Григорий, Как и обещали - направляем Вам пример системы RAG (Retrieval-Augmented Generation) для локальной языковой модели, обученной на данных сайта zerocoder.ru.  
<https://github.com/kpshinnik/rag_zerocoder>  
  
RAG Система с YandexGPT через Ollama  
<https://github.com/kpshinnik/rag_ollama>  
  
Там же подробные инструкции по использованию.  
  
Спасибо за проявленный интерес и будем рады видеть Вас на наших будущих мероприятиях!