Final Report regarding Benchmarking LLMs Project

Cong Chen (cc4887)
Longxiang Zhang (lz2869)
Ruolan Lin (rl3312)
Taichen Zhou (tz2555)
Yichen Huang (yh3550)

December 9, 2023

Contents

1 Introduction

The evolution of Large Language Models (LLMs) represents a monumental shift in the landscape of artificial intelligence. These models, equipped with billions of parameters, have transformed from mere tools of text prediction into sophisticated engines capable of understanding, generating, and even reasoning with human-like prowess. The applications of LLMs have permeated numerous sectors, but their potential in the financial industry stands out as particularly promising—and challenging.

Financial data, by its nature, is intricate, nuanced, and laden with domain-specific jargon. The implications of decisions made based on this data are profound, affecting everything from individual investments to global economic trends. As LLMs are increasingly tasked with roles such as document summarization, risk assessment, and regulatory compliance assistance, the margin for error narrows. It is not enough for an LLM to produce coherent content; it must also be factually accurate, contextually relevant, and devoid of misconceptions.

Recognizing this challenge, Fidelity Investments has embarked on an ambitious initiative. The goal is clear yet complex: to craft an evaluation framework for LLMs tailored to the intricacies of the financial sector. Such a framework seeks to move beyond traditional metrics of text coherence or fluency. It delves deeper, probing for correctness in the face of financial terminologies, sensitivity to nuanced prompts, and the LLM's ability to reason without generating content that sounds plausible but is factually incorrect or baseless.

Our collaboration with Fidelity Investments, facilitated by the Capstone Project initiative, positions us at the forefront of this endeavor. We are not just engineers or data scientists in this journey; we are pioneers, navigating the confluence of finance and AI. Our mandate is to ensure that as financial institutions increasingly rely on LLMs, they do so with a clear benchmark of reliability and accuracy.

This report meticulously documents our journey through this project. It sheds light on our chosen methodologies, elucidates the reasoning behind our decisions, and outlines the frameworks we developed. Detailed are the outcomes for each model, the challenges encountered along the way, and our planned next steps. As we delve deeper into this endeavor, we are ever-aware of its wider significance: setting a benchmark in LLM evaluations that extends its influence beyond the financial sector. Our goal is to ensure that as industries increasingly adopt AI-driven solutions, they do so with a foundation of well-informed trust and meticulousness.

2 Phase I: Initial Setup and Problem Definition

The first step was decisive when we embarked on this ambitious project to design an assessment framework for lifelong learning programs in the financial sector. This phase laid the foundation for the entire project. It was characterized by intense consultations, brainstorming sessions, and careful planning. It was imperative to align our vision, understand the nuances of the challenge, and establish a solid logistical framework before delving into technical details and specific methodologies. Initial consultations were at the heart of this preparatory phase, during which the contours of the project began to take shape.

2.1 Logistical Setup

2.1.1 Tool Selection

In today's dynamic data science landscape, the strategic selection of tools critically shapes a project's path. With this in mind, our team made deliberate choices, synergizing programming languages, APIs, and natural language processing libraries. At the heart of our toolkit is **Python**, a perennial favorite in the data science and machine learning communities for its extensive library ecosystem, offering both flexibility and robust capabilities tailored to our project's needs. Enhancing our language processing tasks, we integrated tools like **OpenAI's API**, **nltk**, **Transformers**, and **spacy**. These libraries and APIs provide advanced functionalities for text analysis and language model interaction, essential for our project's focus on language understanding and processing.

2.1.2 Git and Version Control

To ensure seamless collaboration and maintain a traceable workflow in our data-driven project, we adopted Git, the premier distributed version control system, complemented by GitHub's platform capabilities. Our dedicated GitHub repository centralized all code and documentation. A structured branching strategy was employed, with feature branches for specific tasks, which, post rigorous review via pull requests, were merged into the main branch, ensuring the codebase remained organized and up-to-date. This approach, fortified by GitHub's tools for communication, planning, and documentation, established a robust, transparent, and efficient framework for our project's logistics.

2.2 Problem Definition and Understanding

2.2.1 Project description

Fidelity Investments proposes a project to develop an evaluation framework for LLMs in the context of the financial industry. The aim is to quantify key facets like correctness, sensitivity, and reasoning in LLM-generated content, focusing on financial and regulatory jargon.

