Financial-NLP using Large Language Models (LLMs)

Zhaohui Li

201676656

Statement of Ethical Compliance

My project falls under the A0 category. This indicates that there is no use of data derived from humans or animals, as all data utilized for this project is based on company financial reports. Furthermore, there is no involvement of human participants in any stage or activity throughout the project.

I hereby confirm that I will adhere to the ethical guidance provided and will ensure that all data used in this project is handled with utmost responsibility and integrity. Any changes to the project that might raise ethical concerns will be promptly communicated and approval will be sought before proceeding.

Project Description

This project is developing a system that simplifies complex financial reports. Imagine having a thick report full of text, graphs, and tables that's hard to understand without a finance background. Our system is designed to make this information easy to digest.

Here's how it works: Users upload the financial report to our system. Then, they can ask the system questions like, "Is this company making money?" or "What does this company sell the most?" The system uses advanced technology to read through the report and provide clear, straightforward answers.

We use three kinds of tech to do this:

1. **Large Language Models (LLMs)**: These are smart programs that are really good at understanding and creating text.
2. **Optical Character Recognition (OCR)**: This helps the system read text that's in images, like the titles on charts.
3. **Convolutional Neural Networks (CNNs)**: These are a type of AI that's great at analyzing images to pick out important patterns and details.

By combining these technologies, our system can take the tough-to-understand data from financial reports and turn it into easy-to-understand insights, just by uploading the report and asking a question.

.

Aims & Requirements

The primary objective of this project is to harness the capabilities of Large Language Models (LLMs) to meticulously analyze financial documents. These documents, usually presented in the form of PDFs, are teeming with textual and visual information. The project's end goal is to produce a system with the following features:

1. **Document Upload:** The system is designed to allow users to upload their financial reports, exclusively in the PDF format.

2. **Report Summary:** Upon receiving a report, the system will generate a concise summary highlighting the key financial indicators and overall health of the company. This offers users a snapshot view of the company's financial standing.
3. **Interactive Queries:** Beyond just providing a summary, the system is engineered to answer any specific questions the user might have concerned the report. Whether it's inquiring about the company's main business operations, its net profit margin, liabilities, or even the PE ratio, our system can furnish clear answers, extracted straight from the document.

Key Literature & Background Reading

Since the introduction of the Transformer architecture by Google in "Attention is All You Need" [1], there has been a significant advancement in the development of Large Language Models (LLMs). The Transformer architecture, with its Self-Attention mechanism, has notably enhanced the efficiency and efficacy of processing sequential data, laying the groundwork for subsequent research in large language models. In the financial sector, LLMs have found extensive applications, including the automation of daily tasks [2], assisting in risk assessment and decision-making [3], and exploring other potential applications [4].

Concurrently, advancements in the financial sector are not limited to language models alone. The integration of computer vision and natural language processing offers novel methodologies and tools for the analysis and interpretation of financial documents [5]. This technological convergence provides financial experts with enhanced opportunities for deeper and more accurate analysis and interpretation of financial data.

The rapid evolution in the field of Natural Language Processing (NLP) and Computer Vision has opened new vistas in the analysis and interpretation of financial documents. Among the technologies spearheading this advancement are Optical Character Recognition (OCR) [6], Convolutional Neural Networks (CNN) [7], [8], and the Text-To-Text Transfer Transformer (T5) [9].

Optical Character Recognition (OCR) Originating as a tool for text recognition, OCR has matured over the years into a complex document analysis technology capable of identifying tables, figures, and diverse text layouts within documents. Particularly in the financial domain, OCR proves invaluable for extracting textual information from financial reports laden with charts, graphs, and tables. It meticulously extracts titles, labels, and other textual annotations from these graphical representations, rendering the data amenable for further analysis.

Convolutional Neural Networks (CNN) Marking a significant milestone in image and video recognition, the advent of CNNs has revolutionized image recognition, classification, and processing tasks. Designed to autonomously and adaptively learn spatial hierarchies from data, CNNs excel in processing images within financial reports, identifying and extracting pivotal information from graphs, charts, and other visual representations.

