

**UNIVERSITY OF GREENWICH**  
COMP1682 – Final Year Projects

Proposal

|  |  |
| --- | --- |
| Student name | Lieu Vinh Khoi |
| ID number (00xxxxxxx) | **001391472** |
| Lecturer/Tutor name | Tran Thi Kim Khanh |
| Student submission date | 30 April 2024 |

**Table of Contents**

[**1.** **Overview** 5](#_Toc164200917)

[**2.** **Aim** 5](#_Toc164200918)

[**3.** **Related works / Literature review** 6](#_Toc164200919)

[**3.1** **Evolution and Impact of NLP in Enhancing Educational Technology** 6](#_Toc164200920)

[**3.2** **Solving the Challenge of Information Large in Education through NLP Technology** 6](#_Toc164200921)

[**3.3** **Recent Advances in Text Summarization Techniques and Technologies** 7](#_Toc164200922)

[**3.4** **Advancements in Question Generation from Large Text Documents** 7](#_Toc164200923)

[**4.** **Objectives** 8](#_Toc164200924)

[**4.1** **Research and analysis of requirements:** 8](#_Toc164200925)

[**4.2** **Complete design of necessary components in the system:** 8](#_Toc164200926)

[**4.3** **Complete development and implementation of project requirements:** 9](#_Toc164200927)

[**4.4** **Test and ensure the functions in the system operate effectively:** 9](#_Toc164200928)

[**4.5** **Project framework or methodology used:** 9](#_Toc164200929)

[**5.** **Legal, Social, Ethical and Professional** 9](#_Toc164200930)

[**5.1** **Legal:** 10](#_Toc164200931)

[**5.2** **Social:** 10](#_Toc164200932)

[**5.3** **Ethical:** 10](#_Toc164200933)

[**5.4** **Professional:** 10](#_Toc164200934)

[**6.** **Planning (see Appendix A)** 11](#_Toc164200935)

[**7.** **Initial References** 13](#_Toc164200936)

[**Appendix A - Schedule of Work** 15](#_Toc164200937)

**Table of Figures**

[Figure 1. Gantt Chart of First Phase 15](#_Toc164799463)

[Figure 2. Gantt Chart of Second Phase 15](#_Toc164799464)

[Figure 3. Gantt Chart of Third Phase 15](#_Toc164799465)

[Figure 4. Gantt Chart of Fourth Phase 16](#_Toc164799466)

**List of Tables**

[Table 1. Project Development Plan 11](#_Toc164799591)

**COMP1682 Project Proposal**

**The investigation builds a Python software to generate questions using NLP.**

**Lieu Vinh Khoi**

**BSc (Hons) Computing**

**001391472**

# **Overview**

The project solves the problem of processing large amounts of text data in the field of education. A statement from (Vaswani et al., 2017) that "*With the rapid development of digital technology, traditional methods are no longer effective enough, requiring solutions based on artificial intelligence, especially in the field of Natural Language Processing (NLP).*" helps open up a new approach to NLP to significantly improve the ability to process text data. The project includes building a learning website, especially summarizing text documents and creating test questions for high school students and teachers. Evaluating the solution's effectiveness will be done through metrics such as the accuracy of contextual summaries and the quality of questions generated in the test. The project expects to successfully apply advanced NLP technologies and techniques and significantly improve the learning efficiency of high school students.

**Keywords:** Natural language processing (NLP), text summarization, generate questions.

# **Aim**

The project aims to develop a web system using NLP technology to automatically summarize and create questions from documents in social subjects for high school students and teachers, to partly improve the efficiency of learning and knowledge for students. The project cannot process text across all subjects, has limitations on the input document size, and struggles to maintain stable user access during high-traffic periods, all while lacking complete security measures.

