2024 2nd International Conference on Information and Communication Technology (ICICT)

October 21-22, 2024

Institute of Information and Communication Technology (IICT)
Bangladesh University of Engineering and Technology (BUET)

https://icict.buet.ac.bd

Evaluating the Text Summarization Efficiency of Large Language Models

Mondol Mridul Provakar
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh

Paper ID: 014 Session: 1A Date: 21 Oct 2024

Organised By: IICT, BUET
Technical Co-sponsor: IEEE Bangladesh Section IEEE BDS
Financial Co-sponsors: BdREN BdREN bKash দুয়ার

Outline of Presentation

- Introduction
- Motivation and Objective
- Literature Review
- Dataset Description
- Methodology
- Results
- Conclusion
- Future Works
- References

Introduction

Large Language Models:

Definition: A type of artificial intelligence (AI) that are trained on massive amounts of text data.

Characteristics:

- **Large size:** Trained on billions or even trillions of words of text data.
- **Generative capabilities:** Can generate new text, including translations, summaries, question-answering etc.
- **Domain adaptation:** Can be fine-tuned to perform well on specific tasks or domains. Example: medical chatbot, legal document summarizer etc.

Introduction (Cont'd)

Challenges for Large Language Models:

- **Cost:** Training and inferencing can be computationally expensive.
- **Generalization:** Can have difficulty generalizing to new situations or domains.
- **Fairness:** Can be used to create unfair or discriminatory systems.
- **Privacy:** Can be used to infer sensitive information about people.

Introduction (Cont'd)

Table 1: Best Large Language Models in 2024[1]

LLM	Developer	Popular Apps	# of Parameters (billions)	Access
GPT-3	OpenAI	Microsoft, Duolingo, Stripe, Zapier, Dropbox, ChatGPT	175	API
Gemini	Google	Bard & Nano (some queries)	1.8 - 3.25 (varies)	API
PaLM 2	Google	Bard, Docs, Gmail, and other Google apps	340	API
Llama 2	Meta	Undisclosed	7, 13, and 70	Open Source
Claude 2	Anthropic	Slack, Notion, Zoom	Unknown	API
Falcon	Technology Innovation Institute	Undisclosed	1.3, 7.5, 40, and 180	Open Source
MPT	Mosaic	Undisclosed	7 and 30	Open Source
Mixtral	Mistral AI	Undisclosed	46.7	Open Source

Introduction (Cont'd)

Table 2: Performance of llama2-7b Gemma-7b on Various Benchmarks[2]

Capability	Benchmark	Description	Gemma(7B)	Llama-2(7B)
General	MMLU	Representation of questions	64.3	45.3
Reasoning	BBH	Multi-step reasoning tasks	55.1	32.6
Reasoning	HellaSwag	Commonsense reasoning	81.2	77.2
Math	GSM8K	Basic arithmetic	46.4	14.6
Math	MATH	Challenging math problems	24.3	2.5
Code	HumanEval	Python code generation	32.3	12.8

Introduction (Cont'd)

Table 3: Model Configuration for Llama2 and Gemma[3]

Configuration	Llama2	Gemma
Vocabulary Size	32,000	256,000
Context Length tokens	4096	8192
Hidden Size	4,096	3,072
Number of Hidden Layers	32	28
Number of Attention Heads	32	16

Motivations & Objectives

Motivation:

- AI innovations are transforming various industries.
- Large Language Models (LLMs) offer promising opportunities in article summarization.
- Current LLMs show potential but face challenges in summarizing news article.

Objective:

- Assess proficiency of LLMs, specifically Llama2-7b and Gemma-7b, in news article summarization.
- Evaluate model performance using metrics like BLEU Score, Rouge-1, Rouge-2, Rouge-L and BERTScore.
- Identify strengths and limitations of LLMs in article summarization.

Literature Review

Paper Title	Authors	Contribution
A deep reinforced model for abstractive summarization[4]	Paulus et al. (2017)	The authors developed a neural network model for abstractive summarization that integrates intra-attention mechanisms and combines supervised word prediction with Reinforcement Learning that addressed the challenges of handling lengthy documents and reducing repetitive outputs.
Neural document summarization by jointly learning to score and select sentences[5]	Zhou et al. (2018)	A novel neural network framework was proposed for extractive summarization that integrates sentence scoring and selection into a unified process, leveraging a hierarchical encoder to sequentially extract sentences leading to significant performance improvements on the CNN/Daily Mail dataset.

