Tech Rules — Umrah AI Search Kiosk (ICHS) — OpenAI version (locked)

0) Hard constraints

- **Reliability > completeness** (public kiosk).
- **Text-only** responses.
- **No agentic runtime** (no tool-calling loops; no web browsing tools in production kiosk).
- **No raw question text stored** during the public event (hashed only + analytics).
- **No Docker dependency** assumed for kiosk runtime (to avoid day-of surprises).

---

1) Stack (locked)

Frontend

- **React + Vite + TypeScript**
- TailwindCSS
- Arabic RTL support (layout + typography)
- Video playback for Tayyib loops (WebM alpha preferred; MP4 fallback)

Backend

- **Python 3.11 + FastAPI**
- Uvicorn
- Local SQLite for analytics/events
- Local vector store: **Chroma embedded** (on-disk)

OpenAI API

- **Responses API** for generation (`POST /v1/responses`) — recommended for new projects. (OpenAI Platform)
- **Embeddings API** for retrieval. (OpenAI Platform)

---

2) Model choices (locked)

Generation model

- **`gpt-4o`** (primary). (OpenAI Platform)

- Optional fallback for cost/latency: `gpt-4o-mini` (only if QA passes AR/FR quality). (OpenAI Platform)

Embeddings model

- `text-embedding-3-large` (best multilingual retrieval). (OpenAI Platform)

---

3) Repo structure (monorepo)

```
umrah-kiosk/
  apps/
    kiosk-frontend/
    kiosk-backend/
  packages/
    shared-schema/
  data/
    offline_pack/
    rag_corpus/
    chroma_index/
  assets/
    tayyib_loops/
    branding/
  docs/
    PRD.md
    DESIGN.md
    TECH_RULES.md
    TODO.md
```

---

4) Configuration (env vars only)

Required

- `OPENAI_API_KEY` (server-side only; never in frontend) (OpenAI Platform)
- `OPENAI_MODEL=gpt-4o`
- `OPENAI_EMBED_MODEL=text-embedding-3-large`
- `PUBLIC_QR_BASE_URL` (placeholder now; real domain later)
- `KIOSK_IDLE_TIMEOUT_SEC` (e.g., 60)
- `APP_LANG_DEFAULT` (EN/AR/FR)

Local storage paths

- `SQLITE_PATH=./data/analytics.sqlite`

- `CHROMA_PATH=./data/chroma_index`
- `OFFLINE_PACK_PATH=./data/offline_pack/offline_pack.json`

---

5) Data handling rules

Logging (public event)

Store only:

- `session_id` (UUID)
- `language` (EN/AR/FR)
- `mode` (ask/guide/pose)
- `rating_1_5`
- `thumb` (optional)
- `latency_ms`
- `route_used` = `offline | rag | fallback`
- `hashed_query` = SHA-256(normalized_query + salt)
- `timestamp`
- `time_on_screen_ms` (session duration)

Never store:

- raw question text
- personal identifiers

Offline Answer Pack

Must be **fully trilingual**.
Schema per entry:

- `id`
- `question_variants[]` (per language)
- `answer.direct`
- `answer.steps[]`
- `answer.mistakes[]`
- `tags[]`
- `last_updated`

---

6) RAG rules (Chroma local)

Ingestion (build-time only)

- Curate Saudi official sources → ingest → chunk → embed → store in Chroma.
- No live ingestion at venue.

Chunking defaults

- Chunk size: ~400–800 tokens
- Overlap: ~80–120 tokens
  Metadata per chunk:
- `source_title`
- `source_url`
- `section_heading`
- `lang`
- `approved_by` (ICHS stakeholder + date)

Retrieval

- Retrieve `top_k=5` always; frontend decides how many to display:
  - Vertical default show 3
  - Horizontal default show 5

Confidence + clarifier

- If retrieval confidence below threshold:
  - ask **one** clarifier (buttons preferred), then rerun
  - else safe fallback (no hallucination)

---

7) Generation rules (Responses API)

Endpoint

- `POST /v1/responses` ([OpenAI Platform])

Output format contract (strict)

Backend must return structured JSON:

- `answer.direct` (string)
- `answer.steps[]` (strings)
- `answer.mistakes[]` (strings)
- `sources[]` = { `title`, `url`, `snippet`, `relevance` }
- `confidence` (0–1)

- `refinement_chips[]` (strings)
- `route_used`
- `latency_ms`

Style constraints (system prompt rules)

- Informational, official-tone, not “fatwa”
- Avoid madhhab comparisons
- Short, kiosk-readable blocks
- Arabic output in Fusha; Arabic UI RTL

Timeouts + retries

- Model call hard timeout: **8s**
- Retry: **max 1** on transient network errors
- On failure: offline pack or safe fallback

---

8) UI implementation rules (non-negotiable)

- **No scrolling** in Ask/Guide main panels.
- Limited scrolling allowed only inside Sources drawer/panel.
- Language chosen on Home; locked for session.
- End-of-session prompt: **1–5 star rating**.
- Tayyib always visible:
  - Hero on Home
  - Compact panel/card in Ask/Guide
- Tayyib media:
  - Prefer **WebM with alpha** for transparency; MP4 fallback supported via matched panel background.

---

9) Backend API (locked)

`POST /api/ask`

Input:

- `lang`
- `query`
- `session_id`

Output:

- structured answer JSON (see section 7)

`POST /api/guide`

Input:

- `lang`
- wizard answers

Output:

- `checklist_sections[]`
- `qr_url` = `${PUBLIC_QR_BASE_URL}/share#d=<payload>`

`POST /api/feedback`

Input:

- `session_id`
- `rating_1_5`
- `time_on_screen_ms`
- optional `mode_stats`

`GET /api/health`

- returns OpenAI reachability + Chroma status + disk warnings

---

10) QR share rules (stateless, no retention)

- QR URL uses fragment payload: `/share#d=<compressed_payload>`
- Share page renders **checklist only** in the same language
- No server storage, no deletion jobs

---

11) Quality gates (must pass before event)

- Works with internet disabled (Ask falls back to offline pack + safe fallback UI).
- No crashes or stack traces on screen.
- Offline pack present and validated at startup.
- Admin hidden panel: export CSV, health, reset, offline simulation.