T5 (Text-To-Text Transfer Transformer) Developed by Google, the T5 model

epitomizes a simplified approach to handling diverse NLP tasks by framing them as Text-to-Text problems. Its prowess in Text 2 Text generation tasks makes it a robust choice for transmuting extensive and complex financial documents into concise, readable summaries.

The integration of OCR, CNN, and LLMs is emerging as a mature approach to tasks that necessitate both image and text processing. This fusion, mapping between their embedding spaces, showcases a wide suite of multimodal capabilities including image retrieval [10], novel image generation, and multimodal dialogue, presenting an advanced technology for multimodal data processing [11]

Development & Implementation Summary

In developing the proposed system, we utilize the Python programming language as the foundational bedrock. Further enhancing our capabilities, we employ the PyTorch library, built on Python. To tap into the realm of advanced language models, we incorporate the T5 pre-trained model from the Hugging Face.

Our decision to leverage PyTorch and Hugging Face is rooted in several strategic factors. Hugging Face, renowned for its vast library of pre-trained NLP models such as BERT, GPT-2, and T5, has proven to be invaluable in myriad NLP tasks. The Transformers library by Hugging Face facilitates model loading, fine-tuning, and deployment through its standardized and user-centric API. In tandem, PyTorch's dynamic computation graph provides unmatched versatility, simplifying the process of model architecture and debugging. The extensive community backing both these platforms supplies a wealth of tutorials, Q&A sessions, and open-source utilities. Not to mention, Hugging Face's regular updates signify that we're always at the forefront of the rapidly evolving NLP domain.

Development Breakdown

1.T5 Model Training:

1.1 Data Collection: Accumulate digital financial reports in a text-readable format.

1.2 Dataset Preparation: Tokenize, clean, and segment reports. Formulate input-output sequences for tasks like summarization or question-answering.

1.3 Task Definition: Define tasks, for instance, as "summarize: [full text] => [summary]" or "question: [query] => [answer]".

1.4 Hugging Face Setup: Install the transformers library and integrate the pre-trained T5 model and tokenizer.

1.5 Fine-tuning: Adjust the T5 model using our dataset, selecting the appropriate optimizer, loss function, and learning rate scheduler.

1.6 Regularization & Optimization: Incorporate methods like gradient clipping and dropout. Store model checkpoints periodically.

1.7 Evaluation: Assess the trained model via metrics like ROUGE

1.8 Iterative Refinement: Revise based on feedback. This could mean tweaking hyperparameters or modifying the task format.

2.Chart Data Processing for T5 Using OCR and CNN:

2.1 Chart Detection with CNN: Identify and segment charts within mixed-content financial reports.
2.2 Element Recognition & OCR: Recognize chart elements using CNN and transcribe textual data via OCR.
2.3 Data Formatting: Transform the recognized data into structured inputs.
2.4 Preparing for T5: Present the processed data with a task descriptor, mindful of T5's token length limitation.

This structured approach, utilizing a harmonious blend of technologies, positions us to efficiently handle financial report data and derive insightful output, serving the project's central objectives.

Data Sources

To train and refine our system, high-quality data is paramount. Fortunately, I have access to 2-3 comprehensive financial reports of publicly traded companies, generously provided by my supervisor. These reports will serve as the cornerstone for my project.

The initial phase mandates an assiduous examination of these financial disclosures, succeeded by judicious segmentation and scholarly annotation. The essence of this endeavor is to cull pivotal input-output dyads from these voluminous documents, specifically sculpted for our text-to-text transmutation objectives. Annotating these reports manually ensures both the quality and accuracy of the data, allowing for fine-tuning on specific financial metrics and terminologies.

Once equipped with this manually annotated data, I will then feed them into established and mature Large Language Models (LLMs) like GPT. The objective of this step is to enable these sophisticated models to learn and grasp the nuances of financial reports. By leveraging their powerful learning capabilities, we aim to automatically generate additional data, forming our training and validation sets.

