### **Finetuned LLM Model Comparison: Phi-3 vs. LLaMA for Multi class task**

#### **1. Introduction**

Both the Phi-3 and LLaMA models have been fine-tuned for multi-class tasks. This documentation focuses on their performance metrics, efficiency, and overall suitability for the task.

#### **2. Data Overview**

* **FinancialPhraseBank Dataset**:  
  + **Introduction**: Within the realm of finance and economic texts, annotated datasets are notably rare, with many being exclusively reserved for proprietary purposes. To address the issue of insufficient training data, scholars from the Aalto University School of Business introduced in 2014 a set of approximately 5000 sentences.
  + **Purpose**: This collection aimed to establish human-annotated benchmarks, serving as a standard for evaluating alternative modeling techniques.
  + **Annotators**: The involved annotators (16 people with adequate background knowledge on financial markets) were instructed to assess the sentences solely from the perspective of an investor, evaluating whether the news potentially holds a positive, negative, or neutral impact on the stock price.
  + **Structure**: The FinancialPhraseBank dataset captures the sentiments of financial news headlines from the viewpoint of a retail investor. Comprising two key columns, namely "Sentiment" and "News Headline," the dataset effectively classifies sentiments as either negative, neutral, or positive.
  + **Significance**: This structured dataset serves as a valuable resource for analyzing and understanding the complex dynamics of sentiment in the domain of financial news.

**3. Architecture**

* Meta-Llama/Meta-Llama-3-8B:  
  + **Size:** 8 billion parameters, making it a larger model with potentially greater capabilities.
  + **Open source:** Based on the Llama architecture and available for research and commercial use.
  + **Performance:** Demonstrates strong performance across various tasks like text generation and question-answering.
  + **Versatility:** Suitable for a wide range of applications, thanks to its size and open nature.
  + **Context window:** ~8K tokens
  + **Data trained on:** A new mix of publicly available online data.
* **Microsoft/Phi-3-mini-4k-instruct:**
  + **Size:** 3.8B parameters, making it a very small model.
  + **Efficiency:** Designed for efficiency and can run on low-resource devices like smartphones.
  + **Instruction-tuned:** Fine-tuned to follow instructions, making it particularly good at tasks like code generation and summarization.
  + **Performance:** Despite its small size, it performs surprisingly well in certain tasks, especially when compared to other models of similar size.
  + **Context window:** 4096 tokens (approximately 3200 words)
  + **Data trained on:** Not publicly disclosed, but likely a mix of web text and code.

#### **4. Dataset and Preparation**

* **Phi-3 Model**:
  + **Dataset Split**: The dataset was split into training and test sets with 300 samples each, stratified by sentiment.
  + **Evaluation Data**: Residual examples not included in the train or test sets were used for evaluation, and sampled to ensure equal representation of sentiment classes.
* **LLaMA Model**:
  + **Dataset Split**: Similar stratified split into training and test sets, ensuring balanced representation of sentiments.
  + **Evaluation Data**: A balanced set of residual examples was used for evaluation.

#### **5. Model Fine-Tuning Process**

* **Phi-3 Model**:
  + **Training**: Used a prompt-based approach where the texts were transformed into prompts that included the expected sentiment.
  + **Monitoring**: Training progress was monitored using TensorBoard.
* **LLaMA Model**:
  + **Training**: Employed a similar prompt-based transformation for the input data.
  + **Monitoring**: Also utilized TensorBoard for tracking training metrics.
* **Stratified Splitting**: Ensured balanced representation of sentiment classes in both training and test datasets.
* **Prompt Generation**: Facilitated the transformation of raw text into a format suitable for fine-tuning.
* **Balanced Evaluation**: Used repeated sampling to maintain a balanced evaluation dataset.

##### **Training and Logging**

TensorBoard was utilized to log the training progress, providing a visual representation of the model’s performance metrics over time. This tool is essential for monitoring and diagnosing the training process.

