FIT3161: Advanced Computer Science Project 1
Project Proposal including Literature Review
Project Supervisor: Dr. Soon Lay Ki

Group: MCS15
Ooi Yi Sen – 30720699
Chan Wai Han - 31555373
Nawwaf Ali - 31261949
Yeonsoo Kim - 29584612

Word Count: 10191 words

1.0 Introduction

Since the outbreak of COVID-19, scientists have carried out experiments and research on COVID-19. Vaccines, antibodies, oral antiviral medicine, and cell-based treatments were among the focus areas of their research. The outcome of their research has been sought by the public as well as the governments to control the outbreak. In addition, the general public has not shied away from contributing their thoughts and experiences too. These have resulted in a large burden for the readers seeking to maintain up-to-date knowledge on COVID-19. In fact, some of the information provided is understandable by the general public. This project is hence motivated to design and implement a Question-Answering system that provides a list of recommended and ranked answers to the questions related to COVID-19.

What is covid-19?

A virus is an entity that infects living organisms. It requires a host to survive and reproduce (Ryan, 2022). Coronavirus (Covid-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARs-CoV-2) (Kumar, 2021) This virus originated in Wuhan city of China and then spread all around the world (Kumar, 2021). Even though the virus emerged in an Asian country, China, It spread rapidly to other countries of the world. A total of 195 countries and territories exaggerated with the virus infection (Kumar 2021).The Mortality rate of the virus was calculated by taking into account multiple factors such as age, sex, and their overall health.

Currently there are many ways to diagnose COVID-10 infections such as physical symptoms, specific laboratory tests, imaging techniques, etc. (Kumar, 2021). Physical symptoms may vary in different individuals, but fever is the most common symptom (Kumar, 2021). Molecular Tests such as Antigen, according to WHO, the specimens for this test should be collected from the nasal oropharyngeal routes and the broncho alveolar lavage and expectorated sputum (Kumar, 2021). There are no guaranteed methods to prevent coronavirus currently, every measure or method is preventative.

2.0 Literature review

Q&A system

2.1 Introduction

There are so many kinds of Question Answering Systems Nowadays. In this literature review, we will focus on how the current automatic question answering systems are working so that our team can get the basic ideas on how to construct our own question answer system for the project. As all the question answering systems have their own data processing and answering methods that suit its purpose, we will only focus on the data processing and answering methods that are closely relevant to our project.

Due to the fact that many scientific terms and technical terms are used in the articles, we will simplify the terms and the detailed methods/algorithms so that we can easily understand and identify how the programmes work.

2.2 Question Answering System Approaches

To understand what approaches we should use for our automated Q&A system, we decided to have a look for the Q&A system approaches. According to the article by Dwivedi, S. K., & Singh, V. (2013), there are 3 main approaches in automated- Q&A systems.

Discriminative Model Approach

This approach would be the way this article 'Linguistic kernels for answering re-ranking in question answering systems.' by MOSCHITTI (Moschitti, 2021) writes about moving away from typical approaches that use unsupervised methods that involve computing the similarity between query and answer in terms of leical, syntactic,semantic or logic representation (Moschitti, 2021). Instead they studied supervised discriminative models that learn to select answers from examples of question and answer pairs, where the representation of the pair is implicitly provided by kernel combinations applied to each of its components (Moschitti, 2021).

They found evidence to support the exploitation of advanced linguistic information by using powerful discriminative models such as SVMs and effective feature engineering techniques such as kernel methods in challenging natural language tasks.

A drawback to this approach if we follow Moschitti's article is that we would require a question to have multiple ordered candidates answers assigned to a question as training instances.We can take this article into account should we find it a more feasible approach then the others.

Linguistic Approach

Linguistic Approach is a question answering system that requires understanding of natural language text, linguistics and common knowledge (Dwivedi, S. K., & Singh, V.,2013). So, this Q&A system is often based on Natural Language Processing (NLP) logic and knowledge base due to its characteristics. Linguistic techniques are used in this approach, such as, tokenization, POS tagging and parsing, to pre-process the user's query and get the related answers from the database.

An advantage of the linguistic approach is that it can provide a situation-specific answer to the user. However, this approach has many limitations as it is very time-consuming to build an appropriate and sufficient amount of knowledge bases. Moreover, this approach cannot deal with the domain that is out-of-bound of the system if the knowledge is not stored in the structured database. So, most of the limitations are due to its NLP-based logic.

Statistical Approach

Statistical Approach refers to the use of statistics to learn from examples. It means to collect observations, study and digest them in order to infer general rules or concepts that can be applied to new, unseen observations (*Statistical pattern recognition*., 2019) and this approach is independent of structured query languages and can formulate queries in natural language form. (Dwivedi, S. K., & Singh, V., 2013)

The biggest advantages of this approach is that it can produce the best results between other approaches once it has a sufficient amount of data to train the Q&A system and this approach can deal with the unseen questions based on the statistics learned in the system.

However, the disadvantage of this approach is that it treats each term independently and fails to identify linguistic features for combination of words or phrases.
(Dwivedi, S. K., & Singh, V., 2013)

Pattern Matching Approach

This approach uses the expressive power of text patterns to replace the sophisticated processing involved in other competing approaches. (Dwivedi, S. K., & Singh, V., 2013). To simplify, Pattern Matching Approach uses pre-defined patterns in the back-end system to find the matching patterns from the user's questions.

Advantage of this approach is that it is simple to make a pattern matching based Q&A system compared to the other approaches as it requires relatively short time to make short-medium sized answering systems and as it does not require complex systems to build or maintain.

There are two types of the pattern matching approaches, which are surface pattern based and template based approaches. Most of the patterns matching QA systems use the surface text patterns while some of them also rely on templates for response generation.
(Dwivedi, S. K., & Singh, V., 2013).

Surface Pattern based approach extracts answers from the surface structure of the retrieved documents by relying on an extensive list of patterns. Answers to a question are identified on the basis of similarity between their reflecting patterns having certain semantics.

A Template based approach makes use of preformatted patterns for questions. The focus of this approach is more on illustration rather than interpretation of questions and answers.
(Dwivedi, S. K., & Singh, V., 2013).

There are many classification methods we can use.In the Article 'Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer(BERT)- Based Automatic Classification of Tweets About Eating Disorder: Algorithm Development and Validation Study' (Benítez-Andrades, 2022), writes about methods such as random forest, recurrent neural networks, bidirectional long short-term memory networks and pretrained bidirectional encoder representations.
Bidirectional long short-term memory and bidirectional encoder representations from transformers seem to be the most promising model for natural language processing (Benítez-Andrades, 2022).

Classification Method	Description
Random Forest	● Random forest models are constructed from a set of decision trees, which are usually trained with a method called bagging, to take advantage of the independence between simple algorithms,since error can be greatly reduced by averaging outputs of the simple models. (Benítez-Andrades, 2022). ● Several decision trees are built and fused in order to obtain a more stable and accurate prediction. (Benítez-Andrades, 2022)
Recurrent Neural Network(RNN)	● Type of neural network where a temporal sequence that contains a directed graph made up of connections between different nodes is defined. (Benítez-Andrades, 2022) ● These networks have the capacity to show a dynamic temporal. (Benítez-Andrades, 2022) ● Derived from feedforward neural networks,which have the ability to use memory to process input sequences of varying length. (Benítez-Andrades, 2022)
Bidirectional Long Short-Term Memory	● Bidirectional long short-term memory networks are constructed from 2 long short-term memory modules that, at each time step, take past and future states into account to produce the input. (Benítez-Andrades, 2022)
Bidirectional Encoder Representation from Transformer-Based Models(BERT)	● Bidirectional encoder representation is not a model itself. It can be considered a 'language understanding' model. (Benítez-Andrades, 2022). ● A neural network is trained to learn a language, similar to

	transfer learning in computer vision neural networks, and follows linguistic representation in a bidirectional way, looking at the words both after and before each words. (Benítez-Andrades, 2022)

There are similarities in the sense of ranking the category of tweets mention in the Article 'Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)-Based Automatic Classification of Tweets About Eating Disorder: Algorithm Development and Validation Study' (Benítez-Andrades, 2022) in the sense that we would need properly to label what we would consider a high level answer and what is a low level answer in the same way they label the categories of a tweet.

2.3 Q&A System Structure

Based on the article of Pundge, A. M., Khillare, S. A., & Mahender, C. N. (2016), The Q&A system is made up of four modules, which are Question Processing module, Document Processing Module, Paragraph Extraction Module, and Answer Extraction module.

Question Processing Module

The question processing module pre-process the user's questions into system-friendly queries (Natural Language queries) so that the question can be analysed and the answer type can be determined by the system.

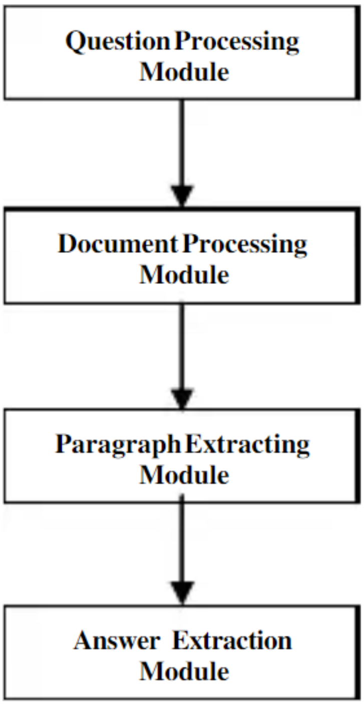

Question Processing Module
Document Processing Module
Paragraph Extracting Module
Answer Extraction Module

Figure 2.1 : Basic Structure of Q&A System
(Pundge, A. M., Khillare, S. A., & Mahender, C. N., 2016)

Document Processing Module

The document processing module is an Information retrieval module that focuses on gathering relevant documents. (Lalithnarayan, C., 2020). It consists of a query generation algorithm and text search engine. The query generation algorithm takes an input the user's question and creates a query containing terms likely to appear in documents containing an answer. This query is passed to the text search engine, which uses it to retrieve a set of documents. (Pundge, A. M., Khillare, S. A., & Mahender, C. N., 2016).

Paragraph Extraction Module

In this module, the documents obtained from the previous step are reduced to produce a concise answer. (Lalithnarayan, C., 2020). 'Passage retrieval' algorithms to reduce the amount of text, which is called 'finding passages' in the documents in scientific terms, are used to break the documents into paragraphs or sentences (passages) and select appropriate passages by using scores for each passage and by selecting the passages with the highest scores.

Answer Extraction Module

The module is the final module of the system, which takes the passages selected from the previous Paragraph Extraction Module as input and finally returns the most precise and reformatted answers to the users. Self-learning Q&A system often updates the system by evaluating the answers by itself.

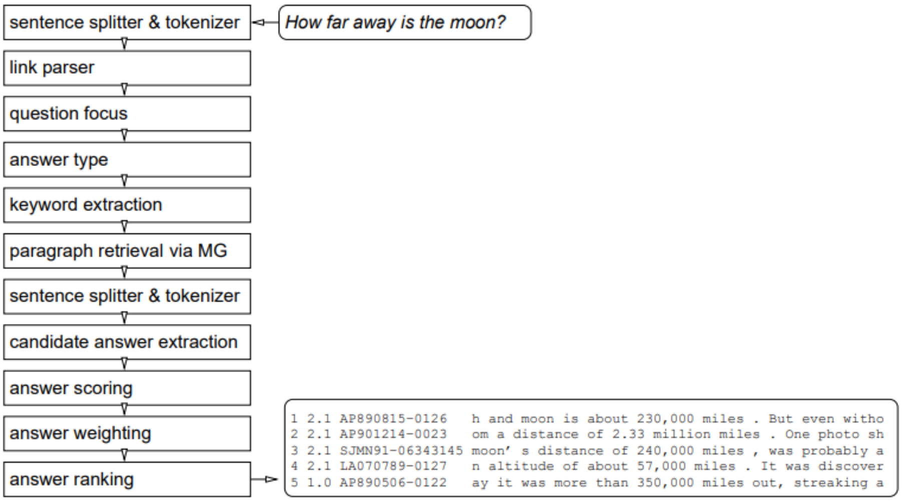
2.4 Real Q&A System's Workings

To have a better understanding of the Q&A system, to decide which approach we should take and to have a detailed plan on how to structure our own Q&A system, we have read a research article of a real Q&A system.

Cooper, R. J., & Ruger, S. M. (2000).
A simple question answering system

Firstly, all the articles available in the database of the programme are stored as a raw content of the article. For example, $ and £ are replaced by words "dollars" and "pounds" to get and store the raw content only.

