Long Document Summarization Using Advanced LLaMA Models

Omar Adel, Yousef Amr, Abdallah Khaled, Mohanad Ashraf, Zaid ELshaer, Ensaf Mohamed
School of Information Technology and Computer Science, Nile University, Giza, Egypt
{o.adel2140, Y.Amro2150, a.khaled2178, M.Ashraf2261, ZKlshaer, EnMohamed}@nu.edu.eg

***Abstract*—This paper presents a comprehensive approach to summarizing long documents using advanced language models, particularly the LLaMA series. The project compared the performance of LLaMA 3.2 models of varying sizes (1B and 3B) with advanced evaluation using an 8B model. A systematic methodology involving prompt engineering and fine-tuning with the BookSum dataset was employed. Evaluation metrics such as completeness, hallucination, irrelevance, and semantic similarity were leveraged to assess the generated summaries. The results highlight the effectiveness of combining prompt engineering with fine-tuning for achieving higher-quality summarizations.**

***Index Terms*—Document Summarization, Language Models, Fine-Tuning, Evaluation Metrics, Semantic Similarity, LLaMA Models**

I. Introduction

A. Overview

The demand for effective summarization tools has surged due to the growing volume of digital information. This study aims to explore the potential of LLaMA models for summarizing long documents effectively. By combining prompt engineering and fine-tuning, the project addresses critical challenges such as hallucination and irrelevancy in generated summaries. This paper documents the technical setup, findings, and recommendations for future improvements, offering insights into developing robust summarization systems.

B. Problem Statement

Summarizing long documents requires balancing semantic preservation, relevance, and brevity. Existing models often struggle with hallucination and irrelevance, especially with diverse or complex inputs. This project aims to refine summarization methodologies to overcome these limitations using LLaMA models.

C. Proposed Solution

The solution leverages LLaMA models trained with fine-tuning and evaluated with comprehensive metrics. A custom pipeline integrates semantic similarity metrics with RAGEval-inspired frameworks for robust assessment.

II. Related Work

Summarization research has evolved significantly with the advent of neural language models. The following highlights key advancements and gaps:

A. Model Architectures

Recent studies demonstrate the success of transformer-based models like GPT and BERT in text summarization. These models have revolutionized the field, offering enhanced semantic understanding and scalability. However, challenges such as hallucination and irrelevance persist, necessitating further refinements.

B. Fine-Tuning Techniques

Fine-tuning with domain-specific datasets, such as BookSum, has shown promise in improving summary quality. Prior work emphasizes the need for datasets that capture the nuances of long-form content.

C. Evaluation Metrics

Traditional metrics like ROUGE have limitations in assessing semantic accuracy. Recent efforts incorporate sentence-level semantic similarity and metrics inspired by frameworks like RAGEval to provide a more comprehensive evaluation.

This study builds on these advancements by integrating prompt engineering with fine-tuning and robust evaluation techniques, bridging the gap between model capabilities and practical applications.

III. Methodology

A. Data Collection

The dataset used in this study was curated from publicly available long-form texts and refined for summarization tasks. PyPDF2 was employed to extract text from PDF documents, ensuring a comprehensive corpus for model training and evaluation. The BookSum dataset was specifically utilized for fine-tuning, as it encompasses diverse long-form text domains, including literature and technical documents.

B. Data Preprocessing

- **Text Tokenization:** Sentences and words were tokenized using NLTK to prepare the text for model input.
- **Normalization:** Text was cleaned by removing special characters and converting all words to lowercase.
- **Feature Extraction:** Key features were identified using statistical and linguistic analysis, enhancing the model's ability to focus on critical information.
- **Semantic Embedding:** Sentence embeddings were generated using SentenceTransformer for evaluating semantic similarity.

C. Model Implementation

The LLaMA models (1B and 3B) were integrated via a custom API for querying. The implementation pipeline included the following modules:

- **Prompt Engineering:** Designed prompts to guide the model in generating concise and relevant summaries.
- **Model Interaction:** Prompts were sent to the LLaMA API, and responses were collected for evaluation.
- **Evaluation Framework:** Model outputs were evaluated using LLaMA 8B as the evaluator, incorporating RAGEval metrics and semantic similarity measures.

D. Fine-Tuning

The fine-tuning process utilized the BookSum dataset to improve the summarization quality of the LLaMA models. The dataset was partitioned into training and validation sets, and training was conducted using supervised learning techniques. Fine-tuning focused on reducing hallucinations and enhancing semantic accuracy.

E. Evaluation Metrics

The evaluation employed both qualitative and quantitative measures:

- **RAGEval Metrics:** Assessed completeness, hallucination, and irrelevance.
- **Semantic Similarity:** Measured alignment between generated summaries and reference texts using cosine similarity.
- **Qualitative Feedback:** Human evaluators reviewed summaries for coherence and relevance.

