ИИ-ассистент поддержки пассажиров на казахском, русском и английском языках

Технологическая задача ТОО «SMSBUS»

Техническое задание

1. Границы пилота

Каналы пилота: чат в приложении SMSBUS и мессенджеры Telegram и WhatsApp. Приложение AlemPay и остальные каналы подключаются после пилота.

Темы пилота: оплата проезда и причины отказа, возврат при ошибочной оплате, транспортные и виртуальные карты, пополнение, льготы, работа приложения.

В пилот не входят: голосовой канал и приём звонков, автоматический возврат денег без оператора, оформление льгот, изменения в платёжной системе.

Исходные показатели поддержки: число обращений в месяц, доля звонков и чатов, число операторов, часы работы линии. Эти данные нужны исполнителям для оценки стоимости.

2. Роли и права доступа

Действие	Пассажир	Оператор поддержки	Редактор базы знаний	Администратор SMSBUS
Диалог с ассистентом	Да	Нет	Тестовый режим	Тестовый режим
Запрос статуса своего платежа	Да	Нет	Нет	Нет
Приём диалога от ассистента	Нет	Да	Нет	Да
Просмотр журнала диалогов	Только свой	Свои диалоги	Все, без личных данных	Все
Изменение базы знаний	Нет	Предложить правку	Да	Да
Аналитика обращений	Нет	Нет	Да	Да
Пользователи, каналы, ключи API	Нет	Нет	Нет	Да

3. User Stories

ID	Роль	Я хочу	Чтобы
US-1	Пассажир	задать вопрос на казахском, русском или английском и получить ответ на том же языке	не ждать оператора и не искать правила на сайте
US-2	Пассажир	узнать, почему не прошла моя оплата	понять, что исправить, и оплатить проезд

ID	Роль	Я хочу	Чтобы
US-3	Пассажир	подать заявку на возврат при ошибке в коде автобуса	вернуть деньги без звонка на линию
US-4	Пассажир	в любой момент позвать оператора	решить вопрос, с которым ассистент не справился
US-5	Оператор поддержки	получать диалог с историей и кратким содержанием	не переспрашивать пассажира
US-6	Редактор базы знаний	менять статьи базы знаний без участия разработчиков	ответы соответствовали действующим правилам и тарифам
US-7	Редактор базы знаний	видеть вопросы, на которые ассистент не ответил	дополнять базу знаний по реальным обращениям
US-8	Администратор SMSBUS	видеть темы обращений, долю решённых без оператора и оценки	управлять качеством поддержки

4. Критерии приёмки (Given-When-Then)

AC-1. Ответ на языке пассажира (US-1). Дано: в базе знаний есть статья по теме вопроса. Когда: пассажир пишет вопрос на казахском, русском или английском, в том числе смешивая казахский и русский. Тогда: ассистент отвечает на языке вопроса не позже чем через 5 секунд и указывает статью, на которой основан ответ.

AC-2. Вопрос вне базы знаний (US-1). Дано: ответа на вопрос в базе знаний нет. Когда: пассажир задаёт такой вопрос. Тогда: ассистент не придумывает ответ, сообщает, что не знает, и предлагает перевод на оператора; вопрос попадает в список неотвеченных.

AC-3. Причина отказа в оплате (US-2). Дано: пассажир авторизован в приложении и у него есть неуспешная оплата за последние 24 часа. Когда: он спрашивает, почему не прошла оплата. Тогда: ассистент запрашивает статус по API, называет причину отказа из справочника и шаги для её устранения.

AC-4. Сервис статусов недоступен (US-2). Дано: API статусов платежа не отвечает дольше 3 секунд. Когда: пассажир спрашивает о своей оплате. Тогда: ассистент сообщает, что не может проверить платёж, не называет предположительную причину и предлагает перевод на оператора.

AC-5. Заявка на возврат (US-3). Дано: у пассажира есть успешная оплата с ошибочным кодом автобуса. Когда: он просит вернуть деньги. Тогда: ассистент собирает номер операции и описание, создаёт заявку для оператора, сообщает пассажиру её номер и срок рассмотрения и не обещает возврат.

AC-6. Перевод на оператора (US-4, US-5). Дано: идёт диалог с ассистентом. Когда: пассажир просит оператора или ассистент дважды подряд не смог ответить. Тогда: диалог передаётся оператору с полной историей и кратким содержанием; в нерабочее время пассажир получает сообщение о часах работы и номер обращения.

AC-7. Посторонние и вредоносные запросы (US-1). Дано: пассажир пишет вопрос не по теме сервиса или просит ассистента нарушить его правила. Когда: сообщение получено. Тогда: ассистент вежливо отказывает, не раскрывает служебные инструкции и возвращает разговор к темам поддержки.

