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を入れないとダメですよね？」という声をよく耳にする
2. 実際は、プロンプトだけで綺麗に回せるケースがたくさんある
3. 内容がパターン化できるなら、むしろRAGより精度が上がることも

問い①：全部参照する？

1. 「AIに全部の情報を見てほしいのかどうか」を考える
2. 全部読んだうえで判断してほしいなら、部分検索のRAGは不向き
3. 全文を読ませるアプローチのほうが自然

問い②：入力上限に収まる？

1. GPTやClaudeは10〜20万字、Geminiなら100万字以上入る
2. 全文を入れてもコストはごくわずか
3. RAGを作り込むより、速くて安定していることが多い

問い③：パターンは何個？

1. パターンが10〜20種類くらいならテンプレート化が最適
2. 業界別に作っておけば使い回しも簡単
3. 精度も運用も申し分ない

3つの問いまとめ

1. 「全部見たい」ならそのままプロンプトへ
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
4. この条件がそろえば、RAGは必要ありません
5. もっとシンプルで強い方法があるから

RAG＝正義ではない

1. RAGは万能ではない
2. 適材適所でこそ輝く技術
3. 「どの状況でもRAGが正解」というわけではない

今日のポイント

1. RAGは「部分を取り出す技術」
2. 全部読む用途には向かない
3. 入力が収まるなら全文投入でOK
4. パターンが少ないならテンプレ化が最善
5. 状況に合わせて賢く選ぶことが大切

RAGを使わない勇気を

1. 勢いでRAGを作って失敗する例は少なくない
2. まずは「本当に必要か」を落ち着いて見極める
3. RAGを使わない判断こそ、成功を近づける大事な一歩