# **Related works / Literature review**

Through the main goal of the application, text summarization from input documents is an important requirement and brings advanced challenges in the application development and research process. In 1979, ANSI gave a brief definition of text summary: "*A brief restatement within a document (usually at the end) of its salient findings and conclusions intended to (Johri et al., 2021) complete the orientation of a reader who has studied the preceding text.*" In this literature review, some issues regarding the ability to create questions from documents are also mentioned and require the necessary research process from previous products.

## **Evolution and Impact of NLP in Enhancing Educational Technology**

A research paper from (Johri et al., 2021) traces the evolution of NLP from basic translation to complex real-time processing, highlighting its potential to transform language learning and understanding, consistent with the project's goal of improving educational effectiveness through text summarization and question generation. A related study from (Khurana et al., 2023) discusses versatile applications of NLP, including machine translation and spam detection, opening up the possibility of achieving linguistic diversity in the system. An article introduces NLP methods and their applications in education, emphasizing the role of extensive text datasets and deep learning (Khensous et al., 2023), an issue related to the project, that helps effectively process text data in educational contexts. Furthermore, the results of research by Khurana et al on the development of NLP provide insight into the complexities of applying NLP in an educational context, helping the project avoid pitfalls and exploiting NLP's potential to meet educational goals. These studies show the impact of growing NLP on education, from its real-time processing capabilities, diverse applications, and potential for text processing in learning, factors that confirm the project can leverage NLP to significantly improve efficiency in the system development process, improving the learning outcomes of high school students using advanced NLP techniques.

## **Solving the Challenge of Information Large in Education through NLP Technology**

In the modern educational landscape, information overload is becoming a major challenge. In research by (Kaouni et al., 2023), platforms such as Moodle and Blackboard often do not provide enough flexibility and personalization for users, reducing learning effectiveness. Another study by (Shaik et al., 2022) points to the challenge of handling written feedback from students, an important source of information for improving content and teaching methods. An article highlights information overload in higher education institutions in Estonia (Lauri et al., 2021), suggesting the need to develop information processing capabilities to support processing and managing information effectively. Natural language processing (NLP) has been proposed as a powerful solution to address these challenges. Research by (Kaouni et al., 2023) and (Shaik et al., 2022) both focused on applying NLP to analyze text response data from students, showing the ability to extract useful information to improve the teaching and learning process. NLP methods such as emotion recognition, entity recognition, text summarization, and topic modeling were used to provide deeper insight into students' opinions and thoughts. The application of NLP in education not only helps solve the problem of information overload but also improves the process of analyzing and understanding student feedback data. These two studies also show that automatically processing text data using NLP helps students save time from document feedback, thereby providing a more efficient and quick. (Lauri et al., 2021) recommend that NLP can play an important role in improving information access, data processing, and natural language interaction, enhancing learning efficiency and management information management for students and teachers. In summary, the challenge of data processing in education is also highlighted and research on NLP also proves the necessity of the project in education, helping to open up new opportunities to deal with the challenge of information overload and improve the learning process.

## **Recent Advances in Text Summarization Techniques and Technologies**

The research contributions from (Khan et al., 2023), (Aswani et al., 2024), and (Shingade et al., n.d.) on Text Summarization offer profound insights that directly influence the project's development, particularly in enhancing the quality and accessibility of educational content. (Khan et al., 2023) focus on the challenges of achieving accurate condensation, reducing bias, and establishing effective evaluation metrics that align to provide precise and unbiased educational summaries. (Aswani et al., 2024) transformer-based models in processing scientific texts help us adopt these advanced NLP techniques to make complex information more digestible for students. Furthermore, the exploration of cloud-based NLP solutions presents an opportunity to leverage cloud technology for efficient large-scale data processing (Shingade et al., n.d.), crucial for the scalability of educational platforms. These collective insights not only shape the technical framework of the project but also highlight the potential of NLP technologies to transform educational experiences by delivering customized, accessible summaries, thereby laying a foundation for future advancements in educational technology.