Literature Review (Cont'd)

Paper Title	Authors	Contribution
Abstractive text summarization with multi-head attention[6]	Li et al. (2019)	A sequence-to-sequence model was introduced with multi-head attention for abstractive summarization to address issues of repetition and information loss by leveraging self-attention layers to capture relevant details and maintain the original structure of input articles.
Deep learning-based extractive text summarization with word-level attention mechanism[7]	Gambhir et al. (2022)	The authors developed the WL-AttenSumm model, a neural network that integrates attention mechanisms and a Convolutional Bi-GRU architecture to improve extractive summarization, achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 32.8%, 11.0%, and 27.5%, respectively, on the CNN/Daily Mail dataset

Literature Review (Cont'd)

Paper Title	Authors	Contribution
Automating customer needs analysis: A comparative study of large language models in the travel industry[8]	Barandoni et al. (2024)	A comparative analysis of large language models (LLMs) was conducted for extracting customer needs from TripAdvisor posts, demonstrating that open-source models like Mistral 7B offer performance comparable to larger proprietary models such as GPT-4, while emphasizing the importance of model size and resource efficiency.
Revisiting zero-shot abstractive summarization in the era of large language models from the perspective of position bias[9]	Chhabra et al. (2024)	The authors investigated the position bias in large language models (LLMs) for zero-shot abstractive summarization revealing that models like GPT 3.5-Turbo, Llama-2, and Dolly-v2 generally exhibit low position bias meaning it treats information from all parts of the text more equally.

Dataset Description

- The CNN/Daily Mail dataset contains over 300,000 news articles from CNN (2007-2015) and the Daily Mail (2010-2015).
- Originally designed for machine reading comprehension and abstractive question answering, now adapted for extractive and abstractive summarization tasks.
- Articles include full text and highlights; average document: 766 words, 29.74 sentences; average summary: 53 words, 3.72 sentences.[10]
- Subset of 500 pairs used for training, 100 pairs each for evaluation and testing.

Methodology

Two methods were utilized on the CNN daily mail dataset:

- **Zero shot prompt engineering**: The approach involves crafting prompts for large language models to generate accurate responses without requiring task-specific training to leverage the models' ability to generalize across different tasks.

- **Supervised fine-tuning**: This process involves taking a pre-trained model and further training it on a specific task or dataset using labeled examples for enabling the model to adapt and enhance its performance for that task through guided learning. Parameter-efficient fine-tuning, using low-rank adapters was applied to fine-tune the model in a supervised manner.

Methodology (Cont'd)

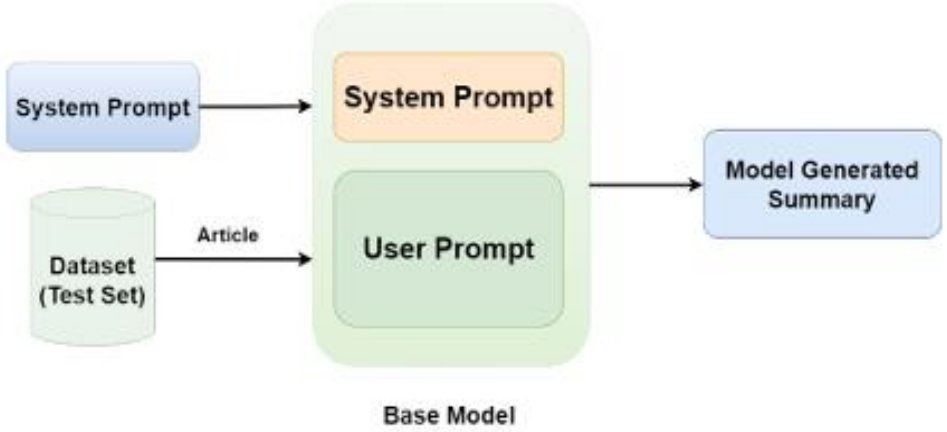

Fig 1: Zero Shot Prompt Engineering

Methodology (Cont'd)

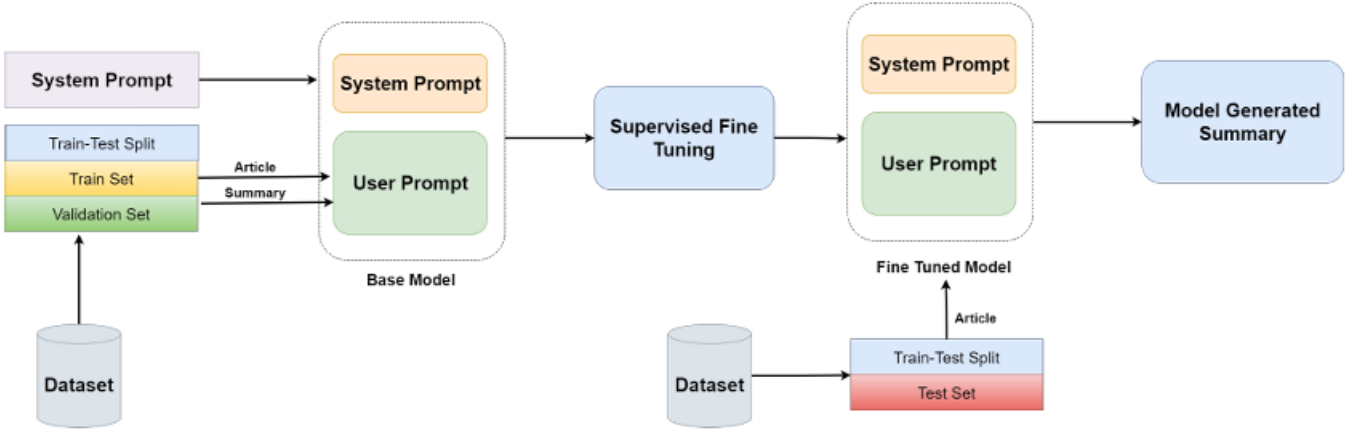

Fig 2: Supervised Fine-Tuning

Methodology (Cont'd)