In depth detail around peft configuration, training arguments & their hyperparameter, sft & lora in the notebook itself

#### 

#### **6. Model Performance Before and After Fine-Tuning**

##### **Phi-3 Model**

* **Zero-Shot Performance Before Fine-Tuning**:
  + **Overall Accuracy**: 75.1%
  + **Accuracy by Sentiment Class**:
    - Positive: 71.3%
    - Neutral: 83.3%
    - Negative: 70.7%
  + **Classification Report**:
    - **Precision**:
      * Positive: 0.94
      * Neutral: 0.60
      * Negative: 0.83
    - **Recall**:
      * Positive: 0.71
      * Neutral: 0.83
      * Negative: 0.71
    - **F1-Score**:
      * Positive: 0.81
      * Neutral: 0.70
      * Negative: 0.76
  + **Confusion Matrix**:
    - [[214, 85, 1], [7, 250, 43], [7, 81, 212]]
* **After Fine-Tuning**:
  + **Overall Accuracy**: 87.4%
  + **Accuracy by Sentiment Class**:
    - Positive: 97.3%
    - Neutral: 83.3%
    - Negative: 81.7%
  + **Classification Report**:
    - **Precision**:
      * Positive: 0.94
      * Neutral: 0.81
      * Negative: 0.86
    - **Recall**:
      * Positive: 0.97
      * Neutral: 0.83
      * Negative: 0.82
    - **F1-Score**:
      * Positive: 0.96
      * Neutral: 0.82
      * Negative: 0.84
  + **Confusion Matrix**:
    - [[292, 5, 3], [14, 250, 36], [3, 52, 245]]
* **Training and Validation Loss**:
  + **Initial Loss**: 1.0515 (Epoch 0)
  + **Final Loss**: 0.5820 (Epoch 3)
  + **Validation Loss**: Increased slightly over epochs indicating possible overfitting.

##### **LLaMA Model**

* **Zero-Shot Performance Before Fine-Tuning**:
  + **Overall Accuracy**: 37.6%
  + **Accuracy by Sentiment Class**:
    - Positive: 11.7%
    - Neutral: 73.3%
    - Negative: 27.7%
  + **Classification Report**:
    - **Precision**:
      * Positive: 0.55
      * Neutral: 0.36
      * Negative: 0.37
    - **Recall**:
      * Positive: 0.12
      * Neutral: 0.73
      * Negative: 0.28
    - **F1-Score**:
      * Positive: 0.19
      * Neutral: 0.48
      * Negative: 0.32
  + **Confusion Matrix**:
    - [[35, 195, 70], [11, 220, 69], [18, 199, 83]]
* **After Fine-Tuning**:
  + **Overall Accuracy**: 85.8%
  + **Accuracy by Sentiment Class**:
    - Positive: 95.7%
    - Neutral: 78.0%
    - Negative: 83.7%
  + **Classification Report**:
    - **Precision**:
      * Positive: 0.93
      * Neutral: 0.81
      * Negative: 0.83
    - **Recall**:
      * Positive: 0.96
      * Neutral: 0.78
      * Negative: 0.84
    - **F1-Score**:
      * Positive: 0.94
      * Neutral: 0.79
      * Negative: 0.83
  + **Confusion Matrix**:
    - [[287, 11, 2], [16, 234, 50], [5, 44, 251]]
* **Training and Validation Loss**:
  + **Initial Loss**: 1.3047 (Epoch 0)
  + **Final Loss**: 0.4964 (Epoch 3)
  + **Validation Loss**: Increased slightly over epochs indicating possible overfitting.

#### **7. Model Efficiency**

* **Phi-3 Model**:
  + **Training Time**: [448/448 20:49, Epoch 3/4]
  + **Resource Utilization**: Efficient in terms of computational resources required for fine-tuning.
  + **Ease of Fine-Tuning**: High adaptability with less data required for effective fine-tuning.
* **LLaMA Model**:
  + **Training Time**: [[448/448 **34:54**, Epoch 3/4]
  + **Resource Utilization**: Comparatively higher resource usage due to model complexity.
  + **Ease of Fine-Tuning**: Requires more data and computational power for fine-tuning.

#### **8. Insights and Benchmarking**

#### **Insights:**

#### **Performance Improvement:**

#### Both models demonstrated significant performance improvements after fine-tuning, with Phi-3 achieving an accuracy of 87.4% and LLaMA reaching 85.8%.

#### Phi-3 showed a substantial increase in accuracy for all sentiment classes, particularly excelling in the positive sentiment category with a 97.3% accuracy.

#### LLaMA exhibited a balanced improvement across all sentiment classes, with the highest accuracy in the positive sentiment category at 95.7%.

