micron
Educator Hub

A memory perspective: The effects of fine-tuning LLMs with high-bandwidth memory 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 attention mechanism allow a model to do?

   A. Translate images into text

   B. Ignore data from specific search engines

   C. Increase the number of layers in the model

   D. Focus on relevant parts of the input sequence and compute the relationship between tokens

2. Why is self-attention important in transformer models?

   A. It allows the model to ignore punctuation

   B. It enables the model to learn relationships between all tokens in a sequence

   C. It reduces the need for training data

   D. It replaces the need for embedding layers

3. What is the primary goal of Natural Language Processing (NLP)?

   A. To manipulate, interpret, and generate human language

   B. To translate programming languages

   C. To compress neural networks

   D. To store large datasets

4. Why has transformer architecture been widely used to train large language models (LLMs)?

   A. Transformers need less training data

   B. Transformers process inputs one at a time

   C. Convolutional neural networks were more accurate but harder to train

   D. Transformers can process all inputs in parallel and perform well across a wide range of NLP tasks

Quiz ideas

5. What are the two main stages of training large language models (LLMs)?

A. Encoding and decoding

B. Pretraining and fine-tuning

C. Classification and regression

D. Tokenization and embedding

6. What is one major challenge in training large language models mentioned in the report?

A. Lack of labeled data

B. Inability to generate multimodal outputs

C. Fitting model parameters into GPU memory

D. Slow inference speed

7. What are the two main components of a typical transformer model?

A. Encoder and decoder

B. Tokenizer and classifier

C. Input and output layers

D. Embedding and normalization

8. What does the decoder in a transformer model do during inference?

A. It tokenizes the input

B. It generates the output sequence based on tokens relationships

C. It normalizes the embeddings

D. It compresses the model weights

Quiz ideas

9. What is the purpose of quantization in large language model training?

A. To eliminate the need for training

B. To improve the model's vocabulary

C. To increase the number of model parameters

D. To reduce memory and computational requirements

10. How does quantization change the way data is represented in a model?

A. It converts text into images

B. It increases the number of bits used per parameter

C. It reduces the precision of data to use fewer bits

D. It removes unused tokens from the vocabulary

Educator Hub

micron