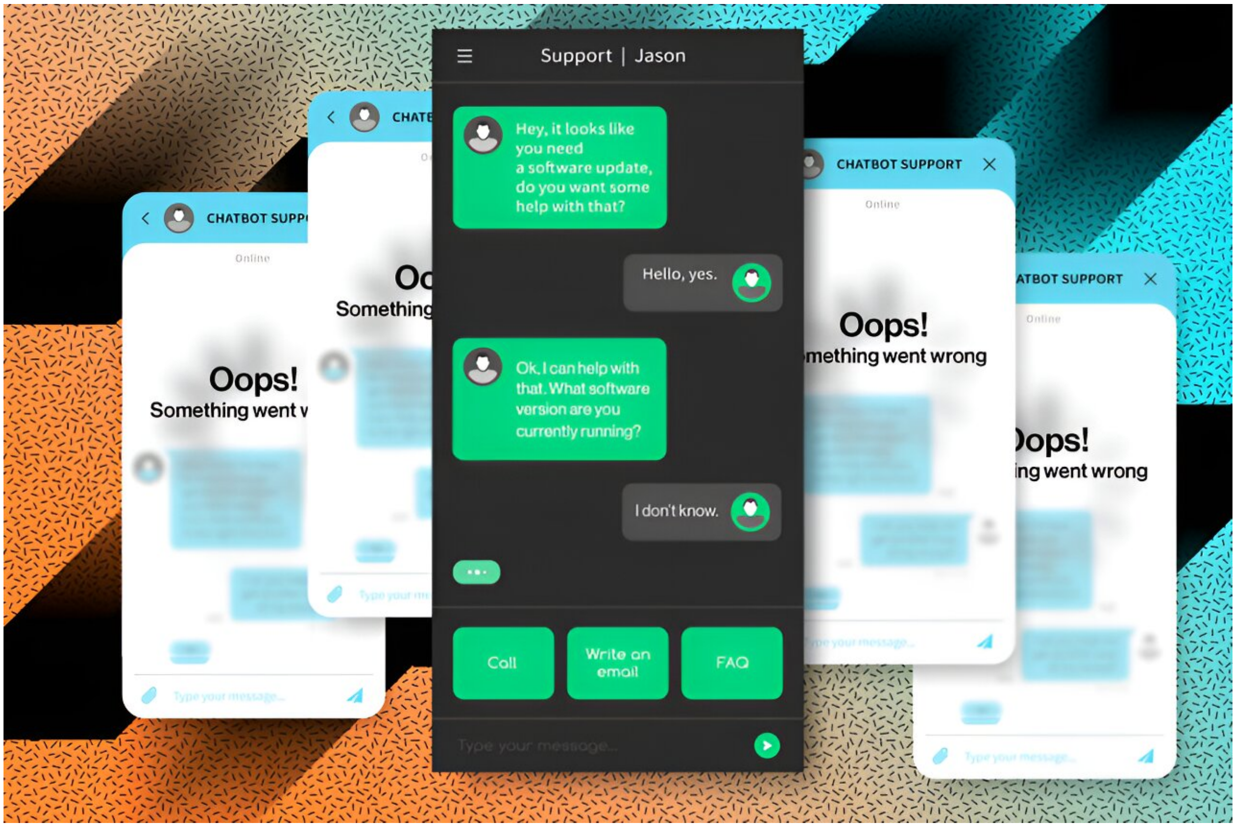
TechXplore

A new way to let AI chatbots converse all day without crashing

February 13 2024, by Adam Zewe



Hey, it looks like you need a software update, do you want some help with that?

Hello, yes.

Ok, I can help with that. What software version are you currently running?

I don't know.

Credit: Christine Daniloff, MIT

When a human-AI conversation involves many rounds of continuous dialogue, the powerful large language machine-learning models that drive chatbots like ChatGPT sometimes start to collapse, causing the bots' performance to rapidly deteriorate.

A team of researchers from MIT and elsewhere has pinpointed a surprising cause of this problem and developed a simple solution that enables a chatbot to maintain a nonstop conversation without crashing or slowing down.

Their method involves a tweak to the key-value cache (which is like a conversation memory) at the core of many large language models. In some methods, when this cache needs to hold more information than it has capacity for, the first pieces of data are bumped out. This can cause the model to fail.

By ensuring that these first few data points remain in memory, the researchers' method allows a chatbot to keep chatting no matter how long the conversation goes.

The method, called StreamingLLM, enables a model to remain efficient even when a conversation stretches on for more than 4 million words. When compared to another method that avoids crashing by constantly recomputing part of the past conversations, StreamingLLM performed more than 22 times faster.

This could allow a chatbot to conduct long conversations throughout the workday without needing to be continually rebooted, enabling efficient AI assistants for tasks like copywriting, editing, or generating code.