2.2.2 Our Understanding

As we delved into the intricacies of this project, our primary objective crystallized: to architect a robust methodology that evaluates content generated by LLMs. This methodology wouldn't merely assess the content's coherence but would also rigorously gauge its accuracy and contextual relevance. To complement this, we envisioned an automated system, finely tuned to measure and score LLM outputs using a predefined set of metrics.

Venturing further into our aspirational goals, we aimed to develop a network graph that intricately maps the relationships between various entities. Such a visual aid would illuminate the intricate interplay between these entities, offering deeper insights into their interactions.

Our system's design was straightforward in terms of input and output. It would ingest content produced by an LLM and, in return, deliver a detailed evaluation of the LLM's response. This evaluation would encompass scores across pivotal dimensions, such as the content's correctness, its reasoning prowess, and the variation evident in the generated narratives.

2.3 Literature Review and Preliminary Research

2.3.1 Academic Paper

A comprehensive understanding of the current state of research on the evaluation of language models, especially with respect to their factual consistency, was essential for our project. We delved into several academic papers to gain insights into prevailing methodologies, challenges, and innovations. Three key papers stood out, providing foundational knowledge that informed our approach:

Factuality Enhanced Language Models for Open-Ended Text Generation by Nayeon Lee *et al.* from the Hong Kong University of Science and Technology and NVIDIA [1]. This paper highlighted the vulnerability of pretrained LMs in generating non-factual information. It introduced the *FACTUALITY PROMPTS* test set and metrics for assessing the factuality of LM outputs, noting the increased factual accuracy in larger models.

Evaluating Factuality in Generation with Dependency-level Entailment by Tanya Goyal and Greg Durrett from The University of Texas at Austin [2]. This research proposed a novel method for evaluating factuality at the level of dependency arcs. By focusing on the semantic relationships in generated content and their alignment with source data, it provided a granular approach to fact-checking.

Evaluating the Factual Consistency of Abstractive Text Summarization by Wojciech Kryściński *et al.* from Salesforce Research [3]. This paper underscored the limitations of current summarization assessment metrics, especially in terms of factual consistency. A weakly-supervised, model-based approach was proposed to verify the factual consistency of summaries and identify discrepancies with source documents.

2.3.2 Industry Examples

The world of academic research is a treasure trove of theoretical foundations and cutting-edge methodologies. Yet, the real test of these concepts lies in their appli-

cation within the industry, where unique challenges and practical constraints come into play. In the financial realm, this interplay between theory and practice is vividly exemplified by two industry giants: Bloomberg and Deutsche Bank. Both have ventured into the domain of LLMs, and their journeys offer invaluable insights.

Bloomberg, a name synonymous with financial data and analytics, has always been at the forefront of innovation [4]. Their exploration into LLMs was no exception. They embarked on a rigorous evaluation journey, scrutinizing LLM outputs across a gamut of tasks, both financial and general. Specific metrics, such as FPB, FiQA SA, Headline, and NER, became their compass, guiding them towards precise evaluations manifested in F1 scores and exact match accuracies. Notably, their internal Sentiment Analysis system became a beacon of their commitment to nuanced evaluations. This exhaustive approach culminated in the birth of *BloombergGPT*, a testament to their dedication to refining LLMs for financial expertise.

On the other side of the spectrum, Deutsche Bank, a stalwart in global banking, charted its unique path in the LLM landscape [5]. Their approach was holistic, viewing LLMs not just as computational entities but as tools with immense potential for the financial industry. Validation accuracy, real-world applicability, and tasks like document summarization became their milestones. Furthermore, they championed the cause of model interpretability and explainability, ensuring that their LLMs were not just black boxes but comprehensible entities. Their rigorous stance against adversarial inputs and their adaptability to dynamic financial scenarios underscored their vision. The outcome? A dedicated LLM named *Large Language Models in Finance*, embodying Deutsche Bank's ethos and commitment to excellence.