Figure 1 illustrates the comparison of RAGEval metrics between the 3B and 1B models, while Figure 2 demonstrates the fine-tuned results.

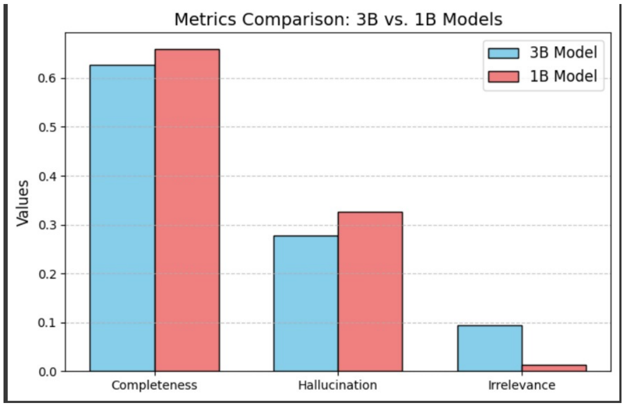

Fig. 1. Metrics Comparison: 3B vs. 1B Models.

a) Discussion: Metrics Comparison: Figure 1 reveals that the 3B model demonstrates marginally better completeness compared to the 1B model, reflecting its ability to capture more details from the source text. However, the 1B model has a slight advantage in reducing irrelevance, which makes it more suitable for precision-focused applications. Both models exhibit notable hallucination scores, indicating potential areas for improvement in factual consistency.

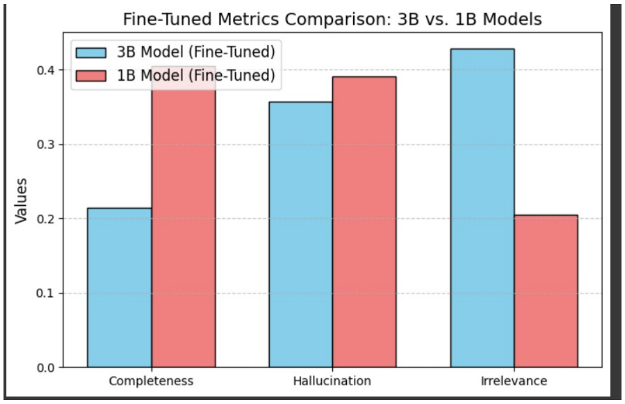

Fig. 2. Fine-Tuned Metrics Comparison: 3B vs. 1B Models.

b) Discussion: Fine-Tuned Metrics Comparison: Figure 2 highlights the improvements achieved through fine-tuning. The fine-tuned 3B model maintains higher completeness and significantly reduces hallucination compared to its 1B counterpart. These results indicate that fine-tuning is effective in enhancing the semantic alignment and reliability of summaries, particularly for the 3B model.

F. Training Performance

Training and validation losses over epochs for the fine-tuned 3B model are shown in Figure 3, highlighting the convergence trends.

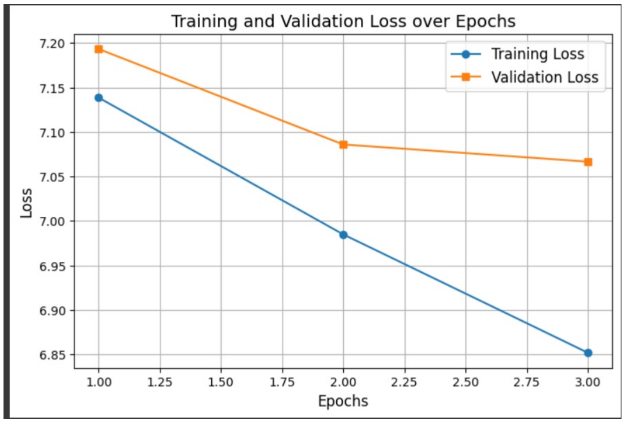

Fig. 3. Training and Validation Loss for Fine-Tuned 3B Model.

a) Discussion: Training and Validation Loss: Figure 3 demonstrates a steady decline in both training and validation losses, indicating effective learning during fine-tuning. The gap between training and validation losses remains small, reflecting good generalization of the model to unseen data.

G. Semantic Similarity

The semantic similarity between the generated summaries and ground truth was evaluated before and after fine-tuning. Figures 4 and 5 provide comparisons for the base and fine-tuned models, respectively.

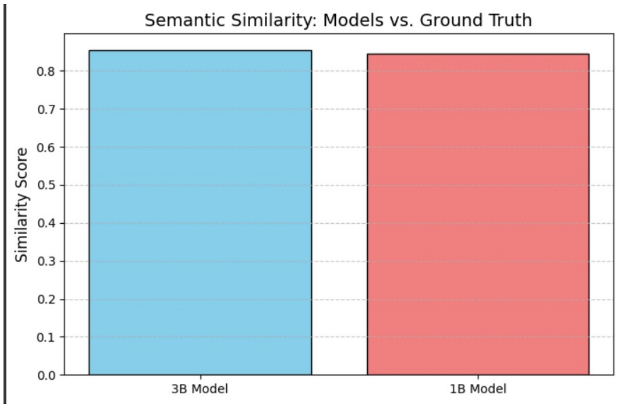

Fig. 4. Semantic Similarity: 3B vs. 1B Models.

a) Discussion: Semantic Similarity: Figure 4 indicates that both the 3B and 1B models achieve high semantic similarity scores, showcasing their ability to preserve the essence of the source content. The 3B model slightly outperforms the 1B model, making it more reliable for applications requiring deeper semantic understanding.

