**Methodology**

**A) System goals & scope**

* **Capabilities**: (1) swim & dryland workout generation, (2) pace/workout analysis from splits/HR/rest, (3) injury & recovery tips, (4) nutrition guidance, (5) general knowledge over swimming docs/records.
* **Safety limits**: non-diagnostic injury guidance; advise medical referral for persistent pain; cite authoritative sources for rules/records/nutrition.

**B) Data & knowledge sources**

1. **Structured CSVs** (profiles, workouts, races, wellness).
2. **Unstructured docs**: technique/injury/nutrition guides, tapering templates, federation rules/records PDFs.
3. (Optional) **Wearables**: per-set splits, HR, SWOLF.
4. **Authority links**: World Aquatics records pages for on-demand citations. [World Aquatics](https://www.worldaquatics.com/swimming/records?utm_source=chatgpt.com)

**C) Retrieval & generation pipeline (local-first)**

1. **Chunking** (e.g., 800 chars / 120 overlap).
2. **Embeddings**: SBERT (all-MiniLM-L6-v2). [arXiv](https://arxiv.org/abs/1908.10084?utm_source=chatgpt.com)
3. **Dense index**: FAISS FlatIP (CPU), with option to move to HNSW/IVF for scale. [arXiv](https://arxiv.org/pdf/2401.08281?utm_source=chatgpt.com)
4. **Sparse index**: BM25 (rank-bm25). [PyPI](https://pypi.org/project/rank-bm25/?utm_source=chatgpt.com)
5. **Fusion**: RRF to combine dense + BM25. [cormack.uwaterloo.ca](https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf?utm_source=chatgpt.com)
6. **Re-rank (optional)**: MiniLM-L6-v2 cross-encoder top-k. [Hugging Face](https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2?utm_source=chatgpt.com)
7. **Prompting**: system+user templates enforcing: (i) cite context, (ii) refuse if not in context, (iii) surface uncertainty.
8. **Local LLM** (optional): GPT4All/Ollama; or operate “context-only” for deterministic responses when needed.

**D) Domain logic (non-LLM)**

* **Workout generator**: parameterized by stroke, goal (aerobic/threshold/VO₂/sprint), volume. Use CSS (or race-pace proxies) to map targeted send-offs/paces; embed taper heuristics for pre-meet weeks. [PubMed+1](https://pubmed.ncbi.nlm.nih.gov/1555562/?utm_source=chatgpt.com)
* **Pace analysis**: compute mean, SD, CV; flag consistency (CV ≤2–4%), mismatch vs. target CSS, rest adequacy; annotate HR if provided; suggest set variations (shorter reps/stricter send-offs). [PubMed](https://pubmed.ncbi.nlm.nih.gov/1555562/?utm_source=chatgpt.com)
* **Injury tips**: shoulder-first (rotator cuff/scapular strength, T-spine mobility, technique checks), with watch-outs for pain-provoking drills; add referral triggers. [PubMed+1](https://pubmed.ncbi.nlm.nih.gov/37515375/?utm_source=chatgpt.com)
* **Nutrition**: pre/during/post templates grounded in ACSM/IOC positions; highlight REDs red flags and referral. [PubMed+2British Journal of Sports Medicine+2](https://pubmed.ncbi.nlm.nih.gov/26891166/?utm_source=chatgpt.com)
* **Records/Rules**: pull from World Aquatics and display source. [World Aquatics](https://www.worldaquatics.com/swimming/records?utm_source=chatgpt.com)

**E) Evaluation plan**

**1) Retrieval quality (offline)**

* **Gold queries**: build a 100–200 query set spanning: technique, injuries, nutrition, taper, records.
* **Metrics**: Recall@k (k = 5, 10), MRR, nDCG@k across: BM25-only, dense-only, hybrid (RRF), and hybrid+rerank. Target: **hybrid+rerank ≥ hybrid ≥ single**. [cormack.uwaterloo.ca](https://cormack.uwaterloo.ca/cormacksigir09-rrf.pdf?utm_source=chatgpt.com)
* **Ablations**: chunk size, overlap, top-k, fusion method, CE re-rank on/off.

**2) Answer quality (human-in-the-loop)**

* **Rubric** (0–5): factuality, citation adequacy, actionability, domain alignment (CSS/taper/nutrition), safety (medical disclaimers).
* **Panel**: 2–3 coaches + 1 physio + 1 nutritionist score blinded outputs; compute average and inter-rater agreement.

**3) Domain correctness checks**

* **Pacing math tests**: synthetic inputs with known CSS → verify prescribed paces/rest logic.
* **Nutrition templates**: verify g·h⁻¹ carb ranges, timing windows, and compliance with consensus statements. [PubMed+1](https://pubmed.ncbi.nlm.nih.gov/26891166/?utm_source=chatgpt.com)
* **Injury guardrails**: ensure red-flag prompts (night pain, neurological signs) always trigger referral language.

**4) Live telemetry (local, privacy-safe)**

* Log anonymized: query type, retrieval hits, chosen k, refusal rate; **no personal data** without consent.

**F) Deployment & data governance**

* **Local-only** by default (free models, on-device FAISS/BM25).
* **PHI/PII minimization**: athlete IDs (e.g., S001) + opt-in mapping to names.
* **Versioned knowledge**: timestamp doc snapshots; re-index on change.
* **Source of truth**: for records/regulations, always cite **World Aquatics** pages to avoid drift. [World Aquatics](https://www.worldaquatics.com/swimming/records?utm_source=chatgpt.com)

**G) Risks & mitigations**

* **Out-of-scope medicalization** → strict disclaimers + referral triggers; keep to coaching-level advice.
* **Stale knowledge** → schedule doc refresh for records/rules; add “last-checked” stamps.
* **Overfitting to docs** → hybrid retrieval + rerank; keep BM25 in the loop for exact-term coverage. [PyPI](https://pypi.org/project/rank-bm25/?utm_source=chatgpt.com)

**Concrete Build Checklist (tl;dr)**

1. Curate docs: technique, injury, nutrition, taper; add **World Aquatics** bookmarks for record lookups. [World Aquatics](https://www.worldaquatics.com/swimming/records?utm_source=chatgpt.com)
2. Index: SBERT embeddings → FAISS; BM25 corpus; enable **RRF** and CE rerank. [cormack.uwaterloo.ca+3arXiv+3arXiv+3](https://arxiv.org/abs/1908.10084?utm_source=chatgpt.com)
3. Domain logic: CSS calculator + pace validator; taper templates; injury/nutrition templates per ACSM/IOC. [British Journal of Sports Medicine+3PubMed+3ScienceDirect+3](https://pubmed.ncbi.nlm.nih.gov/1555562/?utm_source=chatgpt.com)
4. Prompts: enforce citations & uncertainty; add medical-safety rails.
5. Eval: Retrieval metrics + expert rubric; iterate chunk/fusion/rerank.