micron Educator Hub

Top five essential context window concepts in large language models Quiz

Reviewed 2026

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. “*© 2025-2026 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	"© 2025-2026 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) When a model hallucinates, what is it most likely doing? Choose two answers.
   A. AI is probably lying to you
   B. Intentionally misleading the user because a deception feature has been enabled
   C. Misinterpreting the user prompt possibly due to poorly structured training data
   D. Losing track of context due to architectural and hardware limitations

2) You are designing a prompt for a model with a 4,000-token input limit. What's the best approach? Choose two answers.
   A. Use as many words as possible
   B. Make sure you are clear about your intent
   C. Trim unnecessary content for the input
   D. Repeat key phrases to reinforce meaning

3) In the blog's "cocktail party" analogy, what does the model's attention mechanism represent?
   A. How extroverted and charming a model is
   B. The model's ability to prioritize certain inputs over others
   C. Whether a model can generate creative responses
   D. The model's aptitude for translating languages

4) What role does high-bandwidth memory technology (like HBM3E) play in inference with long context?
   A. Its role is to reduce the number of tokens needed
   B. It increases the model's parameter count to support complex reasoning
   C. HBM replaces the need for GPUs by offloading compute tasks to memory
   D. Higher bandwidth enables faster data transfer, allowing models to handle larger context windows

Quiz ideas

5) Which metric for inference is more relevant to memory performance and why? Inter-token latency (ITL) or first token (TTFT)?
   A. ITL because it reflects how fast data can move between HBM and the GPU accelerator
   B. TTFT because it measures how quickly the model starts generating output after receiving a prompt
   C. TTFT because is measures how fast data is read from the HBM
   D. ITL is a metric used to measure GPU performance only

6) Which of the following best explains why generating long *outputs* takes more time than processing long *inputs*? Choose two answers.
   A. Inputs are compressed, but outputs are not
   B. Processing input can be done parallel
   C. Generating output is done token by token
   D. Inputs are tokenized faster than outputs

7) Why is it wrong to assume that 100,000 tokens equals 100,000 words?
   A. Words are split into tokens, and the average ratio is about 0.75 words per token
   B. Tokenization compresses text so there are less tokens per word
   C. Tokens are often but not always longer than words, where the average is 1.75 tokens per word
   D. Token limits are based on characters, not words

8) Why do larger context windows (more than 1 million tokens) raise concerns about energy use?
   A. Energy use is unrelated to memory bandwidth or capacity
   B. Energy use only matters during training, not inference
   C. Bigger context windows reduce energy use because they allow parallel processing
   D. Longer contexts can keep hardware active for more time, increasing power use

Quiz ideas

9) Why do we say attention in transformer-based language models is “weighted”?

   A. All tokens get equal attention under the large language model constitution

   B. Tokens are ranked by frequency, not meaning

   C. The model gives more importance to some tokens than others

   D. Attention weights never change after training

10) A company wants to use a model with a 1 million token context window for a legal document review. What potential challenge could they expect?

   A. With such a large context, the model may not give equal attention to all tokens. The company could break documents into smaller sections or use tools that pull out the most relevant parts.

   B. Putting all documents in at once might lower accuracy. The company could consider outsourcing parts of inference to an overseas model to keep the quality of response acceptable.

   C. To handle such a large input efficiently, the model might further compress tokens, which may distort nuanced legal language.

   D. At this context length, AI is likely to hallucinate and blend unrelated clauses together, making up fictitious legal language. The company can prescribe talk therapy to the model.

Educator Hub

micron