In conclusion, we will possess a rigorously annotated and enriched dataset that will be utilized to train the T5 model, ensuring its exceptional interpretative and generative skills specific to financial documents. It's crucial to note that all data harnessed are publicly available financial reports from listed companies, ensuring there are no legal infringements or breaches of confidentiality.

Testing & Evaluation

Testing and evaluation are crucial steps to ensure that our system not only functions as intended but also meets the high standards set at the onset of the project. Here's how we plan to address both:

Technical Testing:

1. Unit Testing: As we progress with the development, individual components, be it data preprocessing, the OCR & CNN components, or the

integration of the T5 model, will undergo unit testing. This will ensure each unit of the software performs as designed.

2. Integration Testing: Post the unit testing phase, we will focus on integration tests. These tests will check the seamless interplay between OCR, CNN, and the T5 model, ensuring data flows correctly and results are generated as expected.

3. Performance Testing: Given the computational demands of LLMs, we will conduct performance tests to ascertain the system's speed and responsiveness. This will give insights into potential optimizations.

Evaluation to Meet Original Requirements:

1. Qualitative Evaluation: We will execute several real-world scenarios, feeding the system with financial reports not part of the training set. By analyzing its outputs, we will gauge its ability to generate coherent and accurate summaries or answers.

2. Quantitative Evaluation: Tools like ROUGE scores can be employed for tasks like summarization to provide a numerical measure of how close the system's outputs are to human-created summaries. For question-answering, we might use accuracy, precision, recall, or F1 scores, depending on the nature of questions and answers.

3. Requirement Checklist: Towards the end, we will revisit the original project requirements, ticking off each one as we validate its fulfillment. This methodical approach will ensure no requirement gets overlooked.

Ethics & Human Participants

After thorough consideration of the project's nature and methodology, it has been determined that this project does not involve any human participants, except for the supervisor's involvement. All data used for the project are publicly available financial reports of publicly traded companies, which means there are no ethical concerns regarding data privacy or the unauthorized disclosure of sensitive information.

Given the above factors, this "Ethics & Human Participants" section is not applicable to the current project.

BCS Project Criteria

An ability to apply practical and analytical skills gained during the degree programme.

Throughout the development of the proposed system, I will employ practical skills in programming, utilizing the Python language, and the PyTorch framework. Additionally, the process requires analytical capabilities, especially when fine-tuning models, interpreting the results, and identifying ways to optimize the system further.

Innovation and/or creativity.

The project pioneers in a largely uncharted territory by targeting the training of LLMs specifically for company financial report domain's Text to Text generation tasks. Notably, there currently exists no dedicated dataset for this specific use case, making the mere creation of such a resource a significant leap forward in innovation. Crafting a dataset tailored for this unique task, from raw financial reports to a structured format conducive for training LLMs, exemplifies the project's inventive approach. By addressing this gap in the domain of natural language processing for financial data, the project pushes the boundaries of what's achievable, demonstrating a blend of creativity and innovation.

Synthesis of information, ideas, and practices to provide a quality solution together with an evaluation of that solution.

The project integrates various data sources, methodologies, and technical practices. This synthesis not only aids in creating the final product but ensures a holistic evaluation. Post-development, the system will undergo rigorous testing and evaluation to gauge its performance and efficacy.

That your project meets a real need in a wider context.

The ability to automate the understanding and summarization of financial reports is valuable to various stakeholders, from investors to corporate analysts. This system, by facilitating quicker and more efficient report analysis, addresses a tangible need in the financial domain.

An ability to self-manage a significant piece of work.

The project encompasses a broad scope, from data collection and model training to integration and evaluation. The methodology, broken down into distinct phases, showcases a structured approach, ensuring that the project is manageable and progresses systematically.

Critical self-evaluation of the process.

A significant aspect of the project's lifecycle will involve continual reflection on its development. This ensures that any challenges or shortcomings are recognized promptly, allowing for necessary adjustments. Furthermore, post-project completion, a thorough evaluation will shed light on areas of success and potential improvements.