In synthesizing these industry narratives, one thing becomes clear: while academic research lays the groundwork, it's the industry's practical implementations that bring these concepts to life. The endeavors of Bloomberg and Deutsche Bank validate the significance of LLM evaluations, offering a lens into the challenges, triumphs, and nuances that academic corridors might sometimes overlook.

3 Data Wrangling and EDA

The primary objective of our project involves working with raw data extracted from final litigation files, which are primarily in PDF format. Our initial step in this endeavor was to download these PDF files and attempt to extract the text data contained within them. However, we encountered challenges when using the PyPDF2 package for text extraction, as the results were less than satisfactory. We observed numerous word mistakes and line mismatches in the extracted text, which posed a significant obstacle to our goal of creating robust evaluation metrics.

As a result, we turned our attention to Optical Character Recognition (OCR) technology. OCR, which stands for Optical Character Recognition, is a technology that facilitates the electronic or mechanical conversion of images containing typed, handwritten, or printed text into machine-encoded text. This process can be applied to scanned documents, photographs of documents, or even text embedded within images, such as signs and billboards in landscape photos or subtitles from television broadcasts. In our quest to enhance the quality of text extraction from PDF files, we adopted the Pytesseract library, which is a Python wrapper for Google's Tesseract-OCR engine. By applying Pytesseract to the PDF files, we were able to extract the raw text data more effectively than when using PyPDF2. The results obtained through this OCR approach exhibited improved accuracy and fidelity, addressing the word mistakes and line mismatches that had previously hindered our progress. This advancement marked a significant milestone in our project, as it enabled us to work with cleaner and more reliable text data.

After applying OCR to extract text from our documents, one of the most common issues we encountered was the presence of unnecessary line break characters, often represented as '\n'. These line breaks, while seemingly innocuous, could potentially impact the performance of our Large Language Model in the subsequent steps, particularly when generating summaries. To address this concern, we implemented a solution to manage and eliminate these extra line breaks, thus enhancing the overall quality of our text data for the subsequent summarization process.

Additionally, we recognized that the identifications appearing on every page, such as the example you provided, "\nCase 1:10-cv-11665 Document1 Filed 09/29/10 Page 2 of 18\n," could also adversely affect the quality of the generated summaries. These identifiers are often extraneous and carry no informational value for our summarization task. To remedy this issue, we leveraged regular expressions to automatically detect and remove these page identifiers whenever they appeared within the text files.

4 Initial Modeling and Framework Design

Before delving into an exploration of the framework that utilizes assisted large language models for assessing the factual accuracy of large language models, we first present a conventional approach for benchmarking purposes. Initially, we selected a range of traditional metrics commonly used to assess the similarity between two summaries, serving as baseline models. In this context, we opted for established metrics, including Recall-Oriented Understudy for Gisting Evaluation (ROUGE), cosine similarity, Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Bilingual Evaluation Understudy (BLEU).

During this phase, our dataset consisted of a limited set of 31 reference summaries, each matched with its corresponding source text. Given that we only have high-quality summaries crafted by humans for each source document, it becomes essential to generate some lower-quality summaries as well.

In the process of selecting large language models for generating summaries, our study involved evaluating three prominent models: 'facebook/bart-large-cnn,' sourced from Hugging Face; ChatGPT 4, a well-established model; and Claude, which is considered a potential "ChatGPT killer." All of these models were employed to produce summaries for the evaluation phase of our research.

4.1 Metrics Introduction

4.1.1 ROUGE

ROUGE is a metric that evaluates text summarization by comparing overlapping n-grams between generated and reference text. There are many types of ROUGE scores:

ROUGE-N: measures the number of matching n-grams between the generated text and the reference text.

ROUGE-L: evaluates the longest common subsequence between the generated text and the reference text.

We finally choose to use ROUGE 2 which focuses on matching pairs of adjacent words in the text. The reason is that this level of granularity can capture the relationships, such as some semantic and syntactic information, between words in generated summaries, making it suitable for assessing the fluency and coherence of the generated text. We also use F1 score to ensure both precision and recall are considered, which benefits in evaluating the quality of summaries without favoring one aspect over the other.

4.1.2 Cosine Similarity

Cosine similarity is a metric for measuring the similarity between two vectors, often used to assess the similarity between texts in a multi-dimensional space. It is very important to choose the embedding for Cosine Similarity since we need to embed each summary into vector first then calculate the cosine similarity value based on the vectors. Here we choose the classical TF-IDF embedding.

