Large Language Model in Financial Regulatory Interpretation

Zhiyu Cao
School of Business
Stevens Institute of Technology
Hoboken, New Jersey, USA
zcao17@stevens.edu

Zachary Feinstein
School of Business
Stevens Institute of Technology
Hoboken, New Jersey, USA
zfeinste@stevens.edu

2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr) | 979-8-3503-5483-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CIFER62890.2024.10772991

***Abstract*—This study explores the innovative use of Large Language Models (LLMs) as analytical tools for interpreting complex financial regulations. The primary objective is to design effective prompts that guide LLMs in distilling verbose and intricate regulatory texts, such as the Basel III capital requirement regulations, into a concise mathematical framework that can be subsequently translated into actionable code. This novel approach aims to streamline the implementation of regulatory mandates within the financial reporting and risk management systems of global banking institutions. A case study was conducted to assess the performance of various LLMs, demonstrating that GPT-4 outperforms other models in processing and collecting necessary information, as well as executing mathematical calculations. The case study utilized numerical simulations with asset holdings – including fixed income, equities, currency pairs, and commodities – to demonstrate how LLMs can effectively implement the Basel III capital adequacy requirements.**

***Index Terms*—Large Language Models, Prompt Engineering, LLMs in Finance, Basel III, Minimum Capital Requirements, LLM Ethics**

I. Introduction

Over the last decade, Artificial Intelligence (AI) and Natural Language Processing (NLP) models have been developed to process massive amounts of financial text, providing professional investment suggestions and aiding in financial decision-making. Research highlights that machine learning algorithms, including Long Short-Term Memory (LSTM) [1], Convolutional Neural Networks (CNN) [2], Support Vector Machines (SVM) [3], Random Forest (RF) [4], and the Bidirectional Encoder Representations from Transformers (BERT) [5], [6], exhibit strong performance in handling financial texts. Moreover, several financial service firms also focus on developing AI tools to provide powerful support for participants in the financial markets by offering deep market insights and predictive analysis. A prime example is Bloomberg's AI platform, which analyzes data from global stock markets, news reports, and social media trends to predict the future trajectory of specific stocks or sectors.

Large Language Models (LLMs), epitomized by tools like ChatGPT, have marked a revolutionary shift in artificial intelligence. Recently, LLMs such as GPT-3.5 and GPT-4 have demonstrated a remarkable ability to comprehend and generate human-like text. FinBERT, an NLP tool adapted to the finance domain, has been shown to outperform other NLP models in identifying discussions related to environmental, social, and governance (ESG) texts [7]. Meanwhile, a retrieval-augmented LLM framework for financial sentiment analysis has been introduced [8]. The efficacy of LLM-based chatbots for personal finance advisement has also been assessed [9]. BloombergGPT, a model that leverages Bloomberg's vast domain-specific dataset, demonstrates that it outperforms existing models on general financial tasks [10]. Given the impressive performance of LLMs, there is a natural impetus to explore the application of LLMs in financial regulatory interpretation.

Despite their groundbreaking advancements, Large Language Models (LLMs) also exhibit several limitations, primarily due to their nascent stage of development. First, preliminary models such as GPT-3.5, LLaMA-7B, and text-focused LLMs like Claude-3, are constrained by their inability to perform complex mathematical calculations or code analysis. Second, LLMs are highly sensitive to variations in prompts and document loading methods, which may result in significantly divergent outcomes for the same topic. If prompts are overly simplistic or lack precise direction, LLMs frequently generate inaccurate results. Third, the considerable number of parameters within these models makes pre-training and application both challenging and costly, necessitating extensive datasets and significant computational resources.

