Article Generator using Three Open-Source LLMs

1. Introduction

The rapid growth of artificial intelligence has enabled the development of **AI-driven content generation**. This project aims to create an **Article Generator Chatbot** that utilizes three different **open-source Large Language Models (LLMs)** to generate high-quality articles on various topics. The chatbot is assessed based on **fluency, coherence, factual accuracy, and computational efficiency** to determine the most suitable LLM for article creation.

LLMs Used:

1. **GPT-Neo 125M** (by EleutherAI)
2. **Mistral 7B** (by Mistral AI)
3. **Falcon 7B** (by TII UAE)

2. Objectives

- Develop an **AI-powered chatbot** capable of generating coherent and high-quality articles.
- Evaluate and compare different **open-source LLMs** for their performance.
- Assess models based on **fluency, coherence, factual accuracy, and efficiency**.
- Identify the **most suitable LLM** for article creation.
- Implement a structured evaluation methodology using **metrics like precision, recall, and confusion matrix**.

3. System Architecture

3.1 Components

1. **User Interface (UI):** Provides an interactive interface for inputting article topics.
2. **Data Processing Module:**
    - Tokenization & Preprocessing of input text.
3. **LLM Selection & Text Generation:**
    - The system uses different LLMs to generate article content based on the input prompt.
4. **Evaluation Module:**
    - Assess generated articles using **quality metrics**.

4. Methodology

4.1 Model Selection & Setup

Each of the three models was separately integrated into a chatbot system for generating articles. The implementation process involved:

- **Loading the pre-trained models and tokenizers**.
- **Generating text using prompt-based input**.
- **Setting parameters** like max_length, temperature, top_p, and no_repeat_ngram_size to control output quality.
- **Evaluating model-generated articles** based on fluency, relevance, creativity, and factual correctness.

4.2 Evaluation Criteria

- **Fluency & Grammar**: Ensuring the generated text is grammatically correct and readable.
- **Relevance to Prompt**: Checking if the content remains on-topic.
- **Creativity & Coherence**: Evaluating logical flow and engagement.
- **Handling of Long Articles**: Assessing performance on lengthy content.
- **Factual Accuracy**: Measuring correctness of generated information.
- **Computational Cost**: Evaluating hardware requirements and efficiency.

5. Implementation

5.1 Technologies Used

Component	Technology
Programming Language	Python
Machine Learning Library	Transformers (Hugging Face)
Evaluation Metrics	Precision, Recall, Confusion Matrix
LLMs Used	GPT-Neo 125M, Mistral 7B, Falcon 7B
Deployment Framework	Streamlit (Optional)

5.2 Installation & Setup

torch
transformers
streamlit (if UI is needed)

Install using: pip install -r requirements.txt

6. Results & Discussion

6.1 Performance Evaluation

The models were assessed using the following **key metrics**:

Metric	GPT-Neo 125M	Mistral 7B	Falcon 7B
Fluency & Grammar	Moderate	Excellent	Very Good
Relevance to Prompt	Sometimes off-topic	Highly relevant	Accurate & precise
Creativity	High but repetitive	Best for storytelling	Good but less than Mistral
Handling Long Articles	Struggles beyond 300 tokens	Very strong coherence	Handles long-form well
Factual Accuracy	Prone to hallucinations	Moderate	Most factually reliable
Computational Cost	Low (125M parameters, CPU-friendly)	High (7B parameters, GPU needed)	High (7B parameters, GPU needed)
Overall Suitability	Short articles	Best for general article writing	Best for fact-based content

6.2 Model Strengths and Weaknesses

GPT-Neo 125M

- Lightweight and runs on CPU.
- Suitable for short articles and summaries.
- Struggles with long-form content and coherence.
- Can go off-topic and generate inaccuracies.

Best For: Quick, short-form content such as blog intros and summaries.

Mistral 7B

- Performs well in structured, engaging long-form articles.
- Produces coherent writing with strong logical flow.
- Requires a strong GPU for optimal performance.
- Can sometimes generate creative but inaccurate information.

Best For: Blog articles, storytelling, and general-purpose writing.

Falcon 7B

- Excels in factual and structured content.
- Strong at technical and research-based articles.
- Less creative compared to Mistral.
- Requires high computational power (GPU).

Best For: News articles, research writing, and structured factual content.

7. Conclusion & Future Scope

7.1 Conclusion

The study found that **Mistral 7B** performed the best for **long-form article writing**, **Falcon 7B** was more reliable for **factual accuracy**, and **GPT-Neo 125M** was useful for **quick, short articles**. Based on the evaluation:

🏆 **Winner: Mistral 7B** – **Best Overall for Article Writing**

- Handles **long-form writing best**
- Balanced between **creativity & coherence**
- Ideal for **blog writing, storytelling, and general-purpose articles**

🎯 **If accuracy is a priority**, Falcon 7B is better. 🎯 **If a lightweight option is needed**, GPT-Neo works for **short articles**.

7.2 Future Enhancements

- **Fine-tune Mistral 7B** on domain-specific data.
- **Combine Falcon & Mistral** for a mix of creativity + accuracy.
- **Use BLEU, ROUGE, and Perplexity scores** for deeper performance evaluation.
- **Integrate voice-based interaction** for enhanced usability.
- **Develop a GUI using Streamlit** for real-time user interaction.

8. References

1. EleutherAI GPT-Neo: https://huggingface.co/EleutherAI/gpt-neo-125M
2. Mistral 7B: https://huggingface.co/mistralai/Mistral-7B
3. Falcon 7B: https://huggingface.co/tiiuae/falcon-7b
4. Hugging Face Transformers: https://huggingface.co/transformers/
5. Scikit-learn for Evaluation Metrics: https://scikit-learn.org/