4.1.3 METEOR

METEOR is a machine translation evaluation metric that assesses the quality of machine-generated translations by measuring the similarity and fluency of the generated text compared to the reference text. It takes into account several factors, including unigram precision, recall, stemming, synonymy, and word order, to provide a more comprehensive assessment of translation quality.

4.1.4 BLEU

BLEU score is usually used alongside metrics like ROUGE and METEOR to assess the text generation. It emphasizes precision in n-grams between the generated text and the reference text.

4.2 Result Analysis

We employ conventional evaluation methods to assess the three aforementioned large language models. In the case of each model, we initiate the process by inputting source documents into the model to generate summaries. Subsequently, we conduct a comparative analysis between the generated summaries and manually crafted reference summaries by professionals. The evaluation scores are then calculated using established approaches. The results table is presented below.

Model	ROUGE2 F1 Score	Cosine Similarity	METEOR	BLEU
Facebook Bart	**0.294**	0.527	0.298	**0.162**
ChatGPT 4	0.235	**0.541**	**0.353**	0.161
Claude	0.213	0.499	0.291	0.113

In the presented table, we have exclusively considered the average scores obtained from the evaluation of all 31 generated summaries. An examination of these results reveals variations in the performance of different models across various evaluation methods. According to the ROUGE and BLEU metrics, our analysis consistently identifies 'Facebook Bart' as the model achieving the highest average score. This implies that 'Facebook Bart' exhibits superior performance in terms of summary quality according to these particular metrics. However, the landscape changes when employing the 'Cosine Similarity' and 'METEOR' methods. Under these metrics, 'ChatGPT 4' emerges as the prevailing model, yielding significantly higher METEOR scores in comparison to the other two models.

In addition, we have included box plots illustrating the performance of the three large language models for each score. These visual representations offer a more comprehensive perspective of the data. Upon examining the plots, we observe that the conclusions align with our previous findings. Specifically, the ROUGE graph underscores the superior performance of the 'Facebook Bart' model, while the METEOR graph highlights the significantly higher scores achieved by the 'ChatGPT 4' model in comparison to the other two. Furthermore, the Cosine Similarity and BLEU graphs indicate a relatively minimal distinction between the 'Facebook Bart' and 'ChatGPT

4' models.

