SYQ1

- Good report!
- Update the verb tenses and the GANTT chart

1 Introduction

1.1 Overview

Purpose of Legal Contracts

In the modern world, contracts are ubiquitous and covers practically every aspect of our daily lives, ranging from things on a personal level such as employment and mortgages to matters of a national level such as international treaties and agreements. Contracts are enforceable by law, contracting parties that break the agreement can be held accountable, hence, are obligated to fulfill their promises and agreements. The concept of ~~contractarianism behind legal contracts~~ keeping our promises allows us to put our trust and faith in each other, enabling us to treat each other with respect and dignity. It is the means by which a healthy and ordered society thrives in ~~what would otherwise be~~ chaotic anarchy. The alternative is Contracts merely add an element of accountability and serve as witnesses to promises made,

The growing threat of Cyber Frauds

Ever since the outbreak of COVID-19, the number of cyber fraud cases has been skyrocketing. According to the South China Morning Post, within half a year after the pandemic outburst, the Hong Kong Police had intercepted over 3 billion HKD scammed from victims that had been conned through the internet and phone, which was an increase of 150% in numbers compared to the entirety of its previous year [citation from SCMP]. And the number of cases has only been increasing ever since, the number of scams recorded in Hong Kong in 2021 was 124% of its previous year [1], and in just the first half of 2022, the number of deception cases has increased from 8,699 cases to 12,326 cases, a 41.7% increase in numbers [2], ~~and of which~~ Furthermore, 70% of the cases in both years combined ~~are~~ were internet related.

Among the various types of cyber fraud, contract scam is one of the most prominent types in our society, and it can happen in job employment, business deals, ~~and so on~~ etc.

People with less legal knowledge are much easier targets for frauds, as legal terms and conditions are often confusing in longer contracts, and with misrepresentations, it may be easier for ~~frauds~~ fraudsters to persuade or coerce these people into signing the contract. Worst still, some may not even bother to read through a contract before signing, a low level of awareness is what enables scammers to pull off their schemes.

split long sentence

Legal Contract Chatbot

Due to the drastic increase in deception cases in recent years, we ~~will be~~ are developing a software application that primarily focuses on ~~solving~~ helping people identify and avoid contract-related scams.

We ~~will~~ are develop ing a mobile application and a website that allows users to get a better understanding of a contract, be it a physical or a digital document. Users can ask for clarifications on unclear concepts, and the chatbot will give unbiased, factual, and logical answers to every question asked. Hence, it is essentially a tool that facilitates ~~users in~~ the understanding ~~a~~ legal contract, and it is also a means to counter fraud ing.

1.2 Objectives

The goal of this project is to build a "Legal Chatbot" system for legal consultation that enables users to find general legal knowledge and customized legal documents at any time.

The main objectives to achieve the goal are:

1. **Develop a Data Extraction System to extract useful data from files.**
2. **Develop a Preprocessing System to organize a dataset for subsequence training.**
3. **Fine-tune and train the GPT-2 Model to predict the next word of a given input conversation string.**
4. **Develop a Dialogue System to maintain the conversation flow.**
5. **Develop a database that will store the information of users.**
6. **Develop a user interface (UI) for users to have consultations with the chatbot.**

The biggest challenge we have face have been extracting the data from files in different formats and training GPT-2. To extract the data, we studied different tutorials on the internet and tried our best to maximize the availability of different file formats. Regarding the training of GPT-2, we have put much effort into doing research about machine learning and we have had many meetings with MPhil students to have a concrete and solid understanding of the aspects of machine learning.

1.3 Literature Survey

The human language is very intricate, for instance, there are countless ways to order words in a sentence, words can have different meanings, and each language has its own peculiarities and ambiguities. Therefore, we will be incorporating natural language processing (NLP), a branch of artificial intelligence that enables computers to comprehend and synthesize human texts at high efficacy, into our chatbot.

Building an NLP model from scratch is not only difficult but also cost-demanding and time-consuming as it requires a large data set for training before it becomes usable. Hence, a common approach to integrating NLP features into software and applications is fine-tuning existing pre-trained NLP models to specialize in a particular field. Below, we introduce two famous NLP models that are possible candidates for our project. Also, in order to understand chatbots better, two commonly used chatbots will be discussed.