In response to these challenges, this study introduces the potential of LLMs in the interpretation of financial regulations. Our approach includes several key strategies:

First, we conduct a performance comparison among LLMs by manually examining a dataset to assess the capabilities of GPT-3.5, GPT-4, Gemini-1.5, and Claude-3. This assessment focuses on their ability to collect and analyze contextual information within financial regulatory documents. Through accuracy comparisons, we identify the most effective model. Second, our comparative analysis of document loading methods examines the efficacy of GPT-4 when analyzing financial regulatory documents uploaded as PDF files versus images. Our findings indicate that GPT-4 demonstrates greater precision in interpreting images than PDFs, particularly when documents contain a mix of mathematical equations, textual explanations, and footnotes, which typically present challenges for LLMs. Third, we develop an engineering method for

prompt design. By incorporating key elements into prompts, we guide LLMs to analyze documents more accurately. Contrasting the performance of naive prompts with those crafted using our engineering method underscores the critical role of deliberate prompt design in the accuracy of information collection. Finally, through comprehensive case studies, we validate the application of LLMs in financial regulation. By employing proper prompt design and document loading methods, alongside selecting an appropriate LLM, we demonstrate accurate computation of the minimum capital requirements from the 'Minimum Capital Requirements for Market Risk' section of the Basel III framework.

The remainder of this paper is organized as follows: Section II provides an overview of related work concerning the application of LLMs in the finance field. Section III details a comprehensive framework for applying LLMs to financial regulation documents, including algorithms, prompt design, and document loading methods. Section IV presents our dataset, comparative analysis across different LLMs, and a comprehensive case study. Section V examines the ethical considerations surrounding the application of LLMs in financial regulation, addressing critical aspects such as data privacy, transparency, and fairness. Finally, Section VI concludes the paper by highlighting potential limitations and proposing avenues for future research.

II. Literature Review: Large Language Models in Finance

The integration of LLMs into the financial sector represents a significant and emerging area of research. Numerous studies have focused on developing and applying domain-specific LLMs to enhance financial tasks such as sentiment analysis, textual analysis, and stock market predictions. Notably, FinBERT [7], which adapts the BERT framework specifically for the finance domain, has been shown to outperform other NLP models in identifying discussions related to financial texts. Additionally, PIXIU [11] proposes a comprehensive framework that represents the first financial LLM based on fine-tuning LLaMA with instructional data, enabling the model to execute various financial tasks effectively. InvestLM [12] and FinGPT [13] have each contributed uniquely to areas such as market analytics and predictive accuracy. BloombergGPT [10] leverages Bloomberg's vast domain-specific dataset, including news, market forecasts, and regulatory information. Furthermore, a retrieval-augmented LLM framework for financial sentiment analysis was introduced [8], and the efficacy of LLM-based chatbots for personal finance advisement was assessed [9]. A decision framework that aids financial professionals in selecting the most suitable LLM solutions based on their specific needs around data, computing power, and performance objectives was provided [14]. Additionally, insights into Natural Language Processing techniques within the framework of financial regulation were offered [5].

Despite the substantial focus on the development and application of domain-specific LLMs to enhance various financial tasks, their use in interpreting financial regulation documents remains relatively unexplored. This research addresses this gap and contributes to the existing literature in two distinct ways.

First, we investigate the innovative application of Large Language Models as analytical tools for interpreting complex financial regulations. By designing appropriate prompts and employing the correct document loading methods, we guide LLMs to distill verbose and intricate regulatory texts, such as the Basel III capital requirements, into a concise mathematical framework that is then translated into actionable code. Experimental results, discussed in subsequent sections, demonstrate the feasibility and accuracy of our method. This approach has the potential to streamline the implementation of regulatory mandates within the financial reporting and risk management systems of global banking institutions. Second, the Basel III standards necessitate advanced internal systems for risk assessment and management, substantial reporting obligations, and rigorous compliance protocols. These standards pose significant challenges for banks in maintaining higher operational and administrative capacities, particularly in identifying and interpreting regulatory requirements and integrating compliance workflows. Considering the limited resources available to small and medium-sized financial firms, our approach offers a potential solution by improving the identification of essential information and enhancing the efficiency of regulatory compliance processes.

III. Framework

Financial regulation documents are inherently complex, containing dense legal terminology, textual descriptions, mathematical formulations, and numerous footnotes. The high sensitivity of large language models presents challenges in achieving perfect accuracy in information retrieval from such documents.