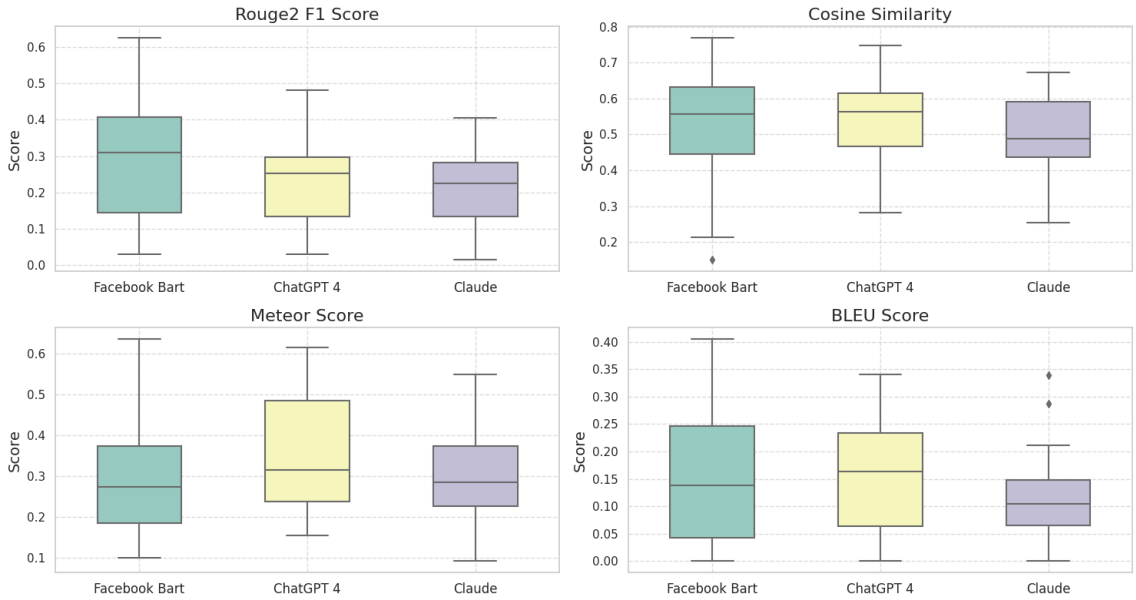


Figure 1: Box-plot for Scores

4.3 Limitations

Upon a comprehensive analysis of the implementation process and the examination of the results, it becomes evident that the baseline traditional approaches exhibit numerous limitations.

4.3.1 Data Gaps and Omissions

While a summary may receive a high score according to traditional approaches, it may still not qualify as a comprehensive summary due to the omission of vital information.

Reference Summary	Generate Summary	Score
FINRA found the firm failed to reasonably supervise for potentially manipulative trading from 2013 through 2019 and SageTrader also failed to establish and implement AML policies and procedures reasonably expected to detect and cause the reporting of suspicious activity and violated FINRA Rules 3310(a) and 2010. As a result, SageTrader consented to a censure and a $100,000 fine to FINRA. Additionally, SageTrader also had to pay a total fine of $775,000 which was paid to FINRA and eight exchanges, and the firm was also required to review and revise its supervisory system for detecting potentially manipulative trading by the firm's customers.	SageTrader has been a FINRA member since May 2006. FINRA found that the firm failed to reasonably supervise for potentially manipulative trading from 2013 through 2019. The AWC imposed a censure and a fine of $775,000 (paid to FINRA and eight exchanges) The firm considered two of the four customers to be high AML risk. The firm's market participant identifiers (MPIDs) were used to route orders to anAlternative trading system and to exchanges. SageTrader failed to establish and implement AML policies and procedures reasonably expected to detect and cause the reporting of suspicious activity.	0.769

For instance, consider the following example: We employed the Cosine Similarity method to assess a summary generated by the 'Facebook Bart' model. Despite achieving the highest Cosine Similarity score among all 93 generated summaries, this particular summary fails to encompass crucial details, such as FINRA Rules 3310(a) and the

$100,000 fine. These details hold significant importance within the context of the source document, particularly within the financial domain.

4.3.2 Threshold Determination Dilemmas

In the context of each traditional approach, the selection of an appropriate threshold is imperative to assess the quality of a summary. A threshold serves as a specific criterion that determines whether a summary is deemed satisfactory or not. To illustrate, in the case of the ROUGE2 score, designating a threshold value of 0.7 allows us to classify summaries as either "good" if their score exceeds 0.7 or "bad" if it falls below this threshold.

Nevertheless, the task of threshold selection presents considerable challenges. On one hand, despite the scores falling within the [0,1] range, their distributions exhibit significant disparities. For instance, the ROUGE2 score's highest 5% value threshold is 0.481, while the Cosine Similarity score's equivalent threshold is 0.721. Consequently, a simplistic approach of selecting a high threshold like 0.9 is infeasible due to the divergence in score distributions.

On the other hand, determining an appropriate threshold based on the distribution of scores for each traditional approach necessitates gathering scores from multiple models, which may introduce bias and one-sidedness. Additionally, reference summaries frequently approach the maximum score of 1 in most conventional methods. Therefore, selecting a threshold based on score distribution, even if relatively low, implies significant dissimilarity between the considered "good" summaries and the reference itself.

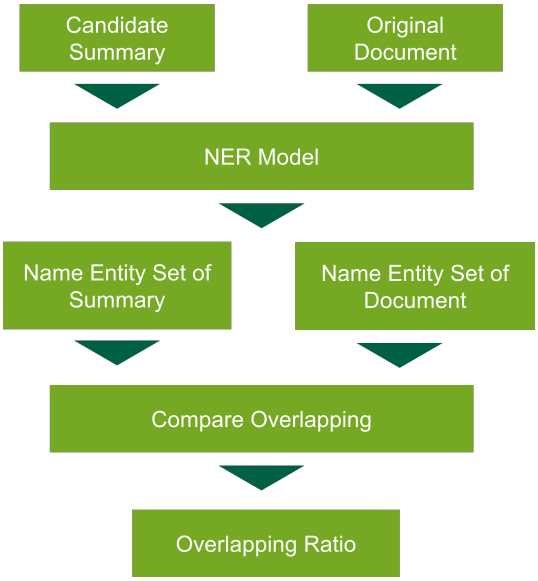
4.3.3 The Necessity of Reference Summaries