1.3.1 Popular Models in NLP

GPT2 and BERT are both models that are state-of-the-art in the NLP field. Starting with the former, Generative Pre-Trained Transformer 2 (GPT-2) is a large transformer-based language model developed by OpenAI, the model has 1.5 million parameters and is trained on a dataset consisting of 8 million web pages [3]. The objective of GPT2 is to predict the next word of a sentence or phrase and has shown its capability of generating synthetic texts sample of unheard-of quality. This model outperforms most other language models, in that it does not require domain-specific training datasets such as Wikipedia, news, or literature and is beginning to learn language tasks including reading comprehension, translation, and summarizing without the aid of task-specific training data [3].

BERT stands for Bidirectional Encoder Representations from Transformers, it was developed by Google AI Language researchers that are used for common language tasks including named entity recognition and sentiment analysis. Similar to GPT2, BERT has an enormous training dataset of 3.3 billion words, of which around 2.5 billion words were contributed from Wikipedia and around 800M words from Google's BooksCorpus [4]. Contrary to GPT2, BERT uses a masked language model (MLM), a new approach for pre-training an NLP. MLM

2. Methodology

2.1 Design

The Design Phase of the project started in mid-August, and ~~we will continue working on~~ it has included the following aspects:

2.1.1 Gathering Relevant Documents

In order to maintain the reliability, usefulness, relevancy, and unbiasedness of the chatbot, we have to design the QA dataset carefully and have consultations with professionals instead of gathering second-hand information from online directly. We decided to use the existing dataset which has around 2 million QA pairs provided by Prof Song. We will continuously work closely with the collaborator, Albert So, who chairs the Hong Kong Mediation and Arbitration Centre, so as to generate a more robust and diverse QA dataset (in terms of vocabulary and grammar) and provide sufficient challenge to the AI model which is conducive for AI to deal with unseen and complex inquiries in dialogue.

2.1.2 Design Data Extraction System

While most text documents are in Docx and TXT file format, to relieve the constraints, we may design a system that can also extract data from JPG and PDF files because some documents such as conversational and handwritten data may be screenshots. This design not only allows us to gather more valuable training data but also provides convenience to users since users can upload a wider range of files. The system will be able to check the extension of an input file and extract text from it using corresponding methods. For the traditional Docx and TXT files, the system will be able to extract text directly. For image files, the system will be able to parse the file using Optical Character Recognition (OCR) technology [10] which used deep learning together with computer vision to scan images and convert the information into machine-readable form. Due to the complexity of OCR, we will first focus on Docx-to-TXT conversion and try image-to-TXT conversion if time allows.

2.2 Implementation

The Implementation Phase will include the following aspects:

2.2.1 AI model Development

We will use Python to build the AI model not only because of its simplicity and consistency but also its great support in machine learning in terms of size of community and availability of libraries. Open-source libraries such as transformers will be used for training a GPT-2 model. To extract text from text files, the python-docx library will be used. OCR libraries such as PyPDF2 or PyTesseract will be studied for image-to-text conversion if the progress of our project runs smoothly.

2.2.2 App Development (Frontend)

We will use react native or flutter to develop the app interface since these two cross-platform programming languages allow the chatbot app to be used in ios and android devices at the same time. Apart from that, we can use the packages of these two programming languages to make the chatbot app more comprehensive. For example, we can use the firebase cloud messaging of flutter to notify the user about the messages sent by the chatbot.

2.2.3 App Development (Backend)

We will use PHP language to develop the backend of the chatbot app which we will mainly develop the authentication API (eg. login, register, logout). Regarding the database, we will use MySQL to develop the database for storing the basic information of users and the feedback from users for future maintenance and improvement.

2 Project Planning

Good!

