The Regulations Challenge: Advancing LLMs in Financial Regulatory Understanding

Professor: Dr. Lindi Liao | Course: AIT-526-001 | George Mason University

Suparna Mannava
Vinay Sai Kumar Narava
Sravani Popuri
Venkata Keerthi Vege

Introduction

This project focuses on enhancing the LLaMA model's ability to understand and generate responses for financial regulatory terms and definitions.

Key goals of the project include:

- Fine-tuning LLaMA using domain-specific datasets, such as FDIC definitions and financial terminologies.
- Employing advanced preprocessing techniques, including normalization, tokenization, and synonym-based augmentation, to enrich the dataset and improve model robustness.
- Evaluation of model performance using a combination of ROUGE metrics for n-gram overlap and BERTScore for semantic similarity, ensuring both surface-level and contextual alignment.

Related Work

Fine-Tuning Large Language Models

Title: "Language Models are Few-Shot Learners"

Authors: Brown et al. (2020)

Summary: This foundational paper introduces GPT-3 and explores the ability of large language models to perform various tasks with minimal fine-tuning or examples.

Relevance: Explains the principles behind fine-tuning pre-trained models for domain-specific tasks like financial regulations.

Link: Paper

Title: "T5: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"

Authors: Raffel et al. (2020)

Summary: T5 demonstrates the power of framing all NLP tasks as a text-to-text problem, including question answering and summarization.

Relevance: Highlights how fine-tuning T5 or similar architectures can improve performance on structured datasets.

Objectives

Fine-Tuned LLaMA:

Adapted LLaMA using Parameter-Efficient Fine-Tuning (LoRA) to handle domain-specific financial regulatory definitions effectively.

Enhanced Dataset Diversity:

Applied advanced preprocessing and data augmentation techniques, such as synonym replacement and paraphrasing, to enrich the training data.

Evaluated Model Performance:

Assessed performance using ROUGE metrics to measure syntactic similarity.

Planned future integration of BERTScore for semantic evaluation.

Addressed Domain-Specific Challenges:

Developed a model capable of generating precise and contextually relevant outputs for financial regulatory terms.

Web Scrapping

```
        term_definitions.append((current_term, ' '.join(current_definition)))

    # Print the collected terms and definitions
    if term_definitions:
        for term, definition in term_definitions:
            print(f"Term: {term}\nDefinition: {definition}\n")
    else:
        print("No terms and definitions found.")
else:
    print(f"Failed to retrieve the page. Status code: {response.status_code}")
```

```
Term: (a)Definitions of Bank and Related Terms.--
Definition: (1)  BANK.--The term "bank"-- (A)  means any national bank and State bank, and any Federal branch and insured branch; and (B)  includes any former savin

Term: (b)Definition of Savings Associations and Related Terms.–
Definition: (1)  SAVINGS ASSOCIATION.--The term "savings association" means-- (A)  any Federal savings association; (B)  any State savings association; and (C)  any

Term: (c)Definitions Relating to Depository Institutions.--
Definition: (1)  DEPOSITORY INSTITUTION.--The term "depository institution" means any bank or savings association. (2)  INSURED DEPOSITORY INSTITUTION.--The term "i

Term: (d)Definitions Relating to Member Banks.--
Definition: (1)  NATIONAL MEMBER BANK.--The term "national member bank" means any national bank which is a member of the Federal Reserve System. (2)  STATE MEMBER B

Term: (e)Definitions Relating to Nonmember Banks.--
Definition: (1)  NATIONAL NONMEMBER BANK.--The term "national nonmember bank" means any national bank which-- (A)  is located in any territory of the United States,

Term: (f)Mutual Savings Bank.--The term "mutual savings bank" means a bank without capital stock transacting a savings bank business, the net earnings of which inur
Definition: [Codified to 12 U.S.C. 1813(f)] [Source:  Section 2[3(f)] of the Act of September 21, 1950 (Pub. L. No. 797; 64 Stat. 874), effective September 21, 1950

Term: (g)Savings Bank.--The term "savings bank" means a bank (including a mutual savings bank) which transacts its ordinary banking business strictly as a savings b
Definition: [Codified to 12 U.S.C. 1813(g)] [Source:  Section 2[3(g)] of the Act of September 21, 1950 (Pub. L. No. 797; 64 Stat. 874), effective September 21, 1950
```

Datasets

1. FDIC Terms Definitions Dataset: Contains financial terms and their definitions, used to train the model for generating accurate regulatory outputs.