#### **Efficiency:**

#### Phi-3: Being lightweight with approximately 125 million parameters, Phi-3 is highly efficient, making it suitable for deployment in resource-constrained environments.

#### LLaMA: With around 65 billion parameters, LLaMA is more resource-intensive but offers high capacity for handling complex and large-scale text generation tasks.

#### **Fine-Tuning Time:**

#### Phi-3: Completed fine-tuning more quickly due to its smaller size, which is advantageous for scenarios requiring rapid model updates and deployments.

#### LLaMA: Took longer to fine-tune given its larger architecture, but the process resulted in a model capable of capturing intricate patterns in the data.

#### **Context Window:**

#### Phi-3: Supports a context window of 512 tokens, sufficient for most financial news sentiment analysis tasks.

#### LLaMA: Offers a larger context window of 2048 tokens, beneficial for understanding longer and more complex documents.

#### **Model Capacity:**

#### Phi-3: Its compact design makes it a versatile choice for applications needing quick, real-time sentiment analysis without heavy computational overhead.

#### LLaMA: Its extensive parameter count and deeper layers allow for a more nuanced understanding of text, making it ideal for applications requiring high accuracy and detailed analysis.

#### **Benchmarking:**

#### **Accuracy Comparison:**

#### Phi-3: Post fine-tuning, Phi-3 outperformed LLaMA slightly with an accuracy of 87.4% vs. LLaMA’s 85.8%, highlighting its effectiveness despite being a smaller model.

#### LLaMA: Showed robust performance improvements but was marginally less accurate than Phi-3 after fine-tuning.

#### **Precision, Recall, and F1-Score:**

#### Phi-3: Achieved high precision and recall across all sentiment categories, particularly excelling in the positive sentiment class with an F1-score of 0.96.

#### LLaMA: Also performed well, with balanced precision and recall, and notable F1-scores, especially in the positive sentiment class at 0.94.

#### **Confusion Matrix Insights:**

#### Phi-3: Demonstrated fewer misclassifications in the positive sentiment class, indicating strong performance in identifying positive sentiments accurately.

#### LLaMA: Showed slightly higher misclassifications in the neutral sentiment class compared to Phi-3 but managed to maintain overall strong performance.

#### **Training and Validation Loss:**

#### Phi-3: Showed consistent training and validation loss reduction, with slight overfitting observed in later epochs.

#### LLaMA: Similar trend with effective loss reduction, indicating robust learning capability despite a more complex architecture.

#### **Resource Utilization:**

#### Phi-3: Optimal for environments with limited computational resources due to its smaller size and faster training times.

#### LLaMA: Best suited for high-resource environments where computational capacity is not a constraint, allowing it to leverage its extensive parameter set for superior performance on complex tasks.

#### 

#### **9.Scope for Improvements in Both Models**

1. **Contextual Understanding**:
   * **Current Performance**: Both models can benefit from better understanding the context in which sentiments are expressed.
   * **Recommendation**: Integrate additional contextual signals and external knowledge bases to enhance the models' understanding of context-specific sentiments.
2. **Continuous Learning**:
   * **Current Performance**: Static models may become outdated as language and sentiments evolve.
   * **Recommendation**: Implement continuous learning or online learning frameworks that allow the models to update with new data and trends over time.
3. **Cross-Domain Generalization**:
   * **Current Performance**: Performance may vary significantly across different domains (e.g., product reviews vs. social media posts).
   * **Recommendation**: Use domain adaptation techniques to improve the models’ ability to generalize across various domains and contexts.
4. **Robustness to Noisy Data**:
   * **Current Performance**: Performance can degrade with noisy or unstructured data.
   * **Recommendation**: Develop robust data preprocessing pipelines and noise-robust training techniques to handle unstructured or noisy text more effectively.

Implementing these improvements can significantly enhance the performance, efficiency, and robustness of both the Phi-3 and LLaMA models, making them more versatile and powerful tools for multi class task tasks in various applications.

#### **10. Conclusion**

Both models have their strengths and are suitable for different use cases. Phi-3 is efficient and performs well with less data, making it ideal for quick deployment scenarios. LLaMA, on the other hand, offers potentially higher performance but at the cost of increased resource usage, making it suitable for more detailed and large-scale sentiment analysis tasks.