รายงานการปรับปรุงพร้อมเพิ่มประสิทธิภาพระบบ RAG และการตอบของ Chatbot (Optimization Result — รอบที่ 2)

(ไฟล์บันทึกผลการดำเนินงานสำหรับรอบการปรับปรุงครั้งที่ 2 — ต่อยอดจาก `optimize_result_1.md`)

รอบนี้เน้น **คุณภาพการค้นคืน (Retrieval) และการตอบของ Chatbot** โดยยึดหลัก **Vanilla RAG เดิม** ไม่เพิ่มสถาปัตยกรรมหนัก (ไม่มี agentic/graph/multi-hop) และ **ไม่แก้ไข/ลบ/ฝัง Vector Storage ใหม่เด็ดขาด** (ใช้ embedding เพื่อ retrieve ตรวจสอบเท่านั้น) การเปลี่ยนแปลงทั้งหมดอยู่ที่ชั้น query-time ในไฟล์ `backend/rag_engine.py` และ `backend/config.py`

---

1. ปัญหาที่ตรวจพบเพิ่มเติม (Root-Cause Findings)

จากการอ่านผลจริงใน `verify_rag_answers.txt` พบ bug ที่ซ่อนอยู่แม้ในเคสที่ "ผ่าน":

1. **คะแนนความเกี่ยวข้องหลอกตา (Misleading Similarity)** — ขั้นตอน rerank เขียนทับค่า `distance` ทำให้ค่า `similarity` ที่โชว์ในแหล่งอ้างอิงจริง ๆ คือ **อันดับของ rerank** (1.0, 0.8, ...) ไม่ใช่ความเกี่ยวข้องเชิงความหมายจริง จึงเป็นเหตุให้เคสนอกขอบเขต (เบาหวาน) แสดง chunk ที่ไม่เกี่ยวข้อง (AOM, ไซนัส, Decolgen) เป็น `similarity: 1.0`
2. **knob กรองความเกี่ยวข้องถูกทิ้งไว้เฉย ๆ** — `SIMILARITY_THRESHOLD` ถูก import แต่ไม่เคยถูกใช้งานในเส้นทางค้นคืนเลย
3. **บังคับดึงครบทุกเล่ม** — source coverage บังคับให้มีอย่างน้อย 1 chunk ต่อ Guideline แม้เล่มนั้นจะไม่เกี่ยวข้อง ทำให้เกิดการอ้างข้ามหัวข้อ/ข้ามช่วงวัย
4. **เลขหน้าวารสารทำให้สับสน** — context ที่ส่งให้โมเดลมีทั้ง `Page` (หน้า PDF) และ `Journal` (หน้าวารสาร) ซึ่งเป็นเลขคนละชุดของ chunk เดียวกัน เป็นต้นเหตุที่โมเดลอ้างเลขหน้า AAFP ไม่ตรงกับไฟล์ PDF
5. **การสกัด Ref ภายนอกเปราะบาง (bug เดิม)** — regex เดิมจับได้เฉพาะ Ref ที่มีวลี "อ้างอิงจาก" เท่านั้น ทำให้ Ref จริง เช่น `[Ref: ความรู้นอกคู่มือ - สมาคมโรคเบาหวานฯ https://...]` **หลุดหายไปเงียบ ๆ** (URL ไม่ขึ้นให้ผู้ใช้)

---

2. สรุปรายการแก้ไข (Key Optimizations)

2.1 ค้นคืนและคะแนนความเกี่ยวข้อง (Retrieval Integrity)

- **แยกคะแนนจริงออกจากอันดับ rerank**: rerank ไม่เขียนทับ `distance` อีกต่อไป — แหล่งอ้างอิงที่แสดงจึงเป็น **ความเกี่ยวข้องเชิง vector จริง** (เช่น 0.74, 0.68) ไม่ใช่อันดับปลอม 1.0
- **เปิดใช้งาน relevance gating** : นำ `SIMILARITY_THRESHOLD` (ปรับ default เป็น **0.66**) มาใช้ตรวจว่า context "เกี่ยวข้องต่ำ/นอกขอบเขต" หรือไม่ (วัดจริง: in-scope ~0.69–0.79, out-of-scope ~0.61)

2.2 การกรองแหล่งอ้างอิงให้ซื่อตรง (Honest Sources)

- เพิ่ม knob `SOURCE_MIN_SIMILARITY` (0.60) — **ตัด chunk ที่เกี่ยวข้องต่ำออกจากรายการแหล่งอ้างอิง** ที่แสดงให้ผู้ใช้ (กันแผง Ref รก เช่น Decolgen 27%)
- **เคสนอกขอบเขต**: ไม่แสดง Guideline เป็นแหล่งอ้างอิงเลย (กันเข้าใจผิดว่าอ้างจากคู่มือ) และแทรกหมายเหตุใน context บอกโมเดลว่าอย่าฝืนอ้าง Guideline ที่ไม่ตรง ให้ระบุว่าเป็นความรู้นอกคู่มือ + แนบ URL แทน
- ป้องกันการกรองจนหมด: ถ้าไม่ใช่เคสนอกขอบเขตแต่ถูกกรองหมด จะคง chunk ที่เกี่ยวข้องสุด 1 อันไว้เสมอ

