🟦 Sprint Review for Sprint 2

🟫 Introduction

This document encapsulates the review of Sprint 2 for the project aimed at developing a sophisticated Q&A platform. This platform leverages the Furhat robot, augmented by data derived from user given websites and empowered by the capabilities of GPT-4, to facilitate dynamic and insightful user interactions.

🗓️ Sprint 2 Achievements and Completed Requirements

Objective Accomplished

- **Established a robust foundational codebase**: The team successfully laid down the primary architecture for the project, ensuring a scalable and secure infrastructure.
- **Commenced data scraping and preprocessing**: Advanced web scraping techniques were employed to extract relevant textual and multimedia information from targeted websites. Substantial progress was made in cleaning and structuring this data for LLM training.

Completed Requirements

Requirement	User Story	Estimation	Priority	Jira Issue
Domain-Specific Language Model function Must provide accurate information of Melbourne Connect to user when asked.	U3.1: As a user interested in services offered at Melbourne Connect, I want the robot to provide detailed information (such as room's information, location and provided services) from websites of Melbourne Connect, so that I can clearly understand the overview of Melbourne Connect without navigating through those websites.	LARGE	MUST HAVE	☑ C2QK-3: Must provide accurate information of Melbourne Connect to user when asked. 已完成
Domain-Specific Language Model function	U3.2: As a user, I want the chatbot to automatically extract and summarize the	LARGE	SHOULD ...	☑ C2QK-8: Must be able to summary the information of the webs

Must be able to summary the information of the website clearly and quickly.	main content of a website I provide, so that I can quickly grasp what the website is about without reading all the content.			ite clearly and quickly. 已完成
Domain-Specific Language Model function Must be able to recommend information based on user's interests and website summary.	U3.3: As a user looking for entertainment or news, I want the chatbot to recommend and summarize articles, blog posts, stories, or provide me with the latest headlines based on my interests, so that I can easily find engaging or informative content.	LARGE	MUST HAVE	☑ C2QK-9: Must be able to recommend information based on user's interests and website summary. 已完成
Domain-Specific Language Model function Must be able to extract information from various careers websites and provide a clear summary job list.	U3.4: As a job seeker, I want the chatbot to compile job listings from various websites, so that I can find opportunities that match my skills and preferences more easily.	LARGE	MUST HAVE	☑ C2QK-16: Must be able to extract information from various careers websites and provide a clear summary job list. 已完成
Domain-Specific Language Model function Must be able to extract academic information from scientific websites.	U3.5: As a user doing research, I want to ask the chatbot specific questions about content found on academic or scientific websites, so that I can gather information efficiently for my studies or work.	LARGE	MUST HAVE	☑ C2QK-11: Must be able to extract academic information from scientific websites. 已完成
Furhat Robot interaction Must be able to avoid answering unrelated questions with the website.	U2.5: As a user, I want the robot to avoid answering questions that are unrelated to website navigation content or involve sensitive information, and to inform me when my queries are outside the scope of available information, to maintain professionalism and set clear expectations.	HIGH	MUST HAVE	☑ C2QK-13: Must be able to avoid answering unrelated questions with the website. 已完成

Requirements Development Progress

- **Data Extraction Metthods**:
  - **Progress**: Completed

    - **Details**: WebScraper methods are implemented to be able to extract information from the user given web url, hence to be used in the RAG pipeline for the LLM.
    - **Challenges**: Had inital troubles extracting information on the website that is displayed using javescript, tackled with advanced extraction functions.
- **Domain-Specific Language Model Functions**:
    - **Progress**: Partial completion.
    - **Details**: Efforts were focused on developing functions that provide accurate information and summarizing web content.
    - **Challenges**: Encountered complexity in preprocessing and structuring data for GPT-4 compatibility.
- **Furhat Robot Interaction**:
    - **Progress**: Initial development phase.
    - **Details**: Work commenced on enabling the robot to filter and sort information based on user criteria.
    - **Challenges**: Integrating user-defined criteria filtering with real-time interaction posed significant technical hurdles.
- **UI Interactive Interface**:
    - **Progress**: Conceptual phase.
    - **Details**: Privacy policies and user data protection mechanisms were outlined.
    - **Challenges**: Balancing user-friendliness with security requirements requires further brainstorming.

Data Collection and Preparation

- Achieved significant milestones in web scraping and data cleaning methodologies. The structured data is now in a preliminary stage, ready for LLM training.

Infrastructure Deployment

- Tested on suitable websites for data extraction and set up development environments successfully.

Technology Utilized

- Utilized Python (langchain, BeautifulSoup, Scrapy) for web scraping and Python (Pandas, NumPy) for data cleaning and preprocessing. The RAG framework was selected for implementation in upcoming Sprints.

📈 Sprint 2 Insights

- **Strengths**: The team demonstrated strong technical skills in data scraping and preprocessing. The foundational codebase and infrastructure setup were completed efficiently.
- **Areas for Improvement**: The integration of complex user-defined criteria in Furhat Robot interactions and balancing user interface design with privacy concerns need further exploration.
- **Lessons Learned**: The importance of a flexible approach in dealing with unexpected technical challenges was highlighted. Early testing of components facilitated the identification of integration challenges.

💡Recommendations for Future Sprints

- **Prioritize Integration Challenges**: Focus on resolving integration issues between the Furhat Robot and the language model functions.
- **Enhance Data Quality for LLM Training**: Increase efforts in data cleaning and preparation to ensure high-quality inputs for GPT-4 training.
- **User Privacy and Security**: Expedite the development of user interface elements that address privacy concerns without compromising user experience.

⏳Next Steps

- Begin detailed work on completing the Domain-Specific Language Model functions with an emphasis on integrating user feedback for refinement.

- Address the technical challenges identified in Furhat Robot interaction, focusing on real-time data processing.
- Finalize UI designs that incorporate privacy and security features, ensuring user trust and compliance with data protection regulations.