2.1 Distribution of Work

Task	Ming Kei	Ki Cheung	Chun On
Do the literature survey	○	○	●
Design the Data Extraction System	●	○	○
Design the Preprocessing System	●	○	○
Design the AI Model	●	○	○
Design the Dialogue System	●	○	○
Design the User Interface	○	○	●
Design the Database	●	○	○
Develop the Data Extraction System	●	○	○
Develop the Preprocessing System	●	○	○
Develop the AI Model	●	○	○
Develop the Dialogue System	●	○	○
Develop the User Interface	○	○	●
Develop the Database	○	●	○
Test the Data Extraction System	○	○	●
Test the Preprocessing System	○	●	○
Test the AI Model	○	●	○
Test the Dialogue System	○	●	○
Test the User Interface	○	○	●
Test the Database	○	●	○
Do the Integration	○	●	○
Write the reports	○	○	●
Prepare for the presentation	○	●	○
~~Design~~ the project ~~poster~~ video trailer	○	○	●

Make

● Leader ○ Assistant

3.2

~~2.2~~ GANTT Chart

Task	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Do the literature survey										
Analyze the GPT-2 and OCR technoglogy										
Design the Data Extraction System										
Design the Preprocessing System										
Design the AI Model										
Design the Dialogue System										
Design the User Interface										
Design the Database										
Develop the Data Extraction System										
Develop the Preprocessing System										
Develop the AI Model										
Develop the Dialogue System										
Develop the User Interface										
Develop the Database										
Test the Data Extraction System										
Test the Preprocessing System										
Test the AI Model										
Test the Dialogue System										
Test the User Interface										
Test the Database										
Do the Integration										
Write the reports										
Prepare for the presentation										
~~Design~~ the project ~~poster~~ video trailer										

Make

↑ Update the chart

4 3 Required Hardware & Software

4.1 Hardware

Personal Computer Windows / macOS	For the development and testing of our app
Smartphones Android / iOS	For the testing of our mobile app

4.2 Software

Github [citation]	To store and keep track of our work with version control
MySQL [citation]	For our database
JavaScript (React Native) [citation]	For our mobile and web app
Python [citation	For our AI model

5 References

Check the CT website for the new 2022 IEEE style

[1] The Government of the Hong Kong Special Administrative Region. (2022, Jan 27) "Law and order situation in Hong Kong in 2021 (with photo)," The Government of the Hong Kong Special Administrative Region - Press Releases. [Online]. Available: https://www.info.gov.hk/gia/general/202201/27/P2022012700577.htm . [Accessed: 10-Sep-2022].

[2] The Government of the Hong Kong Special Administrative Region. (2022, July 25) "Law and order situation in the first half of 2022" The Government of the Hong Kong Special Administrative Region - Press Releases. [Online]. Available: https://www.info.gov.hk/gia/general/202207/25/P2022072500503.htm [Accessed: 09-Sep-2022].

[3] OpenAI (2019, February 14) "Better Language Models and Their Implications,". [Online]. Available: https://openai.com/blog/better-language-models/. [Accessed: 10-Sep-2022].

[4] B. Muller, (2022, March 2) "BERT 101 State Of The Art NLP Model Explained," Huggin Face. [Online]. Available: https://huggingface.co/blog/bert-101#2-how-does-bert-work. [Accessed: 09-Sep-2022].

[5] J.V. Stegeren. (2020, August 4) "A comparison of GPT-2 and BERT,". [Online]. Available: https://judithvanstegeren.com/blog/2020/GPT2-and-BERT-a-comparison.html . [Accessed: 10-Sep-2022].

[6] A. Sharma. (2021, Aug 11) "10 Reasons No-Code AI Chatbot Is Going to Be Big in 2022,". Medium. [Online]. Available: https://chatbotslife.com/10-reasons-no-code-ai-chatbot-is-going-to-be-big-in-2022-a06289319f9e [Accessed:10-Sep-2022]

[7] J. Wouters. (2022, Aug 14) "WHICH ONE IS BETTER? ManyChat vs Chatfuel,". Chatimize. [Online]. Available: https://chatimize.com/manychat-vs-chatfuel/ [Accessed:10-Sep-2022]

[8] J. Wouters. (2021, Nov 9) "Botsify Review,". Chatimize. [Online]. Available:https://chatimize.com/reviews/botsify/ [Accessed:10-Sep-2022]

[9] Entrepreneur Store (2020, August 19) "Utilize Chatbots for any Purpose with This Custom Platform,". Entrepreneur. [Online]. Available:https://www.entrepreneur.com/growing-a-business/utilize-chatbots-for-any-purpose-with-this-custom-platform/354877 [Accessed:10-Sep-2022]

[10] Y. Perwej, S.A. Hannan, A.M.A.M. Asif, and A. Mane, "An overview and applications