The actual question processing is executed as a long pipeline of perl modules which use XML, which is mark-up entities or to communicate other information between the modules.


sentence splitter & tokenizer
How far away is the moon?
link parser
question focus
answer type
keyword extraction
paragraph retrieval via MG
sentence splitter & tokenizer
candidate answer extraction
answer scoring
answer weighting
answer ranking

```
1 2.1 AP890815-0126   h and moon is about 230,000 miles . But even witho
2 2.1 AP901214-0023   om a distance of 2.33 million miles . One photo sh
3 2.1 SJMN91-06343145 moon’ s distance of 240,000 miles , was probably a
4 2.1 LA070789-0127   n altitude of about 57,000 miles . It was discover
5 1.0 AP890506-0122   ay it was more than 350,000 miles out, streaking a
```

Figure 2.2 : Data Flow of the simple question answering system (Cooper, R. J., & Ruger, S. M.,2000)

So, the actual question will be separated into parts based on the pre-defined patterns or characters such as a question mark or an exclamation mark and will be formatted into XML codes so that the question can be easily digested by the system.
If I take the example of a question from the article, “How far away is the moon?” will be separated into “how”, “far”, “away”, “is”, “the”, “moon”, “?” and the separated parts will be formatted into XML codes with multiple additional data contained for the question and this process is called ‘Link Parser’ in this system.

```
<sentence><t n="1">How</t> <t n="2">far</t> <t n="3">away</t> <t n="4">is</t>
<t n="5">the</t> <t n="6">moon</t><t n="7">?</t><parse><pos n="2" pos="a"/><pos
n="4" pos="v"/><pos n="6" pos="n"/><link name="Xp" l="0" r="7"/><link name="Wq"
l="0" r="2"/><link name="PF" l="2" r="4"/><link name="MVp" l="2" r="3"/><link
name="SIs" l="4" r="6"/><link name="Ds" l="5" r="6"/><link name="RW" l="7"
r="8"/></parse></sentence>
```

Figure 2.3 : Example XML codes of the simple question answering system (Cooper, R. J., & Ruger, S. M.,2000)

Third process is called “Question Focus”, in which the system identifies the critical keywords of a asked question. We can easily know that each ‘when’, ‘who’, ‘where’, ‘whom’ and ‘why’ is a critical keyword of a question. For example, if a user asks “Where is Monash University?”, then, “where” and “University” will be the critical keywords of the question. Then, we can define those critical keywords in the system as back-end logics and add more data to the question. When the system finds the critical keywords, an extra XML element will be added to the existing XML codes, which notifies the system what words are critical. However, the article says that we need to be careful in deciding the critical keywords as there could be questions like “In what city is the US Declaration of Independence located?”, which the question does not start with easy critical keywords.

Next process is called “Answer Type”, in which the system identifies what kind of answers the system should provide to the user. The process is highly dependent on the “Question Focus” process, which is the previous process as the algorithm of this process must be based on the question focus that is identified in the previous process. The article mentions that answer type must be decided with detailed algorithms as the question focus will not simply indicate the purpose of the question. When the answer type is decided, an additional answer type XML element will be added to the original XML codes like the third process.

Then, the system will look for the related articles (as this programme is based on news articles database) in the database, using the answer type and question focus found in the above stages. The look-up of articles will be based on the weightings of the keywords and the most-likely subset articles will be chosen as the result.

The candidate answers from the previous process will be splitted and tokenized for a further processing and the question's answer concept is looked up in WordNet and all of its hyponyms are found. A regular expression is then built by taking the disjunction of those hyponyms and any region of text that matches that regular expression is marked up as a candidate answer. This process is called "Candidate Answer Extraction".
The problem with this process is that all the hyponyms could not be related to the question and questions which require descriptions cannot be easily answered using this process as they must be based on very complex Natural Language Processing (NLP).

Lastly, when the candidate answers are chosen in the system, the answer will be scored based on heuristics methods, which are:

(i) score_comma_3_word
If a comma follows the candidate answer then this score is the number of the three words following the comma that appear in the question
(ii) score_punctuation
Scores one if a punctuation mark immediately follows the candidate and zero otherwise
(iii) score_same_sentence
Computes the number of question words that are in the same sentence as the candidate answer
(iv) score_description_before
If the answer concept being looked for is a description then this score is the number of words immediately preceding the candidate answer that appear in the question
(v) score_description_in
Similar to score question before but counting question words that appear in the candidate answer.

After candidate answers going through these heuristics, all the candidate answers will be paired with (id, score) and this will be added as additional XML elements again. Then, the all answer scores will be combined into a final score based on the weight of each heuristics and using the final score of each candidate answers, the final answer that has the highest final score will be provided to the user.

2.5 Conclusion

After having the literature review on the Q&A system, we could get a detailed idea on how to structure our Q&A system and what kind of algorithms and approaches we can use for our project. We found out that all the Q&A systems follow the same step, which is pre-processing the user's query, retrieving the relevant documents stored in the database and finally getting the appropriate answers and returning the answer to the user.

We also found out that most of the Q&A system makes use of ‘Candidate answers’ and chooses the most correct answer among multiple possible answers, not returning an answer directly from a single document.

Moreover, after reading the research article of a real Q&A system, Cooper, R. J., & Ruger, S. M. (2000, November). *A simple question answering system*, we could know that the theories and structures discussed in the other articles are actually implemented in a real Q&A system.

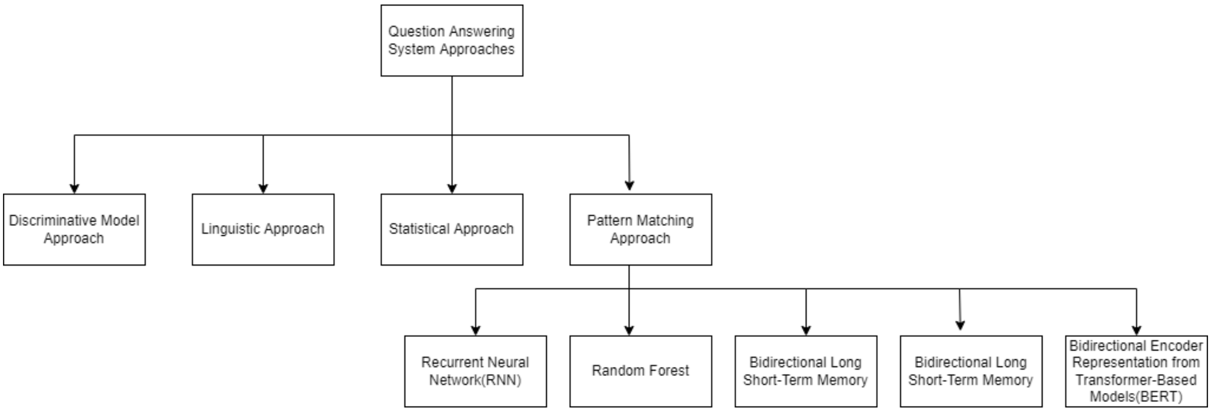

Question Answering System Approaches
Discriminative Model Approach
Linguistic Approach
Statistical Approach
Pattern Matching Approach
Recurrent Neural Network(RNN)
Random Forest
Bidirectional Long Short-Term Memory
Bidirectional Long Short-Term Memory
Bidirectional Encoder Representation from Transformer-Based Models(BERT)

Figure 2.3 : Summary of Q&A System approaches

3.0 Project Management Plan

3.1 Project overview

Coronavirus disease is an infectious disease caused by the SARS-CoV-2 virus. Many people might not have the appropriate knowledge, acceptance, and perception of Covid. Our business goal is to provide a Q&A system for this virus. Our team believes that our Q&A system can provide useful information related to covid, and ensure the information is accessible to everyone. Our main business objective is to provide useful information relating to Covid. Our social objective of this project is to improve everyone's knowledge, acceptance, and perception of Covid. Our human objective in this project is to create more job opportunities and also the development of human resources.

Ultimately, the main objective of our project is to provide an automatic COVID-19 answering website system that will be open to the public. The system should be able to categorize questions into expert-level information and general consumer-level information. Specifically, the technical objectives of this project are:

1. To acquire, pre-process and propose a suitable representation for the question-answering dataset on COVID19.
2. To investigate and identify suitable algorithms for developing a question-answering system for COVID19-related information.
3. To develop a web application as the front end for the question-answering system.

3.2 Project Scope

3.2.1 Project Scope

In-scope:

- Access to the latest information about Covid. Users must be able to search for information about Covid, read the user guide on how to use our system, and see how to apply some of the information in real-life
- Access to reliable databases. So that our system can provide accurate information to our user's
- Access to technical tools, languages, and software for our project. (e.g: Python, Java)
- Able to provide accurate information when our users key in their enquires
- Access via web
- A "Live Support" feature to help users that are facing any trouble or for users to enquire more detailed questions for Covid

Out-of-scope:

- Other features suggested by the users if they add value to the business
- Detailed search option for finding COVID-19 information
- Automated update on latest COVID-19 information

Constraints:

- Our primary constraint is that our project has to be completed by November 2022.
- We will be racing against time and on the other hand, we also have to get approval from our supervisor for the algorithm that we are using.
- By November 2022, we would be focusing solely on trying to answer general questions about covid.
- Our team might not have enough time to create a Q&A system that can answer in-detailed questions.

3.2.2 Product characteristics and requirements

Project management-related deliverables:

1. Project Scope Statement
   - To make sure that the team is on the right track in the development of the program.

2. Work Breakdown Structure
   - To make sure that all the team members equally contribute to the project and to identify who is responsible for which part of the project.

3. Requirement Traceability Matrix
   - To use the Requirement Traceability Matrix as a guideline & a roadmap in the development of the program. Our team prioritized the tasks of the project so that we can finish the project in time by identifying which tasks are urgent and optional

Product-related deliverables:

1. Software codes of the system, including both the front-end website UI and the back-end Python system.
   - This deliverable will be the main deliverable of our project, which will contain all the logics and algorithms to process users' queries and provide a simple user-friendly interface for consumers to easily access the COVID Q&A system.

2. Database server system
   - We are planning to store all the 'Answer's in the database which will be used to answer the users' queries after going through the logic of the back-end system.

3. Design Rationale documents
   - To help other developers to know the logic of our program and to explain our choices in selecting the algorithms & software.

REQUIREMENTS TRACEABILITY MATRIX

Project Name:	Tell me about Covid-A Question-Answering System about COVID					
Project Manager Name:	Kim Yeon Soo					
Project Description:	The Epidemic Question Answering (EPIC-QA) track challenges teams to develop systems capable of automatically answering ad-hoc questions about the disease COVID-19, its causal virus SARS-CoV-2, related corona viruses, and the recommended response to the pandemic.					

ID	***Priority***	***Requirements***	***Assumption(s) and/or Customer Need(s)***	***Category***	***Source***	**Status**
1	**High**	Datafication of possible expert-level questions on COVID-19	Gathered many possible questions by automated / manual research on COVID-19	Required / Functional	Project Supervisor	Open
2	**High**	Datafication of possible consumer-level questions on COVID-19	Gathered many possible questions by automated / manual research on COVID-19	Required / Functional	Project Supervisor	Open
3	**High**	Detailed analysis on the possible expert-level questions on COVID-19	Req ID. 1 is complete and the data for questions are ready	Required / Functional	Project Supervisor	Open
4	**High**	Detailed analysis on the possible consumer-level questions on COVID-19	Req ID. 2 is complete and the data for questions are ready	Required / Functional	Project Supervisor	Open
5	**Low**	Development of efficient automatic methods to update the latest COVID-19 information in the system	-	Optional / Non-functional	Project Supervisor	Open
6	**Medium**	Adding a detailed search option for COVID-19 information	Req ID 5. should be completed	Usability / Non-functional	Project Supervisor	Open
7	**High**	Configure a Database server that stores the latest COVID-19 information	-	Required / Functional	Project Supervisor	Open
8	**High**	Development of an algorithm that automatically provide expert-level answers for users' questions	Req ID. 1,3 should be completed	Required / Functional	Project Supervisor	Open
9	**High**	Development of an algorithm that automatically provide consumer-level answers for users' questions	Req ID. 2,4 should be completed	Required / Functional	Project Supervisor	Open
10	**Low**	A chat component to help users that are facing any trouble or for users to enquire	-	Availability, Usability / Non-functional	Project Supervisor	Open

	🟩	more detailed questions for Covid				
11	**Medium**	Adding support for web & Make Web UI	The back-end system must be completed	Required, Availability, Usability / Functional	Project Supervisor	Open

Table 3.1: Requirement Matrix Table

3.2.3 Product user acceptance criteria

1. The project meets all ‘Required’ requirements within 2022.
2. Project supervisor accepts the final product.
3. More than 7 users out of 10 users are satisfied with the contents of consumer-level answers
4. More than 5 users out of 10 users are satisfied with the contents of expert-level answers
5. Each inquiry process takes a maximum of 10 seconds (related to Time complexity)

3.3 Project Organisation

3.3.1 Process Model

There are many software development models, such as the iterative, spiral, V model, and others (Stoica, Mircea, et al., 2013). The models that will be analyzed are Agile and Predictive, due to their usage in recent years.

