Calvin AI/ML Handout   April 7, 2025   Name:					
7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-dimensional (i.e., 4096/32) keys, queries, and values.  Suppose we have an input of 100 tokens and we're computing the model's prediction of the next token. What are the shapes of the following activation (not parameter) matrices, i.e., the outputs on this sequence, for a single self-attention head from the LLaMA 7B model?  Assume no bias terms are used, and nothing clever is done to avoid storing or computing data that would be masked. Ignore batching: assume a batch size of 1 and don't write that dimension.  Description Rows Columns Input to this head Queries  Keys  Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin Al/ML Handout   April 7, 2025  Name:  Let's consider Meta Al's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Calvin AI/ML Handout   April 7, 2025	Name:			
the shapes of the following activation (not parameter) matrices, i.e., the outputs on this sequence, for a single self- attention head from the LLaMA 7B model?  Assume no bias terms are used, and nothing clever is done to avoid storing or computing data that would be masked. Ignore batching: assume a batch size of 1 and don't write that dimension.  Description Rows Columns Input to this head Queries  Keys  Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	7B variant has an embedding dimension of 4096. It uses m		9		
batching: assume a batch size of 1 and don't write that dimension.    Description	the shapes of the following activation (not parameter) mat	-			
Input to this head  Queries  Keys  Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	——————————————————————————————————————		nat would be masked. Ignore		
Queries  Keys  Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Description	Rows	Columns		
Keys  Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Input to this head				
Values  Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Queries				
Self-Attention Scores  Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Keys				
Self-Attention Weights  Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Values				
Output of this head (after projection back to embedding space)  Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Self-Attention Scores				
Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md  Calvin AI/ML Handout   April 7, 2025  Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Self-Attention Weights				
Calvin AI/ML Handout   <i>April 7, 2025</i> Name:  Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-					
Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Model card: https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md				
Let's consider Meta AI's LLaMA model, the first widely used open-weights LLM. According to the Model Card, the 7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-					
7B variant has an embedding dimension of 4096. It uses multi-head attention: each of its 32 heads computes 128-	Calvin AI/ML Handout   April 7, 2025	Name:			
dimensional (i.e., 4096/32) keys, queries, and values.	·	1 0	8		
Suppose we have an input of 100 tokens and we're computing the model's prediction of the next token. What are the shapes of the following <i>activation</i> (not parameter) matrices, i.e., the outputs on this sequence, for a <i>single self-attention</i> head from the LLaMA 7B model?	the shapes of the following activation (not parameter) mat	-			
Assume no bias terms are used, and nothing clever is done to avoid storing or computing data that would be masked. Ignore batching: assume a batch size of 1 and don't write that dimension.			nat would be masked. Ignore		
Description Rows Columns	Description	Rows	Columns		
Input to this head	Input to this head				

Description	Rows	Columns
Input to this head		
Queries		
Keys		
Values		
Self-Attention Scores		
Self-Attention Weights		
Output of this head (after projection back to embedding space)		

Before you leave, pick a couple of these questions to react to:

- 1. What was the most important concept from today for you?
- 2. What was the muddiest concept today?
- 3. How does what we did today connect with what you've learned before?
- 4. What would you like to review or clarify next time we meet?
- 5. What are you curious, hopeful, or excited about?

Before you leave, pick a couple of these questions to react to:

- 1. What was the most important concept from today for you?
- 2. What was the muddiest concept today?
- 3. How does what we did today connect with what you've learned before?
- 4. What would you like to review or clarify next time we meet?
- 5. What are you curious, hopeful, or excited about?