Những điểm đã làm “tốt nhất có thể” trong phạm vi yêu cầu

Ngày tạo: 08/01/2026

1) Agentic Retrieval đúng pipeline

- Planner (LLM) → quyết định need_subquery, tạo 1–3 subqueries, chọn keyword / vector / hybrid, đề xuất filters (source/updated_at/doc_id…).
- Retrieval:
- keyword → BM25 (multi_match + boost title + exact title.keyword khi có)
- vector → kNN (embedding query)
- hybrid → chạy keyword + vector riêng, RRF fuse ở app (ổn định nhất)
- Merge + Dedup: dedup theo doc_id:chunk_id, giữ TOP_N=8
- Evidence check: nếu yếu → refine tối đa 1 vòng (tránh tốn chi phí)
- Answer: LLM trả lời chỉ dựa trên chunks, kèm citations dạng: [doc_id:chunk_id] title (updated_at) (url nếu có)

2) Hybrid search “ổn định”

- Không dùng DSL hybrid phụ thuộc version.
- BM25 + kNN → RRF tại app: ổn định, dễ tuning, dễ debug.

3) Vector search “chịu lỗi” tốt hơn

Trong searchVector() có 2 query shape:

- Shape #1: bool(filter) + must(knn)
- Nếu fail (tuỳ cluster/plugin): fallback Shape #2: query.knn + post_filter
- ⇒ Giảm rủi ro “cluster của bạn không support knn trong bool”.

4) Debug/Logging đúng chuẩn bạn cần

POST /chat với debug=true trả:

- planner_output_initial, planner_output_refined
- evidence_initial, evidence_refined
- executed_queries (kèm stage: initial/refined)
- retrieved_chunks (top N, đã RRF + dedup)
- final_answer

5) Guardrails + “không bịa”

- Nếu không có chunks hoặc evidence quá yếu: trả đúng câu “Không tìm thấy trong tài liệu hiện có.” + gợi ý từ khoá.
- Không leak system prompt (prompt chỉ nằm trong code server).
- Enforce citations: nếu LLM trả lời mà không có citation, server tự append nguồn tham khảo từ top chunks.

6) Power Apps

- Có sẵn OpenAPI file để import Custom Connector: powerapps/custom-connector/openapi.yaml
- Có hướng dẫn: powerapps/custom-connector/README.md

7) Diagnostics endpoint

- GET /diagnostics: best-effort kiểm tra kết nối OpenSearch + LLM + Embeddings
- Rất hữu ích khi triển khai on-prem (đỡ mò lỗi config).

Chạy ngay

Docker Compose

```
cp .env.example .env
docker compose up --build -d
```

- Node backend: http://localhost:3000
- FastAPI proxy (optional): http://localhost:8001

Test nhanh

```
curl -X POST http://localhost:3000/chat \ -H "Content-Type: application/json" \ -d '{ "message": "Quy trình reset mật khẩu VPN như thế nào?", "history": [], "filters": {"source":["policy","sop"]}, "debug": true }'
```

Diagnostics

```
curl http://localhost:3000/diagnostics
```

Gợi ý quan trọng để “chạy tốt” trong môi trường của bạn

Embeddings

- Nếu LLM server của bạn không có /embeddings: set EMBEDDINGS_ENABLED=false → hệ thống tự fallback keyword.
- Nếu có embeddings: đảm bảo EMBEDDING_MODEL + EMBEDDING_DIM đúng (ví dụ 768).

OpenSearch auth

- OPENSEARCH_AUTH_MODE=none|basic|apikey (đã hỗ trợ đủ).

Tuning keyword fields

- OPENSEARCH_KEYWORD_FIELDS=title^3,title.keyword^6,content (mặc định).
- Nếu bạn hay hỏi đúng “tên SOP/policy”, có thể tăng boost title.keyword^10.

Nếu bạn muốn đẩy thêm 1 nấc production-ready (vẫn giữ nhẹ, dễ chỉnh), có thể cập nhật theo hướng:

- cache theo doc_id:chunk_id + session,
- thêm source boosting theo intent (policy/sop/tech/wiki),
- thêm “structured citations” tách riêng field (để Power Apps render đẹp),
- thêm ACL filter (theo user/role) nếu bạn có trường phân quyền trong index.