The Agile model has been widely used in the IT industry for approximately 20 years now (Kaur, Jajoo, et al,.2015). Its high usage is no surprise as this adaptive approach (Stoica, Mircea, et al., 2013) can be used on small scale development projects, large scale software development projects, testing projects, and software maintenance projects, with Serrador and Pinto‘s article (2015) stating that project success improves through the implementation of Agile. The scrum framework implements roles, ceremonies, meetings, and artifacts (Kaur, Jajoo, et, al,. 2015). Its benefits are heavily correlated with the agile manifesto values (Gustavsson, 2016). Despite emphasizing minimal documentation and close interaction with clients, developers may be stressed over the short iteration time frame, consequently causing adverse effects on the product (Stoica, Mircea, et al., 2013).

The Predictive approach is most appropriate for projects that have a detailed plan and have a complete list of characteristics and tasks. Furthermore, such an approach depends on the requirement analysis, and it also emphasizes documentation in orientation and clarification of the project (Stoica, M., Mircea, M., & Ghilic-Micu, B. 2013). The advantage of using predictive modeling is that everything is easy to plan in detail because the initial user requirements would be fixed (Mario Špundak, Mixed Agile/Traditional Project Management Methodology – Reality or Illusion?. 2014). The disadvantage of this modeling is that requirements may arise after initial requirement gathering and it will be difficult to return to the design stage (Stoica, M., Mircea, M., & Ghilic-Micu, B. 2013).

Each model has its benefits and drawbacks. Hence, it is important to identify factors such as the type of project, the importance of frequent client communication, and the volatility of user requirements.

Despite being unique in their own way, agile and predictive models have their benefits and drawbacks when implemented during projects. The agile model is flexible in the sense that it only focuses on future tasks that are clear (Stoica, Mircea, et al., 2013). On the other hand, the nature of predictive models are having constraints that are well defined (Špundak, 2014). Essentially, projects that have product requirements would benefit more from a predictive model, while an agile model would be more suitable for projects with less detailed planning. The lack of planning in an agile model is offset by frequent and open communication with clients, leading to minimal documentation, despite potentially tiring developers due to multiple meetings and iterations (Stoica, Mircea, et al., 2013). On the other hand, predictive models emphasize documentation and well-defined requirements before the project begins, consequently allowing developers to be unconcerned about meetings with clients (Stoica, Mircea, et al., 2013). This also means that predictive models can be planned in detail from the start of the project, therefore being more robust than an agile model as the project can be run smoothly without any changes (Špundak, 2014).

Nevertheless, the usage of predictive models is very situational, ideally when team members are unable to come to a consensus on different approaches, inexperienced members, or if the project manager does not frequently contact the team members (Špundak, 2014). On the other hand, an agile model is able to be implemented on various types of projects, such as small-scale development projects, large-scale software development projects, testing projects, and software maintenance projects (Gustavsson, 2016). Regardless, both models have their own advantages. An agile model has increased productivity and speed for projects due to the nature of the project being flexible. The team would have better cooperation due to well-defined product requirements (Gustavsson, 2016). On the other hand, a predictive model consists of structured design and documentation which is easy to use. Besides that, project coordination is clear as the stages that are carefully implemented have expected results and an evaluation process (Stoica, Mircea, et al., 2013). The management styles of the two models are different as well, with predictive models implementing a command and control management style with formal communication, while an agile model uses a more modern approach that embraces informal communication and leadership (Stoica, Mircea, et al., 2013).

The agile and predictive models are clearly unique in their own way. Regardless, our team cannot use both methodologies to approach our project. After much discussion among our supervisors and our group members, we have come to a consensus on the methodology that would be used for our project. The approach that would best suit our project would be the predictive life cycle model. The primary reason why we chose this model was that all the project requirements would be given beforehand. There would not be any changes or emergencies, and the project team would not have to consult the client multiple times. In addition, based on our findings we have concluded that agile SDLC is better suited for small and medium-scale projects, and traditional SDLC is for projects on a larger scale (Stoica, M., Mircea, M., & Ghilic-Micu, B. 2013). Hence, we have decided to use traditional SDLC because our project is considered a large-scale project. In a nutshell, the predictive model approach would fit our project better, and our team would also apply some agile methodologies to our project. We plan to implement frequent scrum meetings to keep track of the progress of each team member.

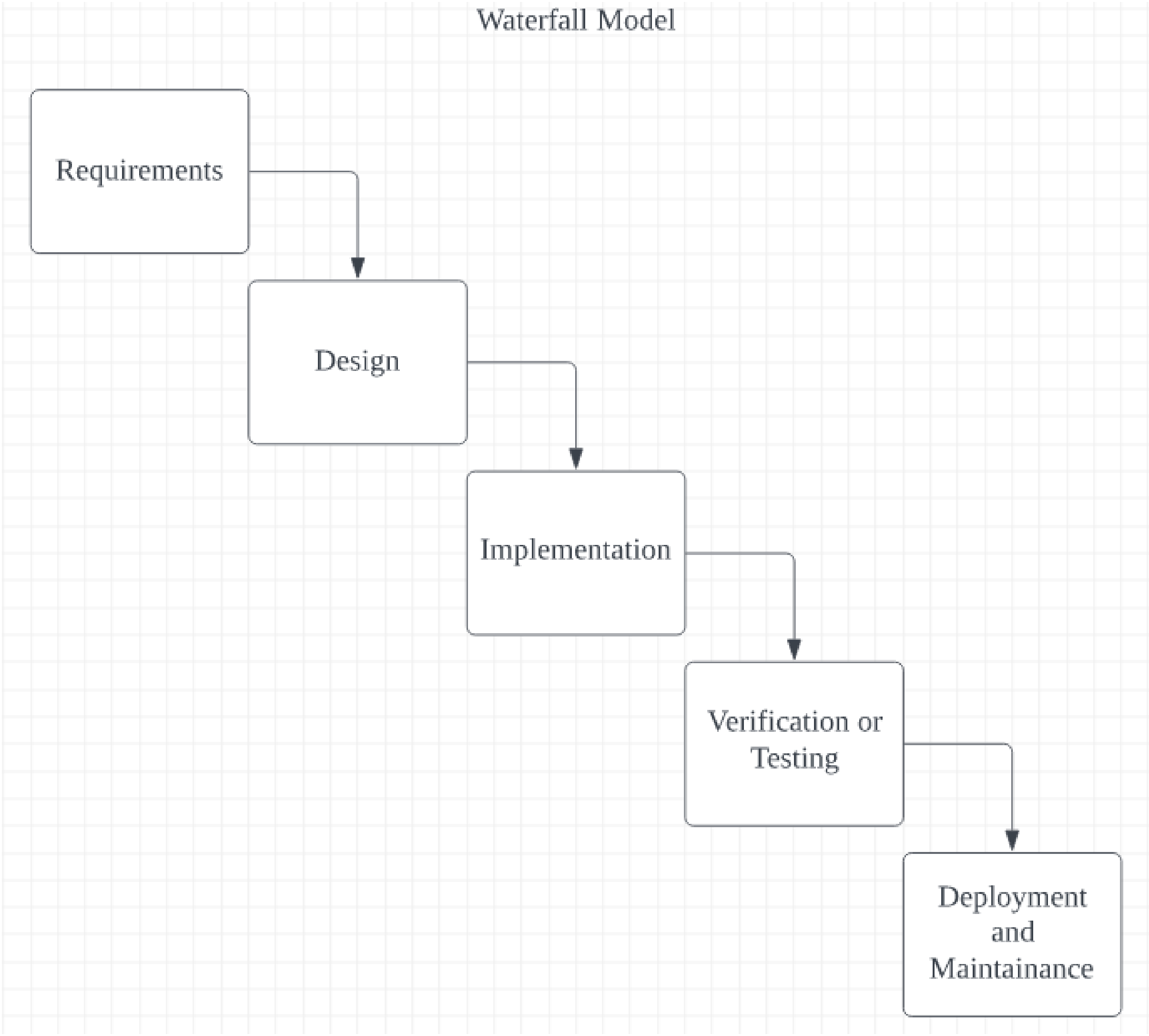


Figure 3.1 Waterfall Model

The picture above shows the predictive methodology that our group will be using for our project. The waterfall model is also known as the linear sequential life cycle model. The model is fairly simple to understand, the phases do not overlap each other. In this model, each phase must be done only then the next phase can start (SDLC - Waterfall Model, n.d.).

3.3.2 Project Responsibilities

The table below shows the responsibility of each team member. To maximize our work efficiency, our team agreed to assign a single role to each team member.

Roles	**Responsibilities**	**Person In Charge**
Project Manager	Ensures the completion of the project	Yi Sen
Technical Lead	Take charge of the technical team	Wai Han

Evaluation and Correction	Evaluates the code and ensure its correctness	Nawwaf Ali
Quality Assurance	Ensure our resources are used efficiently and effectively	Yeonsoo Kim

Table 3.2: List of Project Roles

3.4 Management Process

In this section, we will discuss the two project management techniques that our group uses to identify risk.

3.4.1 Risk Management

Risk can be defined as the possibility of loss or injury (Manage risk,n.d.). Project risk is inevitable when it comes to any project, hence our goal is to always minimize the potential negative risks and try to maximize our potential positive risks. Risk management will then be applied to meet our project goal. In this section, we will be discussing the two risk identification techniques which are brainstorming and the SWOT analysis.

Brainstorming

Brainstorming is a risk management technique that provides a free environment for everyone on the team to pitch in and participate. This method can provide a greater sense of the project risk ownership and as a whole, the team will be more committed to managing risks throughout the duration of the project (Ghazaryan, 2017). Our group would run a brainstorming session via Zoom meeting and during the brainstorming session, we would:

1. Verbally identify the risks in turn
2. Identify the quality of the risks
3. Filter the priority of the risks based on the probability of occurrence

After the brainstorming session we would come up with mitigation strategies which are shown in our risk register in the appendix. Our team would ensure the identified list of risks is always up to date (Ghazaryan, 2017).

SWOT analysis

What is SWOT analysis? SWOT analysis is a framework that is used to identify the negative and positive risks that are present in a project. The acronym for SWOT stands for strengths, weaknesses, opportunities, and threats. To have a good SWOT analysis, we will first create a list of questions to answer for each quadrant. Then, we continue by identifying the strengths and weaknesses of each team member of the team and all the identified strengths and weaknesses will be contributed when we are considering the opportunities and threats quadrants. By using the opportunities quadrant, we would be able to meet our project goals (Kenton, 2021).

These are the risk identification techniques that our team would apply. The risks that we identified in the brainstorming session, will be brought over when we are performing the SWOT analysis. All the risks will then be documented in a risk register, which is attached in Appendix A.

3.4.2 Stakeholder Analysis and Communication plan

Having a stakeholder analysis matrix are as important as trying to complete the project. By having a stakeholder analysis matrix, we can get our project into shape quickly because we can get our project scope from our strongest stakeholders at an early stage, and the remaining stakeholders would be more likely to support you and contribute their opinions to improve the quality of the project. In addition, we can also build a stronger relationship with our stakeholders we will have a greater chance of succeeding in our project.

Stakeholder Analysis Matrix						
Stakeholder Name	**Contact Person**	**Impact**	**Influence**	**Tasks**	**Risks**	**Mitigation Strategy**
Project Supervisors	Dr.Soon Lay Ki	High	High	- Provide project requirements - Provide feedback on our progress	Miscommunication between the supervisor and the project team, causing the work produced is not per requirement	Project team needs to communicate well with the project supervisor and also update the progress with supervisor
Project Manager	Yi Sen Ooi	High	High	-Represents the team and communicates with the project supervisor - In charge to ensure project outcomes is as per requirement - Allocates the tasks evenly	Scope creep might happen due to pressuring team members to take tasks that is outside the project scope	Project team needs to communicate well with the project manager based on the project scope given by supervisor
Technical Lead	Wai Han Chan	High	High	- Guides the technical development to ensure it's on the right track - Oversee the work of all the developers's	Not clear with the requirements and guides the technical development into a wrong path	Project manager should ensure its clarity of the requirements during the planning the phase
Evaluation and Correction	Nawwaf Ali	High	High	-Evaluates the code and ensure its correctness	Not clear with the requirements which leads to inaccurate evaluation	Project manager should ensure its clarity of the requirements during the planning phase
Quality Assurance	Yeonsoo Kim	High	High	-Ensure our resources are used efficiently and effectively	Excessive quality control which results in late submission for our project	Document all the requirements so that the Quality Assurance has a clear understanding of the requirements

Table 3.3: Stakeholder Analysis Matrix Table

In all of our meetings, we will have three roles, the leader, the timekeeper, and the participants. These roles are not fixed and may change depending on the situation. The leader is responsible for leading and managing the meeting. The timekeeper is responsible for keeping track of the meeting minutes and finally, the participants are responsible for participating in the meeting. The meeting minutes template is included in Appendix B. In nature, most of the communication will be informal but for our supervisor meetings and emails, it will be more formal and structured.