In all conventional methodologies, the inclusion of a reference is indispensable as it enables the comparison of generated summaries against the reference to determine their quality. Omitting a reference renders these methods unable to calculate a score for evaluation. However, in real-world scenarios, these models are deployed to aid in the evaluation of AI-generated summaries. Consequently, if we supply a human-authored reference, the role of the model becomes redundant, as the reference alone suffices to assess the quality of the summaries.

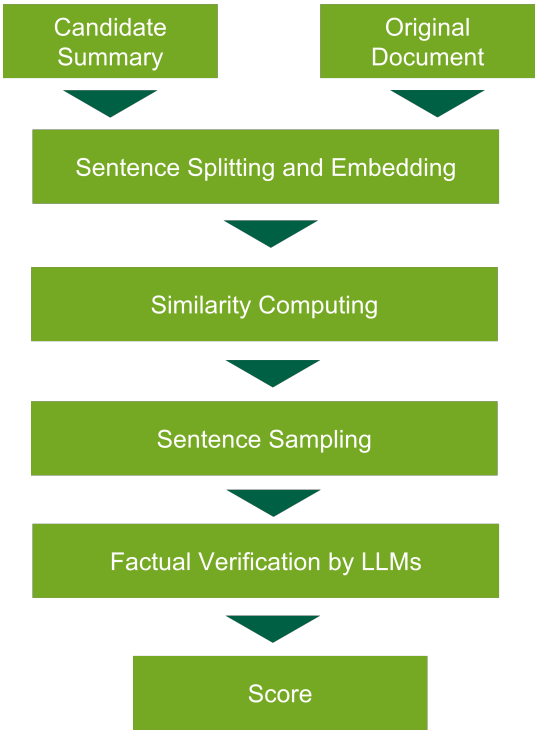
5 Our Methodologies

5.1 Framework Design

In this section, we will introduce our two methodologies for actuality checking from different aspects. The following figures show their frameworks.



(a) The workflow of our Name Entity Comparison method



(b) The workflow of our Sentence-Level Factuality Checking Based on LLMs method

Figure 2: Framework Overview

5.1.1 Name Entity Comparison

According to Figure 2a, firstly, extract name entities related to financial topics from the summary and original document. Then, compare the overlapping of extracted name entities from the summary and original document. The ratio of the overlapping name entities represents the level of factual consistency between the summary and the original document.

5.1.2 Sentence-Level Factuality Checking Based on LLMs

As shown in Figure 2b, in our sentence-level actuality checking based on LLMs method, the candidate summary and the original document are split into lists of sentences. Then, Our system will convert those two lists into two embedding matrices. After that, the similarities between sentences of summary and sentences of text are calculated based on their embeddings. Finally, our system will check if each sentence from the summary

can be obtained from the top-k-related sentences from the text with the help of LLMs. The final score is the ratio of sentences in the summary that are regarded as consistent with the original text by LLMs. A higher score means a higher level of consistency in factuality.

5.2 Experiment Results

A good metric for factuality checking should have the following three properties:

- **Quality Discrimination**: It should be able to tell a good summary from a bad one (The scores of them should be as different as possible).
- **Factual Accuracy Measurement**: It should be able to discern varying degrees of factual distortion (Given any two summaries according to the same document, the worse one should be scored lower).
- **Detail-Oriented Assessment**: It should be able to make evaluations based on the detail of the summary (Give the reason why it makes such an evaluation result).

To show our methods reach the above three principles, we implement two experiments.