## **Advancements in Question Generation from Large Text Documents**

In the field of educational technology, multiple choice question (MCQ) creation has emerged as a key area of innovation, with important implications for enhancing teaching and assessment methods. Articles by (Madri and Meruva, 2023), (Nwafor and Onyenwe, 2021), and (Bitew et al., 2022) similarly emphasize the key role and challenges of automated systems in MCQ generation, a central theme of the project objectives. (Madri and Meruva, 2023) emphasize the importance of precision in automated MCQ generation, addressing the need for sophisticated techniques that can be seamlessly integrated into educational environments, reflecting the ambitions of the project to develop an effective tool to create high-quality educational content. (Nwafor and Onyenwe, 2021) expands on this discussion by validating an NLP-based system for Computer-Based Tests (CBTE), consistent with the project's focus on leveraging NLP to generate questions insightful and contextually adaptable to different educational materials. (Bitew et al., 2022) address the practical challenges involved, such as digital literacy and the complexity of creating distractions, which are critical to MCQ effectiveness. Their suggestion to introduce existing distractors to improve question-generation efficiency offers a potential strategy for the project, highlighting the importance of resourcefulness in enhancing the question-generation process. By combining these insights, the project will benefit from a deep understanding of the current landscape and challenges in automated MCQ generation, helping to develop a more efficient, user-friendly system.

# **Objectives**

## **Research and analysis of requirements:**

1. Study how to create questions and summarize text to understand how they work. [10.0]
2. Research documents on NLP and issues in the field of education to determine the opportunities and feasibility of the project. [10.0]
3. Determine project requirements through surveys and feedback that includes teachers, high school students, and instructors. [5.0]

**Outcome:** Selected technologies and models for adoption, clearly justifying their relevance and effectiveness for the project's objectives.

## **Complete design of necessary components in the system:**

1. Design a project architecture model to define data processing flows in text summarization and question generation. [3.0]
2. Develop use case diagrams to describe user tasks and actions in the system. [4.0]
3. Design the user interface to clearly understand how users will interact with the system. [4.0]
4. Design the database structure to ensure that data is organized logically and efficiently. [2.0]

**Outcome:** Produced comprehensive technical designs and diagrams for all system components, facilitating a clear understanding of system architecture and user interaction.

## **Complete development and implementation of project requirements:**

1. Implement project functions using an approach involving NLP technology to summarize text and generate questions. [15.0]
2. Developed a user-friendly interface for the web application, supporting teachers in uploading documents and displaying summaries and test questions. [12.0]
3. Integrate NLP models with web applications, ensuring accurate processing and output generation. [8.0]

**Outcome:** Developed a web application with integrated NLP technology, offering intuitive interfaces for text summarization and question generation.

## **Test and ensure the functions in the system operate effectively:**

1. Conduct testing of each function to ensure it operates as expected. [5.0]
2. Perform site security and error logic testing to ensure a good user experience. [4.0]
3. Conduct user usability testing surveys to assess ease of use and accessibility for users. [6.0]

**Outcome:** Completed thorough system testing, confirming stability, security, and user-friendliness, ensuring the system meets all operational and user needs.

## **Project framework or methodology used:**

Agile methods are used to ensure rapid adaptation to changes through iterative development iterations. Scrum's structured approach to managing complex tasks in sprints. The project uses advanced Natural Language Processing (NLP) technologies and architectural models because of their effectiveness in processing language, which is essential for summarizing text and generating questions in educational content. NLP-related programming languages and frameworks were chosen for the project to build a learning website because their synchronization would help the project with easier development, scalability, and community support copper is also inherited. This choice ensures a cohesive and efficient web platform, enhancing functionality and user experience in the education sector.

# **Legal, Social, Ethical and Professional**

The project aims to create a web system using Natural Language Processing (NLP) technology to automatically summarize and generate questions from social sector documents for high school students and teachers. This will enhance learning efficiency and knowledge, while also making the learning process more flexible and effective. However, the project needs to consider legal, social, ethical, and professional issues.