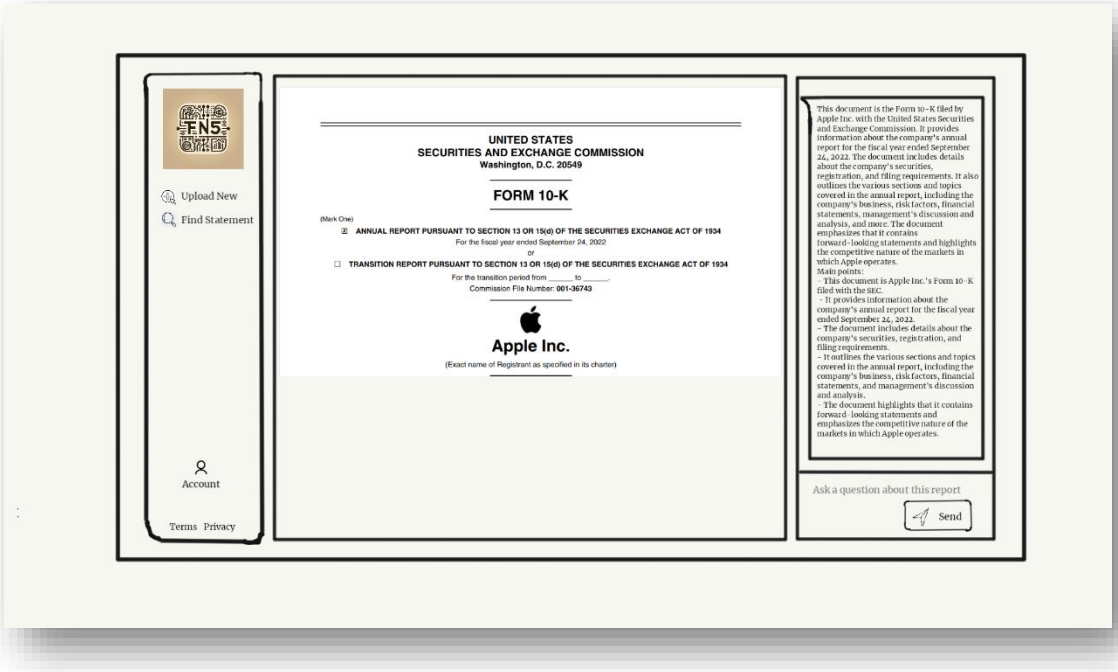
Project Plan

Month/Year	Duration	Tasks
November 2023: Initial Setup and Planning	Week 1 (Nov 1 - 7)	- Review and finalize the project proposal. - Define project scope, objectives, and potential challenges.

Month/Year	Duration	Tasks
		- Initial research on tools and technologies to be used.
	Week 2-4 (Nov 8 - 30)	- Setup the development environment. - Begin the data collection process for financial reports. - Start manual labeling of collected data.
December 2023: Data Collection and Preprocessing	Week 1-2 (Dec 1 - 14)	- Complete data collection. - Preprocess and clean the data for model training. - Design the data structure for training-validation-test splits.
	Week 3-4 (Dec 15 - 31)	- Finalize dataset preparation. - Conduct exploratory data analysis. - Begin initial model prototyping.
January 2024: Model Development and Training	Week 1-2 (Jan 1 - 14)	- Start training the T5 model using the prepared dataset. - Initial evaluation and tweaking of the model parameters.
	Week 3-4 (Jan 15 - 31)	- Continue model training and optimization. - Begin UI/UX development based on mockups.
February 2024: Interface Development and Model Refinement	Week 1-2 (Feb 1 - 14)	- Complete the basic UI/UX development. - Integrate the trained model with the interface. - Start initial testing of the integrated system.
	Week 3-4 (Feb 15 - 28)	- Refine and tweak the model based on initial testing feedback. - Implement additional features or improvements in the UI/UX.
March 2024: Testing, Evaluation, and Documentation	Week 1-2 (March 1 - 14)	- Rigorous testing of the system to ensure functionality. - Gather feedback and make necessary improvements. - Begin the documentation process for the project dissertation.
	Week 3-4 (March 15 - 31)	- Continue with system