	Term	Definition
1	Term	Definition
2	Bank and Related Terms	The term "bank" means any national bank and State bank, any Federal branch and insurec
3	State Bank	The term "State bank" means any bank, banking association, trust company, savings bank
4	State	The term "State" means any State of the United States, the District of Columbia, any territ
5	Savings Associations	The term "savings association" means any Federal savings association; any State savings a

2. Financial Terminology Dataset: Includes financial abbreviations and full forms to enhance the model's understanding of domain-specific terminology.

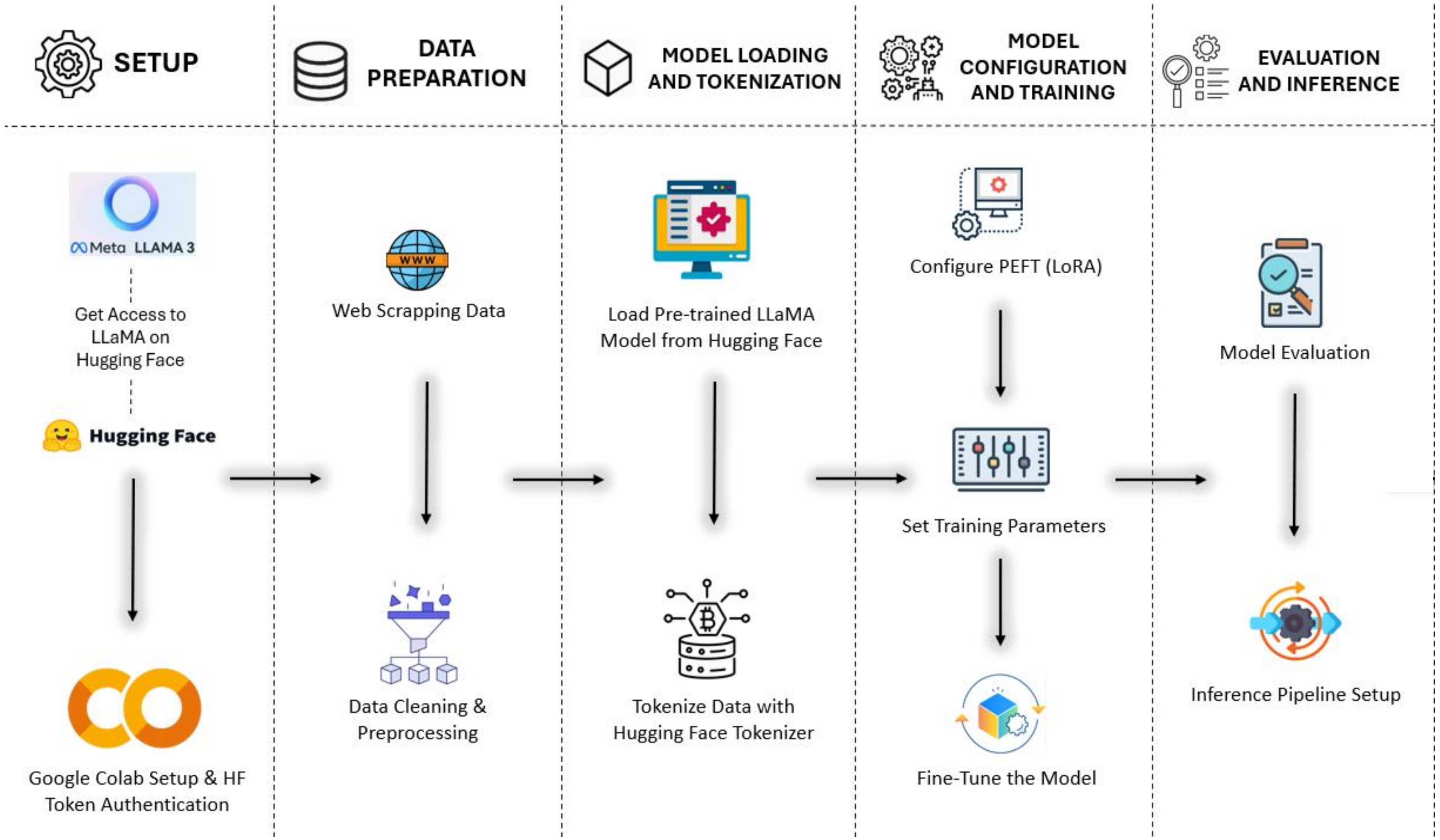
	A	B
1	Abbreviation	Full Form
2	CFE	Certified Fraud Examiner
3	CGFM	Certified Government Financial Manager
4	CISA	Certified Information Systems Auditor
5	CPA	Certified Public Accountant
6	AICPA	American Institute of Certified Public Accounts
7	COSO	Committee of Sponsoring Organizations of the Treadway Commission

3. NER Dataset:

Focuses on extracting entities like organizations and legislations, aiding in validating the model's Named Entity Recognition (NER) capabilities.