## **Legal:**

**- Personal Data Protection:** All data processed within this project complies strictly with GDPR (European Commission, 2018) and other data protection standards, ensuring transparency and security for users.

**- Intellectual Property Rights:** The texts utilized for research and educational purposes adhere to copyright and licensing regulations before being uploaded to the website. We commit to respecting intellectual property rights by verifying the origin and licenses of all utilized materials.

## **Social:**

**- Accessibility:** Developing interfaces and features that are easily accessible to all students and teachers.

**- Social Diversity:** Designing the project with consideration for cultural and social diversity, ensuring fairness and inclusivity.

**- Global Impact:** Strengthening text data processing in education globally, aiming to significantly improve the learning efficiency of high school students and teachers worldwide.

## **Ethical:**

**- User Privacy and Consent:** Implementing measures to verify and regulate generated content, ensuring transparency and safety for users.

**- Negative Content Removal:** Controlling and removing harmful content to protect user experience.

## **Professional:**

**- Testing and quality assurance:** Check and ensure the quality of summarized texts and created questions.

**- Professionalism:** Ensure professional principles during the development process. Research and apply advanced models and algorithms during the development process. Use trained NLP models in text summarization and question generation.

# **Planning (see Appendix A)**

The project under review is a complex project that uses NLP modeling and technology to summarize large-text teaching material in social studies lessons. Therefore, projects need to be approached with appropriate methods and require iterative cycles. For such a challenging project, this is still just the early stages and further developments could change requirements at any time. The project plan will always be monitored to ensure that the requirements will be fully implemented in the final product. Applying Scrum and Agile methods, the project will be monitored weekly and see how much progress has been achieved, so that if adjustments are needed, they will be adjusted most appropriately. The Gantt Chart (Appendix A) will display the work implementation plan of each phase that needs to be completed to ensure on schedule and meets the project process. The project includes 4 main phases with goals:

**- Research and select the appropriate technology for project requirements:** This phase focuses on understanding and selecting appropriate technology for the project, including NLP tools and methods.

**- Complete design of necessary components in the system:** Comprehensive design of necessary components of the system, including architectural model, usage diagram, user interface, and database structure.

**- Complete development and implementation of project requirements:** Develop and implement project requirements using NLP technology for text summarization and question generation, user interface development, and NLP model integration.

**- Test and ensure the functions in the system operate effectively:** Test each system function for functionality, security, and logic errors, and conduct a usage survey to assess ease of use and access.