2.3 ความแม่นยำเลขหน้าอ้างอิง (Page-Citation Accuracy)

- **ตัดเลขหน้าวารสาร (`Journal`) ออกจาก context และ rerank** — โมเดลเห็นเลขหน้าเดียวคือหน้า PDF ที่ใช้เปิดไฟล์จริง ลดการอ้างเลขหน้าผิดเล่ม
- เสริมกฎใน prompt: ต้องดึงเลขหน้าจากฟิลด์ `Page:` ของ chunk ที่หยิบมาใช้จริง และเลขหน้าต้องคู่กับ source เดียวกันเสมอ

2.4 การสกัดอ้างอิงภายนอกให้ทนทาน (Robust External Ref)

- เปลี่ยนตรรกะ: ถือว่า `[Ref: ...]` ใด ๆ ที่มี **URL** หรือมีคำบ่งชี้ "นอกคู่มือ/นอกเอกสาร/ความรู้ทั่วไป" เป็นอ้างอิงภายนอกทันที ไม่ผูกกับวลี "อ้างอิงจาก" อีกต่อไป — URL ภายนอกจึงแสดงให้ผู้ใช้ได้ครบทุกครั้ง

2.5 ปรับ Prompt เสริมช่องว่างจาก Feedback (Answer Quality)

- **ขนาดยาเป็นช่วง Min–Max เต็มช่วง + ความถี่ยืดหยุ่น** เช่น "Amoxicillin 500 mg วันละ 2–3 ครั้ง ตามความรุนแรง" และคำนวณยาน้ำเป็น mL เมื่อทราบความแรง
- **ห้ามตอบว่า "ไม่มีข้อมูลใน Guideline" ถ้าข้อมูลอยู่ใน Context จริง** และให้ตอบยาตามอาการ (Symptomatic) ครบทุกอาการ
- **เทียบหลาย Guideline** (โดยเฉพาะ AOM / pharyngitis) ระหว่าง URI เด็ก 2562 กับ AAFP 2022 ให้ครบทั้งขนาดยาและระยะเวลา
- แยกข้อมูลนอกคู่มือออกจาก `[Ref: Guideline]` ให้ชัด ไม่ปนจนผู้ใช้แยกไม่ออก
- เลิกใช้ Emoji ในข้อความ error (สอดคล้องปัญหา encoding บน Windows)

---

3. ผลการตรวจสอบ (Verification Results)

ทดสอบแบบ **ไม่แตะ Vector Storage** — ใช้ retrieve จริงเพื่อวัดผล และ unit test ตรรกะล้วน

ชุดทดสอบ	รายละเอียด	ผล
Unit tests (20 เคส)	best-similarity, gating, กรอง source, coverage floor, สกัด Ref ภายนอกหลายรูปแบบ, ไม่มี Journal leak, หมายเหตุ weak-context	**ALL PASS**
Live retrieve — in-scope	ผู้ใหญ่ pharyngitis / เด็ก 3 ขวบ / เด็ก 2 ขวบ / Centor / Azithromycin → best sim 0.69–0.79, **weak=False**, แสดง source ปกติ พร้อมคะแนนจริง	**PASS**
Live retrieve — out-of-scope	เบาหวาน type 2 → best sim 0.61, **weak=True**, **ไม่แสดง Guideline source เลย** (เดิมเคยโชว์ AOM/Decolgen เป็น 1.0)	**PASS**
End-to-end — out-of-scope	จัดเป็นประเภทที่ 5 ถูกต้อง, sources เหลือเฉพาะ URL ภายนอกจริง (`https://www.dmthai.or.th`), ไม่มี noise จาก Guideline	**PASS**
Import ทุกโมดูล	`rag_engine`, `config`, `patient_summary`, `main`	**OK**

---

4. ไฟล์ที่แก้ไข (Changed Files)

- `backend/rag_engine.py` — rerank/distance, build_context, source builders + gating, prompt, error strings
- `backend/config.py` — เพิ่ม `SOURCE_MIN_SIMILARITY`, ปรับ default `SIMILARITY_THRESHOLD` = 0.66 (env override ได้ทั้งคู่)

ไม่แตะ: `rag/qdrant_db/` (Vector Storage), `rag/pipeline.py`, ไฟล์ chunk/embedding — ตามข้อกำหนด retrieve-only

5. ข้อสังเกต / รอบถัดไป (Observations & Next Steps)

- ค่า threshold ปัจจุบันตั้งจากค่าที่วัดจริง (in-scope ~0.69–0.79 vs out-of-scope ~0.61) หากเพิ่มคู่มือ/โรคใหม่ ควรวัดซ้ำและปรับผ่าน env
- ปัญหาที่เหลือใหญ่สุดคือ **ความตรงของเลขหน้า PDF ระดับ chunk** ซึ่งเป็นเรื่องของ data/chunking ต้อง re-embed จึงจะแก้ที่ต้นตอได้ — อยู่นอกขอบเขตรอบนี้ (ห้ามแตะ embedding) จึงเลือกลดความกำกวมที่ป้อนให้โมเดลแทน (ตัด Journal page)
- โครงสร้าง sources ที่ส่งออก (`source/page/heading/similarity/type/url`) ยังคง schema เดิม ไม่กระทบ frontend และการบันทึกลง SQLite