Uniandes at the Regulations Challenge Task: A Scalable Framework for Legal Text Understanding in Regulatory and Financial Contexts.

Santiago Martínez Carrión[1], Juan Manuel Castañeda[1], Rubén Francisco Manrique[2]
[1]Dept. of Systems and Computing Engineering
Universidad de los Andes, Bogotá, Colombia
[2]Assistant Professor, Dept. of Systems and Computing Engineering
Universidad de los Andes, Bogotá, Colombia
{s.martinez1,jm.castaneda,rf.manrique}@uniandes.edu.co

Abstract

This study presents a comprehensive approach to developing a domain-specific large language model (LLM) for regulatory and financial text interpretation. A specialized corpus was constructed through large-scale scraping of financial and regulatory documents across domains such as compliance, licensing, and financial reporting. The data was preprocessed using GPT-4o-mini with prompt engineering to retain critical information and remove noise. We further pre-trained a LLaMA-3.1-8B model on the curated corpus and fine-tuned it using an instruction dataset covering nine tasks from the Coling 2025 Regulations Challenge (Wang et al., 2024), including acronym expansion, regulatory question-answering, and XBRL-based financial analytics, employing QLoRA to reduce memory requirements. The model exhibits a slight improvement from baseline answering complex regulatory questions (detailed QA) and expanding acronyms. This study demonstrates the potential of domain-specific LLMs in regulatory text interpretation and lays the groundwork for future research in specialized NLP evaluation methodologies.

1 Introduction

The rapid growth and increasing complexity of regulatory and financial documentation have created a pressing need for tools capable of extracting, analyzing, and responding to nuanced queries with precision and contextual relevance. While large language models (LLMs) have demonstrated exceptional capabilities in natural language understanding, their general-purpose nature often renders them inadequate for domain-specific applications. Addressing this gap, we present a domain-specific LLM designed specifically for regulatory and financial texts, equipped to tackle diverse and intricate tasks with slightly enhanced accuracy and contextual awareness.

To develop this model, we began by constructing a comprehensive corpus sourced from publicly available financial and regulatory documents including legal statutes, compliance guidelines, and financial reports. Recognizing the inherent noisiness of web-sourced data, we implemented a preprocessing pipeline. We first used a subset of our data and trained a TF-IDF model, which we used to score documents to ignore very noisy entries. Then, using prompt engineering with GPT-4o-mini, we refined the corpus by filtering out irrelevant content while retaining key information critical for downstream tasks. This preprocessing approach allowed us to create our dataset, tailored to the unique demands of regulatory language modeling.

The core of our methodology involved fine-tuning the LLaMA-3.1-8B model, presented in (Grattafiori et al., 2024), leveraging its capabilities as a foundational LLM. Notably, the model presents a strategic balance between computational efficiency and performance. While many state-of-the-art large language models require extensive computational resources—often demanding high-end GPU clusters or cloud computing infrastructure—the LLaMA-3.1-8B model offers a more accessible alternative.

With a modest parameter count (for modern LLM standards) of 8 billion, the model strikes a balance between computational complexity and inferential capabilities. This design allows for potential local deployment on high-range computational hardware, such as workstations with 32-64 GB of RAM and modern consumer grade GPUs. However, it is crucial to acknowledge the trade-offs: while the reduced infrastructural footprint enables broader accessibility, it may inherently limit the model's capacity to match the absolute performance of larger, more computationally intensive models.

To optimize computational efficiency and scalability we employed Quantized Low-Rank Adap-

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 326–334

tation (QLoRA), a parameter-efficient fine-tuning technique. QLoRA allowed for substantial memory savings while maintaining model performance (Dettmers et al., 2023). However, the lack of standardized evaluation benchmarks for regulatory NLP tasks posed a significant challenge, leading us to rely on qualitative analyses and comparisons with the base LLaMA-8B model to assess improvements. These qualitative assessments demonstrated notable gains in task performance, particularly in Named Entity Recognition and Question-Answering.

