🏗 Architecture Decision Record

Architecture Decision Record (ADR) 001: Adopting RAG Architecture and LangChain.

1. Decision Record Number

ADR-001

2. Decision Title

Adopting RAG Architecture and LangChain.

3. Status

Accepted

4. Background

Given that we need to answer questions based on any web page provided by the user, we have decided to adopt the RAG (Retrieval-Augmented Generation) architecture. Furthermore, as LangChain is the most comprehensive RAG tool library available, we have decided to use it as our development framework.

5. Decision Details

- Implement the RAG architecture to dynamically retrieve information from various web sources in real-time as the basis for generating responses.
- Utilize LangChain due to its extensive set of tools and features specifically designed for RAG implementations, which support seamless integration with existing systems and provide advanced functionalities for retrieval and generation.

6. Other Options Considered

- Utilizing GPT-4 without external retrieval capabilities.
- Adopting other retrieval frameworks such as Llama Index.

7. Decision Rationale

- The RAG architecture improves the system's ability to provide accurate and contextually relevant answers by leveraging external knowledge sources dynamically.
- LangChain specifically offers robust support for integrating diverse data sources and fine-tuning the retrieval process, which is crucial for enhancing the quality of the generated responses.

8. Impact

- Expected improvements in the accuracy and relevance of responses provided to users.
- Increased complexity in the system architecture, requiring more resources for implementation and maintenance.
- Enhanced adaptability of the system to handle various types of queries by accessing a broader range of information sources.

9. Next Steps

- Begin the integration process of LangChain with our existing systems.
- Develop a prototype to test the effectiveness of the RAG architecture in real-world scenarios.
- Plan for iterative improvements based on feedback and performance metrics from the prototype.

10. References

1. LangChain
2. What is RAG? - Retrieval-Augmented Generation Explained - AWS (amazon.com

---

Architecture Decision Record (ADR) 002: Using GPT-4 as our chatbot.

1. Decision Record Number

ADR-002

2. Decision Title

Using GPT-4 as our chatbot.

3. Status

Accepted

4. Background

Considering that we need to handle multimodal inputs while ensuring the stability and accuracy of the robot's responses, we have decided to adopt GPT-4 as our chat robot.

5. Decision Details

- GPT-4 has the capability to accept image inputs, allowing it to process visual data effectively. This feature is particularly useful for addressing challenges in assisting users with navigation tasks that rely on web-based maps.
- To ensure the completeness and accuracy of the question-answering process, GPT-4 is currently the most worthwhile model to use.
- Additionally, our clients have provided us with the corresponding API key to ensure that we can complete our development.

6. Other Options Considered

- Utilizing GPT-3.5 as chatbot.
- Utilizing open-source model such as Llama-70B.

7. Decision Rationale

- GPT-4 offers superior performance in terms of understanding and generating human-like text due to its advanced training on a diverse range of data and tasks.
- Its ability to process multimodal inputs, such as images, provides a significant advantage in interactive environments where visual context is crucial.

- The provision of an API key by our clients facilitates a secure and tailored implementation that meets specific operational needs.

8. Impact

- Implementation of GPT-4 is expected to enhance the user experience by providing more accurate, reliable, and context-aware responses.
- The system's ability to handle complex and varied queries will be significantly improved, leading to higher customer satisfaction.
- Dependence on advanced AI technology may increase operational costs and necessitate ongoing technical support and updates.

9. Next Steps

- Finalize the contractual agreements for using GPT-4 with the API provider.
- Begin integration of GPT-4 into the existing chatbot infrastructure.
- Conduct extensive testing to ensure the system's performance and reliability across different scenarios and input types.
- Train the support team on managing and troubleshooting the new system.

10. References

- None