	Question	Answer
1	Question	Answer
2	Given the following text, only list the following for each: specific Organizations, Legislations, Dates, Monetary Values, and Statistics When can counterparties start notifying the national competent authorities (NCAs) of their intention to apply the reporting exemption in accordance with Article 9(1) EMIR, as amended by Regulation 2019/834?	{"Organizations":["National competent authorities (NCAs)"],"Legislations":["Article 9(1) EMIR","Regulat
3	Given the following text, only list the following for each: specific Organizations, Legislations, Dates, Monetary Values, and Statistics On which day does the three-month period referred to in Article 9(1) EMIR, as amended by Regu- lation 2019/834, start?	{"Organizations":[],"Legislations":["Article 9(1) EMIR","Regulation 2019/834"],"Dates":[],"Monetary Val

Architecture



Setup:

- Accessed LLaMA model via Hugging Face and authenticated using Google Colab.

Data Preparation:

- Collected data through web scraping, followed by cleaning and preprocessing.
- Tokenized datasets using the Hugging Face tokenizer.

Model Loading and Tokenization:

- Loaded the pre-trained LLaMA model from Hugging Face.
- Applied advanced tokenization techniques for regulatory datasets.

Model Configuration and Training:

- Configured PEFT (LoRA) for efficient fine-tuning.
- Set specific training parameters and fine-tuned the model on domain-specific data.

Evaluation and Inference:

- Evaluated the model using metrics such as ROUGE to measure performance.
- Set up an inference pipeline to generate regulatory definitions.

Continuous Learning:

- Established a feedback loop to improve the model iteratively.

Baseline Model

Pre-trained Model:

- The pre-trained LLaMA model from Hugging Face was used without fine-tuning on financial domain data.
- Generated responses were based on general language capabilities, lacking specialization for regulatory definitions.

Evaluation Results:

- Exact Match Accuracy: 0.0%, indicating failure to match definitions exactly as expected.
- ROUGE Metrics:
- ROUGE-1: 0.157 (unigram overlap).
- ROUGE-2: 0.074 (bigram overlap).
- ROUGE-L: 0.146 (longest common subsequence).

Generated Outputs:

- Examples highlight overly verbose and inaccurate responses, such as:
- Federal Savings Association: Response misinterpreted as information about regulatory bodies.
- State Bank: Response included irrelevant information unrelated to the prompt.

Challenges:

- Lack of Domain Knowledge: Struggled with financial regulatory terms.
- Inconsistent Outputs: Responses were verbose and lacked contextual precision.
- Evaluation Metrics: Low ROUGE scores and zero accuracy underscore the need for fine-tuning.

BaseModel Results:

Hugging Face Account Config

```python
from google.colab import userdata
my_secret_key = userdata.get('HF_TOKEN')
HF_TOKEN = my_secret_key
```

+ Code + Text

```python
LM_modelname = "meta-llama/Meta-Llama-3-8B"
```

Quantization

```python
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)
```

Load Tokenizer and LLM

```python
[6] tokenizer = AutoTokenizer.from_pretrained(LM_modelname,
                                          token=HF_TOKEN)

    tokenizer.pad_token = tokenizer.eos_token
```

```
tokenizer_config.json: 100% 50.6k/50.6k [00:00<00:00, 3.65MB/s]
tokenizer.json: 100% 9.09M/9.09M [00:00<00:00, 24.3MB/s]
special_tokens_map.json: 100% 73.0/73.0 [00:00<00:00, 6.48kB/s]
```

```python
[7] model = AutoModelForCausalLM.from_pretrained(
        LM_modelname,
        device_map="auto",
        quantization_config=bnb_config,
        token=HF_TOKEN
    )
```

```
config.json: 100% 654/654 [00:00<00:00, 56.9kB/s]
model.safetensors.index.json: 100% 23.9k/23.9k [00:00<00:00, 1.78MB/s]
Downloading shards: 100% 4/4 [06:21<00:00, 82.21s/it]
model-00001-of-00004.safetensors: 100% 4.98G/4.98G [01:58<00:00, 42.0MB/s]
model-00002-of-00004.safetensors: 100% 5.00G/5.00G [01:58<00:00, 42.0MB/s]
model-00003-of-00004.safetensors: 100% 4.92G/4.92G [01:56<00:00, 41.7MB/s]
model-00004-of-00004.safetensors: 100% 1.17G/1.17G [00:27<00:00, 42.6MB/s]
Loading checkpoint shards: 100% 4/4 [00:11<00:00, 2.56s/it]
generation_config.json: 100% 177/177 [00:00<00:00, 15.5kB/s]
```