Communication Plan						
Type of communication	**Objectives**	**Method of Communication**	**Frequency**	**Recipients**	**Person Responsible**	**Duration**
Emails	Set a meeting time with the project supervisor	Gmail	When it is required	Project supervisor	Project Manager	N.A
Direct Messages	Informal conversation	Slack	When it is required	Project supervisor	Project Manager	N.A
Team meeting	Working on the project together	Discord	Weekly	Project Team	Project Team	15 - 120 minutes
Supervisor meeting	Update project supervisor with our progress	Zoom	When it is required	Project supervisor	Project Manager	Not fixed

Table 3.4: Communication Plan Table

3.4.3 Monitoring and Controlling Mechanisms

Review and Audit Mechanisms

For the review, we would be performing software quality assurance. In this section, our team will be focusing more on the software process rather than the product itself. Quality Assurance is a set of rules where it is designed to ensure the project team follows these procedures which are defined beforehand. For the Audit Mechanisms, our team would be assessing our work to ensure all the process was followed (Hamilton, 2022).

Version Control

Our team decided to use Git as our version control, in a large-scale project version control plays a vital role because there will be heavy coding work involved. We will be sharing a repository on GitHub, where we clone the HTTPS link to our local repository. By doing so, each team member will be able to have edit access.

Feature	**Description**
Git Add	Adds the changes in the working directory to the staging directory
Git Commit	Capture the project's current staged changes
Git Push	To upload your work from the local repository into the remote repository
Git Pull	Fetch and download the work from the remote repository to your local repository
Git Merge	Merge all the past development branches into a single master branch

Table 3.5: List of Git Features

All the features that were mentioned in the table above are the features that we will be using for our version control. The primary reason why we use GitHub as our version control is that it provides easy project management and it also increases code safety. Another advantage of using GitHub is that multiple team members can work on the same project together on different devices.

Quality Assurance

For our project, we will always have a code review before any member wants to push the work into GitHub. We will come out will a code review checklist before our meeting, and during the meeting, we will ensure all the items in the checklist are passed then only we will allow the rest of the members to pull the code into their local repository. We can ensure code safety, and if any member accidentally messes up their code, the rest of the members will not get affected. The table below shows a sample code review checklist that our team will be using.

No.	Item	Result	Remarks
1	Does the program terminate?		
2	Does the program return the desired output?		
3	Are there documentation for all the functions?		
4	Are all the variable names properly named according to their purpose?		

Table 3.6: Sample Code Review Checklist

Documentation

Our group decided to document all of our work so that we can easily manage, control, and deliver the project effectively. The benefit of having documentation is that the project tasks are more traceable and it also helps us to evenly distribute the work to all team members. We will constantly update all of our documents at all times so that we ensure that the integrity of our document is maintained

Documentation	Purpose
Project Scope	Scope management
Requirements traceability matrix	Scope management
Risk Register	Risk management
Work Breakdown Structure	Task management
Gantt Chart	Timeline management
Kanban Board	Task management
Version Control Log	Keep track of the progress made
Code Documentation	Other developers and understand our work easier
Meeting Minutes	Time and Task management
RACI Matrix	Task management

Table 3.7: List of Documentation Table

Training

Our team has agreed that we will be having training sessions during our upcoming semester break to empower our knowledge for our project because we do not have any experience in making a Q&A System, and this is the main objective of our project. To ensure that we are ready for the implementation stage in the upcoming semester, we planned to carry out:

1. Read more research papers that are related to our project
2. Research more on Q&A algorithms, specifically on how they function and what are the possible ways we can tackle them
3. Read more online articles about Covid-19 so that we can provide accurate responses
4. Enroll in online courses that are related to our project

By going through our planned training, we are confident that we will acquire all the necessary knowledge that is needed to complete this project.

3.5 Schedule and Resource Requirements

In this section, we will discuss the schedule and all the resource requirements for this project.

3.5.1 Schedule

Work Breakdown Structure (WBS)

The work breakdown structure is a method that our team uses to manage our large-scale project. This method is also known as the divide and conquer method because it can get things done much faster and more efficiently. The main objective of this using WBS is to make our project more manageable. Our WBS is included in Appendix C (What Is Work Breakdown Structure in Project Management?, n.d.).

Gantt Chart

The Gantt Chart is used to manage the progress and schedule of each team member. Our Gantt Chart is the whole duration of all the assessments that we had this semester. Our team's Gantt Chart is included in Appendix D.

Kanban Board

The Kanban Board divides the tasks into three different sections. The first section is the To-Do section and the second section is the Doing section and the final section is the Done section. We will be using the Kanban Board to keep track of tasks that we have completed, tasks we need to complete, and tasks that we are doing. Our Kanban Board is included in Appendix E.

RACI Matrix

The RACI Matrix is used to identify all the key roles and responsibilities for a project. The purpose of this chart is to balance the workload for all project members. The acronym for RACI is responsible, accountable, consulted, and informed. Each role represents the levels of involvement of each member of the team (*Olivia Montgomery, P. and Kumar,R*, 2020). This is shown in Appendix F.

3.5.2 Resource Requirements

The number of personnel required to complete this project is four-person, along with one project supervisor. The computer time includes the time used to implement the algorithm, code the website out, and also set up our database to store all the information about Covid-19.

All the resource requirements including the software requirements and hardware requirements are listed down below.

Software Requirements

1. HTML/CSS source code is used for the website development
2. Python backend development will be based on python language
3. SQL Developer for the database server of the system
4. GitHub project used to store our source code
5. Testing scripts that include sample consumer questions and sample expert questions
6. Database schema and a simple UML diagram of the source code
7. Software Documentation on the workings of the program

Hardware Requirements

This is the minimum hardware requirement specification for this project. Where the CPU uses fast single and multi-threaded performance. The GPU will be used to speed up the whole process for our algorithm. The RAM is used to read and write data. The Solid State Disk is required to store all the information for this project.

Hardware	**Specifications**
CPU	AMD Ryzen 5 Quad-Core up to 3.7 GHz
GPU	AMD Radeon Vega 8
RAM	8 GB
Display	15.6 inch FHD (1920x1080)
Solid State Disk	512 GB SSD

Table 3.8: List of Hardware Requirements Table

4.0 External Design

4.1 Website & UI

When identifying and designing a user interface there has to be a set way in which we approach this, Kelden Lin's article extracts a few common UI principles from multiple other sources to help us identify and design a great user interface. Firstly a great UI has to have consistency in the way you design the elements in it. These elements should look the same if their behaviors are also similar (Lin, 2016). The second principle is Familiarity/design disciplines, this essentially means that the UI design can and should follow already existing design disciplines within the platform they are in (Lin, 2016), basically saying that we do not need to always make something new and complex. This makes the user interface easier to use as the user already will be familiar with the design should they have any prior experience with an interface that follows the same design discipline. The third principle is to reduce the cognitive load, this principle fundamentally means the interface design should make up for human's limited capacity for information processing(Lin, 2016). Grouping together elements that have certain relations as well as making sure that each screen is for a set purpose rather than multiple elements with different roles appearing on the same screen overloading the user with too much information.

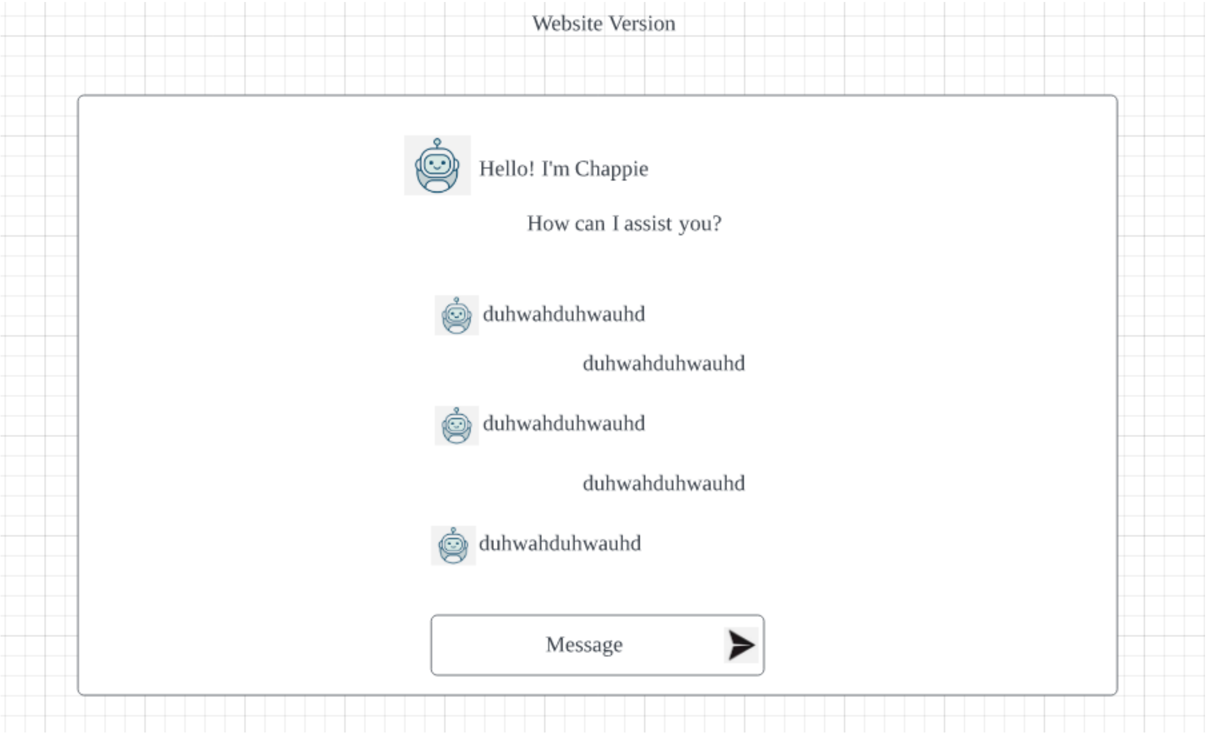
Since we are going to be designing a web application, the design we are choosing falls under Graphical User Interface(GUI) which is a type of User Interface(UI) that uses a tactile UI input with a visual UI output (Indeed Editorial Team, 2021). Each UI has its own advantages as well as disadvantages mentioned in the article written by Alan blog(n.d) but GUI is the most suitable one for our implementation.

Advantages of GUI (taken from Alan Blog(n.d.):

- that it is suitable for non-technical users.
- The complexity of actions is hidden from users.
- Immediate visual feedback
- Enhanced by attractive visual
- Enables the use of multiple input devices such as keyboard, mouse, etc…

Disadvantages (taken from Alan Blog(n.d.):

- That it requires power and memory resources
- Can overwhelm users with the growing amount of control elements
- Might have low discoverability.
- Hidden commands need to be searched for



Website Version

Hello! I'm Chappie

How can I assist you?

duhwahduhwauhd

duhwahduhwauhd

duhwahduhwauhd

duhwahduhwauhd

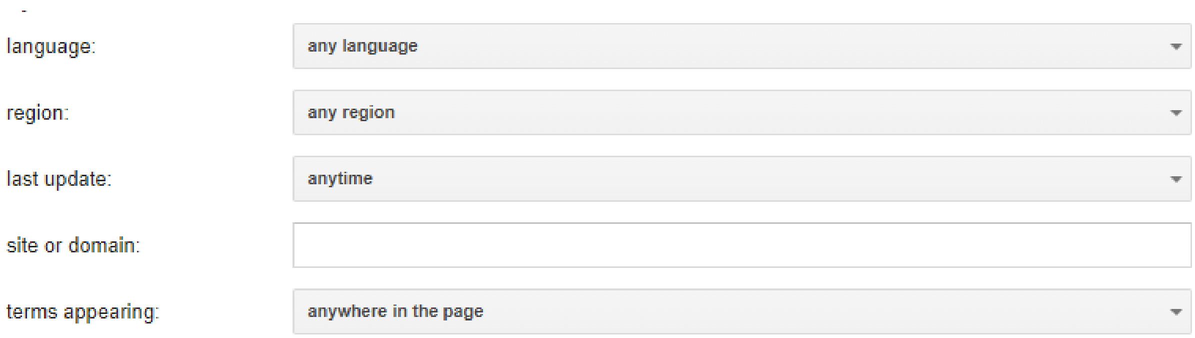
duhwahduhwauhd

Message

Figure 4.1: Sketch of our Web Chatbot

This is our initial planning on how the chatbot UI will be designed. The chatbot would generate an automated message to greet the user whenever they come into the website. Then, the user can input their question via the textbox provided. After the user finishes their query, there will be an upload button for the user to upload their question.

