**Техническое задание (ТЗ)**

«Мультиарендный AI-чатбот для продаж (Telegram/WhatsApp), диалог без кнопок»

**Цель документа.** Зафиксировать целевую архитектуру, функциональные требования, API, модели данных, интеграции и план работ по доведению текущего проекта до логически завершённого продукта. Документ — пошаговая инструкция для Replit Agent и разработчиков.

**1. Цели и KPI**

* Единый код → «ферма» ботов (тенанты).
* Полный отказ от кнопок: живой диалог текст/голос/картинки.
* Каталоги/КП/счета и медиа — Google Drive/Sheets per tenant.
* Мультиагентность и мультиязычность (RU/KZ auto).
* ИИ-экспертность: ответы строго на основе каталога/материалов (без «воды»).
* Воронка из 3 уточняющих вопросов.
* SLA: среднее время ответа < 2с (без учёта ASR/LLM), конверсия в лид > 8%.

**2. Область работ**

Три слоя: вход (вебхуки TG/WA, ASR, vision), оркестрация (NLU, FSM, LLM), backend (Catalog/Leads/Docs/Admin).

* TG/WA вебхуки, роутинг по tenant.
* ASR/TTS: voice/audio/document(audio/\*); ответ голосом при голосовом входе.
* Image: OCR/Vision для поиска по каталогу.
* NLU без кнопок: intents (search/add/show\_cart/quote/invoice/checkout/callback/address/help/stop) + сущности (sku/qty).
* FSM-воронка (3 вопроса), слот-филлинг.
* LLM responder с guard-rails (только поверх данных каталога/текстов).
* Хранилище: PostgreSQL (Drizzle, multitenant), Redis — опционально.
* Файлы: Google Drive per tenant, Google Sheets (или Postgres) для прайсов/лидов.
* Admin: агенты/интеграции/промпты/файлы; импорт каталогов; аналитика.

**3. Архитектура**

**3.1 Сервисы**

* **Gateway:** /webhook/tg/:tenantKey, /webhook/wa — нормализация входа (text/voice/audio/image), resolve tenant.
* **ASR/TTS:** Whisper-стиль транскрипция; сменный TTS.
* **NLU:** парсинг интентов/сущностей; RU/KZ; fallback в LLM.
* **Dialog Orchestrator:** FSM + воронка + корзина + команды-слова.
* **Catalog API:** FTS + pg\_trgm (опечатки), карточки, остатки/цены, cross-sell.
* **Orders/Leads:** корзина→заказ/лид, статусы, уведомления.
* **Docs:** генерация КП/счёта (DOCX→PDF), загрузка в GDrive, выдача ссылок.
* **Admin:** мультиарендность, интеграции, тексты/промпты, импорт, роли, аудит.

**3.2 Данные (PostgreSQL, multitenant)**

* tenants(id, key, title, tg\_token?, wa\_phone\_id?, wa\_api\_key?, gdrive\_folder\_id, locale, currency, price\_policy, …)
* catalogs(id, tenant\_id, sku, name, brand, category, unit, price, stock, pack\_qty, warehouse, lead\_time\_days, photo\_url, description, attrs jsonb, searchable tsvector)
* customers(id, tenant\_id, chat\_id, channel, phone, name, type b2b|b2c, bin, email)
* carts, cart\_items, orders(total/status/delivery/payment/meta)
* leads(need\_by\_date, comment, file\_url)
* texts(key,value) — промпты/скрипты/локализация
* users(role,email,hash), audit(actor, action, payload, ts)

**3.3 Интеграции**

Google Drive/Sheets; KaspiQR/Halyk/Stripe; CRM (Bitrix/Amo); 2GIS.

**4. Функциональные требования**

**4.1 Диалог без кнопок**

Ввод: текст/голос/фото. Вывод: текст; голос (TTS) — если вход голосом; медиа/ссылки по необходимости.  
Команды-в-словах: «найди …», «добавь … 5 шт», «корзина», «оформить заказ», «КП», «счёт», «адрес», «перезвон», «стоп».  
RU/KZ автоязык (детект → ответ на языке входа).

**4.2 ИИ-экспертность и воронка**

3 уточняющих вопроса (назначение/объём/бюджет → срок/условия).  
LLM-ответ только на данных каталога/текстов (запрет на выдумку цен/наличия).  
2–3 варианта + экспертный комментарий + CTA; cross-sell/up-sell.

**4.3 Документы и хранилище**

КП/счёт (DOCX→PDF) по шаблонам, загрузка в GDrive, ссылка клиенту/менеджеру.  
Каталоги/медиа/шаблоны — в GDrive папке тенанта.  
Импорт каталога из CSV/XLSX (до 50к), отчёт о дубликатах/ошибках.

