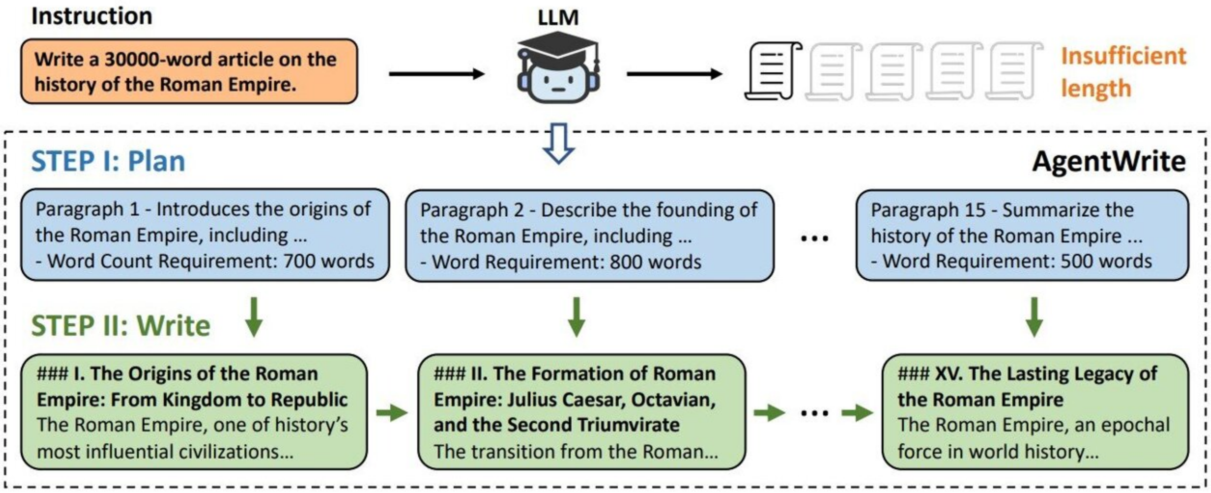
AI researchers introduce an LLM capable of generating text outputs of up to 10,000 words

August 16 2024, by Bob Yirka



Instruction
Write a 30000-word article on the history of the Roman Empire.
LLM
Insufficient length
STEP I: Plan
AgentWrite
Paragraph 1 - Introduces the origins of the Roman Empire, including ...
- Word Count Requirement: 700 words
Paragraph 2 - Describe the founding of the Roman Empire, including ...
- Word Requirement: 800 words
...
Paragraph 15 - Summarize the history of the Roman Empire ...
- Word Requirement: 500 words
STEP II: Write
### I. The Origins of the Roman Empire: From Kingdom to Republic
The Roman Empire, one of history's most influential civilizations...
### II. The Formation of Roman Empire: Julius Caesar, Octavian, and the Second Triumvirate
The transition from the Roman...
...
### XV. The Lasting Legacy of the Roman Empire
The Roman Empire, an epochal force in world history...

As existing LLMs fail to generate long enough output, AgentWrite adopts a plan-thenwrite pipeline to obtain a sufficient length output with off-the-shelf LLMs. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2408.07055

A team of AI researchers at Tsinghua University, working with a colleague from Zhipu AI, has developed a large language model (LLM) called LongWriter that they claim is capable of generating text output of up to 10,000 words. The group has written a paper describing their efforts and new LLM, which is available on the *arXiv* preprint server.

As LLMs have become mainstream, many have noticed that they are not capable of generating very long answers, such as full books or manuscripts—the current limit appears to be approximately 2,000 words. The researchers suggest this is because they are all trained on short documents. In their new effort, they have found that if LLMs are changed slightly and then trained using much longer documents, they are able to produce longer documents.

To test their idea, the research teams first trained a 9-billion parameter LLM using a conventional dataset, which included documents that were mostly less than 2,000 words long. As expected, when queried, it was not able to create texts longer than 2,000 words long.

Next, the team modified a traditional LLM using a pipeline they named AgentWrite to decompose training material into subtasks as it was processed. They then assembled a dataset they named "LongWriter-6k," which is a dataset that holds 6,000 written documents ranging in length from 2,000 to 32,000 words. They then trained the modified LLM using the new dataset LongWriter-6k and found that doing so increased the word length of documents it could produce to approximately 10,000 words.

In reviewing the newly produced long documents generated by the LLM, the team found them to be coherent and useable in a variety of contexts. They have posted the open-source code for their model on GitHub—a move that will allow others to build on what the team in China has done. They also posted a video showing LongWriter producing a 10,000-word tourist guide for people traveling in China.

The researchers acknowledge that there are ethical considerations that must be considered now that it has been found that LLMs can generate entire research papers, books, manuscripts or perhaps even movie scripts.

More information: Yushi Bai et al, LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs, *arXiv* (2024). DOI: 10.48550/arxiv.2408.07055

Github: github.com/THUDM/LongWriter

© 2024 Science X Network

Citation: AI researchers introduce an LLM capable of generating text outputs of up to 10,000 words (2024, August 16) retrieved 16 August 2024 from https://techxplore.com/news/2024-08-ai-llm-capable-generating-text.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.