Blog Generator Application using LLM

Chandu Priya Matavalam
dept. Electronics & Communication
NIT Calicut
Kozhikode, India
chandupriya0506@gmail.com

Sainatha Rudra
dept. Electronics & Communication
NIT Calicut
Kozhikode, India
rudrasainatha@gmail.com

Nikhil Pastham
dept. Electronics & Communication
NIT Calicut
Kozhikode, India
nikhil@gmail.com

***Abstract*—The rapid growth of digital content creation necessitated efficient and automated blog generation solutions across multiple domains. This project introduces a Blog Generation Application developed using the Llama-2-7B-chat-GGUF model from Hugging Face, integrated with the Streamlit framework. The application enables users to input a topic, select a specific job profile, and specify the desired word count, generating a tailored blog in realtime. The Llama-2-7B-chat-GGUF model, renowned for its capability to produce high-quality and contextually relevant text, serves as the core engine for content generation. A comprehensive methodology is employed, incorporating a structured prompt template and interactive user interface through Streamlit. Validation of the generated blogs was conducted through qualitative and quantitative studies, demonstrating their coherence, relevance, and adherence to specified word counts. The results affirm the effectiveness and efficiency of the application in meeting user requirements and highlight the potential of automated text generation technologies in revolutionizing content creation processes.**

***Index Terms*—LLM, Llama-2, Pretrained Model, Large language model Meta.**

I. Introduction

Digital transformation has increased demand for automated content generation, especially blog writing. This project addresses this growing need by creating a Blog Generation Application using Hugging Face's advanced Llama-2-7B-chat-GGUF model. The app helps professionals, researchers, and content creators create customized blogs quickly. After entering a topic, job profile, and word count, the Llama-2-7B-chat-GGUF model generates a coherent and contextually relevant blog in realtime. The Streamlit framework gives the app an intuitive and interactive user interface, improving the user experience. This project streamlines content creation across multiple blogs to save time and maintain quality.

A. Motivation

Significant advancements have been observed in the field of content generation due to the emergence of natural language processing (NLP) techniques. These technological advancements have provided significant capabilities to applications such as blog generation, facilitating more efficient and streamlined processes for creating content. The motivation for undertaking this project arises from the increasing demand for automated solutions in the realm of content creation, which is required to meet the requirements of various domains and audiences.

B. Problem Definition

Traditional methods of blog writing are time-consuming and require substantial human effort. Additionally, maintaining consistency and quality across multiple blogs can be challenging. Hence, there is a need for an automated solution that can produce coherent and relevant blogs tailored to specific job profiles and topics

C. Objective

- Provide users with the ability to input a topic and select a job profile.
- Generate a blog of specified word count tailored to the selected job profile.
- Maintain consistency and quality in the generated content

D. Background

- **Model Name**: Llama-2-7B-chat-GGUF
  The Llama-2-7B-chat-GGUF model is a large language model trained on a vast dataset, capable of generating coherent and contextually relevant text. It utilizes advanced natural language processing techniques to understand and respond to user prompts effectively.
- **Model Details**: Meta developed and publicly released the Llama 2 family of large language models (LLMs), a collection of pre-trained and fine-tuned generative text models ranging in scale from 7 billion to 70 billion parameters. The fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in the human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.
- **Variations**: Llama 2 comes in a range of parameter sizes — 7B, 13B, and 70B , as well as pre-trained and fine-tuned variations.
- **Input**: Models can take input text only.
- **Output**:Models can generate text only.
- **Model Architecture**: Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety.
- Llama -2 (LLMs) is based on the transformer neural network architecture. The transformer architecture was

first introduced in the paper “Attention is All You Need“ [3] by Google Brain in 2017. LLMs/GPT models use a variant of this architecture called de' decoder-only transformer' shown in Fig 1.The only purpose of these models is to receive a prompt (an input) and predict the next token/word that comes after this input.

- Overview of the (decoder-only) Transformer model : The **Input** is a prompt (often referred to as context) fed into the transformer as a whole.The **output** depends on the goal of the model. For LLM models, the output is a probability distribution of the next token/word that comes after the prompt. It outputs one prediction for the complete input.Next, it is essential to understand the key components that make up the decoder-only transformer architecture:
  1.Embedding: the input of the transformer model is a prompt. This prompt needs to be embedded into something that the model can use.
  2.Block(s): This is the main source of complexity. Each block contains a masked multi-head attention submodule, a feedforward network, and several layer normalization operations. Blocks are put in sequence to make the model deeper.
  3.Output: the output of the last block is fed through one more linear layer to obtain the final output of the model (a classification, a next word/token etc.)