Table 4: Hyperparameters used for Llama2 and Gemma model for supervised fine-tuning

Hyperameters	Quantity
lora_config (r)	16
lora_alpha	64
lora_dropout	0.1
train_batch_size	4
optimizer	paged_adamw_32bit
learning_rate	2e-5
num_train_epochs	5

Results

Evaluation Matrices:

- **BLUE Score**: It calculates how many n-grams (word sequences) in the candidate summary match those in high-quality human-written references, with more matching n-grams indicating greater similarity
- **ROUGE-1**: Measures the overlap of unigrams (single words) between the generated summary and reference text.
- **ROUGE-2**: Assesses the overlap of bigrams (pairs of consecutive words), providing a more precise comparison.
- **ROUGE-L**: Uses the Longest Common Subsequence (LCS) to evaluate the coherence and structure between the generated summary and the reference.
- **BERT Score**: It is designed to evaluate the coherence and quality of summaries by leveraging BERT (Bidirectional Encoder Representations from Transformers) embeddings to measure the similarity between the generated summary and the reference text incorporating precision, recall, and F1 score for a comprehensive assessment.

Results (Cont'd)

Table 5: Evalutation Metrices for LLMs in (%)

LLM	Method	BLEU	Rouge -1	Rouge-2	Rouge-L	BERTScore (Precision)	BERTScore (Recall)	BERTScore (F1)
Llama2	Zero-Shot	21.36	32.39	12.34	22.59	85.98	87.91	86.92
Llama2	Fine-Tune	**22.43**	**32.44**	**12.60**	**22.65**	**86.06**	**87.95**	**86.99**
Gemma	Zero-Shot	20.01	28.52	10.59	20.01	84.80	86.27	85.78
Gemma	Fine-Tune	21.02	28.89	10.78	20.35	85.33	87.31	86.03

Conclusion

- Llama2 and Gemma demonstrate strong performance in text summarization especially when fine-tuned with higher scores in BLEU, ROUGE, and BERT Score.
- Fine-tuning significantly boosts Llama2's summarization abilities improving coherence and relevance.
- Gemma performs well but requires further refinement, particularly in improving ROUGE score.
- Both models show potential for improvement with larger and more diverse training datasets.

Future Works

- A diverse range of LLMs will be evaluated to identify strengths and weaknesses across different domains.
- Innovative fine-tuning strategies, such as progressive layer freezing and differential learning rates, will be explored to improve model performance.
- Augment training datasets with diverse data to enhance models' ability to generate contextually relevant summaries.

References

[1]. Data Science Dojo, “Best Large Language Models (LLMs) in 2024.”https://datasciencedojo.com/blog/best-large-language-models/. Accessed: 12/04/2024.
[2]. K. LLC, “Gemma vs. llama vs. mistral: A comparative analysis with a coding twist,” Medium, 2024. Accessed: 12/04/2024.
[3]. T. A. Dream, “Google gemma open source llm: Everything you need to know,” The AI Dream, 2024. Accessed: 12/04/2024
[4]. R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” arXiv preprint arXiv:1705.04304, 2017.
[5]. Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, “Neural document summarization by jointly learning to score and select sen tences,” arXiv preprint arXiv:1807.02305, 2018.

References (Cont'd)

[6] J. Li, C. Zhang, X. Chen, Y. Cao, P. Liao, and P. Zhang, "Abstractive text summarization with multi-head attention," in 2019 international joint conference on neural networks (ijcnn). IEEE, 2019, pp. 1–8.
[7]. M. Gambhir and V. Gupta, "Deep learning-based extractive text sum marization with word-level attention mechanism," Multimedia Tools and Applications, vol. 81, no. 15, pp. 20829–20852, 2022.
[8] S. Barandoni, F. Chiarello, L. Cascone, E. Marrale, and S. Puccio, "Automating customer needs analysis: A comparative study of large lan guage models in the travel industry," arXiv preprint arXiv:2404.17975, 2024.
[9] A. Chhabra, H. Askari, and P. Mohapatra, "Revisiting zero-shot ab stractive summarization in the era of large language models from the perspective of position bias," arXiv preprint arXiv:2401.01989, 2024.
[10] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang et al., "Abstractive text summarization using sequence-to-sequence rnns and beyond," arXiv preprint arXiv:1602.06023, 2016.

Thank You