In addition to the chatbot, we are planning to add additional pages on the project website to provide the latest COVID-19 information so that the users of the website can search for the COVID-19 information for those who do not want to use the chatbot to get the information. We are currently planning to add search options and automatic updates of the information to the system, however, those features are not our priority and the features will only be added if we have sufficient time to add the features.



language:	any language
region:	any region
last update:	anytime
site or domain:	
terms appearing:	anywhere in the page

Figure 4.2: Advanced search options of Google

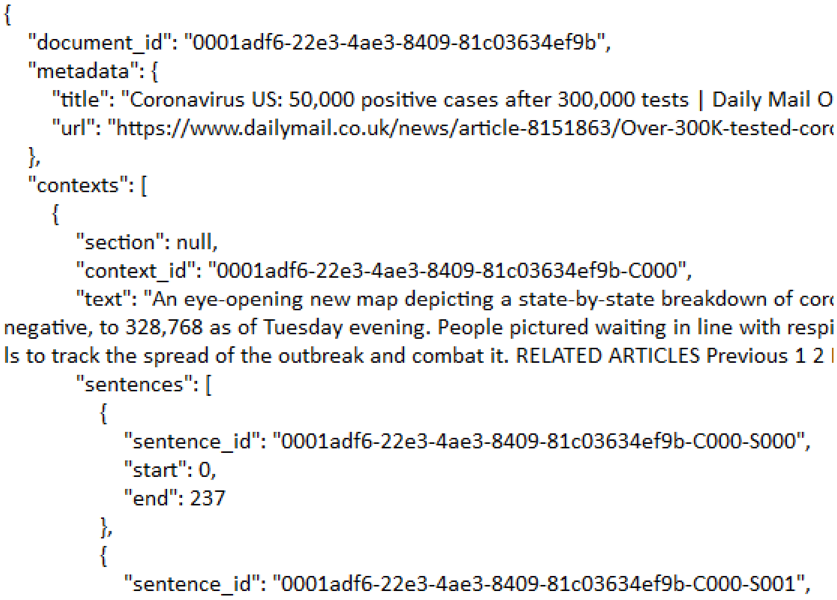
4.2 Datasets

We are planning to use datasets provided by ‘Epidemic QA at TAC2020’ (Epidemic QA at TAC 2020.) which was a public project which was making a program that automatically answers the users’ questions.

List of Datasets

- Collection of COVID-19 research articles
- Collection of COVID-19 consumer articles
- Collection of consumer-level questions
- Collection of expert-level questions

All of the datasets are in JSON format and all of the data in each dataset have a unique ID with relevant fields of the data.



```
{
   "document_id": "0001adf6-22e3-4ae3-8409-81c03634ef9b",
   "metadata": {
      "title": "Coronavirus US: 50,000 positive cases after 300,000 tests | Daily Mail O
      "url": "https://www.dailymail.co.uk/news/article-8151863/Over-300K-tested-cor
   },
   "contexts": [
      {
         "section": null,
         "context_id": "0001adf6-22e3-4ae3-8409-81c03634ef9b-C000",
         "text": "An eye-opening new map depicting a state-by-state breakdown of cor
negative, to 328,768 as of Tuesday evening. People pictured waiting in line with respi
ls to track the spread of the outbreak and combat it. RELATED ARTICLES Previous 1 2
         "sentences": [
            {
               "sentence_id": "0001adf6-22e3-4ae3-8409-81c03634ef9b-C000-S000",
               "start": 0,
               "end": 237
            },
            {
               "sentence_id": "0001adf6-22e3-4ae3-8409-81c03634ef9b-C000-S001",
```

Figure 4.3: Snapshot of a Consumer Article Data

The datasets already have all the information which is necessary for our program. As the datasets are pre-processed in system-friendly format, we are planning to store all the JSON datasets into a separate Database server as our program will be based on the Web application and it is easier to use and manage the data by connecting the web application to a database server, rather than storing the datasets directly into a web server.

5.0 Methodology

5.1 Software and Hardware to build Answer ranking system

To create this system, We will be using SQL developer for the database server of the system and the GitHub project will be used to store all the source codes and will allow easier cooperation between team members. The Programming languages we will be using will HTML/CSS source codes for website development and Python 3.8 minimum which has access to various libraries we can use for the machine learning aspect of this project, some of those libraries are as follows:

- Numpy
- Pandas
- PyTorch
- Pandas
- Matplotlib

We will also be making use of visualisation tools such as database schema and UML diagrams to support the software documentation on the source codes of the programs to give better understanding of the programs.
As mentioned in Part 4.2 we will be using data sets from ‘Epidemic QA at TAC2020’ (Epidemic QA at TAC 2020.).

In terms of the hardware specification, we would be using the best laptop from Michael’s article (n.d.) where he recommends budget laptops for computer science students. The minimum hardware specifications for the laptop that would be our development platform is as follows:

- CPU: AMD Ryzen 5 Quad-Core up to 3.7GHz
- RAM: 8GB
- GPU: AMD Radeon Vega 8
- Storage: 512GB SSD
- Display: 15.6 inch FHD(1920 x 1080)

5.2 Basic Model to the Answer ranking system

When the user asks a question using the system, the user must specify the level of answers he/she wants for later use in the database.

Level	Question	Answer
Low	How is covid-19 transmitted?	Through the air,if an infected person sneezes near you and
High	How is covid-19 transmitted?	Direct person-to-person respiratory transmission is the primary means of transmission of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It is thought to occur mainly through close-range contact (ie, within approximately six feet or two meters) via

		respiratory particles; virus released in the respiratory secretions when a person with infection coughs, sneezes, or talks can infect another person if it is inhaled or makes direct contact with the mucous membranes. Infection might also occur if a person's hands are contaminated by these secretions or by touching contaminated surfaces and then they touch their eyes, nose, or mouth, although contaminated surfaces are not thought to be a major route of transmission.

Table 5.1: Table classifying high and low level answer

This is a simple diagram that shows what will be stored in the database and what processes the database will go through. So, the answers to the questions will be stored in the database and the system will use the stored answers to respond to the users. Every 'Answer' entity will be linked with 'Document' entities as the references of the answer.

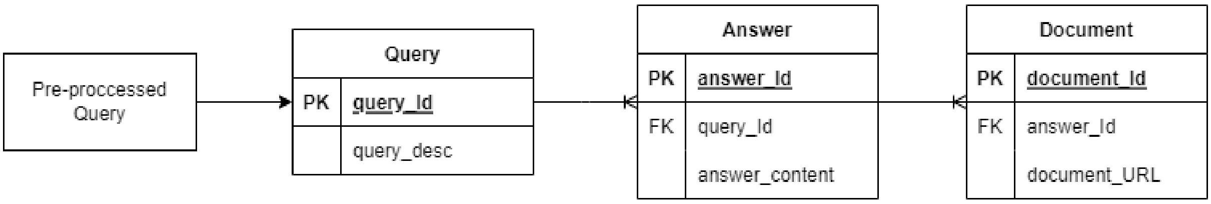


Figure 5.2: Diagram of database

When the user inputs a question, the program will pre-process the question using algorithms and will make the question into a specific query that is already stored in the database. For example, if the user asks 'What is the number of COVID-19 positive patients?', the program must pre-process the question into a short system-friendly query like 'covid patients number'. Then, the program will send the query (or query ID directly) to the database to find an 'Answer' that has the matching query.

5.3 Approaches that could be taken to develop to the Answer ranking system

The first approach that could be used in our project is the 'Linguistic Kernel' approach to the question answering system, Which uses multiple multiple Kernel methods to get structured data. Kernel Methods refer to large class learning algorithms based on inner product vector spaces,among which support vector machines (SVMs) are well-known algorithms(Moschitti, 2011). The paper 'Linguistic kernels for answer re-ranking in question answering systems' talks about the use of kernel functions to exploit syntactic/semantic structures for relation learning from questions and answers(Moschitti, 2011). There process in more detail was: (i) model sequence kernels for words and part-of-speech tags that capture basic lexical semantics and syntactic information, (ii) apply tree kernels to encode deeper syntactic information and more structured shallow semantics and (iii) analyse the proposed shallow semantic kernels in terms of both accuracy and efficiency (Moschitti, 2011). This could be the steps we would be taking if we were to take this approach. This approach is written further in the literature review.

Another approach we could use is from the article ‘Hands-On Question Answering Systems with BERT: Applications in Neural Networks and Natural Language Processing’. Most probably use an altered version of this approach which will allow us to rank these answers. BERT extracts tokens from the question and passage and combines them together as an input (Sabharwal, 2021). It starts with a CLS token that indicates the start of a sentences and uses SEP separator to separate the question and passage, BERT then creates two segment embeddings one for the question and the other for the passage (Sabharwal, 2021).After this embedding are input into the BERT model it will use softmax to generate probability distribution for the start and end index over an input text sentence that defines a substring which is an answer (Sabharwal, 2021). This approach however might have a drawback in the sense that the variety of datasets we can use are limited, as it requires the question and the answer for the question to be provided.

5.4 Version control

Version control,also known as source control, is the practice of tracking and managing changes to software code (Atlassian BitBucket, n.d).Version control systems are software tools that help software teams manage changes to source code over time (Atlassian BitBucket, n.d). We wrote more about this in part 3.4.3 of this project proposal. As mentioned earlier in part 3.4.3 version control, we will be using GitHub as our version control system, as GIT has many benefits that is common to version control systems such as a complete long-term change history of every file, branching, merging and finally traceability (Atlassian BitBucket, n.d).

6.0 Test Planning

Software testing is defined as the process of evaluation and verification to ensure that a software product meets its requirements. Software testing is pivotal as it ensures that bugs are prevented, development costs are reduced, and overall performance is improved (*What is Software Testing?*, n.d.). We have briefly thought about how we would go about ensuring that our software product runs correctly. The front end will not be as important as it is just about the user interface. On the other hand, the most important part of our software product would actually be the backend code, where most of the logic would be carried out. We have also thought about the different types of testing methods that we may potentially use.

The first testing method that we would implement is the black-box testing method. It is defined as the testing of a software application's functionalities with its internal code structure, implementation details, and internal paths being unknown (Hamilton, 2022). Through the implementation of this testing method, we will be able to identify the system's response to expected and unexpected user actions (*Black Box Testing*, n.d.). Our team has managed to identify the few types of black-box testing that may be beneficial to our software product.

The first method would be random testing, which is also known as monkey testing (*Random Testing*, n.d.). It essentially uses random inputs to test the system's reliability and performance. However, despite it saving time and effort compared to other test efforts, our team decided that this would be used as a last resort as it is a method used when there is insufficient time to write and execute tests.

Besides that, equivalence testing is another black-box testing method that we may use. It is a method that divides the input domain of a program into multiple classes of data, where test cases can be derived (Qualitance, 2014). Essentially, all input values from a specific partition will make the program behave in the same way. Through equivalence testing, we will be able to achieve minimum test coverage. Besides, the general test execution time is decreased, while the set of test data is reduced, as it is a process-oriented testing method. However, not all inputs that are necessary may be covered. Not only that, the test engineer may assume that the output for all data set is correct (*Equivalence Partitioning Technique*, n.d.).

On the other hand, another type of testing method that our team identified is the white-box testing method, also known as clear box testing, open box testing, transparent box testing, code-based testing, and glass box testing. It is essentially a software testing technique where the internal structure, design, and coding are tested for the verification of the flow of input to output while improving the design, usability, and security (Hamilton, 2022). It is the counterpart to black-box testing, as it tests the inner working of an application and revolves around internal testing, opposite to black-box testing.

