✅ Что нужно доработать (даже с хорошей моделью)

Направление Что сделать

🔍 Поиск (RAG) Добавить фильтры, chunking по смыслу, reranker

🧠 Контекст Передавать в prompt: цель, ограничения, что игнорировать

⚖️ Отбор результатов Оценивать релевантность с помощью reranker или LLM-фильтра

📦 Метаданные Добавить к каждому чанку: год, отдел, категория

🧠 Агент-обёртка Сделать планировщика, который разбивает запрос на подзадачи

🧩 Память (memory) Чтобы агент помнил, что уже нашёл, куда смотрел, что уже отвечал

🔌 Tools/Functions Возможность делать SQL-запросы, вызывать API, запускать скрипты

📈 Отладка + логгирование Смотреть, что именно было найдено, как LLM использует контекст