**Task 1 – Article Generator Chatbot Using Open-Source LLMs**

**Report on the three LLMs, assessing their performance, and identify which LLM is most appropriate for article creation.**

**Table Of Contents**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Title | Page No. |
|  | Aim | 1 |
|  | Introduction | 1 |
|  | Methodology | 1 |
|  | Large Language Models (LLMs) | 2 |
|  | Performance Evaluation | 3 |
|  | Model Overview | 4 |
|  | Overall Recommendation | 4 |
|  | Conclusion | 5 |

**Aim:**

Use three different open-source LLMs to build an article generator chatbot. Report on the three LLMs, assess their performance, and identify which LLM is most appropriate for article creation

**Introduction:**

The purpose was to assess which of the available open-source large language models-LLMs-included Ollama3, Gemini, and GPT2-could potentially be used best to develop the article generator chatbot. Integration of these different models into the same chatbot and testing its capabilities in all topics for various articles helped create coherent, relevant, and qualitative content.

**Methodology:**

Approach –

* Frameworks and Tools: The chatbot was built and deployed using Visual Studio Code. The interface of the chatbot was designed to easily switch between the 3 LLMs for direct comparison.
* Integration: Each LLM was integrated into the chatbot using its respective API or pre-trained model implementation.
* Testing: The models were testing using topics with different levels of difficulty.
  + Easy: How Technology Makes Everyday Life Easier.
  + Moderate: The Importance of Cybersecurity in the Modern World.
  + Hard: The Rise of Electric Vehicles: Revolutionizing Transportation in the 21st Century.
* Evaluation Criteria:
  + Resources (CPU, Memory) consumed.
  + Time required to generate response
  + Relevance of the response considering the topic.

**Large Language Models (LLMs):**

Models selected for this specific task are –

* Ollama3
* Gemini
* GPT-2

1. Ollama3: Ollama3 8B is one of the new open-source large language models boasting 8 billion parameters for performance in applications which demand good natural language generation quality. Such architecture is quite perfect, especially at producing coherent, even detailed content - articles, reports, or pieces of creative writing. And so, an added advantage from transformer architecture and increased support for long tokens, this allows Ollama3 for pretty versatile operations. Its lightweight design provides a balance between computational efficiency and output quality, making it well-suited to applications where resource constraints exist. Although it does well out of the box, fine-tuning can further enhance its relevance for domain-specific tasks. Overall, Ollama3 8B is a reliable option for content generation, delivering robust performance without the overhead of larger models.
2. Gemini: Gemini is a new advanced type of large language model developed by Google DeepMind to process and create human-like text. Unlike standard models, the multimodal ability to understand various forms of data-including text, images, audio, and video-allows it to complete almost any conceivable type of job: from understanding natural language to creating content. The model is optimized in three sizes: Ultra, Pro, and Nano, targeting different computational needs and applications. Gemini's architecture is built upon Transformer decoders, which it improves for better efficiency and scalability. Its presence in Google services like Search, Ads, and Google Workspace is a testament to its important place in AI capability development.
3. GPT-2 (small): The OpenAI-developed GPT-2 "small" model is a transformer-based language model with 124 million parameters. It was trained using a huge corpus of text data in English in the self-supervised manner, which enables the model to predict the next word in a sequence. This training makes GPT-2 generate coherent and contextually relevant text on the basis of an input prompt. It is available at Hugging Face and can be used for any kind of natural language processing such as generation, summarization, or translation.

**Performance Evaluation:**

* Performance Evaluation for Easy Topic - How Technology Makes Everyday Life Easier