AC-8. Личные данные в диалоге (все роли). Дано: пассажир прислал в чате номер банковской карты или ИИН. Когда: сообщение сохраняется в журнал. Тогда: эти данные в журнале замаскированы и не передаются внешним сервисам.

AC-9. Обновление базы знаний (US-6). Дано: редактор изменил статью и опубликовал её. Когда: прошло не более 10 минут. Тогда: ассистент отвечает по новой редакции, а прежняя редакция сохранена в истории изменений.

AC-10. Доступ к чужому платежу (US-2). Дано: пассажир авторизован. Когда: он называет номер операции другого пользователя. Тогда: ассистент не сообщает сведений об этой операции.

5. Структура API

Базовый путь `/api/v1`, формат JSON, авторизация по токену, описание в OpenAPI 3.0. Состав методов исполнитель может уточнить на этапе проектирования.

API ассистента (разрабатывает исполнитель):

Метод	Путь	Назначение
POST	`/conversations`	Начать диалог: канал, язык, идентификатор сессии
POST	`/conversations/{id}/messages`	Отправить сообщение пассажира и получить ответ
POST	`/conversations/{id}/handoff`	Передать диалог оператору
POST	`/conversations/{id}/feedback`	Оценка ответа пассажиром
POST	`/refund-requests`	Создать заявку на возврат для оператора
GET, POST, PUT	`/kb/articles`	Статьи базы знаний на трёх языках и история изменений
GET	`/analytics/topics`	Темы обращений, доля решённых без оператора, оценки
GET	`/analytics/unanswered`	Вопросы, на которые ассистент не ответил
GET	`/health`	Состояние сервиса

API заказчика (предоставляется в тестовом окружении):

Метод	Путь	Назначение
GET	`/payments/{operation_id}/status`	Статус платежа и код причины отказа для авторизованного пассажира

Пример запроса и ответа:

```
POST /api/v1/conversations/c-10482/messages
{"text": "Неге төлем өтпеді?", "channel": "app_smsbus"}

200 OK
{
  "conversation_id": "c-10482",
  "language": "kk",
  "answer": "Төлем өтпеді, себебі ...",
  "sources": ["kb-017"],
  "payment_checked": true,
  "handoff_suggested": false
}
```

Ошибки возвращаются в едином виде `{"code": "...", "message": "...", "request_id": "..."}` с кодами 400, 401, 403, 404, 422, 429 и 500.

6. Нефункциональные требования

Категория	Требование
Производительность	Первый ответ не позже 5 секунд (p95); не менее 50 одновременных диалогов
Надёжность	Доступность не ниже 99,5% в месяц; при сбое ассистента пассажир видит контакты поддержки; сбой не влияет на приём оплаты
Качество ответов	Ответы только по базе знаний заказчика со ссылкой на статью; набор из не менее 300 проверочных вопросов на трёх языках прогоняется перед каждым обновлением
Языки	Казахский (кириллица, а также смешанная казахско-русская речь), русский, английский; язык определяется автоматически
Безопасность	Номера карт, ИИН и телефоны маскируются в журналах; TLS 1.2 и выше; ролевая модель доступа; данные диалогов не используются для обучения сторонних моделей
Размещение и хранение	Журналы диалогов хранятся на сервере; срок хранения 12 месяцев
Интеграция	Встраивается в приложения готовым модулем чата без полной переработки приложений; мессенджеры подключаются через официальные API
Стоимость эксплуатации	Исполнитель указывает стоимость одного диалога
Передача результата	Исходный код, документация API, инструкция по развёртыванию и ведению базы знаний; права на результат определяются договором

7. Цели пилота

Показатели измеряются 30 дней по журналам диалогов и выборочной проверке ответов сотрудниками заказчика.

Метрика	Цель
Обращения, закрытые без оператора	не менее 50%
Ответы, признанные верными при выборочной проверке	не менее 95%
Ответы о деньгах и льготах не по базе знаний	0
Верно определённый язык обращения	не менее 98%
Время первого ответа (p95)	не более 5 секунд
Оценка пассажиров	не ниже 4,3 из 5

8. Сроки и этапы

Разработка и интеграция занимают 8 недель, затем 30 дней наблюдения за пилотом.

1. Анализ обращений и проектирование — 2 недели: темы, база знаний на трёх языках, контракт API.
2. Ядро ассистента и набор проверочных вопросов — 2 недели.
3. Интеграция со статусами платежа, заявки на возврат, перевод на оператора — 2 недели.
4. Кабинет поддержки, подключение каналов и запуск пилота — 2 недели.