  Self-Attention Mechanism: Self-attention makes the transformer powerful. The intuition of self-attention is that the mechanism allows the model to focus on (attend to) the most relevant parts of the input. A single self-attention mechanism is called a head. The head works as shown in Fig 2. First, the input is fed into three separate linear layers. Two of those (the queries Q and the keys K) are multiplied, scaled, and turned into a probability distribution using a softmax activation function.Finally, the output is multiplied with values V. This thus gives V * the importance of each of the tokens in V. A key observation is that the learnable parameters in the head are the three linear layers. Fig 2 gives an overview of the operations done in a head and an overview of how multi-head attention works.
- Multi-Head Attention: Multi-head attention is nothing more than several individual heads stacked on top of one another. The input to all heads is equivalent. However, each head has its own weights. After forwarding the input through all the heads, the output of the heads is concatenated and passed through a linear layer which brings the dimensionality back to the dimension of the initial input.
- Masked Self-Attention: In the decoder-only transformer, masked self-attention is nothing more than sequence padding. The 'masking' term is a left-over of the original encoder-decoder transformer module in which the encoder could see the whole (original language) sentence, and the decoder could only see the first part of the sentence which was already translated. As such, they called it 'masking'.
- **Block**:Each block contains a multi-head attention submodule, a feedforward network, 2 layer-normalization operations, and 2 skip connections.The **feedforward network** is simply a multi-layer perceptron. In the original paper, the proposed feedforward module consisted of (1) a fully connected layer; (2) a ReLU activation; (3) another fully connected layer; and (4) a dropout layer.
- The **'add & norm'** blocks get the output from the multi-head attention/feedforward submodule and add it to the input into those modules. After that, a layer normalization operation is performed. Adding the input and output of a submodule together is known as a skip-connection. As blocks can be put in sequence, the skip connections help tremendously reduce the problem of vanishing or exploding gradients. In other words, skip connections are necessary to ensure proper backpropagation of the gradients.
- **Positional Embedding**:Transformers take in a complete prompt at once (in contrast to RNNs) and embed this as one big Tensor. As such, transformers do not know which word is at what position in the sentence. This is

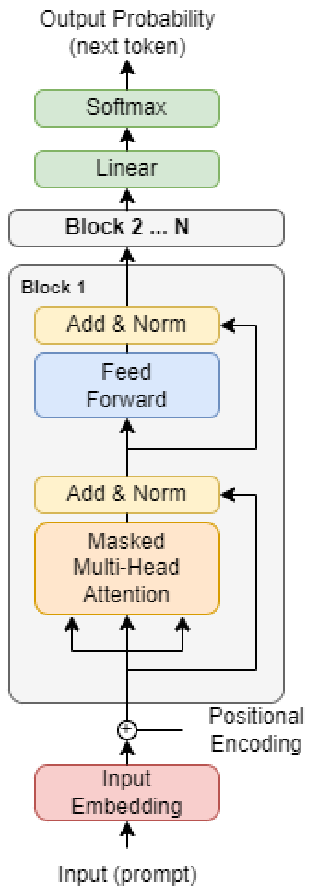


Fig. 1: Transformer Architecture [3].

problematic as the following two sentences mean entirely different things, only dependent on the order of the words:
(1).The boy chased the bird with a butterfly net.
(2).The bird chased the boy with a butterfly net.
To that end, a positional embedding is added to allow the model to deduce which word is where.

- **Output**:After the prompt is forwarded through all the blocks sequentially, the output is forwarded through one final linear layer. This final linear layer maps the output of the model back to the size of the 'vocabulary'.The output of the model is a probability distribution
- **Streamlit Framework**:Streamlit is an open-source Python library that enables rapid development of web applications for data science and machine learning projects. It provides a user-friendly interface for interacting with machine learning models and displaying results in real-time.
- **Previous Work**:For several years, researchers have focused on automated text generation. Several models and techniques, including GPT-3, BERT, and LSTM, have been used to create applications for content creation, summarization, and translation. However, the Llama-2-7B-chat-GGUF model provides distinct benefits in terms of response quality and coherence.