"Now, with this method, we can persistently deploy these large language models. By making a chatbot that we can always chat with, and that can always respond to us based on our recent conversations, we could use these chatbots in some new applications," says Guangxuan Xiao, an electrical engineering and computer science (EECS) graduate student and lead author of a paper on StreamingLLM now posted to the *arXiv* preprint server.

Xiao's co-authors include his advisor, Song Han, an associate professor in EECS, a member of the MIT-IBM Watson AI Lab, and a distinguished scientist of NVIDIA; as well as Yuandong Tian, a research scientist at Meta AI; Beidi Chen, an assistant professor at Carnegie Mellon University; and senior author Mike Lewis, a research scientist at Meta AI. The work will be presented at the International Conference on Learning Representations held May 7–11 in Vienna.

A puzzling phenomenon

Large language models encode data, like words in a user query, into representations called tokens. Many models employ what is known as an attention mechanism that uses these tokens to generate new text.

Typically, an AI chatbot writes new text based on text it has just seen, so it stores recent tokens in memory, called a KV Cache, to use later. The attention mechanism builds a grid that includes all tokens in the cache, an "attention map" that maps out how strongly each token, or word, relates to each other token.

Understanding these relationships is one feature that enables large language models to generate human-like text.

But when the cache gets very large, the attention map can become even more massive, which slows down computation.

Also, if encoding content requires more tokens than the cache can hold, the model's performance drops. For instance, one popular model can store 4,096 tokens, yet there are about 10,000 tokens in an academic paper.

To get around these problems, researchers employ a "sliding cache" that bumps out the oldest tokens to add new tokens. However, the model's performance often plummets as soon as that first token is evicted, rapidly reducing the quality of newly generated words.

In this new paper, researchers realized that if they keep the first token in the sliding cache, the model will maintain its performance even when the cache size is exceeded.

But this didn't make any sense. The first word in a novel likely has nothing to do with the last word, so why would the first word be so important for the model to generate the newest word?

In their new paper, the researchers also uncovered the cause of this phenomenon.

Attention sinks

Some models use a Softmax operation in their attention mechanism, which assigns a score to each token that represents how much it relates to each other token. The Softmax operation requires all attention scores to sum up to 1. Since most tokens aren't strongly related, their attention scores are very low. The model dumps any remaining attention score in the first token.

The researchers call this first token an "attention sink."

"We need an attention sink, and the model decides to use the first token as the attention sink because it is globally visible—every other token can see it. We found that we must always keep the attention sink in the cache to maintain the model dynamics," Han says.

In building StreamingLLM, the researchers discovered that having four attention sink tokens at the beginning of the sliding cache leads to optimal performance.

They also found that the positional encoding of each token must stay the same, even as new tokens are added and others are bumped out. If token 5 is bumped out, token 6 must stay encoded as 6, even though it is now the fifth token in the cache.

By combining these two ideas, they enabled StreamingLLM to maintain a continuous conversation while outperforming a popular method that uses recomputation.

For instance, when the cache has 256 tokens, the recomputation method takes 63 milliseconds to decode a new token, while StreamingLLM takes 31 milliseconds. However, if the cache size grows to 4,096 tokens, recomputation requires 1,411 milliseconds for a new token, while StreamingLLM needs just 65 milliseconds.

"The innovative approach of StreamingLLM, centered around the attention sink mechanism, ensures stable memory usage and performance, even when processing texts up to 4 million tokens in length," says Yang You, a presidential young professor of computer science at the National University of Singapore, who was not involved with this work.

"This capability is not just impressive; it's transformative, enabling StreamingLLM to be applied across a wide array of AI applications. The performance and versatility of StreamingLLM mark it as a highly promising technology, poised to revolutionize how we approach AI-driven generation applications."

Tianqi Chen, an assistant professor in the machine learning and computer science departments at Carnegie Mellon University who also was not involved with this research, agreed, saying "Streaming LLM enables the smooth extension of the conversation length of large language models. We have been using it to enable the deployment of Mistral models on iPhones with great success."

The researchers also explored the use of attention sinks during model training by prepending several placeholder tokens in all training samples.

They found that training with attention sinks allowed a model to maintain performance with only one attention sink in its cache, rather than the four that are usually required to stabilize a pretrained model's performance.

But while StreamingLLM enables a model to conduct a continuous conversation, the model cannot remember words that aren't stored in the cache. In the future, the researchers plan to target this limitation by investigating methods to retrieve tokens that have been evicted or enable the model to memorize previous conversations.

More information: Guangxuan Xiao et al, Efficient Streaming Language Models with Attention Sinks, *arXiv* (2023). DOI: 10.48550/arxiv.2309.17453

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: A new way to let AI chatbots converse all day without crashing (2024, February 13) retrieved 6 May 2024 from https://techxplore.com/news/2024-02-ai-chatbots-converse-day.html