In the first experiment, we compared the similarity between a good summary and the original text with the similarity between a poor summary and the original text, to identify the differences. The results are shown in Figure 3a and Table 1. According to the result, our LLM score returns the highest differences in score between the golden summary and the bad summary, meaning that our method can discern the similarity between the golden summary and the original document. The percentage in Figure 3a and Table 1 is defined as the ratio difference between the matrices score of the good summary and the matrices score of the bad summary and the matrices score of the good summary as shown in the following equation.

$$\text{Percentage a} = \frac{\text{Score(Good Summary)}}{\text{Score(Good Summary)} - \text{Score(Bad Summary)}}$$

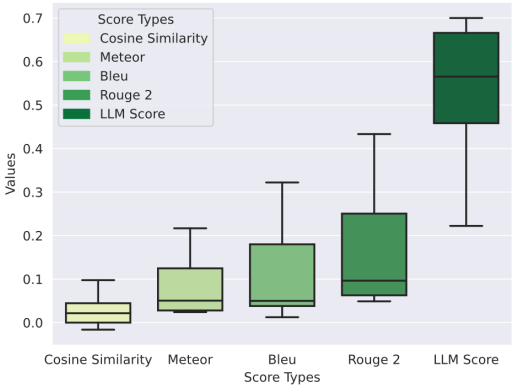
In the second experiment, we modified the falsified level, as defined below, for a selected sample. Subsequently, we compared the resulting score with the score from the previous experiment, where a lower falsification rate was applied.

$$\text{Falsified Level} = \frac{\text{Number of Sentences with false facts}}{\text{Number of Sentences in the summary}}$$

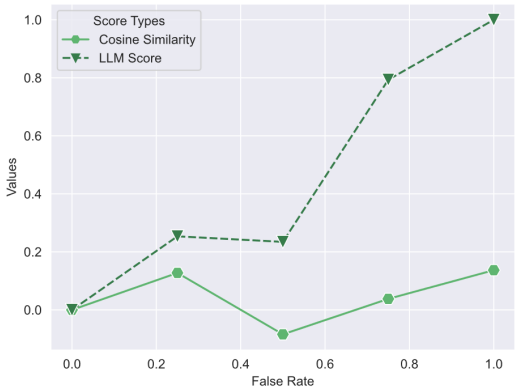
The percentage in Figure 3b is defined as the ratio difference between the matrices score of the summary and the matrices score of the previous summary with a lower falsified level and the matrices score of the previous summary.

$$\text{Percentage b} = \frac{\text{Score(Previous Summary)}}{\text{Score(Previous Summary)} - \text{Score(Summary)}}$$

Figure 2 and Table 1 reveal that common metrics fail to effectively distinguish between original and falsified summaries. In contrast, our methods demonstrate a clear capacity to discern these differences. Figure 3 shows that cosine similarity proves inadequate in differentiating between the original and falsified summaries, whereas our methods exhibit a distinct advantage in highlighting these discrepancies.



(a) Distributions of similarities between the golden summary and original document according to different metrics.



(b) Different metric results trends with the falsified ratio increases

Figure 3: Visualization results regarding the comparison between different metrics

Metric	Good Summary Score	Falsified Summary Score	Percentage
LLM Score	0.6897	0.3318	**53.00%**
Cosine Similarity	0.8832	0.8382	5.12%
Meteor	0.3063	0.2811	8.24%
Bleu	0.1015	0.0900	10.46%
Rouge 2	0.1512	0.1258	16.03%

Table 1: Differences between metric results between good summary and bad summary according to different metrics

6 Challenges

6.1 Computational Challenge

In our quest to optimize our model's performance, we strategically implemented a range of advanced Large Language Models (LLMs). This included the deployment of Llama-2-7B and RWKV, as well as the utilization of OpenAI's models, notably GPT-3.5-Turbo and GPT-4. Our comprehensive testing revealed that GPT-4 outperformed its counterparts in terms of efficiency and accuracy. However, we faced budgetary constraints as our resources were limited to GCP credits. This limitation imposed a cap on our access to the OpenAI API, restricting the extent to which we could test our data. Despite these financial restrictions, we endeavored to maximize the utility of our available resources to achieve the most effective results within the given constraints.