This paper details the methodologies and challenges encountered in developing this domain-specific regulatory language model. By combining advanced preprocessing techniques with task-specific fine-tuning strategies, our work highlights the potential of tailored LLMs in addressing the unique challenges of regulatory text interpretation and establishes a foundation for future research in this critical area. All code, prompts and implementation details can be found in this repository: https://github.com/smartinez1/COLING-2025-Regulations-Challenge_Uniandes

2 Related Work

In the evolving landscape of large language models, their application to specialized domains such as regulatory and financial text analysis has gained significant attention. These domains present unique challenges due to the complexity and specificity of the language used, which often surpasses the capabilities of general-purpose models. Tailored approaches are thus essential to effectively address the specific challenges of these domains.

Li et al. (Li et al., 2024) developed the LegalQA dataset, enhancing LLM performance through retrieval-augmented generation (RAG) with expert-curated question-answer pairs. While this dataset performs well in legal question-answering, it falls short in covering the diverse tasks addressed in our study, such as Named Entity Recognition (NER) and XBRL Analytics.

The Regulations Challenge at COLING 2025, led by Wang et al. (Wang et al., 2024), provides a benchmark to assess the readiness of LLMs for financial regulations. Their framework, which includes nine tasks and corresponding datasets, is a valuable tool for evaluating LLM performance in legal and financial contexts. While our tasks differ, their methodology has greatly influenced ours, emphasizing the need for thorough evaluation.

Mavi et al.'s work (Mavi et al., 2023) on retrieval techniques for semi-structured domains aligns with our data preprocessing efforts, where we use frequency-based methods to curate high-quality datasets. Similarly, Pipitone's LegalBench-RAG (Pipitone and Alami, 2024) supports retrieval techniques, ensuring scalability and adaptability across regulatory contexts. Our approach uses TF-IDF for document retrieval, aligning with the emphasis on precise retrieval in specialized domains, while differing on the specific tasks and data used.

Dahan and Wu's studies (Dahan et al., 2023; Wu et al., 2024) emphasize the critical need to mitigate hallucination and ensure data reliability, particularly when guiding non-expert users. In our model, these insights are incorporated through task-specific prompt design, which should enhance the model's practical utility by ensuring reliable and accurate responses.

Our study addresses gaps in previous research by developing a domain-specific LLM that integrates frequentist preprocessing with task-specific fine-tuning. This method shows promising results in managing cross-domain tasks and complex financial data, providing a robust alternative for tackling the challenges of regulatory and financial text analysis.

3 Dataset Creation

The challenge tasks aim to assess the ability of large language models (LLMs) to generate accurate responses to questions related to regulatory texts, focusing on their performance across the following nine specific areas:

- **Abbreviation Recognition**: Identifying and expanding domain-specific abbreviations
- **Named Entity Recognition (NER)**: Detecting and classifying entities in regulatory texts
- **Question-Answering (QA)**: Providing accurate responses to regulatory queries
- **Link Retrieval**: Identifying relevant regulatory document references
- **Certificate Analysis**: Processing certification-related queries (CFA, CPA)
- **XBRL Analytics**: Analyzing eXtensible Business Reporting Language data
- **CDM Processing**: Working with Common Domain Model data

- **Financial Mathematics**: Solving financial calculations and problems
- **License Compliance**: Analyzing software license requirements

Table 1 shows the evaluation metrics for each task.

Task	Evaluation Metric
Abbreviation	Accuracy
Definition	BERTScore
NER	F1 Score
QA	FActScore
Link Retrieval	Accuracy
Certificates	Accuracy
XBRL	FActScore
CDM	FActScore
Licensing	Accuracy

Table 1: Evaluation Metrics by Task.

The FactScore metric is defined in (Min et al., 2023).

3.1 Data Sources Overview

The dataset used in this work was created using scrapers. All sources scraped come from the Coling 2025 Regulations Challenge. For each task, a set of target domains and corresponding candidate sources for data extraction are defined; however, additional sources were also permitted. In this work, data was exclusively extracted from the sources recommended by the challenge. Table 2 shows the target domain, and Table 3 summarizes the suggested sources for each task.

Task	Domains
Abbreviation	EMIR, SEC, FDIC, Federal Reserve
Definition	EMIR, SEC, Federal Reserve
NER	EMIR
QA	SEC, FDIC, Federal Reserve
Link Retrieval	EMIR, SEC, Federal Reserve
Certificates	CFA, CPA
XBRL	Financial reports
CDM	Regulatory frameworks
Licensing	Open-source software

Table 2: Summary of Tasks and Domains [1].