**4.4 Админ-панель**

Tenants (токены TG/WA, локаль/валюта, GDrive/Sheets, CRM/оплата).  
Expert Scripts (промпт ниши, тональность, приоритеты сравнения).  
Funnel (конструктор 3 вопросов, валидации, маппинг к заказу/КП).  
Catalogs (импорт, валидация, индексация tsvector/trgm).  
Integrations (статусы, тест-кнопки). Security (login/roles/audit). Analytics (диалоги/лиды/конверсия).

**5. API (черновик)**

* POST /webhook/tg/:tenantKey, POST /webhook/wa
* GET /api/catalog/search?q=&tenant=&limit=
* GET /api/catalog/sku/:sku?tenant=
* POST /api/cart/add {tenant, chat\_id, sku, qty}
* GET /api/cart?tenant=&chat\_id=
* POST /api/order/checkout {tenant, chat\_id, delivery, payment}
* POST /api/lead {tenant, ...}
* POST /api/docs/quote {tenant, lead\_id}, POST /api/docs/invoice {tenant, order\_id}
* Admin: GET/POST/PATCH /api/tenants, POST /admin/catalog/import, GET /admin/leads, GET /admin/orders, POST /admin/texts

**6. NLU и правила ответа**

**Intents:** search, add\_to\_cart, show\_cart, remove/clear, checkout, quote, invoice, delivery/info, promo, address, callback, help, stop, unknown.  
**Entities:** sku, name, qty, unit, date, address, phone.  
**Правила:** кратко, экспертно, язык запроса; цены/наличие — только из Catalog API; при голосовом входе — добавлять TTS-аудио-ответ.

**7. Безопасность**

Row-level мультитенантность, проверки tenant на каждом запросе.  
Секреты в ENV; GDrive доступ per tenant (service account).  
Рейт-лимит; маскирование PII; бэкапы БД.

**8. Развёртывание**

**ENV минимум:** DATABASE\_URL, TG\_WEBHOOK\_BASE или TG\_TOKEN\_\* (до tenants), WA\_VERIFY\_TOKEN, WA\_ACCESS\_TOKEN/360dialog, OPENAI\_API\_KEY/ANTHROPIC\_API\_KEY, TTS\_PROVIDER\_KEY?, GOOGLE\_JSON, GDRIVE\_ROOT\_FOLDER, PROMO\_URL, TWO\_GIS\_URL, ADMIN\_EMAIL.  
**CI/CD:** Drizzle миграции, health-checks, структурные логи.

**9. План работ (с принтами)**

**Спринт 1 (7–10 дней):**

1. Postgres + Drizzle (tenants, catalogs, leads, customers, carts) и DAO.
2. Вебхуки TG/WA, нормализация входящих, определение tenant.
3. ASR: voice/audio/document(audio/\*); TTS для ответов на голос.
4. NLU: парсер интентов/сущностей; слот-филлинг воронки.
5. LLM-ответ с guard-rails (RU/KZ).
6. Admin: страница Tenants (токены, GDrive/Sheets), логин (passport).
7. GDrive: загрузка КП/файлов и выдача ссылок.

**Спринт 2 (10–14 дней):**

1. Docs: КП/счёт DOCX→PDF→GDrive.
2. CRM (Bitrix/Amo): лиды при checkout/lead.
3. Payments: KaspiQR/Halyk/Stripe статусы; отправка QR в WA как изображение.
4. Catalog Import UI, индексация (tsvector/trgm).
5. Analytics: диалоги/лиды/конверсия.

**Спринт 3 (10+ дней):**  
Фото-поиск (Vision), персональные цены B2B tiers, клонирование бота (копия тенанта), автоперевод RU↔KZ, очереди/уведомления/KPI.

**10. Критерии приёмки (DoD)**

* Текст/голос/фото в TG/WA; ответ голосом при голосовом входе.
* Импорт каталога; поиск с опечатками; cross-sell.
* КП/счёт — в GDrive со ссылками.
* 2+ тенанта с изолированными данными.
* Админ-панель: агенты/интеграции/промпты/импорт; login/roles/audit.
* LLM-ответы без выдумки цен/наличия.

**11. Инструкция для Replit Agent (шаги)**

1. Создать/применить миграции Drizzle для таблиц из §3.2.
2. Перейти на Telegram webhooks: POST /webhook/tg/:tenantKey.
3. Нормализовать вход: text/voice/audio/document(image|audio).
4. Добавить ASR+TTS; включить режим «ответ голосом при голосовом входе».
5. Реализовать NLU+FSM (3 слота); LLM responder с guard-rails.
6. Admin: /api/tenants + UI; логин (passport).
7. Docs: /api/docs/quote|invoice (DOCX→PDF→GDrive).
8. CRM/Payments — плагины per tenant.
9. E2E тесты: импорт/поиск/заказ/КП/голос.

**12. Риски и ограничения**

Галлюцинации LLM → строгие guard-rails.  
Ограничения WA медиа → документы/QR как ссылки/изображения.  
Стабильность ASR/TTS → кэш/таймауты/fallback.