micron Educator Hub

Analyzing LLM performance: The impact of high-bandwidth memory on model inference Quiz

Reviewed 2025

micron

Copyright guidelines

By using any content provided by the Micron Educator Hub, you acknowledge that Micron Technology, Inc. (“Micron”) is the sole owner of the content and agree that any use of the content provided by the Micron Educator Hub must comply with applicable laws and require strict compliance with these Guidelines:

1. Credit shall be expressly stated by you to Micron for use of the content, including any portion thereof, as follows:
    a. “*© 2025 Micron Technology, Inc. All Rights Reserved. Used with permission.*”
2. You may not use the content in any way or manner other than for educational purposes.
3. You may not modify the content without approval by Micron.
4. You may not use the content in a manner which disparages or is critical of Micron, its employees, or Micron’s products/services.
5. Permission to use the content may be canceled/terminated by Micron at any time upon written notice from Micron to You if You fail to comply with the terms herein.
6. You acknowledge and agree that the content is provided by Micron to You on an “as is” basis without any representations or warranties whatsoever, and that Micron shall have no liability whatsoever arising from Your use of the content. Micron shall ensure that the content does not violate any statutory provisions and that no rights of third parties are infringed by the content or its publication. Otherwise, liability of the parties shall be limited to intent and gross negligence.
7. You acknowledge and agree that the content is the copyrighted material of Micron and that the granting of permission by Micron to You as provided for herein constitutes the granting by Micron to You of a non-exclusive license to use the content strictly as provided for herein and shall in no way restrict or affect Micron’s rights in and/or to the content, including without limitation any publication or use of the content by Micron or others authorized by Micron.
8. Except for the above permission, Micron reserves all rights not expressly granted, including without limitation any and all patent and trade secret rights. Except as expressly provided herein, nothing herein will be deemed to grant, by implication, estoppel, or otherwise, a license under any of Micron’s other existing or future intellectual property rights.

How to cite sources from the Micron Educator Hub

- Micron is committed to collaborate with educators to make semiconductor memory education resources available through the Micron Educator Hub
- The content in the Micron Educator Hub has been identified by Micron as current and relevant to our company
- Please refer to the table on the right for proper citation

Use case	How to cite sources
Whole slide deck or whole document Description: User uses the whole slide deck or whole document AS IS, without any modification	No additional citation required
Full slide or full page Description: User incorporates a full slide or a full page into their own slide deck or document	"© 2025 Micron Technology, Inc. All Rights Reserved. Used with permission."
Portion of a slide or portion of a page Description: User copies a portion of a slide or a portion of a page into a new slide or page	This is not allowed

Quiz ideas

1. What does the term “inference” mean when talking about large language models (LLMs)?

A. Training the model to understand new data

B. Storing the model’s responses for future use

C. Measuring how many tokens the model can generate

D. The process of generating a response based on user input and the model’s previous training

2. When evaluating the performance of an LLM during inference, which two metrics are commonly used?

A. Throughput and latency

B. Accuracy and vocabulary size

C. Training time and memory usage

D. Token count and model depth

3. Which of the following best describes how throughput is measured during inference?

A. By how quickly the model learns new information

B. By how many tokens the model can process per second

C. By the accuracy of the model’s response

D. By how much memory the model uses during training

4. What does the term “autoregressive” mean in the context of LLM token generation?

A. Each token is generated independently of others

B. Tokens are generated in parallel using multiple GPUs

C. Each token depends on the previously generated tokens

D. Tokens are generated based on fixed rules, not learned patterns

Quiz ideas

5. Which of the following best describes a major challenge in generative inference for large language models?

A. Limited access to training datasets

B. High memory requirements that exceed single-GPU capacity

C. Inability to tokenize long input sequences

D. Lack of support for autoregressive decoding

6. Why is inference in large language models fundamentally constrained?

A. Because GPUs are optimized for training, not inference

B. Because CPUs cannot handle parallel processing

C. Because GPUs have limited memory capacity

D. Because model weights are stored in cloud storage

7. What kinds of strategies help overcome GPU memory constraints during inference?

A. Using more training data and larger models

B. Applying techniques to optimize memory usage and distribute computation

C. Removing attention mechanisms from the model

D. Increasing the number of output tokens

8. Why does it take more computing power to process longer inputs in a language model?

A. The model stores each word separately

B. The model compares every word to every other word, and the number of comparisons (operations) increases with the square of the input length, which is known as quadratic scaling

C. The model uses more punctuation in longer inputs

D. The model switches to a slower mode for long inputs

Quiz ideas

9. If something “scales quadratically” with input size, N, what does that mean?

A. It grows slowly as the input gets bigger (N)

B. It doubles when the input doubles (2*N)

C. It increases by the square of the input size (N^2)

D. It stays the same no matter the input (constant at 1)

10. What does the batch size represent in the context of language model inference?

A. The number of GPUs used to run the model

B. The number of output tokens generated per second

C. The number of concurrent requests or clients the model can handle at once

D. The amount of memory required to store the model weights

Educator Hub

micron