Table 1. Project Development Plan

|  |  |  |  |
| --- | --- | --- | --- |
| **Task Name** | **Duration** | **Start** | **Finish** |
| **The investigation builds a Python software to generate questions using NLP.** | **92 days** | **Tue 1/16/24** | **Wed 5/22/24** |
| **Research and select the appropriate technology for project requirements** | **25 days** | **Tue 1/16/24** | **Mon 2/19/24** |
| Study how to create questions and summarize text to understand how they work. | 10 days | Tue 1/16/24 | Mon 1/29/24 |
| Research documents on NLP and issues in the field of education to determine the opportunities and feasibility of the project | 10 days | Tue 1/30/24 | Mon 2/12/24 |
| Determine project requirements through surveys and feedback that includes teachers, high school students, and instructors | 5 days | Tue 2/13/24 | Mon 2/19/24 |
| Complete the technology options, adoption models, and reasons for choosing them. | 0 days | Mon 2/19/24 | Mon 2/19/24 |
| **Complete design of necessary components in the system** | **13 days** | **Mon 2/26/24** | **Wed 3/13/24** |
| Design a project architecture model to define data processing flows in text summarization and question generation. | 3 days | Mon 2/26/24 | Wed 2/28/24 |
| Develop use case diagrams to describe user tasks and actions in the system. | 4 days | Thu 2/29/24 | Tue 3/5/24 |
| Design the user interface to clearly understand how users will interact with the system. | 4 days | Wed 3/6/24 | Mon 3/11/24 |
| Design the database structure to ensure that data is organized logically and efficiently. | 2 days | Tue 3/12/24 | Wed 3/13/24 |
| Complete technical drawings and detailed designs for every system component. | 0 days | Wed 3/13/24 | Wed 3/13/24 |
| **Complete development and implementation of project requirements** | **35 days** | **Thu 3/14/24** | **Wed 5/1/24** |
| Implement project functions using an approach involving NLP technology to summarize text and generate questions. | 15 days | Thu 3/14/24 | Wed 4/3/24 |
| Developed a user-friendly interface for the web application, supporting teachers in uploading documents and displaying summaries and test questions. | 12 days | Thu 4/4/24 | Fri 4/19/24 |
| Integrate NLP models with web applications, ensuring accurate processing and output generation. | 8 days | Mon 4/22/24 | Wed 5/1/24 |
| Complete website system, including friendly user interface and effective NLP model integration. | 0 days | Wed 5/1/24 | Wed 5/1/24 |
| **Test and ensure the functions in the system operate effectively** | **15 days** | **Thu 5/2/24** | **Wed 5/22/24** |
| Conduct testing of each function to ensure it operates as expected. | 5 days | Thu 5/2/24 | Wed 5/8/24 |
| Perform site security and error logic testing to ensure a good user experience. | 4 days | Thu 5/9/24 | Tue 5/14/24 |
| Conduct user usability testing surveys to assess ease of use and accessibility for users. | 6 days | Wed 5/15/24 | Wed 5/22/24 |
| Complete detailed testing, ensuring the system is stable, secure, and easy to use for end users. | 0 days | Wed 5/22/24 | Wed 5/22/24 |

# **Initial References**

*Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is All You Need. In Advances in Neural Information Processing Systems (pp. 5998-6008).*

*European Commission. (2018). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). EUR-Lex - Access to European Union Law.* [*https://eur-lex.europa.eu/eli/reg/2016/679/oj*](https://eur-lex.europa.eu/eli/reg/2016/679/oj)

*ANSI, American National Standard for Writing Abstracts Z39-14, ANSI, New York, 1979.*

*Aswani, S., Choudhary, K., Shetty, S., Nur, N., 2024. Automatic text summarization of scientific articles using transformers—A brief review. J. Auton. Intell. 7.*

*Bitew, S.K., Hadifar, A., Sterckx, L., Deleu, J., Develder, C., Demeester, T., 2022. Learning to reuse distractors to support multiple choice question generation in education. IEEE Trans. Learn. Technol.*

*Johri, P., Khatri, S.K., Al-Taani, A.T., Sabharwal, M., Suvanov, S., Kumar, A., 2021. Natural language processing: History, evolution, application, and future work. Presented at the Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020, Springer, pp. 365–375.*

*Kaouni, M., Lakrami, F., Labouidya, O., 2023. Integrating Artificial Intelligence and Natural Language Processing in E-Learning Platforms: A Review of Opportunities and Limitations. Presented at the 2023 7th IEEE Congress on Information Science and Technology (CiSt), IEEE, pp. 56–63.*

*Khan, B., Shah, Z.A., Usman, M., Khan, I., Niazi, B., 2023. Exploring the landscape of automatic text summarization: a comprehensive survey. IEEE Access.*

*Khensous, G., Labed, K., Labed, Z., 2023. Exploring the evolution and applications of natural language processing in education. Rev. Rom. Informatică Și Autom. 33, 61–74.*

*Khurana, D., Koli, A., Khatter, K., Singh, S., 2023. Natural language processing: State of the art, current trends and challenges. Multimed. Tools Appl. 82, 3713–3744.*

*Lauri, L., Virkus, S., Heidmets, M., 2021. Information cultures and strategies for coping with information overload: Case of Estonian higher education institutions. J. Doc. 77, 518–541.*