This section describes a systematic algorithm designed for the analysis of financial regulation documents. The process begins with the efficient loading of relevant documents, followed by prompt engineering to define the overall problem-solving process. For complex financial tasks, achieving accuracy typically requires multiple iterative steps. Each task is broken down into smaller, manageable objectives, with each being addressed through carefully crafted prompts designed to secure accurate outcomes.

Although LLMs provide preliminary insights and help identify specific data locations, initial results may not always be accurate. We manually verify all key information. If the retrieved information is incorrect or unavailable, we activate an additional mechanism to locate accurate data sources. Following this, we manually upload relevant documents, aiding the LLMs in refining their analysis. Finally, we use the LLM for mathematical calculations to determine the correct outputs.

Figures 1 and 2 present a high-level illustration of the proposed architecture. We will elucidate prompt engineering, document loading methods, and the detailed process in the rest of this section.

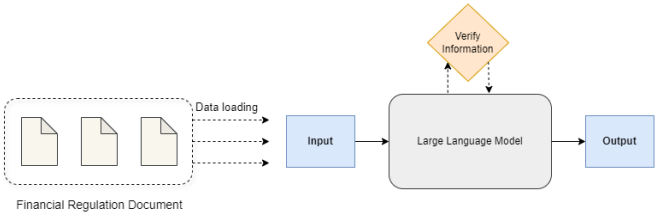

Fig. 1. Visual representation of the process for interpreting financial regulation documents.

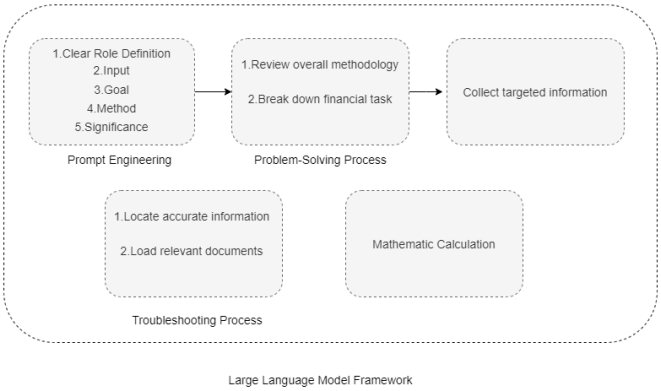

Fig. 2. Detailed schematic of the Large Language Model used for financial document analysis.

A. Document Loading Methods

Initially, we utilized the plug-in feature within GPT-4 to process PDF files from the Basel III framework. While this method proved adequate for simple tasks, it faltered when handling more complex content, resulting in errors in LLM outputs.

Recognizing the limitations inherent to the PDF format—which primarily prioritizes presentation over content extraction—we sought alternative document loading techniques. Our exploration led to a significant finding: converting PDFs to images before uploading markedly improves GPT-4's analytical capabilities. Unlike PDFs, images circumvent the complexities tied to parsing diverse formats and layouts. This simplification allows the LLM to focus more on visual data extraction, which is inherently less prone to the errors typically introduced by the varied and intricate structures of PDFs. By adopting an image-based analysis approach, GPT-4 achieves a more reliable interpretation of mathematical formulas, charts, and tables, which often present challenges in PDFs due to their layered configurations.

In Section IV, we will compare the efficacy of PDF and image loading methods in enabling LLMs to accurately identify correlations between sensitivities.

B. Prompt Design

Research by [15]–[17] highlights that prompt engineering is a relatively new discipline that focuses on developing and optimizing prompts to effectively utilize Large Language Models (LLMs), particularly in natural language processing tasks. This practice has emerged as an essential skill for effective communication and interaction with LLMs. The design of prompts significantly influences the performance of these models. Our study articulates key principles of prompt design to guide LLMs in the interpretation of financial regulation documents.