Task	Sources	Scraper Depth
Abbreviation	EUR-LEX, ESMA	4
Definition	EUR-LEX	4
NER	EUR-LEX	4
QA	FDIC, Fed Reserve	4
Link Retrieval	eCFR, SEC	4
Certificates	CFA, CPA Exam	2
XBRL	XBRL Int'l	1
CDM	CDM Docs	4
Licensing	OSI	1

Table 3: Primary Data Sources for Regulatory Tasks

3.2 Corpus Collection

A recursive scraping methodology was utilized to construct a comprehensive document corpus. The process began by extracting all HTML text and downloading any text document from the provided source links, then recursively extracting and processing additional links found within these sources. This iterative approach continued up to a defined maximum depth. The depth was determined manually, depending on how the pages were structured and it ranged between 1 and 4. We also developed source-specific scrapers which were used to enrich the dataset in a finer level. These relied on each of the specific sources' web structure or API availability. They can be found within our repository in the "scraper" folder.

We also implemented a score-based filtering to eliminate potentially noisy documents obtained during the scraping process. The following paragraphs provide a detailed explanation of this strategy.

3.3 Relevance Filtering Pipeline

A subset of documents obtained on the first scraping round was used to build a simple BoW (Bag of Words) representation of each document with a TF-IDF (Term Frequency-Inverse Document Frequency) weighting schema for further similarity analysis. We manually checked the examples looking for noisy data, keywords within useful data, and other patterns. This similarity analysis served as the foundation for relevance scoring.

[1]Abbreviations: EMIR - European Market Infrastructure Regulation; SEC - U.S. Securities and Exchange Commission; FDIC - Federal Deposit Insurance Corporation; CFA - Chartered Financial Analyst; CPA - Certified Public Accountant; XBRL - eXtensible Business Reporting Language; CDM - Common Data Model. For more information, see the challenge details at https://coling2025regulations.thefin.ai/.

The following steps were applied to this subset data:

- **Stopword Removal:** Common words, such as "the," "and," and "in," that do not carry significant meaning in the context of regulatory texts, were removed. This reduces noise and focuses the model on more meaningful content.
- **Stemming:** Words were reduced to their root forms (e.g., "running" to "run") to ensure that variations of a word are treated as the same, improving the model's ability to generalize.
- **Tokenization:** The text was split into individual words or tokens, which are the basic units for further analysis.
- **Composite Terms:** Some terms in the text were composite phrases, such as "market abuse" or "financial stability," which are important for regulatory contexts. These multi-word expressions were modified by removing spaces (e.g., "marketabuse") so they could be treated as single tokens in the model.
- **Dictionary and BoW representation:** A dictionary was constructed to map unique tokens (words) to numeric identifiers. This dictionary was used to convert the preprocessed documents into a BoW representation, where each document is represented by a set of words and their frequencies.
- **TF-IDF weighting:** A invert index was built to evaluate the importance of each word using the TF-IDF schema.

Using the trained TF-IDF representation, the remaining documents in the corpus were scored based on similarity scores. This was achieved by employing a positive query and a negative query. The positive query was constructed by selecting keywords from relevant documents and further enriched using GPT to include additional legal domain-related terms. The negative query was created manually by identifying keywords found in irrelevant data, such as javascript artifacts, social media names, error pages' names, etc. The keywords used can be found in our repository, in the file data_processing.py, as POS_QUERY and NEG_QUERY.

The final score for document i is calculated by subtracting the negative similarity score from the positive similarity score. The positive similarity score is the cosine similarity between the positive query vector Q_{pos} and the document vector D_i, while the negative similarity score is the cosine similarity between the negative query vector Q_{neg} and the document vector D_i. The formula for the final score S_i is:

$$\begin{aligned} S_i = \text{cosine\_similarity}(Q_{\text{pos}}, D_i) \\ -\text{cosine\_similarity}(Q_{\text{neg}}, D_i) \end{aligned} \quad (1)$$

Figure 1 shows the document scores obtained for the first scraping round, before applying the threshold. A mean slightly above 0 is evidenced, indicating that most documents were moderately relevant to the initial subset. On the other hand, Figure 2 reveals the variation in relevance scores across different sources.