*Madri, V.R., Meruva, S., 2023. A comprehensive review on MCQ generation from text. Multimed. Tools Appl. 82, 39415–39434.*

*Nielsen, J., 1994. Usability engineering. Morgan Kaufmann.*

*Nwafor, C.A., Onyenwe, I.E., 2021. An automated multiple-choice question generation using natural language processing techniques. ArXiv Prepr. ArXiv210314757.*

*Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., Galligan, L., 2022. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. IEEE Access 10, 56720–56739.*

*Shingade, B., Matha, Y., Kolambkar, V., Kasar, S., Palve, R., n.d. A Survey of Text Summarization Using NLP.*

# **Appendix A - Schedule of Work**

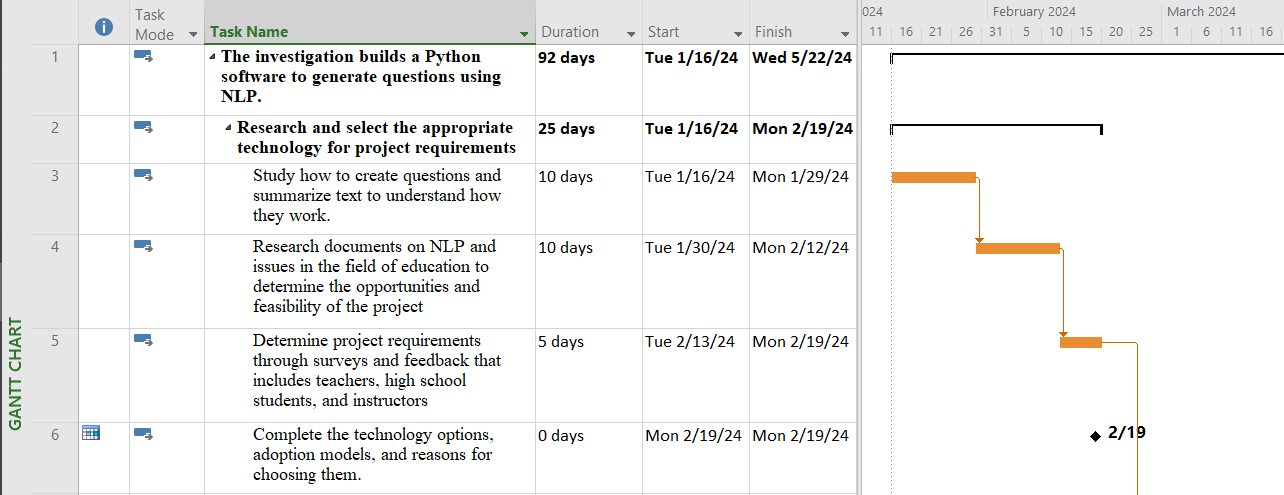
**

Figure 1. Gantt Chart of First Phase

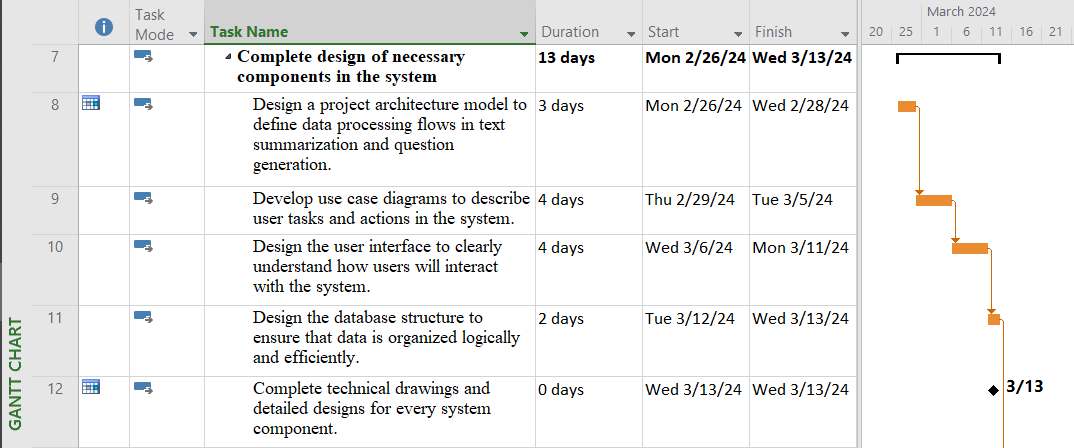


Figure 2. Gantt Chart of Second Phase

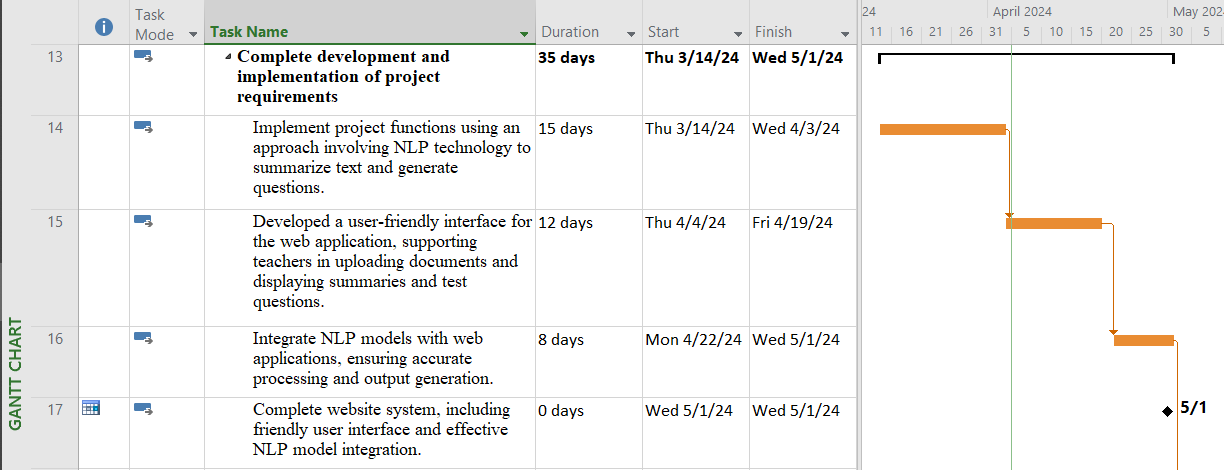


Figure 3. Gantt Chart of Third Phase

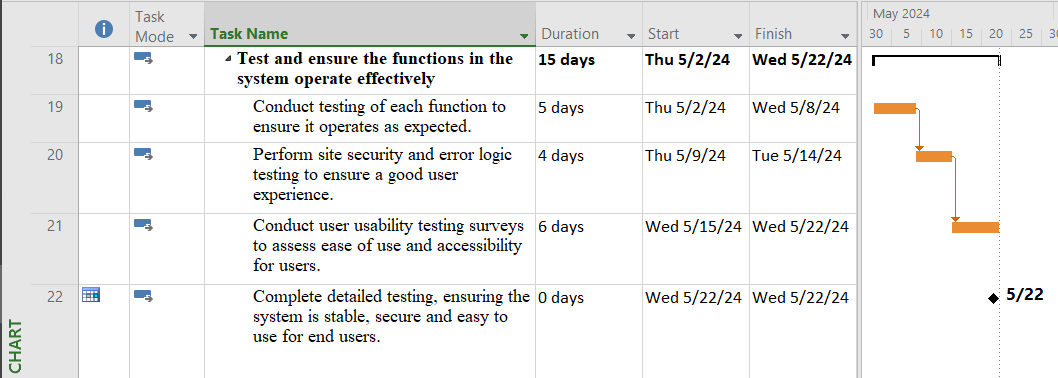


Figure 4. Gantt Chart of Fourth Phase