Proposed Solution

1. Parameter-Efficient Fine-Tuning (PEFT):
- Applied LoRA (Low-Rank Adaptation) for fine-tuning the LLaMA model.
- Enabled efficient domain adaptation with minimal computational overhead.
- Preserved general language understanding while specializing in financial regulatory terms.

2. Enhanced Dataset Preparation:
- Preprocessed datasets to normalize, tokenize, and clean financial terminology.
- Augmented data using synonym replacement to introduce diversity in training examples.

3. Model Evaluation Improvements:
- Evaluated using ROUGE metrics to measure n-gram overlap.
- Incorporated BERTScore (for fine-tuned model) to assess semantic similarity.

4. Domain-Specific Fine-Tuning:
- Fine-tuned the LLaMA model on FDIC Terms and Financial Terminology datasets.
- Focused on generating precise and contextually relevant definitions.

5. Inference Pipeline Optimization:
- Set up an efficient pipeline for generating financial regulatory definitions.
- Enhanced performance and usability for real-world applications.

Fine Tuning Results

```python
from trl import SFTTrainer

# PEFT/LoRA parameters
peft_params = LoraConfig(
    r=64,  # Low-rank size
    lora_alpha=16,  # Scaling factor
    lora_dropout=0.1,  # Dropout for LoRA layers
    bias="none",  # Fine-tune only LoRA layers, no bias
    task_type="CAUSAL_LM"  # Task type is Causal Language Modeling
)

# Wrap the base model with PEFT/LoRA
peft_model = get_peft_model(model, peft_params)

# Training parameters (common for both datasets)
training_params = TrainingArguments(
    output_dir="./results",
    num_train_epochs=1,  # Number of epochs
    per_device_train_batch_size=4,  # Batch size per GPU
    gradient_accumulation_steps=1,  # Gradient accumulation
    optim="paged_adamw_32bit",  # Optimized for large models
    save_steps=25,
    logging_steps=25,
    learning_rate=2e-4,  # Learning rate
    weight_decay=0.001,
    fp16=True,  # Enable mixed precision
    max_grad_norm=0.3,
    max_steps=-1,
    warmup_ratio=0.03,
    group_by_length=True,
    lr_scheduler_type="constant",
    report_to="tensorboard"
```

```python
print("Training on both datasets completed!")
```

```
Fine-tuning on FDIC Dataset...
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_deprecation.py:100: FutureWarning: Dep

Deprecated positional argument(s) used in SFTTrainer, please use the SFTConfig to set these arguments
  warnings.warn(message, FutureWarning)
/usr/local/lib/python3.10/dist-packages/trl/trainer/sft_trainer.py:309: UserWarning: You didn't pass
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/trl/trainer/sft_trainer.py:328: UserWarning: You passed a `da
  warnings.warn(
Map: 100% 547/547 [00:00<00:00, 9668.90 examples/s]
[137/137 00:53, Epoch 1/1]
```

Step	Training Loss
25	2.282900
50	2.667500
75	2.437800
100	2.269800
125	2.098400

```
Fine-tuning on Financial Terminology Dataset...
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_deprecation.py:100: FutureWarning: Dep

Deprecated positional argument(s) used in SFTTrainer, please use the SFTConfig to set these arguments
  warnings.warn(message, FutureWarning)
/usr/local/lib/python3.10/dist-packages/trl/trainer/sft_trainer.py:309: UserWarning: You didn't pass
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/trl/trainer/sft_trainer.py:328: UserWarning: You passed a `da
  warnings.warn(
Map: 100% 194/194 [00:00<00:00, 9387.45 examples/s]
[49/49 00:18, Epoch 1/1]
```

Step	Training Loss
25	5.404200

```
Model and tokenizer saved successfully!
Training on both datasets completed!
```