First, the prompt must incorporate the following key elements:

1) **Clear Role Definition**: Provides specific instructions that guide the model's behavior and outline a macro-level approach to problem-solving. For us, the role is for an LLM to act as a specialized financial regulations interpreter.
2) **Input**: Describes the document or question that we want the model to process and provide a response for. For us, the input is the Minimum Capital Requirements for Market Risk" from Basel III issued by the Basel Committee on Banking Supervision.
3) **Goal**: Specifies the details of the desired output. For us, the goal is to transform complex financial regulations into clear mathematical representations while maintaining the integrity of the regulation's core principles.
4) **Method**: Outlines a general strategy for addressing the problem, wherein the model reads the document and enriches its response based on the contained information. For us, the method is to understand legal terminology, offer straightforward explanations, and summarize the key elements to complete tasks.
5) **Significance**: Highlights the overarching concept and importance of the task, analogous to providing tips" within the prompt to guide effective outcomes. For us, the significance is to bridge the gap between dense regulatory texts and practitioners with no legal background.

In our prompt design, it is crucial to avoid ambiguity, bias reinforcement, overfitting, lack of context, and unrealistic dependency on model limitations [17]. These considerations ensure the effectiveness of our prompts in applications related to financial tasks.

In Section IV, we will compare the efficacy of naive prompts with detailed prompts developed using our method to enable LLMs to accurately identify risk buckets, risk weights, and correlations.

IV. Case Studies

A. Dataset

In accordance with the Basel III framework, we conducted simulations on over 40 different asset holdings to evaluate the capability of LLMs in interpreting financial regulation documents. The document used for this analysis includes a comprehensive section on 'Minimum Capital Requirements for Market Risk',[1] which outlines the required capital reserves that banks must maintain to mitigate potential losses from market

[1]Bank for International Settlements, *Minimum Capital Requirements for Market Risk*, https://www.bis.org/bcbs/publ/d352.pdf

fluctuations. The document comprises approximately 184,000 tokens and includes legal terms, intricate calculations, and case analyses pertinent to diverse market risks, such as interest rate, equity, foreign exchange (FX), and commodity risks. To illustrate our methodology, Table I presents a typical simulated case, outlining a variety of bank asset holdings, ranging from treasury bonds and futures contracts to equities and currency pairs.

TABLE I
BANK ASSET HOLDINGS

Asset Type	Description	Quantity / Value
U.S. Treasury Bond	5-year	$10,000
U.S. Treasury Bond	10-year	$10,000
Futures Contracts	Gold	600 ounces
Futures Contracts	Crude Oil	2,000 barrels
Equity	Exxon Mobil	10,000 shares
Equity	AT&T	10,000 shares
Currency Pair	Long EUR/USD	100,000 EUR
Currency Pair	Short USD/JPY	10,000,000 JPY

B. Minimal Capital Requirement Calculation

In this case study, we provide a comprehensive analysis of the application of LLMs as analytical tools for interpreting the "Minimal Capital Requirement" section of the Basel III framework. This includes a detailed demonstration of how minimal capital requirements are calculated based on bank asset holdings, as illustrated in Table I.

We begin by uploading images of the relevant sections of the Basel III document, focusing on minimal capital requirements, into GPT-4. A custom-designed prompt is used to guide GPT-4 in extracting methodologies for calculating these requirements. The response from GPT-4 breaks down the task of calculating the Minimal Capital Requirement into the following manageable objectives:

1) **Risk Classification:** Identify and categorize various types of risks associated with asset holdings, including equity risk, FX risk, general interest rate risk, and commodity risk.
2) **Sensitivity Calculation:** Determine the sensitivity of each asset holding to different risk factors, such as changes in market prices, interest rates, and foreign exchange rates.
3) **Aggregation of Risk Positions:** Combine individual risk positions to compile an overall risk profile.
4) **Capital Requirement Calculation:** Compute the minimal capital requirement based on the aggregated risk profile, applying the specific risk weights and methodologies prescribed by the Basel III framework.