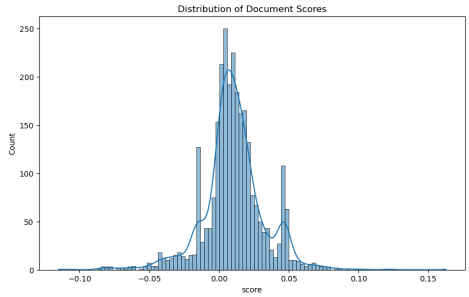

Figure 1: Distribution of document scores from our TF-IDF model.

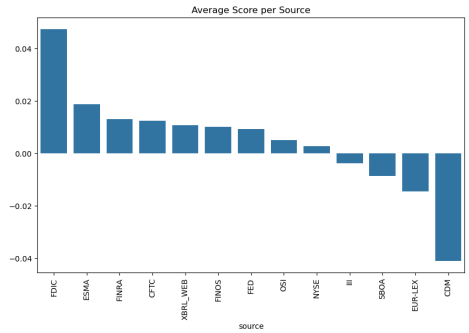

Figure 2: Average TF-IDF score by source.

3.4 Data Cleaning Process

The data cleaning process in this study was designed to ensure that the text data used for model training was of quality, relevant, and properly formatted. This was done using GPT-4o-mini, along with a prompt engineering process.

3.4.1 Token Encoding

To prepare the data for processing by language models, it was important to ensure that each text entry fit within the token limits of the model. The data was examined to calculate the number of tokens in each text, and entries that exceeded the

context window of our cleaning process were truncated.

3.4.2 GPT-4o Based Data Cleaning

This step involved using two custom cleaning prompts to refine the text and ensure that only the most relevant and coherent content remained for model training. The first prompt simply made sure the text in question was written in english and was "relevant" (related to the financial or legal fields). Although we may lose data by filtering out information in other languages, this is necessary because the rest of our pipeline requires the input to be in English. The second cleaning prompt provided to GPT-4o guided the model to:

- Retain factual content, such as laws, regulations, and domain-specific terms, while removing irrelevant sections like social media links, navigation menus, HTML markers, and unnecessary symbols.
- Remove incoherent or irrelevant text and fix issues like unnecessary spaces between letters and words stuck together.
- Remove tabular data, OCR artifacts, and numeric data not relevant to the regulatory or financial domain.

This process involved iterating over a small set of examples (around 20 examples) and manually validating that the model correctly removed noisy elements, while retaining the core information. We then applied this process to the entire corpus to get a clean dataset.

Finally, the cleaned corpus was serialized and stored for future use. The resulting dataset comprised 2 286 documents with diverse textual content.

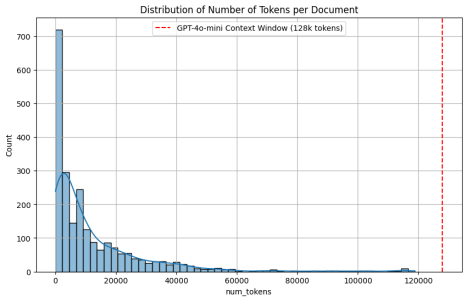

Figure 3: Token length distribution per document in our corpus. The model's context window is displayed at 128k tokens.

For this work, we opted in truncating the contents of the document length in such way that it would fit into the model's context window. We are aware that in doing this we might miss out on valuable information, considering that for some sources (such as EUR-LEX), documents that contained over 30 pages were common.

4 Instruction Dataset Generation

In this section, we present the methodology for creating an instruction dataset specifically designed to optimize large language models (LLMs) for understanding regulatory and financial texts. This dataset aims to improve model performance in specialized tasks such as abbreviation expansion, question-answering, and named entity recognition.

4.1 Overview

We used the cleaned dataset obtained from the previous section as input for subsequent steps. Prompts were applied to each cleaned document to extract and organize all relevant information.

