Calvin AI/ML Handout | *Apr 10, 2026* Name: _____

Let's consider Alibaba's Qwen2.5 model, a widely used open-weights LLM. The 0.5B variant has an embedding dimension of 896. It uses multi-head attention: each of its 14 heads computes 64-dimensional (i.e., 896/14) keys, queries, and values.*

Suppose we have an input of 50 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 Qwen2.5 0.5B 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: https://huggingface.co/Qwen/Qwen2.5-0.5B
* We simplify here. The model actually computes only 2 keys and values for each token, but still 14 queries. See the tech report.

Calvin AI/ML Handout | *Apr 10, 2026* Name: _____

Let's consider Alibaba's Qwen2.5 model, a widely used open-weights LLM. The 0.5B variant has an embedding dimension of 896. It uses multi-head attention: each of its 14 heads computes 64-dimensional (i.e., 896/14) keys, queries, and values.*

Suppose we have an input of 50 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 Qwen2.5 0.5B 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: https://huggingface.co/Qwen/Qwen2.5-0.5B
* We simplify here. The model actually computes only 2 keys and values for each token, but still 14 queries. See the tech report.

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

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

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

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