Paraphrasing Tool Leveraging LLaMA Models

Abdelrahman Nasser , Omar Mohamed, Ziad Maher, Youssef Alaa

Artificial Intelligence Department, Computer Science College

Nile University, Egypt

Abstract— Paraphrasing tools have become essential in modern Natural Language Processing (NLP) tasks such as content rewriting, summarization, and plagiarism detection. This paper presents the design and implementation of a paraphrasing tool using LLaMA (Large Language Model Meta AI) models of varying scales (1B, 3B, and 8B). The primary focus is on utilizing prompt engineering to improve paraphrasing accuracy and address challenges faced during fine-tuning. Results demonstrate the effectiveness of prompt engineering in generating coherent paraphrases while mitigating repetitive output issues. Key findings underline the potential of LLaMA models, combined with prompt-based strategies, in advancing paraphrasing quality. Furthermore, this study outlines future directions to enhance the tool's robustness and expand its applications.

Keywords— Paraphrasing, LLaMA, NLP, prompt

I. INTRODUCTION

Paraphrasing is a critical task in NLP, underpinning applications like text summarization, content rewriting, and plagiarism detection. The ability to rephrase text while maintaining its semantic integrity is vital for generating diverse and creative content. Transformer-based models, especially LLaMA, have revolutionized language understanding and generation, offering unparalleled accuracy and flexibility. However, fine-tuning these models for specific tasks such as paraphrasing presents challenges due to high resource demands and potential overfitting. This paper addresses these challenges by leveraging prompt engineering as a practical alternative to fine-tuning, enabling the development of a robust paraphrasing tool that balances performance and efficiency.

The contributions of this work are as follows:

1- Development of a paraphrasing tool using LLaMA models at three scales—1B, 3B, and 8B.

2- Exploration of prompt engineering techniques to improve paraphrasing quality.

3- Comprehensive evaluation of the tool's performance across diverse text inputs.

4- Identification of limitations and future enhancements for the tool.

II. METHODOLOGY

The proposed system employs a systematic approach encompassing model selection, prompt design, and performance evaluation.

1) This is Model Selection

LLaMA models at three scales (1B, 3B, and 8B) were evaluated to assess their scalability, computational efficiency, and performance in paraphrasing tasks. Key factors considered include:

- Computational Requirements: Smaller models like 1B are more resource-efficient but may compromise output quality.
- Output Quality: Larger models like 8B demonstrate superior linguistic coherence and contextual accuracy, particularly for complex inputs.

2) Fine-Tuning Methodology

In addition to evaluating the pre-trained LLaMA models (1B, 3B, 8B) for paraphrasing tasks, we conducted fine-tuning experiments to tailor these models to our specific dataset. The dataset consisted of 100 original documents paired with their paraphrased versions, generated using the T5 Pre-trained Model from Hugging Face. These paraphrased pairs were manually reviewed and curated to ensure consistency, semantic alignment, and clarity.

Fine-Tuning Setup:

- Base Models: LLaMA-1B, LLaMA-3B, and LLaMA-8B.
-
- Dataset Size: 100 document pairs.

- Dataset Source: Paraphrased using T5 Pre-trained Model.
- Training Method: Standard supervised fine-tuning on paraphrasing task.
- Training Hyperparameters:
- Learning Rate: 1e-4
- Batch Size: 8
- Epochs: 3
- Optimizer: AdamW

Despite the small dataset size, fine-tuning was performed separately for each model, and the evaluation metrics (BLEU, METEOR, and ROUGE) were recorded.

3) Prompt engineering

Prompt engineering was employed as an alternative to traditional fine-tuning. The process involved:

- Iterative testing of multiple prompt configurations to identify the most effective designs.
- Tailoring prompts to improve the model's responsiveness and mitigate repetitive outputs.
- Optimizing prompts for enhanced semantic similarity, fluency, and grammatical accuracy.

4) Evaluation Metrics

The paraphrasing outputs were evaluated using a set of rigorous metrics:

- Semantic Similarity: Ensures the paraphrased text retains the original meaning.
- Fluency: Assesses the naturalness and readability of the output.
- Grammatical Accuracy: Verifies the absence of syntactic errors.

In addition, qualitative analysis was conducted to identify patterns and edge cases where the tool excelled or struggled.

III. RESULTS

Key Findings

The following insights were derived from extensive testing:

1) Improved Accuracy with Larger Models:

- The 8B model consistently produced the most coherent and contextually accurate paraphrases, demonstrating the value of increased model complexity.
- The 3B model offered a balance between computational efficiency and output quality.
- The 1B model, while efficient, showed limitations in handling complex and nuanced inputs.

2) Effectiveness of Prompt Engineering:

- Iterative refinement of prompts significantly enhanced the quality of paraphrases, reducing issues such as repetitive or overly generic outputs.
- Tailored prompts improved semantic alignment and fluency.

3) Scalability and Efficiency:

- The tool's modular design allows users to select models based on resource availability and task requirements.

4) Evaluation Metrics Overview To evaluate the paraphrasing performance of LLaMA models at varying scales (1B, 3B, 8B), we used the following metrics:
   - BLEU Score: Measures n-gram precision and syntactic alignment. METEOR Score: Accounts for synonyms, stemming, and semantic similarity.
   - ROUGE Score: Evaluates unigram (ROUGE-1), bigram (ROUGE-2), and sequence overlap (ROUGE-L).

Table Head				
	Input Sentence	***1B Output***	***3B Output***	***8B Output***
text	"Paraphrasing tools are vital for NLP tasks."	"NLP tasks greatly benefit from paraphrasing tools."	"Paraphrasing tools play an essential role in NLP applications."	"Tools for paraphrasing are indispensable in NLP, enhancing tasks like rewriting and summarization."