4.2 Prompt Design and Customization

Custom prompts were developed for each task to guide the language model (GPT-4o) in generating desired outputs. These prompts were crafted to elicit responses that are both accurate and contextually appropriate. For example, abbreviation expansion prompts were designed to ensure comprehensive extraction of domain-specific acronyms. The task-specific prompt structures were as follows:

- **Abbreviation Expansion:** Prompts aimed to expand domain-specific acronyms into their full forms.
- **Question-Answering:** Prompts generated question-answer pairs relevant to regulatory and compliance issues.
- **Named Entity Recognition:** Prompts identified and listed specific organizations, legislation, dates, monetary values, and key statistics.

Similarly to the data cleaning process, we manually iterated the prompts over a small set of examples (over 50, around 10 per task) to verify the coherence of extracted information and ensure no relevant details were overlooked. Subsequently, we applied the prompts to the entire cleaned dataset.

The specific prompts are available in our repository at `tasks/prompts.py`. We parsed the model's responses to a standardized prompt template:

```
Below is an instruction that
```

```
describes a task.
Write a response that appropriately
completes the request.
### Instruction:
[Task Description]
### Answer:
[Response]
```

We supplemented our data set with existing Hugging Face datasets [2] [3] to incorporate reliable information for the CFA and XBRL data, which could not be extracted through our prompt processing due to the complexity of the task.

Using these sources, including the extracted answer-instruction pairs, we built the final instruction dataset.

4.3 Dataset Summary

Table 4 provides a summary of the tasks included in the instruction dataset, along with the number of examples for each task. The table offers a concise overview of the dataset's composition, highlighting the diversity and scope of the tasks addressed.

Task Name	Number of Examples
Abbreviation Expansion	6518
Common Domain Model (CDM)	10
Financial Mathematics (FM)	1036
Definitions	2279
Link Retrieval	2279
Named Entity Recognition (NER)	2781
Question-Answering (QA)	3087
OSI Abbreviation	131
OSI Question-Answering	219

Table 4: Summary of Instruction Dataset Tasks

5 Training Methodology

The training process was conducted on an NVIDIA A40 GPU. The base model employed was the 8-billion-parameter LLaMA (LLaMA 3.1). Additionally, the associated tokenizer was modified to include a custom padding token, ensuring consistent input formatting throughout the training process.

The dataset was randomized and divided into training and validation subsets, with 95% of the batches allocated for training and the remaining 5% for validation.

[2] https://huggingface.co/datasets/ChanceFocus/flare-cfa

[3] https://huggingface.co/datasets/mirageco/XBRLBench

Step	Training Loss	Validation Loss
500	2.7693	2.7813
1000	2.7211	2.7117
1500	2.7003	2.6780
2000	2.6606	2.6627

Table 5: Training and Validation Loss per Step

5.1 Further Pretraining Process

The training was conducted using the Hugging Face `Trainer` class with the following hyperparameter configuration:

- Batch size: 28 for training and 20 for evaluation per device.
- Gradient accumulation steps: 4.
- Optimizer: AdamW with an 8-bit precision variant.
- Learning rate: 2e-4 with a warm-up of 10 steps.
- Evaluation strategy: Validation performed every 500 steps.
- Number of epochs: 2.
- Floating-point precision: FP16.

The training process was monitored for both training loss and validation loss at regular intervals. Table 5 summarizes the performance metrics recorded during training:

Upon completion of training, the model and tokenizer were saved for downstream tasks. The fine-tuned model showed consistent improvement, as seen in the decreasing training and validation losses. These results suggest that the model adapted well to the fine-tuning dataset without overfitting.

6 Fine-Tuning the Model with Instruction Data

We developed a two-stage fine-tuning approach driven by the differing context window requirements across tasks, with Named Entity Recognition (NER) posing unique computational and contextual challenges. While most instruction-based tasks could be effectively handled within a standard 128-token context, NER requires a much larger context window to capture complex interdependencies and long-range relationships in the text.

We designed a two-part fine-tuning strategy to address these contextual differences:

1. **Initial Fine-Tuning (4 Epochs):** We utilized the pre-trained Llama-3.1-8B model as a base,

fine-tuning it on all tasks excluding NER. With a constrained 128-token context window.