In this report, we particularly focus on the Delta Equity Risk calculation. The same method is applied to other sections, repeating the process as necessary. Regarding the Sensitivity Calculation, through a carefully designed prompt, GPT-4 is capable of identifying relevant sources based on the context, specifically locating information regarding Delta Equity Sensitivities in Paragraph 67. We summarize GPT-4's findings as follows:

Delta Sensitivity Calculation: For an equity k, the sensitivity s_k is defined as the change in the market value of instrument i (V_i) with respect to a 1 % change in the market value of equity k (EQ_k). Mathematically, it is expressed as:

$$s_k = \frac{V_i(1.01 \cdot EQ_k) - V_i(EQ_k)}{0.01}$$

where:

- k is a given equity.
- EQ_k is the market value of equity k.
- V_i is the market value of instrument i as a function of the price of equity k.

In this case study, GPT-4 can calculate the sensitivity of each equity:

- Delta Sensitivity for "Exxon Mobil": $1,100,000.
- Delta Sensitivity for "AT&T": $170,000.

In the Aggregation of Risk Positions step, GPT-4 is utilized to collect and analyze targeted information firstly.

- **Exxon Mobil** is classified under the category *Energy, Oil and Gas* within an advanced economy, and is placed in *Bucket 7*.
- **AT&T** is classified under *Telecommunications* within an advanced economy, and is placed in *Bucket 6*.

Risk Weights and Correlations:

- The risk weight assigned to Bucket 7 (Exxon Mobil) is 40%.
- The risk weight assigned to Bucket 6 (AT&T) is 35%.
- The prescribed correlation between Exxon Mobil and AT&T is set at 15%.

This outcome is the result of GPT-4's comprehensive capabilities, which include the identification of each equity and a thorough understanding of the Basel III framework. GPT-4 efficiently processes tables, textual descriptions, and mathematical formulas to produce these results.

GPT-4 successfully located information regarding Delta Equity risk calculation in Paragraph 51. Here is the response:

Calculation Within Delta Buckets

- **Weighted Sensitivity** (WS_k)**:**

$$WS_k = RW_k \cdot S_k$$

where RW_k is the risk weight for equity k and S_k is the sensitivity of equity k.

- **Risk Position** (K_b) **for Delta Bucket** b**:**

$$K_b = \sqrt{\sum_k WS_k^2 + \sum_k \sum_{k \neq l} \rho_{kl} WS_k WS_l}$$

where ρ_{kl} represents the correlation between equities k and l within the same bucket.

Aggregating Delta Risk Positions Across Buckets

- **Delta Risk Charge Calculation:**

$$\text{Delta} = \sqrt{\sum_b K_b^2 + \sum_b \sum_{c \neq b} \gamma_{bc} S_b S_c}$$

where $S_b = \sum_k WS_k$ for all risk factors in bucket b, and $S_c = \sum_k WS_k$ in bucket c, and γ_{bc} is the correlation between buckets b and c.

The calculated delta equity risk capital requirement, as shown in Table I, is approximately \$461,262.67. This result matches the outcome from our manually simulated dataset.

C. Comparative Analysis across Different LLMs

In this case study, we investigate the capabilities of large language models (LLMs) in interpreting and applying complex financial regulations, based on the 'Minimum Capital Requirements for Market Risk' document from Basel III. To conduct a comprehensive assessment, we divided the document into several sections based on its table of contents, focusing on general interest rate risk, equity risk, foreign exchange (FX) risk, and commodity risks. For each sector, we designed detailed prompts to guide the LLMs in identifying key elements necessary for calculating the Minimum Capital Requirements from provided asset holding cases, aiming to assess the models' ability to identify buckets, risk weights, and correlations. We evaluated the performance of four prominent LLMs: GPT-4, GPT-3.5, Claude-3, and Gemini-1.5-pro, on our manually simulated dataset including 40 different asset holdings. Their accuracy was measured by the number of cases correctly identified, scaled by the total number of cases in the testing sample. The results of this evaluation are presented in Table II.

