Instituto Politécnico de Viana do Castelo

Engenharia Informática

Inteligência Artificial

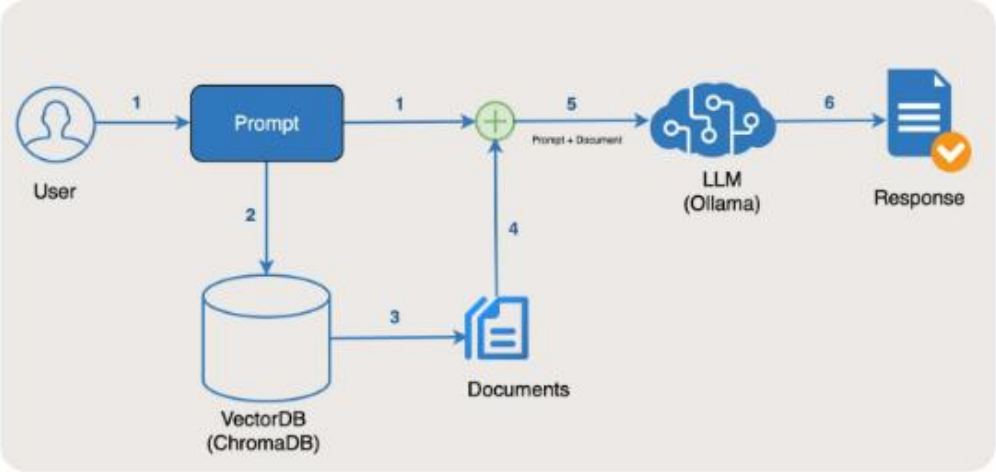
ipvc estg

ENUNCIADO do Trabalho Prático TP 3
Exploring Retrieval Augmented Generation to augment and Large Language Models (LLMs)

Índice

1. Enquadramento

Pretende-se neste trabalho a exploração da Inteligência Artificial Generativa através da implementação dos Large Language Models (LLMs) com Retrieval Augmented Generation para optimizar a exatidão das respostas dos LLMs.



Informação sobre **Large Language Models e RAG Retrieval-Augmented Generation** pode ser consultada, por exemplo nestes links:

- https://aws.amazon.com/pt/what-is/retrieval-augmented-generation/
- https://www.deepchecks.com/glossary/rag-architecture/

- https://medium.com/@samarrana407/learn-pdf-processing-with-googles-gemini-api-cut-costs-with-context-cache-does-gemini-kill-the-6d7fc6588b0c
- https://medium.com/@arunpatidar26/rag-chromadb-ollama-python-guide-for-beginners-30857499d0a0

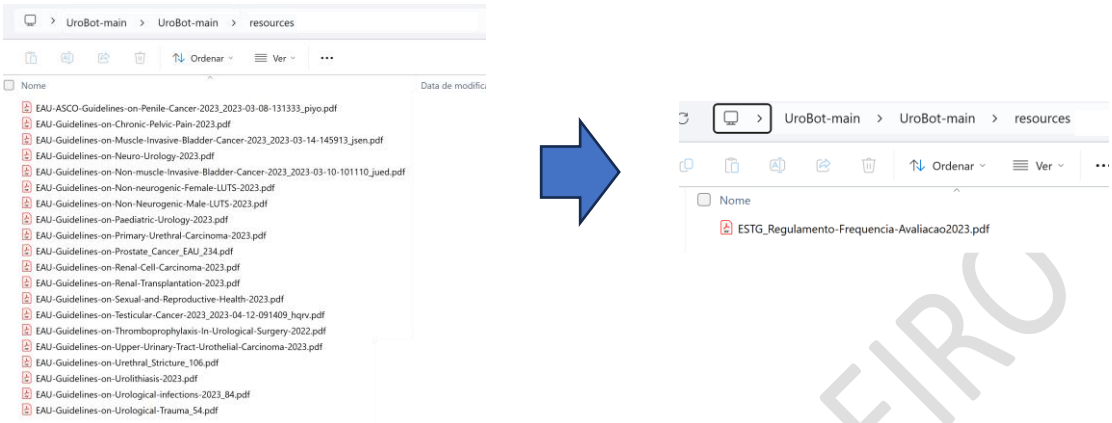
2. Tarefas

Pretende-se que execute as seguintes tarefas:

- Tarefa 1: Instalar localmente a plataforma UROBOT;
- Tarefa 2: Instalar localmente a plataforma LLM Ollama
- Tarefa 3: Adaptar a plataforma UROBOT para, em vez de invocar o ChatGPT, invoque o Ollama.
- Tarefa 4: Execute o Urobot e faça duas questões associadas à urologia para perceber que a informação é obtida dos ficheiros PDF do RAG (na pasta resources) em conjugação com a informação que o LLM (Ollama) contém.
    - Questão 1: what is “Thromboprophylaxis post-surgery” and the Baseline risk of key outcomes;
    - Questão 2: what is “aetiology and Risk groups for stone formation”.
        - NOTA: Pretende-se que a resposta do chatbot seja obtida com a ajuda da informação dos PDFS que estão no RAG uma vez que o LLM (ChatGPT ou OLLAMA) não tem esta informação espeficia.
- Tarefa 5: Como o Urobot usa ficheiros PDF da área da urologia e que estão na pasta resources, o objetivo é agora remover esses ficheiros PDF e colocar o pdf do Regulamento Pedagógico da ESTG disponível em https://www.ipvc.pt/estg/wp-content/uploads/sites/3/2021/02/ESTG_Regulamento-Frequencia-Avaliacao2023.pdf

Desta forma o sistema “Urobot” poderá ser denominado de “ESTG_RegPedagogicoBot” e estar preparado para responder a questões do regulamento pedagógico:

ENUNCIADO do Trabalho Prático TP 3
Exploring Retrieval Augmented Generation to augment and Large Language Models (LLMs)

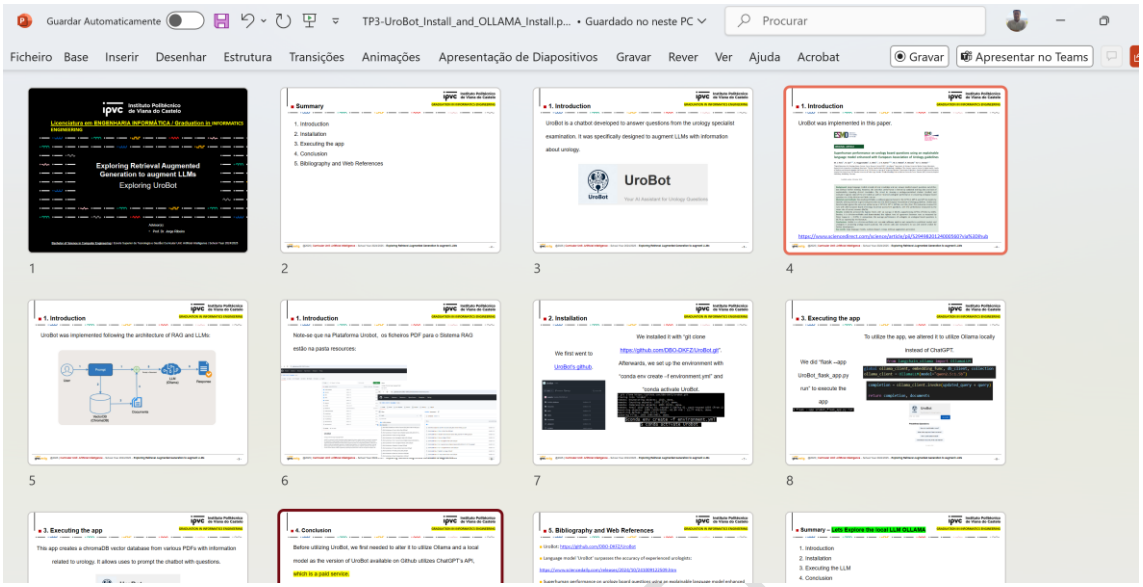


Tarefa 6: Faça as seguintes questões:

- Como posso justificar as faltas?
- O que é a avaliação contínua?

Par ajudar na instalação do Urobot e Ollama, juntamente com este enunciado é disponibilizado um PPT com os passos necessários à sua instalação. No entanto, poderá complementar com outra informação na internet:

ENUNCIADO do Trabalho Prático TP 3
Exploring Retrieval Augmented Generation to augment and Large Language Models (LLMs)



3. Resultados

Deverá em grupo de 2 alunos, ou de forma individual:

- construir um PPT (seguindo a template da UC) com prints/screenshots da instalação do UROBOT e OLLAMA e da sua integração, assim como adaptar o Ollama para que o RAG leia e crie a Base de dados vectorial com a informação do PDF do Regulamento Pedgaógico. Deverá simular as duas questões descritas na tarefa 6.
- Deverá criar um ZIP com o código e o PPT e submeter no link TP3 no moodle da UC