2. **NER-Specific Fine-Tuning (1 Epoch):** Recognizing the inherent complexity of Named Entity Recognition, we performed a specialized fine-tuning step using a substantially expanded 512-token context window. This approach ensures the model can effectively parse and understand the nuanced, extended textual dependencies critical to accurate NER task performance.

6.1 Implementation Details

- **Quantization:** The model was loaded with 4-bit quantization to optimize memory usage and computational efficiency, employing the "nf4" quantization type with mixed precision.
- **Dataset Preparation:** Input data was tokenized and stratified into training and validation sets (95%/5%), with a custom PyTorch dataset class handling token masking and formatting.
- **LoRA Fine-Tuning:** We applied Low-Rank Adaptation (LoRA) with hyperparameters: $r = 16$, $\alpha = 32$, and a dropout rate of 0.05.
- **Training Configuration:** Training was conducted using a batch size of 14, with gradient accumulation over 4 steps, and a learning rate of 2×10^{-4} for 3 epochs (general tasks) and 1 epoch (NER).

7 Results

In this section, we present the findings of our study on the performance of our domain-specific large language model for regulatory and financial text understanding. We compare our model's performance with baseline models including GPT-4o, Llama 3.1 8B, and Mistral Large 2. The full leaderboard results can be found at https://coling2025regulations.thefin.ai/winners.

7.1 Comparison with Baseline Models

7.2 Task-Specific Analysis

7.2.1 Performance Across Tasks

Our model's performance reveals significant variability across different specialized tasks. In Abbreviation Recognition, our score of 0.2748 demonstrates competitiveness with Llama 3.1 8B (0.2320), though still trailing behind GPT-4o (0.3784). The Definition Task presents a similar challenge, with our 0.4688 score positioned below top performers like FinMind-Y-Me (0.5849) and GPT-4o (0.5520).

Model	Final Score	Abbreviation
Our Model	0.43929	0.2748
FinMind-Y-Me	0.54801	0.2095
GPT-4o	0.63567	0.3784
Llama 3.1 8B	0.53572	0.2320
Mistral Large 2	0.62489	0.2230

Table 6: Performance Comparison: Model Scores and Abbreviation

Model	Definition	NER	QA
Our Model	0.4688	0.4302	0.7688
FinMind-Y-Me	0.5849	0.7174	0.8609
GPT-4o	0.5520	0.7108	0.8842
Llama 3.1 8B	0.5130	0.6352	0.8079
Mistral Large 2	0.5338	0.7062	0.8263

Table 7: Performance Comparison: Task-Specific Metrics

Named Entity Recognition (NER) emerged as a critical weakness, with our 0.4302 score substantially lagging behind FinMind-Y-Me (0.7174) and GPT-4o (0.7108), signaling an urgent area for methodological refinement. Conversely, our Question-Answering (QA) performance stands out as a notable strength, scoring 0.7688 surpassed by GPT-4o's 0.8842. However given that our model is much smaller, this demonstrates robust effectiveness in this domain.

7.3 Comprehensive Task Breakdown

The detailed task analysis unveils nuanced performance characteristics across specialized financial domains. Our Certificate-related tasks scored 0.3112, markedly lower than FinMind-Y-Me (0.4701) and GPT-4o (0.6568), suggesting potential improvements through expanded training data and more comprehensive public dataset integration. XBRL Analytics similarly revealed performance limitations, with an average score of 0.3444 indicating the need for enhanced financial reporting language processing capabilities. The Common Data Model (CDM) interpretation, scoring 0.2857, further highlighted structural data processing as a key development area.

7.4 Analysis of Leaderboard Performance

Our final weighted score of 0.43929 secures a second-place position, simultaneously highlighting the model's promising potential and significant improvement opportunities. While Question-

Answering tasks demonstrate our model's inherent strengths, critical development pathways clearly emerge in Named Entity Recognition, XBRL Analytics, and Certificate-related computational tasks. These findings provide a strategic roadmap for future model refinement and targeted performance enhancement.