TABLE II
COMPARISON OF LLMS IN IDENTIFYING KEY ELEMENTS

Model	Buckets (%)	Risk Weights (%)	Correlation (%)
GPT-4	85	100	96.5
GPT-3.5	10	30	0
Claude-3-Opus	82.5	100	97.5
Gemini-1.5-Pro	27.5	75	80

As indicated in Table II, GPT-4 and Claude-3-Opus achieve the highest overall performance among the models evaluated. Both models demonstrate near-perfect accuracies in identifying risk weights and correlations, with scores of 100% and 96.5% for GPT-4, and 100% and 97.5% for Claude-3-Opus, respectively. Additionally, GPT-4 slightly outperforms Claude-3-Opus in identifying buckets, with an accuracy of 85% compared to 82.5%. This suggests that these two models have a strong capability to comprehend and apply the complex rules and guidelines outlined in the 'Minimum Capital Requirements for Market Risk' document. In contrast, GPT-3.5 and Gemini-1.5-Pro show lower performance across all three aspects. GPT-3.5 struggles to identify buckets (10% accuracy) and fails to identify any correlations (0% accuracy). Gemini-1.5-Pro, on the other hand, also struggles with identifying buckets (27.5% accuracy) but performs moderately well in identifying risk weights (75%) and correlations (80%).

Recent research by [18]–[20] suggests that mathematical reasoning poses significant challenges for large language models. In this work, we evaluated the mathematical computation capabilities of GPT-4, GPT-3.5, Claude-3, and Gemini-1.5-pro within the context of complex calculations related to the Minimum Capital Requirements (MCR) sector. We designed five distinct scenarios corresponding to key risk categories outlined in the regulatory document: general interest rate risk, equity risk, foreign exchange (FX) risk, and commodity risks.

TABLE III
ACCURACY OF LLMS IN COMPLEX MATHEMATICAL CALCULATIONS

Model	Accuracy in MCR Calculations (%)
GPT-4	95
GPT-3.5	0
Claude-3-Opus	38
Gemini-1.5-Pro	58

Table III reveals a notable disparity in the accuracy of large language models when performing complex mathematical calculations for Minimum Capital Requirements (MCR). GPT-4 stands out as the most capable model, achieving an impressive accuracy score of 95% across all tested scenarios. In contrast, GPT-3.5, while capable of providing a mathematical framework for the calculations, fails to execute the required complex operations, resulting in an accuracy score of 0%. Claude-3-Opus and Gemini-1.5-Pro demonstrated moderate performance, with accuracy scores of 38% and 58%, respectively. Despite their ability to locate relevant information within the document, these models show significant room for improvement in performing complex mathematical calculations.

D. Document Loading Method

We selected Claude-3-Opus and GPT-4, which have performed well in identifying and interpreting regulatory requirements, to compare the performance of PDF loading and image loading. We conducted tests on these models using our manually simulated dataset to evaluate their ability to discern correlations between sensitivities. This task is inherently complex, as it typically involves intricate elements of Basel III such as mathematical formulas, legal terms, tables, and footnote analysis.

TABLE IV
ACCURACY OF IDENTIFYING CORRELATION FROM DIFFERENT DOCUMENT LOADING METHODS IN CLAUDE-3-OPUS AND GPT-4

Model / Document Type	PDF (%)	IMAGE (%)
Claude-3-Opus	76.5	97.5
GPT-4	68	96.5

Table IV displays the accuracy of Claude-3-Opus and GPT-4 in identifying correlations from documents loaded as PDFs and images. For documents loaded as PDFs, the accuracies recorded were 76.5% for Claude-3-Opus and 68% for GPT-4. Remarkably, when analyzing documents loaded as images,

both models achieved impressive accuracies of 97.5% and 96.5%, respectively. Our troubleshooting process, outlined in Section III, revealed issues when PDF-loaded documents included mathematical formulas, legal terms, tables, and footnotes. This observation suggests that image-based document loading may be particularly effective for LLMs. This could be potentially due to the visual processing capabilities inherent in their architectures, which might better handle layouts and embedded information such as mathematical formulas that are typically difficult to parse in text-based PDF files.