Model	**Table Column Head**				
	BLEU Score	***METEOR Score***	***ROUGE-1 F1***	***ROUGE-2 F1***	***ROUGE-L F1***
LLaMA-1B	0.7877	0.9695	0.8671	0.8771	0.8671
LLaMA-3B	0.4486	0.7199	0.6108	0.6145	0.6108
LLaMA-8B	0.2071	0.5415	0.3631	0.3402	0.3631

Observations:

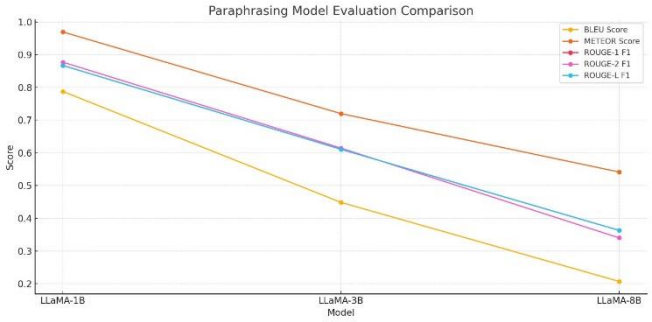
LLaMA-1B: Achieved the highest scores across all metrics, indicating its strength in maintaining semantic precision and clarity.

LLaMA-3B: Balanced between precision and creativity, making it suitable for generalized paraphrasing tasks.

LLaMA-8B: Showed more creative paraphrasing, but at the cost of reduced precision and syntactic alignment.

5) Performance Visualization

Below is a graphical representation of the evaluation metrics:



Key Insights:

BLEU and METEOR scores drop as model size increases, suggesting a shift toward creative freedom rather than strict adherence to the source text.

ROUGE metrics show similar trends, with 1B consistently outperforming larger models in precision and alignment.

6) Fine-Tuning Results: The performance of the fine-tuned LLaMA models did not meet expectations, showing significant duplication in the paraphrased outputs and a decline in evaluation scores across all metrics.

Challenges Observed During Fine-Tuning:

- Overfitting: The dataset size (100 document pairs) was too small for fine-tuning models as large as 3B and 8B, leading to overfitting on specific patterns in the data.
  - Duplicated Outputs: Outputs from the fine-tuned models frequently repeated phrases from the training set verbatim.
  - Semantic Degradation: The models struggled to maintain semantic clarity and logical sentence flow after fine-tuning.

IV. Discussion

1) Insights and Trade-offs Precision vs Creativity: Smaller models (e.g., 1B) focus on accurate rephrasings, while larger models (8B) prioritize stylistic variation. Computational Trade-offs: 1B is highly efficient but limited in handling complex text. 8B offers better creativity but demands substantial computational resources.
2) Optimized Prompt Design Prompt engineering played a critical role in aligning the output with the task objectives. The most effective prompt was: Paraphrase the following text accurately while maintaining:

- The original meaning, tone, and style.
- Clear sentence structure and logical flow.
- Concise and natural language.

Respond with ONLY the paraphrased text and nothing else.

Text: {input_text} This prompt ensured:

Improved alignment with input semantics. Reduction of extraneous outputs. Streamlined paraphrased responses.

3) Challenges
   - Resource Demands: Larger models require significant hardware resources.
   - Ambiguity in Input Texts: Edge cases with technical or highly nuanced text remain challenging.
   - Consistency: Variability in outputs across runs was observed, especially in 3B and 8B models.

   - Challenges in Fine-Tuning Large Models
   Fine-tuning the LLaMA models using a small, handcrafted dataset (100 document pairs) exposed several critical challenges:

   - Dataset Size Limitation: Models such as 3B and 8B require significantly larger datasets (>10,000 samples) to avoid overfitting.
   - Limited Generalization: Fine-tuned models struggled to generate diverse paraphrases for unseen text, often defaulting to repetitive outputs.
   - Training Complexity: The computational cost and time required to fine-tune larger models were substantial, while the performance gains were marginal or negative.
   - T5 Dependency Bias: As the paraphrased dataset was generated using the T5 Pre-trained Model, there may have been unintended biases transferred during fine-tuning.
4) Lessons Learned
   - Instruction Tuning: Instead of fine-tuning on a limited dataset, instruction tuning with larger paraphrasing datasets (e.g., PAWS-X, Quora Paraphrase Dataset) could lead to better outcomes.
   - Dynamic Prompt Engineering: Enhancing prompts dynamically during inference might mitigate the need for fine-tuning altogether in smaller datasets.

V. Conclusion

This study demonstrates that: LLaMA-1B excels in precision-focused paraphrasing tasks. LLaMA-3B offers a balanced trade-off between accuracy and creativity. LLaMA-8B produces diverse paraphrases but struggles with precision metrics.

Future Directions: Explore LoRA Fine-tuning for improving efficiency. Implement dynamic prompts to handle edge

cases. Develop real-time paraphrasing evaluation dashboards for better usability.

ACKNOWLEDGMENTS

We like to express their gratitude to Eng. Ziad Elshaer and Dr. Ensaf for their invaluable guidance and support throughout the development of this project. Their insights and expertise have significantly contributed to the success of this work.

REFERENCES

[1] Vaswani, A., et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.

[2] Touvron, H., et al., "LLaMA: Open and Efficient Foundation Language Models," Meta AI, 2023.

[3] Raffel, C., et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, 2020.

[4] Brown, T., et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, 2020.