8 Conclusions and Future Work

Our research presents an approach to developing a domain-specific large language model (LLM) for regulatory and financial text interpretation, addressing the critical challenges of extracting and analyzing complex regulatory documents. By constructing a data collection and preprocessing pipeline, we successfully created a corpus of 2,286 diverse regulatory documents. The methodology integrated recursive web scraping, TF-IDF-based relevance scoring, and text cleaning techniques using GPT-4o-mini, demonstrating a novel approach to building domain-specific training datasets.

The two-stage fine-tuning strategy utilizing LLaMA-3.1-8B revealed both the potential and limitations of our domain-specific model. While achieving notable strengths in question-answering tasks, the model also exposed critical areas for future improvement, particularly in named entity recognition and XBRL analytics. These insights not only highlight the complexities of developing specialized language models for regulatory domains but also provide a clear roadmap for future research, emphasizing the need for more sophisticated approaches to capturing the nuanced language of financial and regulatory texts.

Our study contributes to the field of domain-specific natural language processing by demonstrating the feasibility and challenges of creating targeted large language models. By providing a transparent methodology, we offer researchers and practitioners a valuable framework for developing more accurate and contextually aware language models in specialized domains, ultimately advancing the capability of AI to understand and process complex regulatory information.

Due to time and resource constraints, we were unable to conduct comprehensive expert validation. Inspired by the work of Chen et al. (Chen et al., 2024), we propose developing a novel methodology to create discriminative small language models specifically designed for autonomous data quality assessment, in close collaboration with domain experts.

Drawing from their "Honest AI" approach, we aim to develop a collaborative framework where specialized small language models (e.g., BERT) are trained with data curated by legal or financial experts. These models will be co-designed to validate data, acknowledge limitations, and provide transparent insights. By integrating expert knowledge throughout the model development process, we can create a scalable and efficient approach to data validation that combines the strengths of AI and human expertise.

The proposed system would:

- Train models to recognize subtle domain-specific nuances
- Develop mechanisms to confidently identify information gaps
- Provide clear indications of potential hallucinations

By enhancing the dataset, further improvements in accuracy and truthfulness could be achieved by building a knowledge base and implementing RAG on top of the fine-tuned model. This would allow for adjustments such as different chunk splitting methods, indexing techniques, and hybrid search implementations. These changes would help the model handle large documents that exceed its context window, a key consideration given the extensive nature of regulatory texts. Additionally, implementing a retrieval strategy to provide better context for answering queries could reduce hallucinations and improve the accuracy and relevance of the responses.

Larger models could improve task performance, particularly for tasks that require structured responses or long sequence retention, such as NER and link retrieval. Bigger models are better at capturing intricate patterns in structured text, as they can memorize more information from training data than smaller models (Tirumala et al., 2022). Using a higher-parameter model with our training data would be a logical next step to assess improvements in these tasks.

References

Xinxi Chen, Li Wang, Wei Wu, Qi Tang, and Yiyao Liu. 2024. Honest ai: Fine-tuning "small" language models to say "i don't know", and reducing hallucination in rag. *Preprint*, arXiv:2410.09699.

Samuel Dahan, Rohan Bhambhoria, David Liang, and Xiaodan Zhu. 2023. Lawyers should not trust ai: A call for an open-source legal language model. *Queen's University Legal Research Paper*. Available at SSRN: https://ssrn.com/abstract=4587092 or http://dx.doi.org/10.2139/ssrn.4587092.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jonathan Li, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2024. Experimenting with legal ai solutions: The case of question-answering for access to justice. *Preprint*, arXiv:2409.07713.

Vaibhav Mavi, Abulhair Saparov, and Chen Zhao. 2023. Retrieval-augmented chain-of-thought in semi-structured domains. *Preprint*, arXiv:2310.14435.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Preprint*, arXiv:2205.10770.

Keyi Wang, Sarah Huang, Charlie Shen, Kaiwen He, Felix Tian, Jaisal Patel, Christina Dan Wang, Kairong Xiao, and Xiao-Yang Liu. 2024. Professional readiness of llms in financial regulations? a report of regulations challenge at coling 2025. *International Workshop on Multimodal Financial Foundation Models (MFFMs) at 5th ACM International Conference on AI in Finance (MFFM at ICAIF '24)*.

Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. 2024. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. *Preprint*, arXiv:2406.03600.