

なんでもかんでもRAGじゃない

よくある相談パターン

- 1 「RAGを入れないといけない」という声が多い
- 2 プロンプトだけで十分に回るケースが多い
- 3 パターン化できるならRAGより精度が上がることも

問い①：全部見てほしいか？

- 1 AIにデータを全部見てほしいのかどうか
- 2 全部読んで判断してほしいなら部分検索のRAGは向かない
- 3 全文を読ませるほうが自然で確実

問い②：入力上限に収まる？

- 1 GPT/Claudeは10～20万字、Geminiは100万字以上という上限
- 2 意外と全部入る。コストはほんのわずか
- 3 複雑なRAGを組むより安く・速く・安定して動く

問い③：パターン数は？

- 1 問い合わせや業務が10～20パターンほどならテンプレート化が最適
- 2 業界別に作ったプロンプトは使い回しが簡単
- 3 精度も運用性も申し分ない

RAGが不要ない条件

- 1 全部見たいならプロンプトに入れる
- 2 上限に収まるなら全文を入れる
- 3 パターンが少ないならテンプレ化する
- 4 この条件がそろえばRAGは不要ない
- 5 もっとシンプルで強い選択肢がある

RAGは万能ではない

- 1 RAGは万能の正義ではない
- 2 適材適所でこそ輝く技術
- 3 どんな状況でもRAGに頼る必要はない

今日覚えておきたいこと

- 1 RAGは部分を取り出す仕組み
- 2 全部読ませたいなら向かない。収まるなら全文投入が合理的
- 3 パターンが少ないならテンプレ化で十分
- 4 状況に合わせて柔軟に選ぶことが鍵

RAGを使わない勇気

- 1 勢いでRAGを導入して失敗する事例は多い
- 2 「本当に必要か」を落ち着いて見極める
- 3 RAGを使わない勇気が成功を近づける