6.2 Ethical and Social Challenge

6.2.1 Data Leakage

One of the foremost ethical and practical concerns in deploying an AI model like GPT-4 for summary evaluation is the potential for data leakage. When submitting various types of summaries for assessment, there's a risk that sensitive or confidential informa-

tion may be inadvertently revealed to the model. In corporate or institutional contexts, it can lead to the inadvertent exposure of proprietary or confidential data. To mitigate this risk, robust data anonymization and redaction techniques should be employed to ensure that personally identifiable information and sensitive content are not accessible to the model during the evaluation process. Strict data handling and access controls are essential to address this challenge effectively.

6.2.2 Model Instability

AI models, including GPT-4, can exhibit instability in their responses and outputs. This instability can manifest as inconsistencies in scoring summaries or providing different evaluations for the same input when the model is queried repeatedly. Such variability can undermine the reliability and trustworthiness of the evaluation framework, making it challenging to provide consistent feedback on summaries. It is crucial to conduct rigorous testing and validation of the model's behavior and consider measures like ensemble methods or additional context to mitigate this instability and ensure more dependable results.

6.2.3 Lack of Interpretability

Another significant challenge when using ChatGPT 4 for summary evaluation is the inherent lack of transparency and interpretability in AI models of this complexity. The model's decision-making process is often considered a "black box," making it difficult to understand why it assigns particular scores or evaluations to summaries. This opacity hinders the ability to provide meaningful feedback and can lead to frustration and skepticism among users. Striking a balance between the model's complexity and interpretability is essential for building trust in the evaluation process and ensuring it aligns with human judgment and values.

6.2.4 Scalability and Resource Consumption

Implementing an AI-based summary evaluation framework can pose significant challenges due to its demanding computational resource requirements. This can make the technology less accessible to individuals or organizations with limited financial resources, potentially exacerbating a digital divide where only well-funded entities can benefit from advanced evaluation tools. This divide raises ethical concerns, particularly in fields like journalism and education, where equitable access to content evaluation tools is essential. To address these challenges, efforts should focus on making the technology more affordable, energy-efficient, and accessible. This can involve optimizing models for efficiency, offering flexible pricing structures, providing educational resources, and fostering collaborations between institutions of varying sizes to ensure that the advantages of AI-powered evaluation tools are distributed more equitably.

6.2.5 Legal and Compliance Issues

Legal and compliance issues can significantly vary based on the nature of the summaries being assessed. They encompass a broad spectrum of potential challenges, including copyright infringement, intellectual property rights violations, and potential breaches of

confidentiality or privacy. To avoid legal complications that could potentially result in litigation, financial liabilities, or damage to the reputation of the framework's operators, it is essential to operate within the boundaries of relevant laws and regulations. Prudent steps include implementing robust content filtering mechanisms to identify and handle copyrighted material appropriately, securing the necessary permissions and licenses for the use of protected content, and ensuring the framework complies with data privacy and protection standards.

7 Conclusion and Future Steps

Our methodology has demonstrably exceeded baseline metrics across the three key properties we initially targeted. This marked superiority in performance not only underscores the effectiveness of our methods but also highlights their enhanced capability in detecting summary falsification within PDF files. Compared to traditional baseline metrics, our approach offers a significant advancement, reinforcing its potential as a more reliable and efficient tool in this domain.

In the existing framework, our focus has been limited to sentence-level analysis. Moving forward, we plan to expand our methodology by incorporating additional packages to extract dependency arcs from summaries. This will enable us to cross-verify whether these dependency arcs are also present in the source text, thereby enhancing the accuracy of our analysis. Currently, we are constrained by the limitations of the OpenAI API, which does not provide the probability distribution of the next token. To overcome this, we intend to deploy larger Large Language Models (LLMs) on cloud servers in the future. This advancement will not only circumvent current limitations but also furnish us with more profound insights and improved analytical capabilities.

References

[1] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc., 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/df438caa36714f69277daa92d608dd63-Paper-Conference.pdf.

[2] Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment. *arXiv preprint arXiv:2010.05478*, 2020.

[3] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online, November 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.emnlp-main.750, doi:10.18653/v1/2020.emnlp-main.750.

[4] Bloomberg. Bloomberg's llm evaluation and implementation, 2023. URL: https://www.arxiv-vanity.com/papers/2303.17564/#S5.

[5] Deutsche Bank. Deutsche bank's exploration into large language models, 2023. URL: https://www.nvidia.com/en-us/on-demand/session/gtcspring23-s51160/.