A method that our team identified as code coverage, which is a measure that describes the degree of the source code of the program being tested. It is able to identify areas of the program that is not exercised by the test cases (Hamilton, 2022). The type of coverage that our team sees viable for our software product is the condition + branch coverage. Branch coverage provides two branches for each decision, regardless of a decision's complication. On the other hand, condition coverage splits the decisions into single conditions. Essentially, instead of having one big decision block with the entire condition, condition coverage creates multiple decision blocks, with each of them consisting of only one condition, which will then be exercised separately. The goal is to achieve 100% condition coverage and 100% branch coverage, which would mean that all outcomes of all conditions and all outcomes of all decisions are exercised. Furthermore, a testing method that our team identified is path coverage, where all paths of the program are tested (*White Box Testing: A Complete Guide With Techniques, Examples, & Tools*, 2022).

Path coverage is better than branch coverage, as it ensures that all paths of the program are traversed at least once. Path coverage is useful for testing complex programs such as our software product, which consists of multiple if and else statements when evaluating a user query to generate an answer. Despite path coverage being able to cover a wide range of conditions, it needs to test all possible paths, such as loops, conditional combinations, and branch statements. Thus, a large number of complex test cases need to be designed, leading to an increase in workload (Yan, 2020).

Furthermore, as our team has decided to use Python as our back-end development tool, we can use the Python unit testing framework, also referred to as "PyUnit" to test our code. Through unit testing, we will be able to make sure that the individual units are working. PyUnit is very useful as we will be able to write test code in a separate file from the production code, which helps in the organization of our source code in keeping it neat. The parts of our software product that we will be testing would be the essential stages, such as the user queries and the system's response. With PyUnit, we will be able to split the tests into their respective test files, and test each responsibility separately, thus easing error handling.

The following table shows the testing methods that we have identified.

Testing Type	**Testing Method**
Blackbox	● Random testing ● Equivalence testing
Whitebox	● Code coverage ● Condition + Branch coverage ● Path coverage
Python Unit Test	● PyUnit

Table 6.1: Table of testing methods

7.0 Conclusion

With the occurrence of the coronavirus pandemic, it is evident that proper knowledge of what we are facing is pivotal to make important decisions in life. As stated in our project overview, our project aims to benefit the public, to essentially provide accurate information when our end users key in their enquiries. The literature review discusses the multiple available tools that can be used to do so. With the multiple approaches and modules available, we will be able to smoothly create our own Q&A system.

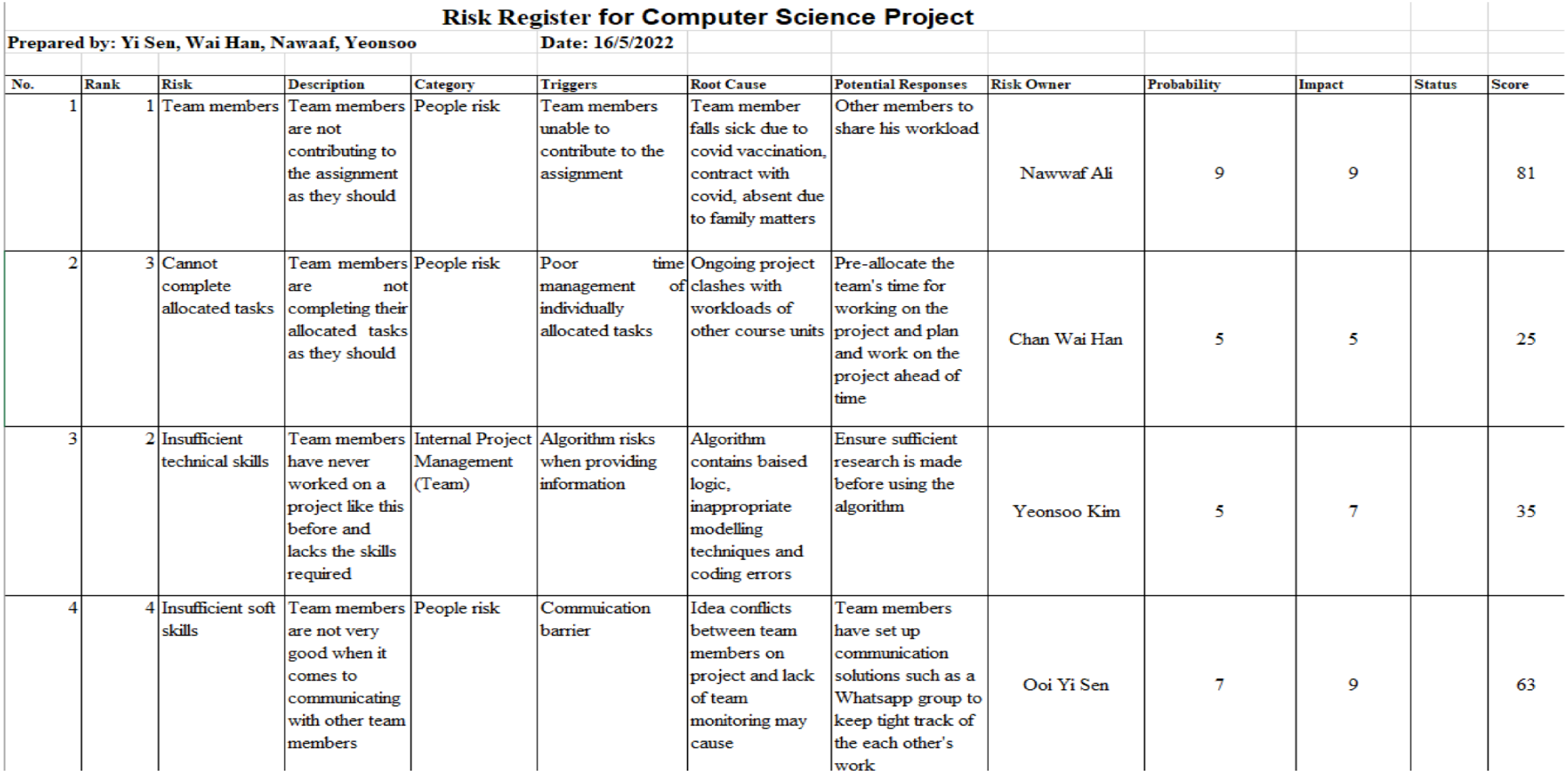
As our team have decided to implement a predictive model approach - the waterfall approach to carry out this project, we have used multiple project management-related deliverables, such as the project scope statement, work breakdown structure, and requirements traceability matrix. On the other hand, the product-related deliverables that we used were software codes of the system, which include both the front-end website UI and the back-end Python system, database server system, and design rationale documents. These deliverables will ensure that we are able to develop our software product smoothly with minimal error. Besides that we have delegated project responsibilities to each team member to maximize work efficiency.

Furthermore, our team carried out risk management to minimize the potential risks and try to maximize our potential positive risks. A stakeholder analysis was also created to have a greater chance of succeeding in our project. In terms of the external design, the GUI is the most suitable for our project as it has advantages that are superior to its disadvantages. Finally, testing methods were identified to verify and validate our software product at the end of our project.

In a nutshell, the end goal of our project is to help the public in these difficult times. More rearch and innovation is certainly required to fully tackle this virus, and we hope taht we will be able to assist in doing so through our software product.

8.0 Appendix

A RISK REGISTER

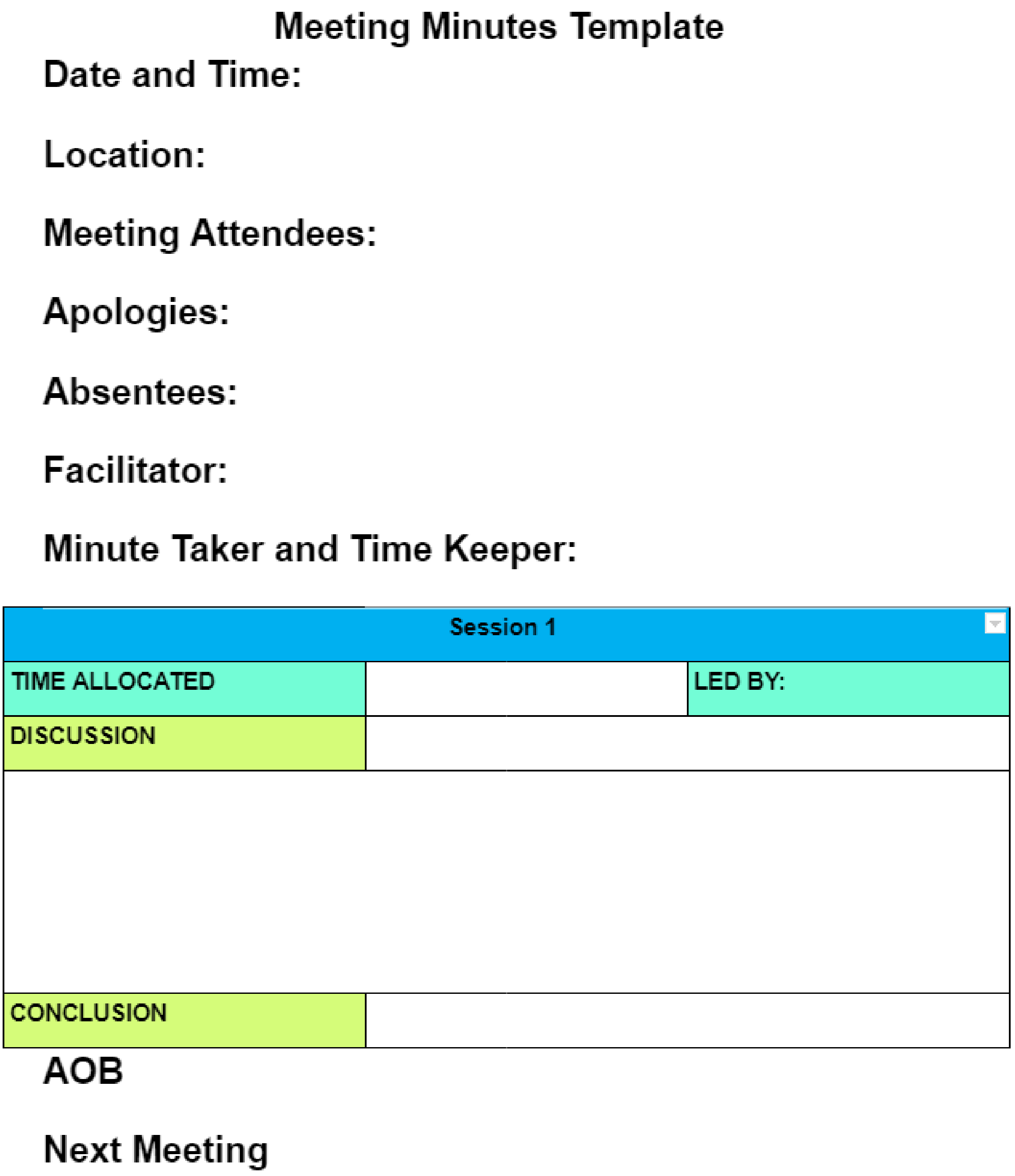


Risk Register for Computer Science Project

Prepared by: Yi Sen, Wai Han, Nawaaf, Yeonsoo   Date: 16/5/2022

No.	Rank	Risk	Description	Category	Triggers	Root Cause	Potential Responses	Risk Owner	Probability	Impact	Status	Score
1	1	Team members	Team members are not contributing to the assignment as they should	People risk	Team members unable to contribute to the assignment	Team member falls sick due to covid vaccination, contract with covid, absent due to family matters	Other members to share his workload	Nawwaf Ali	9	9		81
2	3	Cannot complete allocated tasks	Team members are not completing their allocated tasks as they should	People risk	Poor time management of individually allocated tasks	Ongoing project clashes with workloads of other course units	Pre-allocate the team's time for working on the project and plan and work on the project ahead of time	Chan Wai Han	5	5		25
3	2	Insufficient technical skills	Team members have never worked on a project like this before and lacks the skills required	Internal Project Management (Team)	Algorithm risks when providing information	Algorithm contains baised logic, inappropriate modelling techniques and coding errors	Ensure sufficient research is made before using the algorithm	Yeonsoo Kim	5	7		35
4	4	Insufficient soft skills	Team members are not very good when it comes to communicating with other team members	People risk	Commuication barrier	Idea conflicts between team members on project and lack of team monitoring may cause	Team members have set up communication solutions such as a Whatsapp group to keep tight track of the each other's work	Ooi Yi Sen	7	9		63

B MEETING MINUTES



Meeting Minutes Template

Date and Time:

Location:

Meeting Attendees:

Apologies:

Absentees:

Facilitator:

