивает
Хостинг в РК (Alem Plus, PS Cloud или иной)	Размещение модуля	VPS: 4 vCPU, 8 ГБ RAM, 100 ГБ SSD (вариант А) или сервер с GPU 24 ГБ (вариант Б)	Заказчик предоставляет доступ

3.3. Обязательства заказчика

- в течение 5 рабочих дней передать материалы базы знаний, список URL и паспорт базы знаний;
- предоставить доступ к хостингу и репозиторию, ключ LLM-API (при варианте А), e-mail службы поддержки;
- согласовать эталонные ответы контрольного набора на неделе 2;
- разместить код виджета на тестовой странице сайта к началу недели 5;
- обеспечить участников пилота: 20–30 тестовых пользователей и 2–3 оператора поддержки;
- назначить ответственное лицо для оперативных согласований (Мәдениет Дидар).

3.4. Пилот (MVP): границы

Параметр	Значение
Объект пилота	Веб-виджет ассистента на отдельной тестовой странице сайта Daryn Online; только консультации по базе знаний
Темы	5 тем: прохождение тестов; результаты; вход и восстановление доступа; тарифы и оплата; технические ошибки
Участники	20–30 тестовых пользователей (сотрудники заказчика, учащиеся и родители по приглашению) и 2–3 оператора поддержки
Объём	База знаний ≈ 50 документов; контрольный набор — 300 вопросов; в пилоте — не менее 300 диалогов
Языки	Казахский и русский
Сроки	Разработка MVP — недели 1–4; пилот — недели 5–6; итоговый отчёт — конец недели 6
Критерии оценки	Метрики раздела 9 на контрольном наборе (неделя 4) и в диалогах пилота (недели 5–6)
Не входит в пилот	Реальная эксплуатация для всех пользователей, личные кабинеты, оплата, мессенджеры
Результат	Отчёт о пилоте с метриками и рекомендацией: переходить ли к полной версии и что доработать

4. Пользовательский сценарий

1	Администратор загружает документы базы знаний и запускает индексацию.
2	Пользователь открывает виджет чата и задаёт вопрос на казахском или русском языке.
3	Ассистент находит релевантные фрагменты и отвечает на языке вопроса со ссылкой на источник.
4	Если уверенность низкая или вопрос вне тематики — ассистент сообщает об этом и передаёт обращение оператору.
5	Пользователь оценивает ответ; администратор видит журнал диалогов и неотвеченные вопросы.

5. Функциональные требования

№	Функция	Требование MVP
1	База знаний	Загрузка PDF, DOCX, HTML, CSV-FAQ; переиндексация при обновлении
2	Поиск	Гибридный поиск (векторный + по ключевым словам) по двум языкам
3	Ответ	Ответ на языке вопроса, не длиннее 150 слов, со ссылкой на источник
4	Защита от выдумок	Ответ только по найденным фрагментам; иначе — «не знаю» и передача оператору
5	Передача оператору	Форма обращения с историей диалога
6	Виджет	Веб-виджет чата для встраивания на сайт
7	Панель администратора	Документы, журнал диалогов, неотвеченные вопросы, оценки
8	API	Отправка вопроса, получение ответа, загрузка документов; авторизация по ключу
9	Отчёт о качестве	Метрики на контрольном наборе отдельно по языкам

6. Контроль качества ответов

- **Обоснованность:** каждый ответ проверяется на соответствие найденным фрагментам.
- **Порог уверенности:** при низкой релевантности ответ не формируется — обращение передаётся оператору.

- **Тематика:** вопросы вне тематики платформы и попытки обойти инструкции отклоняются.
- **Языковой паритет:** качество проверяется отдельно по языкам; разница не более 10 п. п.

7. Технологическое решение

Компонент	Решение для MVP
Индексация	Разбиение документов на фрагменты, многоязычные эмбеддинги, pgvector/Qdrant
Генерация ответа	RAG: гибридный поиск + переранжирование + LLM с инструкцией отвечать только по источникам
Оценка качества	Автоматическая оценка на контрольном наборе + выборочная экспертная проверка
API и интерфейс	FastAPI, OpenAPI; веб-виджет и панель администратора
Развертывание	Docker, docker-compose; GPU 24 ГБ при локальной LLM

Безопасность: панель — вход по логину и паролю; доступ к API по ключу; HTTPS; секреты вне кода; ассистент не запрашивает и не хранит персональные данные, предупреждает пользователя об этом; защита от prompt-injection в вопросах пользователя.

8. Нефункциональные требования

- Время ответа ассистента — не более 5 секунд для 90% вопросов.
- Одновременная работа не менее 50 пользователей в MVP.
- Индексация 50 документов — не более 10 минут; данные не теряются при перезапуске; журнал диалогов.

9. Критерии приёмки MVP и оценки пилота

Метрика	Где измеряется	Целевое значение
Доля правильных ответов	Контрольный набор (300 вопросов)	≥ 80%
Доля типовых вопросов, закрытых без оператора	Диалоги пилота	≥ 70%
Доля ответов с выдуманными фактами	Контрольный набор + выборка 100 диалогов пилота	≤ 5%
Ответы со ссылкой на источник	Контрольный набор и пилот	100%
Корректный отказ / передача оператору для вопросов вне тематики	Контрольный набор	≥ 90%
Разница доли правильных ответов между казахским и русским	Контрольный набор	≤ 10 п. п.
Время ответа для 90% вопросов	Пилот	≤ 5 секунд
Число диалогов в пилоте	Пилот	≥ 300
Оценка удовлетворённости участников пилота	Опрос участников	≥ 4 из 5
Персональные данные в базе знаний и тестовых наборах	Проверка наборов	Отсутствуют

10. Этапы разработки и пилота

Этап	Срок	Результат
Подготовка и данные	Неделя 1	Получение базы знаний и паспорта, выбор LLM (А/Б), индексация, набор вопросов
RAG и модели	Недели 2–3	Поиск, генерация ответов, защита от выдумок, согласование эталонов
Виджет, панель и API	Недели 2–3 (параллельно)	Виджет чата, панель администратора, API, Docker
Приёмочные испытания	Неделя 4	Проверка метрик на контрольном наборе, отчёт о качестве, исправления
Пилот	Недели 5–6	Работа виджета на тестовой странице, ≥ 300 диалогов, отчёт о пилоте и рекомендации

11. Бюджет и порядок оплаты

Ориентировочный бюджет MVP и пилота: **от 800 000 до 1 300 000 тенге**. Форма вознаграждения — договорная. В сумму входят работы исполнителя по всем этапам, включая сопровождение пилота. Оплата LLM-API или GPU-сервера и хостинга — за счёт заказчика.

Этап	Доля	Сумма, тенге (от – до)
Подготовка и данные	15%	120 000 – 195 000
RAG и модели	35%	280 000 – 455 000
Виджет, панель и API	25%	200 000 – 325 000
Приёмочные испытания	10%	80 000 – 130 000
Пилот и отчёт	15%	120 000 – 195 000
Итого	**100%**	**800 000 – 1 300 000**

Оплата поэтапная: 30% — аванс после подписания договора; 40% — после приёмочных испытаний (неделя 4); 30% — после завершения пилота и подписания акта по критериям раздела 9.