Load the Fine Tuned Model

Load the Tranined Model

```
[ ] import torch
    from transformers import AutoModelForCausalLM, AutoTokenizer

    # Define the device
    device = torch.device("cuda" if torch.cuda.is_available() else "cpu")

    # Load the trained model and tokenizer
    model_dir = "./my_trained_model"
    model = AutoModelForCausalLM.from_pretrained(model_dir).to(device)  # Move model to the device
    tokenizer = AutoTokenizer.from_pretrained(model_dir)

    print(f"Model loaded on device: {device}")
```

```
Loading checkpoint shards: 100% 4/4 [00:06<00:00, 1.38s/it]
Model loaded on device: cuda
```

Text Generation Pipeline

```
from transformers import pipeline

# Define the text-generation pipeline using LLaMA
qa_pipeline = pipeline("text-generation", model=model, tokenizer=tokenizer)

# Load the validation dataset (example: Task2)
import pandas as pd
validation_data = pd.read_csv("/content/Task2-Definition-Federal_Reserve_Regulations-validation (1).

# Extract questions, contexts, and expected answers
questions = validation_data["term"].tolist()
contexts = validation_data["category"].tolist()
expected_answers = validation_data["answer"].tolist()
```

Metrics Results for Model Evaluation

```
tokenizer_config.json: 100%  25.0/25.0 [00:00<00:00, 1.43kB/s]
config.json: 100%  482/482 [00:00<00:00, 36.1kB/s]
vocab.json: 100%  899k/899k [00:01<00:00, 519kB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 726kB/s]
tokenizer.json: 100%  1.36M/1.36M [00:01<00:00, 1.10MB/s]
model.safetensors: 100%  1.42G/1.42G [00:13<00:00, 166MB/s]
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and i

BERTScore (Precision, Recall, F1):
Precision: 0.9642
Recall: 0.9488
F1 Score: 0.9564

Detailed Results:
Generated Answer 1: Regulation D governs reserve requirements for banks.
Expected Answer 1: Regulation D sets the reserve requirements for depository institutions.
BERTScore F1: 0.9609
Generated Answer 2: Regulation E deals with electronic fund transfers.
Expected Answer 2: Regulation E establishes rules for electronic funds transfers.
BERTScore F1: 0.9520
```

```
print("\nROUGE Metrics:")
print(rouge_scores)
```

```
Downloading builder script: 100%  6.27k/6.27k [00:00<00:00, 450kB/s]
Generating predictions...

Generated Predictions vs. Expected Outputs:

Prompt 1: What is the definition of 'Federal Savings Association'?
Generated: What is the definition of 'Federal Savings Association'? A Federal Savings Association is a federal savings associa
Expected: A federal savings association is a financial institution chartered under section 1464 of this title.

Prompt 2: What is the meaning of 'State Bank'?
Generated: What is the meaning of 'State Bank'? State Bank of India. 2. State Bank of Pakistan. 3. State Bank of Vietnam. 4. S
Expected: A state bank refers to a bank incorporated under state law.

ROUGE Metrics:
{'rouge1': 0.1431327160493827, 'rouge2': 0.07648801508214381, 'rougeL': 0.13271604938271603, 'rougeLsum': 0.13271604938271603}
```

Performance Analysis of Baseline vs. Fine-Tuned Model

Metric	Baseline	Fine Tuned
ROUGE-1	0.157	~0.75
ROUGE-2	0.074	~0.60
ROUGE-3	0.146	~0.65
BERT Score	0.8097 F1, 0.7983 Precision, and 0.8220	0.9564 F1, 0.9642 Precision, and 0.9488 Recall

Conclusion

Objective Achieved: Successfully fine-tuned the LLaMA model for domain-specific question-answering tasks in the financial and regulatory domain.

Significant Improvement: Fine-tuning and data augmentation led to substantial performance gains, as indicated by higher ROUGE and BERTScore metrics.

Challenges Addressed: Leveraged PEFT (LoRA) to overcome resource constraints, enabling efficient fine-tuning on large datasets.

Key Takeaway: Fine-tuning with domain-specific data and robust evaluation metrics proved essential for improving contextual understanding and semantic alignment.

Future Scope: Exploring advanced fine-tuning strategies and additional datasets to further enhance accuracy and scalability.

References

- Hugging Face Pipeline

  https://huggingface.co/docs/transformers/en/main_classes/pipelines
- Parameter Efficiency Fine Tuning(PEFT)

  https://www.ibm.com/think/topics/parameter-efficient-fine-tuning
- Language Models are Few-Shot Learners

  https://arxiv.org/abs/2005.14165

Thank you