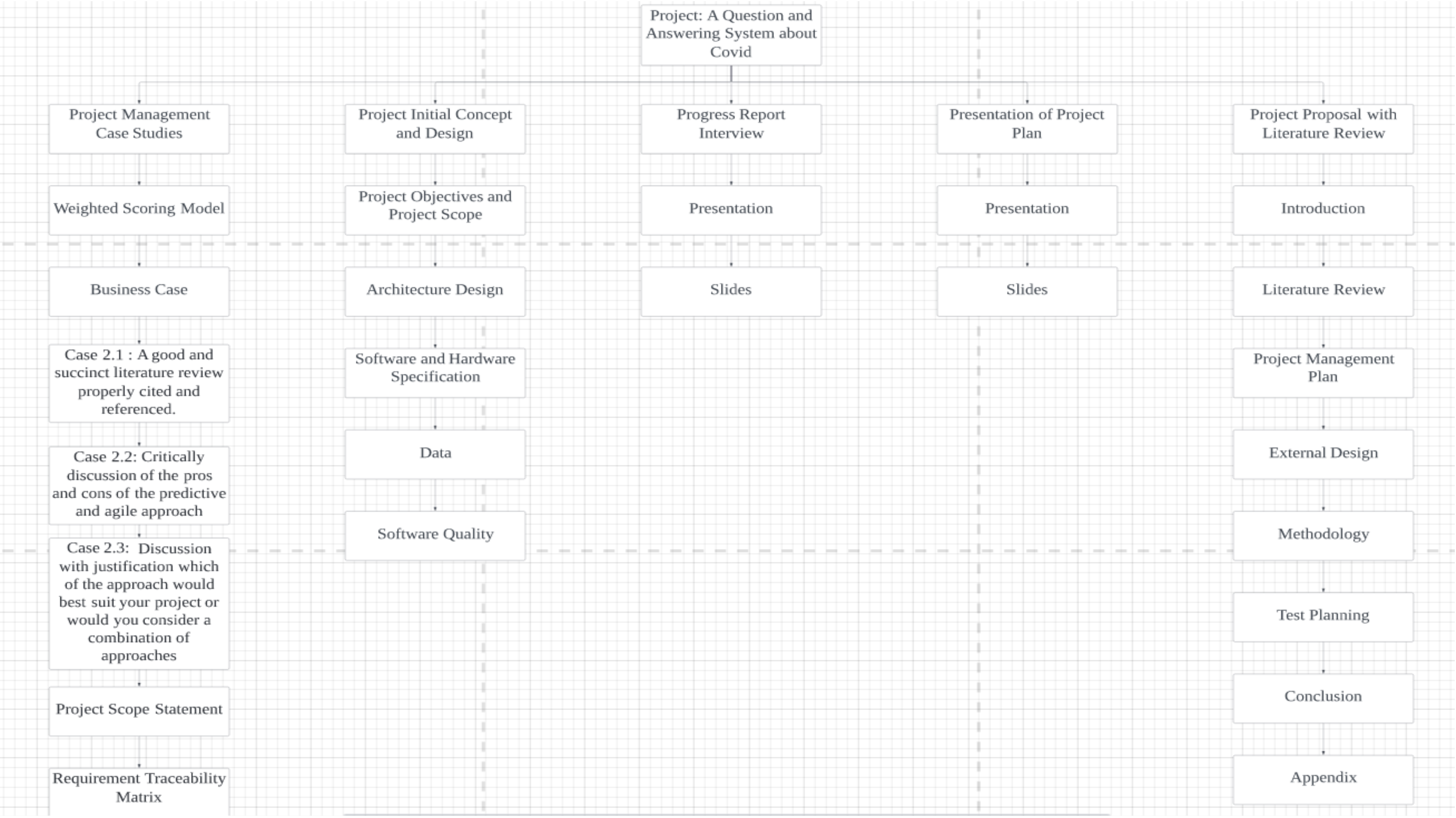
Minute Taker and Time Keeper:

Session 1		
TIME ALLOCATED		LED BY:
DISCUSSION		
CONCLUSION		

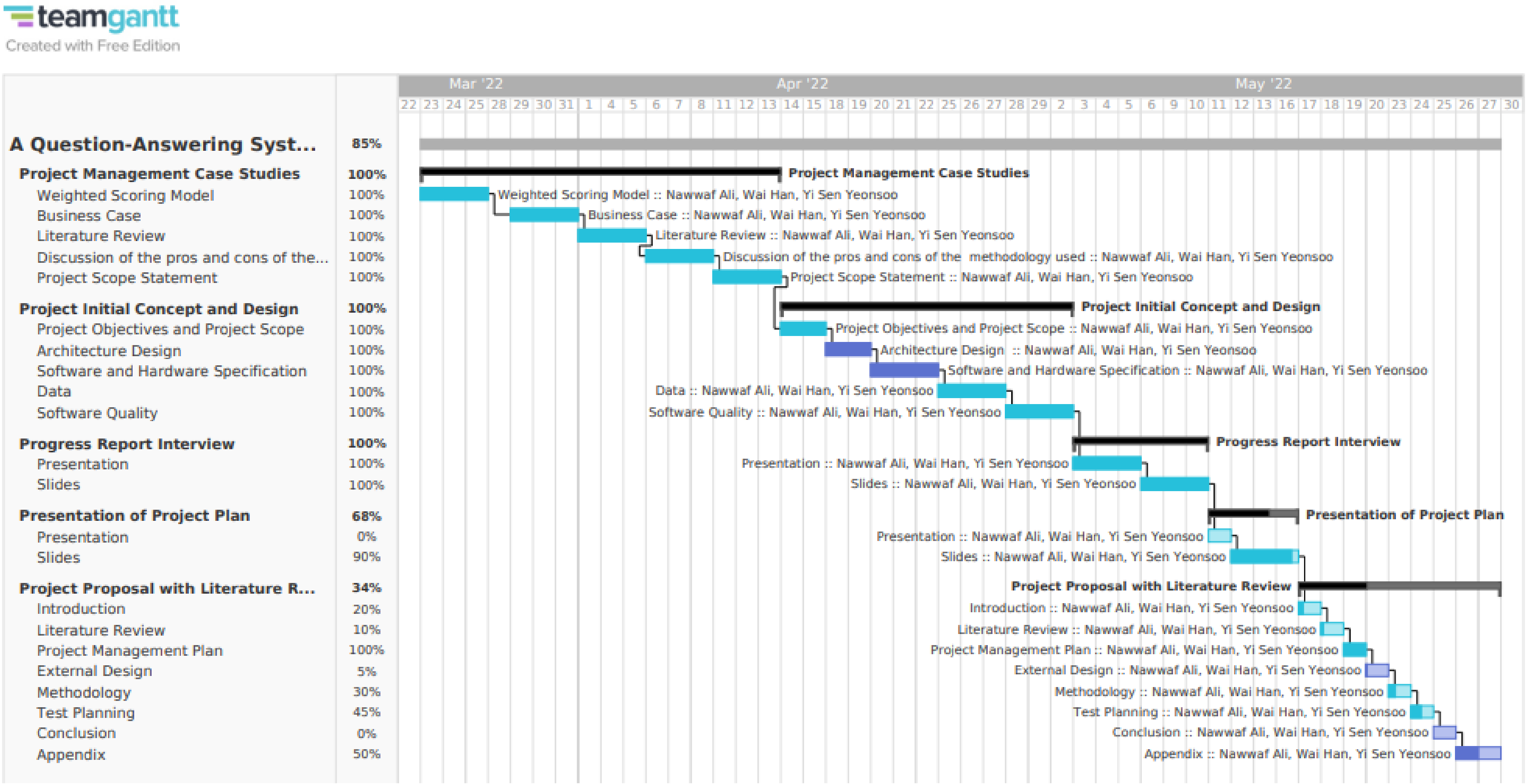
AOB

Next Meeting

C WORK BREAKDOWN STRUCTURE


Project: A Question and Answering System about Covid
Project Management Case Studies
Weighted Scoring Model
Business Case
Case 2.1 : A good and succinct literature review properly cited and referenced.
Case 2.2: Critically discussion of the pros and cons of the predictive and agile approach
Case 2.3: Discussion with justification which of the approach would best suit your project or would you consider a combination of approaches
Project Scope Statement
Requirement Traceability Matrix
Project Initial Concept and Design
Project Objectives and Project Scope
Architecture Design
Software and Hardware Specification
Data
Software Quality
Progress Report Interview
Presentation
Slides
Presentation of Project Plan
Presentation
Slides
Project Proposal with Literature Review
Introduction
Literature Review
Project Management Plan
External Design
Methodology
Test Planning
Conclusion
Appendix

D GANTT CHART

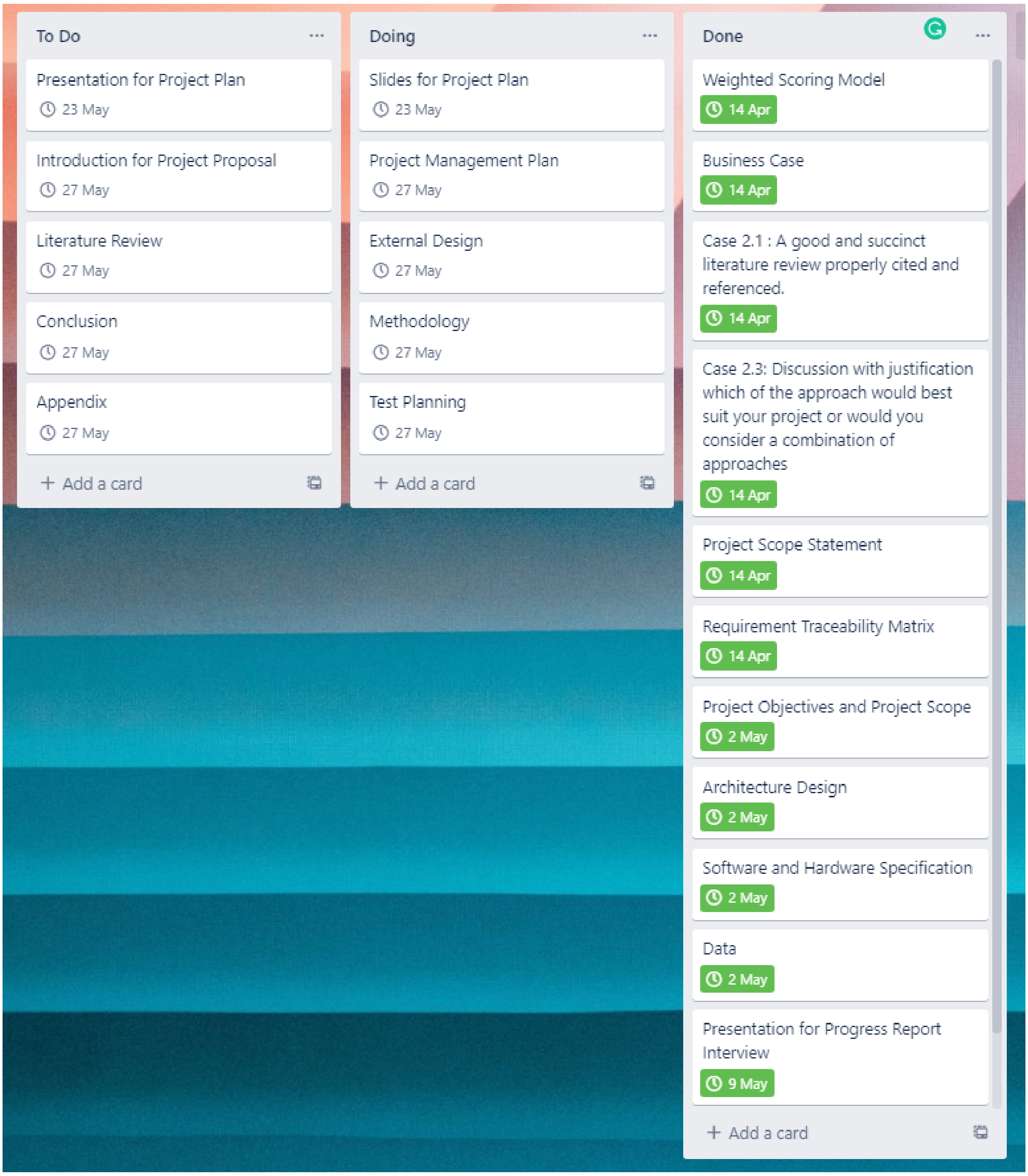


teamgantt
Created with Free Edition

A Question-Answering Syst...	**85%**
Project Management Case Studies	**100%**
Weighted Scoring Model	100%
Business Case	100%
Literature Review	100%
Discussion of the pros and cons of the...	100%
Project Scope Statement	100%
Project Initial Concept and Design	**100%**
Project Objectives and Project Scope	100%
Architecture Design	100%
Software and Hardware Specification	100%
Data	100%
Software Quality	100%
Progress Report Interview	**100%**
Presentation	100%
Slides	100%
Presentation of Project Plan	**68%**
Presentation	0%
Slides	90%
Project Proposal with Literature R...	**34%**
Introduction	20%
Literature Review	10%
Project Management Plan	100%
External Design	5%
Methodology	30%
Test Planning	45%
Conclusion	0%
Appendix	50%