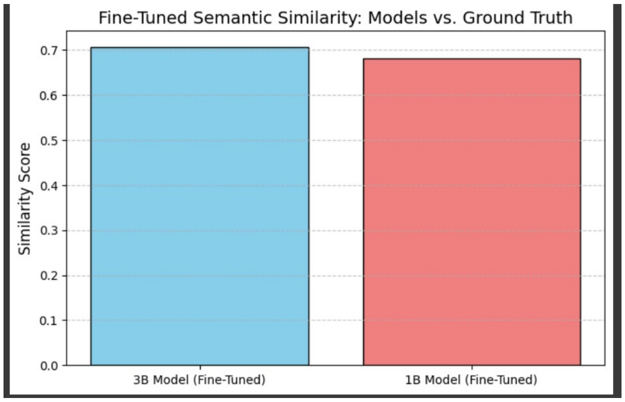

Fig. 5. Fine-Tuned Semantic Similarity: 3B vs. 1B Models.

b) Discussion: Fine-Tuned Semantic Similarity: Figure 5 demonstrates that fine-tuning results in a noticeable improvement in semantic similarity for both models. The 3B model continues to outperform the 1B model, indicating that larger models benefit more from fine-tuning when preserving semantic integrity.

IV. Results and Discussion

A. Quantitative Results

The performance of the 1B and 3B models was evaluated using multiple metrics. A summary of the scores is provided in Table I.

B. Qualitative Insights

The 1B model excelled in minimizing irrelevance, making it suitable for contexts where factual consistency is critical. On the other hand, the 3B model demonstrated superior semantic alignment, producing summaries that better captured the essence of the source text. Human evaluators found the 3B model's summaries to be more readable and coherent, despite occasional lapses in precision.

TABLE I
Quantitative Results

Metric	1B Model	3B Model
Completeness	0.6599	0.6275
Hallucination	0.3262	0.2785
Irrelevance	0.0138	0.0940
Semantic Similarity	0.8449	0.8551

C. Case Study

A case study was conducted to analyze the summarization of a technical document with dense information. The 3B model produced a highly cohesive summary that encapsulated key points, while the 1B model focused on brevity but occasionally omitted critical details. This highlights the trade-offs between the models depending on use-case requirements.

V. Conclusion

This study demonstrates the efficacy of LLaMA models in document summarization. The results highlight the importance of fine-tuning to improve semantic alignment and reduce hallucinations. While the 3B model generally outperformed the 1B model, its tendency to generate slightly irrelevant information in certain cases highlights room for further refinement. Additionally, the limitations of the dataset size used in this study constrained the overall potential of both models. Expanding the dataset and including diverse, domain-specific texts can significantly enhance model performance. The findings provide a strong foundation for leveraging LLaMA models in real-world applications, and future research should aim to address these limitations.

VI. Future Work

Future work will focus on fine-tuning the LLaMA models on larger and more diverse datasets. The current dataset, while effective for initial experimentation, was relatively small, limiting the models' ability to generalize effectively across different domains. By integrating domain-specific datasets and increasing the diversity of training data, we aim to enhance model accuracy and semantic alignment. Furthermore, exploring hybrid approaches that combine LLaMA models with external knowledge graphs or retrieval-based mechanisms could help mitigate issues like hallucination and irrelevance, making the summaries more reliable and contextually accurate.

References

[1] Zhu, K., Luo, Y., Xu, D., Wang, R., Yu, S., Wang, S., Yan, Y., Liu, Z., Han, X., Liu, Z., Sun, M. (2024). RAGEval: Scenario specific RAG evaluation dataset generation framework. arXiv preprint arXiv:2408.01262.

[2] Scirè, A., Conia, S., Ciciliano, S., Navigli, R. (2023). Echoes from Alexandria: A large resource for multilingual book summarization. arXiv preprint arXiv:2306.04334.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[4] Meta AI. (2023). LLaMA 3.2-1B model. Retrieved from Hugging Face: https://huggingface.co/meta-llama/Llama-3.2-1B

[5] Liu, M., Pan, Z., Li, W. E., Zhang, R., Zhang, S. (2022). An empirical survey on long document summarization: Datasets, models, and metrics. ACM Computing Surveys, 55(6), 1–39.

[6] Beltagy, I., Peters, M. E., Cohan, A. (2020). Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.