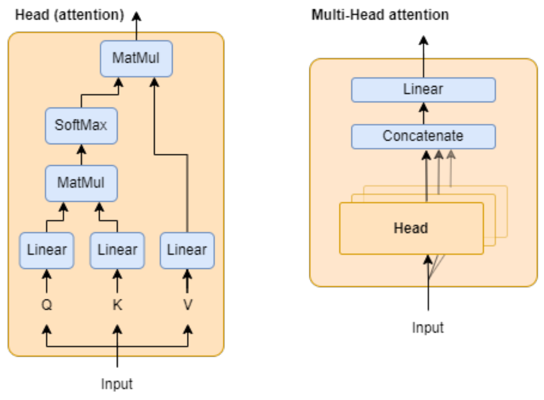


Fig. 2: Multihead Attention [3].

II. Method

A. Procedure

User Input: The user provides a topic, selects a job profile, and specifies the desired word count through the Streamlit interface.
Prompt Generation: A prompt template is constructed using the user's input to guide the Llama-2-7B-chat-GGUF model in generating the blog.
Model Interaction: The prompt is passed to the Llama-2-7B-chat-GGUF model, which generates a blog response based on the provided instructions.
Output: The generated blog is displayed to the user through the Streamlit interface.

B. Justifications

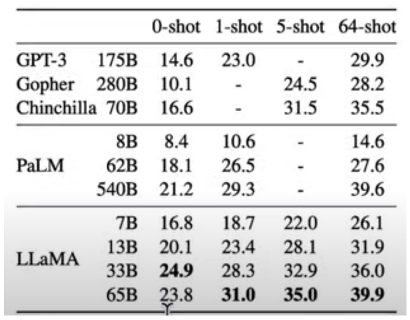
- Prompt Template: The prompt template includes placeholders for the topic, job profile, and word count, ensuring that the generated content is relevant and tailored to the user's requirements.
- Llama-2-7B-chat-GGUF Model: This model (Open source) was chosen due to its ability to produce high-quality and contextually relevant text and faster inference time than other closed models, making it suitable for blog generation.
- It has limited GPU requirements for fine-tuning, smaller transformer architecture and trained on public data.
- Streamlit Interface: Streamlit provides an intuitive and interactive platform for users to input parameters, interact with the model, and view results seamlessly.

C. Validation

Qualitative Study: The generated blogs were evaluated based on their coherence, relevance, and quality. Feedback from users indicated that the blogs produced by the application were well-structured and relevant to the selected job profiles and topics.

Significance in Model Evaluation :

- **Zero-shot** evaluations assess the model's generalization capability and its ability to handle new, unseen tasks without additional training.
- **Few-shot** evaluations provide insights into the model's adaptability and performance when provided with limited examples, mimicking real-world scenarios where acquiring extensive training data may not be feasible.
- Comparison between different open models and closed models' performance for NaturalQuestions in exact match performance in percentage shown in Figure **??**.



		0-shot	1-shot	5-shot	64-shot
GPT-3	175B	14.6	23.0	-	29.9
Gopher	280B	10.1	-	24.5	28.2
Chinchilla	70B	16.6	-	31.5	35.5
	8B	8.4	10.6	-	14.6
PaLM	62B	18.1	26.5	-	27.6
	540B	21.2	29.3	-	39.6
	7B	16.8	18.7	22.0	26.1
LLaMA	13B	20.1	23.4	28.1	31.9
	33B	**24.9**	28.3	32.9	36.0
	65B	23.8	**31.0**	**35.0**	**39.9**

Fig. 3: NaturalQuestions. Exact match performance. [2].

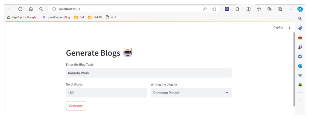
Quantitative Study:

- Word Count: The application successfully generated blogs with the specified word count, demonstrating its capability to produce content of the desired length.
- Response Time: The model's response time was found to be reasonable, allowing for quick generation of blogs.