Mar '22
Apr '22
May '22
22 23 24 25 28 29 30 31 1 4 5 6 7 8 11 12 13 14 15 18 19 20 21 22 25 26 27 28 29 2 3 4 5 6 9 10 11 12 13 16 17 18 19 20 23 24 25 26 27 30
Project Management Case Studies
Weighted Scoring Model :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Business Case :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Literature Review :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Discussion of the pros and cons of the methodology used :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Project Scope Statement :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Project Initial Concept and Design
Project Objectives and Project Scope :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Architecture Design :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Software and Hardware Specification :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Data :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Software Quality :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Progress Report Interview
Presentation :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Slides :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Presentation of Project Plan
Presentation :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Slides :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Project Proposal with Literature Review
Introduction :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Literature Review :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Project Management Plan :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
External Design :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Methodology :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Test Planning :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Conclusion :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo
Appendix :: Nawwaf Ali, Wai Han, Yi Sen Yeonsoo

E KANBAN BOARD



To Do

- Presentation for Project Plan
  23 May
- Introduction for Project Proposal
  27 May
- Literature Review
  27 May
- Conclusion
  27 May
- Appendix
  27 May

+ Add a card

Doing

- Slides for Project Plan
  23 May
- Project Management Plan
  27 May
- External Design
  27 May
- Methodology
  27 May
- Test Planning
  27 May

+ Add a card

Done

- Weighted Scoring Model
  14 Apr
- Business Case
  14 Apr
- Case 2.1 : A good and succinct literature review properly cited and referenced.
  14 Apr
- Case 2.3: Discussion with justification which of the approach would best suit your project or would you consider a combination of approaches
  14 Apr
- Project Scope Statement
  14 Apr
- Requirement Traceability Matrix
  14 Apr
- Project Objectives and Project Scope
  2 May
- Architecture Design
  2 May
- Software and Hardware Specification
  2 May
- Data
  2 May
- Presentation for Progress Report Interview
  9 May

+ Add a card

F RACI Matrix

Task \ Role	Project Manager	Technical Lead	Evaluation and Correction	Quality Assurance
	Ooi Yi Sen	Chan Wai Han	Nawwaf Ali	Yeonsoo Kim
Phase 1: Project Management Case Studies				
Weighted Scoring Model	R	A	C	I
Business Case	I	R	A	C
Literature Review	I	A	R	C
Pros and Cons of the predicteive and agile	A	C	I	R
Justification of approach	C	A	R	I
Project Scope Statement	A	R	I	C
Requirement Traceability Matrix	I	A	C	R
Phase 2: Project Initial Concept And Design				
Project Objectives and Project Scope	I	C	A	R
Architecture Design	C	A	R	I
Software and Hardware Specification	C	R	A	I
Data	R	I	C	A
Software Quality	I	R	A	C
Phase 3: Progress Report Interview				
Presentation	C	A	I	R
Slides	R	A	I	C
Phase 4: Presentation of Project Plan				
Presentation	C	A	I	R
Slides	R	A	I	C
Phase 5: Project Proposal with Literature Review				
Introduction	R	A	C	C
Literature Review	I	R	A	I
Project Management Design	I	A	R	C
External Design	I	I	I	R
Methodology	C	C	R	I
Test Planning	R	A	C	I
Conclusion	A	I	R	C
Appendix	C	A	I	R

9.0 References

Associationforum.org.(2013). *Performance Measurement & Metrics - Association Forum*.https://www.associationforum.org/mainsite/browse/professional-practice-statements/performance-measurement-metrics

Atlassian BitBucket(n.d) *What is version control? https://www.atlassian.com/git/tutorials/what-is-version-control*

Benítez-Andrades, José Alberto, Alija-Pérez, José-Manuel, Vidal, Maria-Esther, Pastor-Vargas, Rafael, & García-Ordás, María Teresa. (2022). *Traditional Machine Learning Models and Bidirectional Encoder Representations From Transformer (BERT)-Based Automatic Classification of Tweets About Eating Disorders: Algorithm Development and Validation Study. JMIR Medical Informatics*, 10(2), e34492–e34492. *https://doi.org/10.2196/34492*

Betterteam.(2022). *Software Project Manager Job Description*. https://www.betterteam.com/software-project-manager-job-description#:~:text=Software%20project%20managers%20are%20in,free%20trial%2C%20no%20card%20required.

Betterteam.(2022). *Technical Lead Job Description*. https://www.betterteam.com/technical-lead-job-description#:~:text=Technical%20leads%20take%20charge%20of,and%20ensure%20overall%20client%20satisfaction.

Black Box Testing. (n.d.). Imperva. https://www.imperva.com/learn/application-security/black-box-testing/

Cooper, R. J., & Ruger, S. M. (2000, November). *A simple question answering system. In TREC.* http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.1041&rep=rep1&type=pdf

Dwivedi, S. K., & Singh, V. (2013). *Research and reviews in question answering system. Procedia Technology*, *10*, 417-424. https://reader.elsevier.com/reader/sd/pii/S2212017313005409?token=E5102E56714A39CA5D993001AD15B7E8F73DACD1481AD7EE5FE938CA9CC4B26083219D554B32280ABEDB6186066AC9C3&originRegion=eu-west-1&originCreation=20220525181349

Epidemic QA at TAC 2020. (n.d.). *Eqidemic QA at TAC 2020.* https://bionlp.nlm.nih.gov/epic_qa/

Equivalence Paritioning Technique (n.d.). Javatpoint. https://www.javatpoint.com/equivalence-partitioning-technique-in-black-box-testing

Ghazaryan, M.(2017). *Brainstorming as a Risk Identification Technique for Scrum*. https://www.macadamian.com/learn/brainstorming-as-a-risk-identification-technique-for-scrum

Gustavsson, T. (2016). *Benefits of Agile Project Management in a Non-Software Development Context : A Literature Review. Project Management Development –*

Practice and Perspectives : Fifth International Scientific Conference on Project Management in the Baltic Countries, CONFERENCE PROCEEDINGS, 114–124. http://kau.diva-portal.org/smash/record.jsf?pid=diva2%3A1266099&dswid=-4055

Sabharwal, Navin, & Agrawal, Amit. (2021). *Hands-On Question Answering Systems with BERT*. Apress L. P. https://link.springer.com/chapter/10.1007/978-1-4842-6664-9_5#citeas

Hamilton, T. (2022). *Code Coverage Tutorial: Branch, Statement, Decision, FSM*. Guru 99. https://www.guru99.com/code-coverage.html#7

Hamilton, T., 2022. *Software Quality Assurance(SQA): Plan, Audit & Review*. Guru99. https://www.guru99.com/software-quality-assurance-test-audit-review-makes-your-life-easy.html.

Hamilton, T. (2022). *What is BLACK Box Testing? Techniques, Example & Types*. Guru 99. https://www.guru99.com/black-box-testing.html

Hamilton, T. (2022). *What is WHITE Box Testing? Techniques, Example & Types*. Guru 99. https://www.guru99.com/white-box-testing.html

Infoentrepreneurs.org.(n.d). *Manage risk*. https://www.infoentrepreneurs.org/en/guides/manage-risk/.

KENTON, W.(2021). *How SWOT (Strength, Weakness, Opportunity, and Threat) Analysis Works*.https://www.investopedia.com/terms/s/swot.asp.

K. Kaur, A. Jajoo and Manisha, *"Applying Agile Methodologies in Industry Projects: Benefits and Challenges,"* 2015 International Conference on Computing Communication Control and Automation, 2015, pp. 832-836. https://doi.org/10.1109/ICCUBEA.2015.166

Lalithnarayan, C. (2020, December 30). *An Introduction to Question Answering Systems.* Engineering Education (EngEd) Program | Section. https://www.section.io/engineering-education/question-answering/

Lin, K.(2016, Oct 19) *Identifying great User Interfaces. https://medium.theuxblog.com/identifying-great-user-interfaces-1a34545ef70c*

Michael, P. (n.d.). *7 Best Budget Laptops for Computer Science Students | 2022.* Media Peanut. *https://mediapeanut.com/best-budget-laptops-for-computer-science-students/*

MOSCHITTI, Alessandro, & QUARTERONI, Silvia. (2011). *Linguistic kernels for answer re-ranking in question answering systems. Information Processing & Management, 47(6), 825–842. https://reader.elsevier.com/reader/sd/pii/S0306457310000518?token=B88315AC26AEE377540C92D62DD61BF3DE6908099C0F0F17E03929EC5B64B635FB8AC5BF74893D2178D6ABC0110295E6&originRegion=eu-west-1&originCreation=20220526194409*

Olivia Montgomery, P. and Kumar,R., 2020. What Is a RACI Chart? Here's Everything You Need To Know.Top Business Software Resources for Buyers - 2022 | Software Advice. https://www.softwareadvice.com/resources/what-is-a-raci-chart/.

Pundge, A. M., Khillare, S. A., & Mahender, C. N. (2016). Question answering system, approaches and techniques: a review. *International Journal of Computer Applications*, *141*(3), 0975-8887. https://www.academia.edu/25420092/Question_Answering_System_Approaches_and_Techniques_A_Review?auto=citations&from=cover_page

Qualitance. (2014). *Black Box Techniques*. Qualitance. https://qualitance.com/blog/black-box-techniques/#:~:text=Equivalence%20partitioning%20is%20a%20black,classes%20for%20an%20input%20condition.

Random Testing. (n.d.). Tutorials Point. https://www.tutorialspoint.com/software_testing_dictionary/random_testing.htm

Ryan, J. Michael, & Nanda, Serena. (2022). COVID-19. Taylor & Francis Group. https://monash.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=TN_cdi_proquest_ebookcentral_EBC6863052&context=PC&vid=MUMUI&lang=en_US&search_scope=Combined-everything&adaptor=primo_central_multiple_fe&tab=allmum_tab&query=any,contains,Covid-19&offset=0

Serrador, P., & Pinto, Jeffrey K. (2015). *Does Agile work? — A quantitative analysis of Agile project success. International Journal of Project Management*, 33(5),1040–1051.https://doi.org/10.1016/j.ijproman.2015.01.006

Špundak, M. (2014). *Mixed Agile/Traditional Project Management Methodology – Reality or Illusion?,Procedia - Social and Behavioral Sciences,Volume 119,2014, Pages 939-948,ISSN 1877-0428. https://doi.org/10.1016/j.sbspro.2014.03.105.*

Statistical pattern recognition. (2019, September 19). Pattern Recognition Tools. https://37steps.com/189/statisticalpr/

Stoica, M., Mircea, M., & Ghilic-Micu, B. (2013). *Software Development: Agile vs. Traditional. Informatica Economica, 17*(4), 64-76. https://www.proquest.com/scholarly-journals/software-development-agile-vs-traditional/docview/1492882301/se-2

Tutorialspoint.com. n.d. *SDLC - Waterfall Model.*: https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.html.

Types of User Interface. (n.d.) Alan Blog. https://alan.app/blog/types-of-user-interface/#graphicaluserinterface

What is Software Testing? (n.d.). IBM. https://www.ibm.com/topics/software-testing#:~:text=Software%20testing%20is%20the%20process,Test%20management%20plan

White Box Testing: A Complete Guide With Techniques, Examples, & Tools. (2022). Software Testing Help.https://www.softwaretestinghelp.com/white-box-testing-techniques-with-example/#:~:text=Path%20coverage%20tests%20all%20the,for%20testing%20the%20complex%20programs.

Wrike.com.(n.d). *What Is Work Breakdown Structure in Project Management?*. https://www.wrike.com/project-management-guide/faq/what-is-work-breakdown-structure-in-project-management/#:~:text=Work%20breakdown%20structure%20(WBS)%20in,a%20large%20project%20more%20manageable.

Yan. (2020). *Detail the path coverage and its advantages and disadvantages of the logical coverage of white box testing*. CodeTD. https://www.codetd.com/en/article/11046704

10.0 Team Member’s Contribution Annex

Task	**Member’s Contribution**
Introduction	Nawwaf Ali
Literature Review on Q&A System	Nawwaf Ali, Yeonsoo Kim
Project Overview	Wai Han
Project Scope	Yeonsoo Kim
Project Organisation	Yi Sen
Management Process	Yi Sen
Schedule and Resource Requirements	Yi Sen
External Design	Yeonsoo Kim
Methodology	Nawwaf Ali
Test Planning	Wai Han
Conclusion	Wai Han