Выбор модели

Цель

Сгенерировать персональные рекомендации по улучшению сна на основе метрик со смарт-часов с балансом:

- качество рекомендаций
- стоимость
- latency (время ответа)

Почему выбрали эту комбинацию моделей

Мы используем двухшаговый пайплайн:

1. Gemini 2.5 Flash/Fast — первичный анализ метрик как "первый эксперт-сомнолог".
2. Mistral — финальная формулировка рекомендаций с учетом контекста из RAG.

Причины:

- Разделение задач: быстрый первичный анализ + отдельная модель на финальный "человеческий" вывод.
- Можно независимо улучшать промпты и/или менять одну из моделей без переписывания всего пайплайна.

Сравнение альтернатив

Альтернативы, которые обычно рассматривают:

- Один "самый сильный" LLM (например GPT-класс): проще архитектурно, но часто дороже и/или медленнее.
- Open-source (Llama/self-host): дешевле на токен, но требует инфраструктуры (GPU/сервер), и качество может проседать на узких задачах.
- Один быстрый LLM без RAG: быстрее, но выше риск "галлюцинаций" и советов без научной базы.

Trade-offs: качество vs стоимость vs latency

- Качество:
  - Усиливается за счет RAG (фактическая база) и строгих ограничений промпта (без диагнозов/лекарств).
- Стоимость:
  - Двухшаговый подход может быть выгоднее, чем постоянно гонять дорогую модель на всю задачу.
- Latency:
  - Два шага обычно медленнее одного, но если первый шаг оптимизирован и/или шаги частично параллелятся, итог остается приемлемым.

Baseline vs optimized версия

Baseline

- Один запрос к одной модели
- Без RAG
- Риск: менее обоснованные советы, хуже привязка к метрикам, больше “общих фраз”

Optimized

- Нормализация метрик
- Первичный анализ Gemini → выделение ключевых тем/проблем
- RAG по статьям/определениям/гайдам
- Финальный промпт + Mistral → итоговая рекомендация
- Post-processing (контроль формата и ограничений)