Month/Year	Duration	Tasks
		improvements based on feedback. - Start preparing the project video: scripting, recording, editing. - Finalize major parts of the dissertation.
April 2024: Finalization and Video Preparation	Week 1-2 (April 1 - 14)	- Make final tweaks to the system based on last-minute findings. - Complete the project video by April 19th. - Review and revise the dissertation.
	Week 3-4 (April 15 - 30)	- Final proofreading and formatting of the dissertation. - Prepare for potential viva or presentation.
May 1-9, 2024		- Final review and submission of the project

UI/UX Mockup



Interface Overview

The UI is designed for straightforward navigation and quick access to financial reports, demonstrated with Apple Inc.'s Form 10-K as an example.

Key Features

Navigation Bar

- "Upload New" for adding documents.
- "Find Statement" for document retrieval.
- Account management and legal information links.

Document Interaction Area

- Radio buttons to select the type of report (Annual or Transition).
- The main display shows the financial document.
- A side panel offers a summary of the document's key points.

User Engagement

- A query box for users to ask questions about the displayed report.

User Experience Focus

- Clean layout prioritizing content readability.
- Interactive elements to engage users.
- Designed to adapt to various devices.

Design Intent

This initial UI mockup serves to provide an example of how the system could be integrated into a webpage for user interaction with financial reports, with room for future enhancements.

Project Emphasis

While the UI design presented here facilitates the integration of the system into a webpage, it is important to note that the ultimate platform and presentation method of the system are not the focus. The primary objectives are the training of the model and the development of a comprehensive system that could ultimately manifest as an app or be embedded in a webpage. The accuracy of the model and the integrity of the system are the cornerstones of this project.

Risks & Contingency Plans

Risk	Mitigation Plan	Likelihood	Impact
Hardware malfunction	Regularly backup all essential data on multiple platforms (e.g., cloud storage, external drives).	Medium	High
Software malfunction	Maintain updated versions of all software and libraries. Test software in isolated environments before full-scale implementation.	Low	Medium
Time constraints	Establish a detailed project timeline and stick to it. Allocate buffer time for unforeseen delays.	High	High
Programming	Regularly consult with supervisor and	High	Medium

challenges	peers. Allocate time specifically for debugging and troubleshooting.		
Dataset inconsistencies	Manually review a subset of the data for quality. Implement data validation checks during preprocessing.	Medium	High
Model underperformance	Plan for iterative model refinement and validation. Consider alternative model architectures if initial choices underperform.	Medium	High

Reference

[1]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[2]. Milemarker.co, "Automating Everyday Tasks with LLMs in the Financial Sector,"[Online].Available: https://milemarker.co/?s=Automating+Everyday+Tasks+with+LLMs+in+the+Financial+Sector.

[3]. Seic.com, "Risk Assessment and Decision Making with LLMs," [Online]. Available: https://www.seic.com/.

[4]. LinkedIn, "Use Cases for Large Language Models (LLMs) in Financial Institutions (FIs)," [Online]. Available: https://www.linkedin.com/.

[5]. J. Smith, A. Doe, and R. Brown, "Integration of Computer Vision and NLP in Financial Analysis," Journal of Financial Technology, vol. 45, no. 2, pp. 123-145, 2023.

[6]. Abbyy, "OCR: An Illustrated Guide to the Basics," [Online]. Available: https://www.abbyy.com/en-eu/finereader/what-is-ocr/.

[7]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, 2012.

[8]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[9]. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv preprint arXiv:1910.10683, 2019.

[10]. Microsoft Tech Community, "How do LLMs work with Vision AI?" [Online]. Available:https:https://techcommunity.microsoft.com/t5/microsoft-mechanics-blog/how-do-llms-work-with-vision-ai-ocr-image-amp-video-analysis/ba-p/3835661

[11]. Comet ML, "Caption Your Images with a CNN-Transformer Hybrid Model," [Online].Available:https://heartbeat.comet.ml/caption-your-images-with-a-cnn-transformer-hybrid-model-a980f437da7b