E. Naive Prompt vs Detailed Prompt

We compare the efficacy of naive prompts with detailed prompts developed using our method, in enabling LLMs to accurately identify risk buckets, weights, and correlations in the GPT-4 Model.

TABLE V
COMPARISON OF NAIVE PROMPT AND DETAILED PROMPT IN IDENTIFYING KEY ELEMENTS WITH GPT-4

Model	Buckets (%)	Risk Weights (%)	Correlation (%)
Naive Prompt	65.5	100	30
Our Detailed Prompt	85	100	96.5

Table V illustrates the effectiveness of naive and detailed prompts in the GPT-4 model. While the naive prompt achieves high accuracy (100%) in simpler tasks such as identifying risk weights, it exhibits considerable limitations in more complex tasks. For example, it only achieves a 65.5% accuracy in bucket identification and 30% accuracy in correlation determination. In contrast, the detailed prompt significantly enhances performance across more complex areas, maintaining 100% accuracy in risk weights and notably improving to 85% and 96.5% in identifying buckets and correlations, respectively. This data underscores the importance of using detailed prompts when addressing tasks of higher complexity to avoid errors.

V. LLM ETHICAL CONSIDERATIONS

The integration of Large Language Models (LLMs) into financial regulatory frameworks necessitates a comprehensive examination of multifaceted ethical dimensions [21], with particular emphasis on data privacy, transparency, and fairness. These critical considerations are expected to become central to the discourse on deploying LLMs within the financial regulatory sphere in the near future.

Firstly, data privacy emerges as a paramount concern, considering the highly sensitive nature of financial information. LLMs trained on financial regulation documents must strictly adhere to robust privacy protection measures, such as differential privacy techniques, to safeguard the confidential information of both individuals and institutions. For instance, the data used to compute market risks for financial institutions often encompasses proprietary information regarding their asset and liability levels. Implementing these rigorous privacy protocols can effectively prevent data breaches that could erode public trust in financial institutions, thereby undermining their credibility and potentially catalyzing widespread economic repercussions.

Secondly, transparency in the application of LLMs within the domain of financial regulation is a critical ethical consideration that warrants careful attention. To maintain public trust and ensure accountability, the development and deployment of LLMs in this context must be characterized by a high degree of transparency. This necessitates clear and comprehensive documentation of the data origins, training methodologies, and decision-making processes employed by these models. Regular publication of evaluation results and performance metrics is essential to facilitate external scrutiny and validate the integrity of LLM-driven financial regulatory systems.

Thirdly, ensuring fairness in LLM-driven financial decision-making is paramount to prevent discriminatory outcomes based on factors such as the geographic location and size of financial institutions. Bias mitigation strategies, including dataset enhancement and adversarial learning, should be employed to promote equitable treatment of all financial institutions, regardless of their regional context or operational scale, in regulatory processes. LLMs must be designed to account for the unique challenges faced by smaller, regional financial institutions, ensuring that they are not unfairly disadvantaged compared to larger, multinational corporations. This may involve incorporating diverse datasets that adequately represent the full spectrum of financial institutions, from local credit unions to global banking conglomerates. Moreover, adversarial learning techniques can help identify and mitigate potential biases that could skew regulatory decisions in favor of certain types of institutions.

VI. CONCLUSION

In this study, we have demonstrated the innovative application of Large Language Models such as GPT-4, Gemini-1.5-pro and Claude 3 in interpreting complex financial regulation documents. While achieving 100% accuracy in extracting information from texts that include legal terms, textual descriptions, mathematical formulations, and extensive footnotes remains a formidable challenge, we have developed a troubleshooting process and we deconstruct a complex financial task into smaller, manageable objectives. This approach has enabled us to achieve precise results through detailed prompting. Additionally, we conducted a comparative analysis of document loading methods and the performance of LLMs in interpreting financial documents. We have also devised a systematic approach to prompt design, which enhances the LLMs' capability to analyze and summarize financial texts effectively.