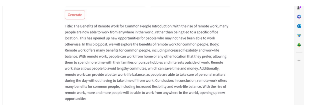
III. Outputs

A. Remote Work

Topic: Remote Work, Blog Style: Common People, Word Count: 150


Generate Blogs
Enter the Blog Topic: Remote Work
No of Words: 150
Writing the blog for: Common People

(a) Input



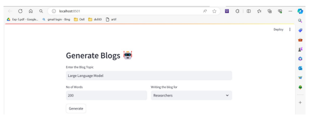
Title: The Benefits of Remote Work for Common People Introduction: With the rise of remote work, many people are now able to work from anywhere in the world, rather than being tied to a specific office location. This has opened up new opportunities for people who may not have been able to work otherwise. In this blog post, we will explore the benefits of remote work for common people. Body: Remote work offers many benefits for common people, including increased flexibility and work-life balance. With remote work, people can work from home or any other location that they prefer, allowing them to spend more time with their families or pursue hobbies and interests outside of work. Remote work also allows people to avoid lengthy commutes, which can save time and money. Additionally, remote work can provide a better work-life balance, as people are able to take care of personal matters during the day without having to take time off from work. Conclusion: In conclusion, remote work offers many benefits for common people, including increased flexibility and work-life balance. With the rise of remote work, more and more people will be able to work from anywhere in the world, opening up new opportunities

(b) Output

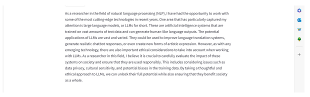
Fig. 4: Generated blog about Remote Work for Common People

B. Large Language Model

Topic: Large Language Model, Blog Style: Researcher, Word Count: 200


Generate Blogs
Enter the Blog Topic: Large Language Model
No of Words: 200
Writing the blog for: Researchers

(a) Input

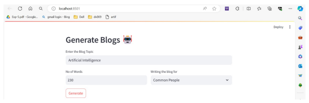

As a researcher in the field of natural language processing (NLP), I have had the opportunity to work with some of the most cutting-edge technologies in recent years. One area that has particularly captured my attention is large language models, or LLMs for short. These are artificial intelligence systems that are trained on vast amounts of text data and can generate human-like language outputs. The potential applications of LLMs are vast and varied. They could be used to improve language translation systems, generate realistic chatbot responses, or even create new forms of artistic expression. However, as with any emerging technology, there are also important ethical considerations to take into account when working with LLMs. As a researcher in this field, I believe it is crucial to carefully evaluate the impact of these systems on society and ensure that they are used responsibly. This includes considering issues such as data privacy, cultural sensitivity, and potential biases in the training data. By taking a thoughtful and ethical approach to LLMs, we can unlock their full potential while also ensuring that they benefit society as a whole.

(b) Output

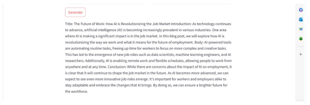
Fig. 5: Generated blog about LLM for a Researcher

C. Artificial Intelligence for Common People

Topic: Artificial Intelligence, Blog Style: Common People, Word Count: 230


Generate Blogs
Enter the Blog Topic: Artificial Intelligence
No of Words: 230
Writing the blog for: Common People

(a) Input



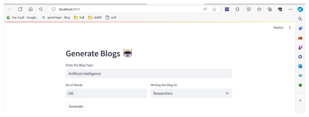
Title: The Future of Work: How AI is Revolutionizing the Job Market Introduction: As technology continues to advance, artificial intelligence (AI) is becoming increasingly prevalent in various industries. One area where AI is making a significant impact is in the job market. In this blog post, we will explore how AI is revolutionizing the way we work and what it means for the future of employment. Body: AI-powered tools are automating routine tasks, freeing up time for workers to focus on more complex and creative tasks. This has led to the emergence of new job roles such as data scientists, machine learning engineers, and AI researchers. Additionally, AI is enabling remote work and flexible schedules, allowing people to work from anywhere and at any time. Conclusion: While there are concerns about the impact of AI on employment, it is clear that it will continue to shape the job market in the future. As AI becomes more advanced, we can expect to see even more innovative job roles emerge. It's important for workers and employers alike to stay adaptable and embrace the changes that AI brings. By doing so, we can ensure a brighter future for the workforce.

(b) Output

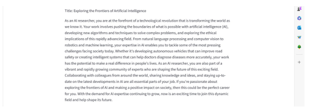
Fig. 6: Generated blog about Artificial Intelligence for a Common People

D. Artificial Intelligence for Researcher

Topic: Artificial Intelligence, Blog Style: Researcher, Word Count: 230


Generate Blogs
Enter the Blog Topic: Artificial Intelligence
No of Words: 230
Writing the blog for: Researchers

