**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-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 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-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

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?