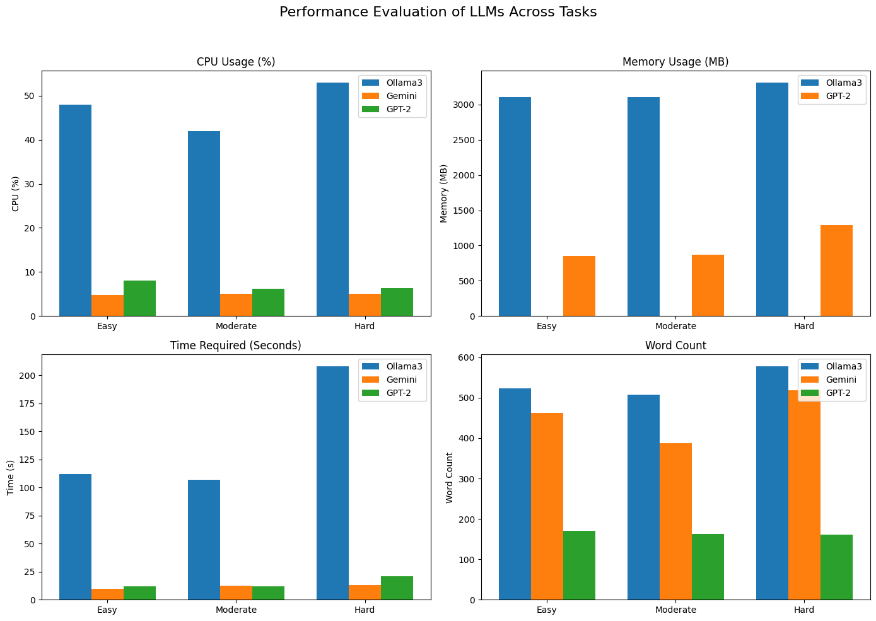
|  |  |  |  |
| --- | --- | --- | --- |
| Metrics \ Models | Ollama3 | Gemini | GPT-2 |
| CPU (%) | 48 | 4.7 | 8 |
| Memory (MB) | 3107 | - | 853 |
| Time required (Seconds) | 112 | 9.45 | 11.98 |
| Word count | 523 | 462 | 170 |
| Relevance | Moderate | High | Moderate |

* Performance Evaluation for Moderate Topic - The Importance of Cybersecurity in the Modern World

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics \ Models | Ollama3 | Gemini | GPT-2 |
| CPU (%) | 42 | 5 | 6.2 |
| Memory (MB) | 3107 | - | 869.2 |
| Time required (Seconds) | 107 | 12.33 | 12.13 |
| Word count | 508 | 388 | 163 |
| Relevance | Moderate | High | Low |

* Performance Evaluation for Hard Topic - The Rise of Electric Vehicle: Revolutionizing Transportation in 21st Century

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics \ Models | Ollama3 | Gemini | GPT-2 |
| CPU (%) | 53 | 5 | 6.3 |
| Memory (MB) | 3312 | - | 1293 |
| Time required (Seconds) | 208 | 13 | 20.86 |
| Word count | 578 | 518 | 161 |
| Relevance | High | Moderate | Low |



**Models Overview:**

* Ollama3
  + Strengths:
    - Highest word count in all tasks, hence detailed answers.
    - High relevance for hard tasks and moderate for others, thus suitable for complex topics.
  + Weaknesses:
    - High CPU usage compared to others, which can strain systems with low processing capacity.
    - Longest time requirement, especially for harder tasks, which means slower processing.
* Gemini
  + Strengths:
    - Most efficient CPU usage, as it has much lower percentages compared to Ollama3.
    - Fastest processing time, which completes tasks very quickly on all levels of difficulty.
    - Consistently provides high relevance for easy and moderate tasks.
  + Weaknesses:
    - Does not have memory usage data, which may be a limitation in assessing resource efficiency.
    - Word count is moderate, which may limit its detail on harder tasks.
* GPT-2
  + Strengths:
    - Moderate CPU usage, which is a balance between performance and resource consumption.
    - Manages memory better than Ollama3, but uses more memory for harder tasks.
  + Weaknesses:
    - Has the lowest word count across all tasks, making it less detailed.
    - Low relevance for medium and hard tasks, so not so great for complex topics.
    - Longer processing time than Gemini for most tasks.

**Overall Recommendation:**

* Ollama3 is capable of producing very in-depth answers but is less feasible because it is highly resource-intensive and requires more time to process.
* Gemini excels in efficiency, being best used for quick, high-relevance outputs for moderately complex topics.
* GPT-2 is suitable for low resource applications where mere minimum detail is sufficient, but it has relevance issues with more complex topics.

**Conclusion:**

If only one LLM had to be picked to generate articles, then **Gemini** would be the choice, based on all these factors. Since Gemini has a word count that is relatively lower compared to Ollama3, the two big advantages it possesses are resource efficiency and processing speed. It consistently requires minimal CPU usage and completes tasks in the shortest time frame, outperforming both Ollama3 and GPT-2 in terms of operational efficiency. This balance between relevance, output quality, and resource consumption establishes Gemini as the best option for generating articles efficiently and effectively.