(a) Input


Title: Exploring the Frontiers of Artificial Intelligence

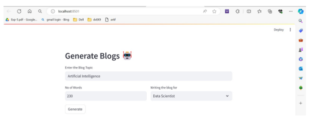
As an AI researcher, you are at the forefront of a technological revolution that is transforming the world as we know it. Your work involves pushing the boundaries of what is possible with artificial intelligence (AI), developing new algorithms and techniques to solve complex problems, and exploring the ethical implications of this rapidly advancing field. From natural language processing and computer vision to robotics and machine learning, your expertise in AI enables you to tackle some of the most pressing challenges facing society today. Whether it's developing autonomous vehicles that can improve road safety or creating intelligent systems that can help doctors diagnose diseases more accurately, your work has the potential to make a real difference in people's lives. As an AI researcher, you are also part of a vibrant and rapidly growing community of experts who are shaping the future of this exciting field. Collaborating with colleagues from around the world, sharing knowledge and ideas, and staying up-to-date on the latest developments in AI are all essential parts of your job. If you're passionate about exploring the frontiers of AI and making a positive impact on society, then this could be the perfect career for you. With the demand for AI expertise continuing to grow, now is an exciting time to join this dynamic field and help shape its future.

(b) Output

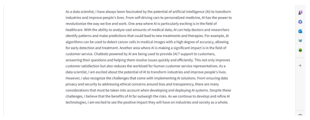
Fig. 7: Generated blog about Artificial Intelligence for a Researcher

E. Artificial Intelligence for Data Scientist

Topic: Artificial Intelligence, Blog Style: Data Scientist, Word Count: 230


Generate Blogs
Enter the Blog Topic: Artificial Intelligence
No of Words: 230
Writing the blog for: Data Scientist

(a) Input


As a data scientist, I have always been fascinated by the potential of artificial intelligence (AI) to transform industries and improve people's lives. From self-driving cars to personalized medicine, AI has the power to revolutionize the way we live and work. One area where AI is particularly exciting is in the field of healthcare. With the ability to analyze vast amounts of medical data, AI can help doctors and researchers identify patterns and make predictions that could lead to new treatments and therapies. For example, AI algorithms can be used to detect cancer cells in medical images with a high degree of accuracy, allowing for early detection and treatment. Another area where AI is making a significant impact is in the field of customer service. Chatbots powered by AI are being used to provide 24/7 support to customers, answering their questions and helping them resolve issues quickly and efficiently. This not only improves customer satisfaction but also reduces the workload for human customer service representatives. As a data scientist, I am excited about the potential of AI to transform industries and improve people's lives. However, I also recognize the challenges that come with implementing AI solutions. From ensuring data privacy and security to addressing ethical concerns around bias and transparency, there are many considerations that must be taken into account when developing and deploying AI systems. Despite these challenges, I believe that the benefits of AI far outweigh the risks. As we continue to develop and refine AI technologies, I am excited to see the positive impact they will have on industries and society as a whole.

(b) Output

Fig. 8: Generated blog about Artificial Intelligence for a Data Scientist

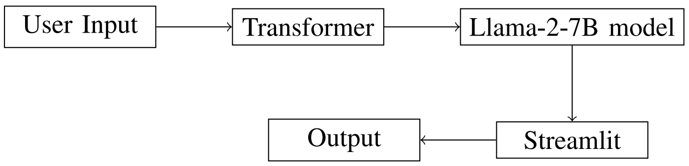


Fig. 9: System Model Flowchart

IV. Conclusion:

The Blog Generation Application developed in this project leverages the capabilities of the Llama-2-7B-chat-GGUF model to automate the process of blog writing. The application allows users to input a topic and select a job profile, generating a coherent and contextually relevant blog tailored to their requirements. The results of qualitative and quantitative studies validate the effectiveness and efficiency of the application in producing quality content which is better than GPT-3 and many other closed models such as PaLM,Gopher etc.

Acknowledgment

We extend our heartfelt gratitude to our course instructor Dr.Sudeep P.V sir. The contributors, researchers, and developers of the Llama-2-7B-chat Model and Streamlit framework, whose work and dedication have significantly enriched the development of our project. Their insights and efforts have been invaluable in advancing the field of natural language processing.

References

[1] Touvron, H., Martin, L., Stone, K. H., Albert, P. J.,et al . "Llama 2: Open foundation and Fine-Tuned chat models". arXiv (Cornell University).Journal article ,2023 . https://doi.org/10.48550/arxiv.2307.09288
[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix,et al. "LLAMA: Open and Efficient Foundation Language Models". 2023, February 27. https://arxiv.org/abs/2302.13971
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones,. et al. "Attention is all you need".2017, June 12. https://arxiv.org/abs/1706.03762