Our work further highlights a pair of promising avenues for future research. Although our framework effectively performs well on our manually simulated dataset, our investigation focused on scenarios involving a limited number of cases. Real-world cases, however, may encompass a more extensive range of assets holding. Thus, future research can integrate more comprehensive, interconnected datasets to derive profound insights. Furthermore, incorporating stress testing by

generating synthetic datasets and designing stress tests using LLMs could further enhance the robustness and regulatory applicability of our framework.

REFERENCES

[1] T. Mikolov, "Advances in neural information processing systems," *(No Title)*, p. 3111, 2013.

[2] S. R. Das *et al.*, "Text and context: Language analytics in finance," *Foundations and Trends® in Finance*, vol. 8, no. 3, pp. 145–261, 2014.

[3] M. Gentzkow, B. Kelly, and M. Taddy, "Text as data," *Journal of Economic Literature*, vol. 57, no. 3, pp. 535–574, 2019.

[4] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine learning for financial risk management: a survey," *Ieee Access*, vol. 8, pp. 203 203–203 223, 2020.

[5] I. Achitouv, D. Gorduza, and A. Jacquier, "Natural language processing for financial regulation," *arXiv preprint arXiv:2311.08533*, 2023.

[6] K. Bochkay, S. V. Brown, A. J. Leone, and J. W. Tucker, "Textual analysis in accounting: What's next?" *Contemporary accounting research*, vol. 40, no. 2, pp. 765–805, 2023.

[7] A. H. Huang, H. Wang, and Y. Yang, "Finbert: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, 2023.

[8] B. Zhang, H. Yang, T. Zhou, M. Ali Babar, and X.-Y. Liu, "Enhancing financial sentiment analysis via retrieval augmented large language models," in *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 349–356.

[9] K. Lakkaraju, S. E. Jones, S. K. R. Vuruma, V. Pallagani, B. C. Muppasani, and B. Srivastava, "Llms for financial advisement: A fairness and efficacy study in personal decision making," in *4th ACM International Conference on AI in Finance*, 2023, pp. 100–107.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[11] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "Pixiu: A large language model, instruction data and evaluation benchmark for finance," *arXiv preprint arXiv:2306.05443*, 2023.

[12] Y. Yang, Y. Tang, and K. Y. Tam, "Investlm: A large language model for investment using financial domain instruction tuning," *arXiv preprint arXiv:2309.13064*, 2023.

[13] H. Yang, X.-Y. Liu, and C. D. Wang, "Fingpt: Open-source financial large language models," *arXiv preprint arXiv:2306.06031*, 2023.

[14] Z. Li, S. Fan, Y. Gu, X. Li, Z. Duan, B. Dong, N. Liu, and J. Wang, "Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 608–18 616.

[15] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas, "Guiding pretraining in reinforcement learning with large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 8657–8677.

[16] G. Marvin, N. Hellen, D. Jjingo, and J. Nakatumba-Nabende, "Prompt engineering in large language models," in *International Conference on Data Intelligence and Cognitive Informatics*. Springer, 2023, pp. 387–402.

[17] L. Giray, "Prompt engineering with chatgpt: a guide for academic writers," *Annals of biomedical engineering*, vol. 51, no. 12, pp. 2629–2633, 2023.

[18] Z. Yuan, H. Yuan, C. Li, G. Dong, C. Tan, and C. Zhou, "Scaling relationship on learning mathematical reasoning with large language models," *arXiv preprint arXiv:2308.01825*, 2023.

[19] S. Imani, L. Du, and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," *arXiv preprint arXiv:2303.05398*, 2023.

[20] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, "Large language models for mathematical reasoning: Progresses and challenges," *arXiv preprint arXiv:2402.00157*, 2024.

[21] J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, "Navigating llm ethics: Advancements, challenges, and future directions," *arXiv preprint arXiv:2406.18841*, 2024.