Project: Development of a RAG supported Chatbot including Website integration

Client: Daniel M. Ringel (dmrcon)
Supplier: Saleh Ahmad

Preamble
The client requires a custom-built RAG-based chatbot and its integration into a website hosted on the client's webhosting supplier (t.b.d.). The chatbot is to have an ongoing conversation with users about the client's research, teaching, and profile. Chatbot responses are streamed and consider the previous conversation in the session. The website will be open to anyone, so no sign-in/sign-up user management is required. The chatbot finds the right content across documents; users do not pre-select a category or a document. There are 5 start questions pre-set that users can click on to query the chatbot with that exact question. Chat session can be reset by user (button). The fully implemented chatbot is to be deployed by the supplier for the client. All materials including well documented code and an overview document that outlines the structure and process of chatbot deployment and configuration is to be delivered.

Skills required
- Proficiency in chatbot development
- Can build state-of-the-art RAG systems including preprocessing of documents, vector database, retrieval, reranking, and output generation.
- Experience in deployment of RAG chatbots on Web hosting platforms.
- Ability to integrate chatbot into a website (and build basic website)

Documents
- 8 research articles of approximately 60 pages each (3 as pdf from journals, 5 as either docx or pdf)
- CV (docx or pdf) 4 pages
- Researcher Profile (docx or pdf) 2 pages
- Teaching Profile (docx or pdf) 3 pages
- Public profile derived from my LinkedIn by supplier (docx or pdf) 4 pages
- Personal profile (docx or pdf) 3 pages

Example of a research paper: https://doi.org/10.1177/00222437221110460

Chatbot Functionality
- Chatbot responses are streamed and consider the previous conversation in the session (maintains context across queries).
- There are 5 start questions pre-set that users can click on to query the chatbot with that exact question.
- Chat session can be reset by user (button).
- Chatbot finds the right content across documents for its response.
- Chatbot supplies links to the relevant documents (on the website) when asked about the source of its answer.
- All documents are static for the moment. Documents need to be curated/preprocessed (formatted) to be chucked / inserted into the RAG database (vector database).
- Option to update documents by rebuilding the database with the supplied code (i.e., update documents and import them again).
- Possible to easily modify system prompts to control behavior.
- Possible to easily switch the genAI used to generate responses (e.g., from llama3.1405b to gpt-4o).

Technical Aspects
- Response generation using llama3.1405b or gpt-4o.
- Vector database needs to be fast (e.g., Pinecone Vector DB or PostgresSQL)
- Vector database should be built with a top performing embedding model and approach.
- Suggest using a Bi-cross model for now.
- Must use a good re-ranker.

- Hugging face suggests the following models (but supplier is free to choose better/more suitable models for retrieval and reranking):
  - **Retrieval**: NV-Embed-v2, bge-en-icl, stella_en_1.5B_v5, NV-Retriever-v1
  - **Retrieval with instructions**: FollowIR-7B, mistral-7b-instruct-v0.2, flan-t5-base, monot5-3b-msmarco-10k
  - **Reranking**: gte-Qwen2-7B-instruct, stella_en_1.5B_v5, NV-Embed-v2, SFR-Embedding-Mistral
  - **Instruct/Output generation**: llama-3.1-405B, GPT-4o
- Platform solutions like cohere possible (if better performance)
- Expect up to 50 conversations per day at peak, on average 10.
- Expect up to 20 concurrent users (conversations) at peak, on average 2.
- Low latency with rapid responses for fluent conversations.
- Streamed responses.
- Backend using, e.g., Flask to run python code (or other).
- Coding in python, html, css, javascript

Client Supplies

- Webhosting (e.g., ionos, Heroku, AWS – listed in order of preference).
- Outlined documents (in document section)
  - o *Initial documents* for development will be supplied within 3 days of contract:
    - 3 Research Papers
    - CV
    - Researcher Profile
  - o *Final documents* will be supplied upon basic demo.
  - o Note: Initial documents will need to be updated with final documents.

Deliverables

- Curated (formatted) documents to be chucked / inserted into the RAG database (vector database).
- Fully implemented RAG chatbot (based on supplied documents)
- Deployed on webhosting platform of client.
- Basic website with integrated chatbot.
- Website includes a small dropdown where links to original documents (in PDF format) are available and can be downloaded. Chatbot supplies links to the relevant documents (on the website) when asked about the source of its answer.
- Website contains 5 start questions pre-set that users can click on to query the chatbot with that exact question.
- Website is deployed and fully operation with active chatbot on client's domain and hosting.
- Website fully functional on both desktops/laptops and mobile devices
- Fully documented code and implementation notes.
- Overview of utilized systems, software, packages, and platforms.

Rights, Confidentiality, Intellectual Property

- Supplier transfers all rights to developed code, content, and integration to Client.
- Supplier agrees to treat all development confidentially and will not disclose their work to third-parties at any time.
- Supplier retains no rights on the implemented chatbot and website nor has any rights on the intellectual property (documents) supplied by the client.
- Supplier may not use chatbot, website, or content for their own promotional purposes.
- Supplier may not violate third-party intellectual property rights with developed chatbot.
- Open-source code and models permissible when cited and where terms of use are not violated.

Payment

- Flat rate payment totaling USD400
- Payment in 3 milestones (when milestone reached, demonstrated, and accepted):
  - o ***Milestone 1***: chatbot backend + Files Preprocessing (USD 175)

- ***Milestone 2***: Complete website in html, css, javascript (USD 75)
- ***Milestone 3***: Deployment on web (USD 50)
- ***Milestone 4***: Update vector database with final documents (USD 100)