Calvin AI/ML Handout | *April 9, 2025*

Name: _____

1. Sketch a plausible self-attention matrix that an autoregressive Transformer LM might compute when given the sequence "1 a 2 b 3" (assume each gets its own token). Which attention weights must be 0?

	key for `1`	key for `a`	key for `2`	key for `b`	key for `3`
query for `1`					
query for `a`					
query for `2`					
query for `b`					
query for `3`					

2. Suppose, as part of post-training, a model is prompted with "a JSON list of 1 through 3", two samples are generated, and we get: `[1, 2, 3]` and `[1, 2, 3, 4)`.

   a. How might we obtain a label that the first sample is better than the second?
   b. How might we use that signal to update the model?

Calvin AI/ML Handout | *April 9, 2025*

Name: _____

1. Sketch a plausible self-attention matrix that an autoregressive Transformer LM might compute when given the sequence "1 a 2 b 3" (assume each gets its own token). Which attention weights must be 0?

	key for `1`	key for `a`	key for `2`	key for `b`	key for `3`
query for `1`					
query for `a`					
query for `2`					
query for `b`					
query for `3`					

2. Suppose, as part of post-training, a model is prompted with "a JSON list of 1 through 3", two samples are generated, and we get: `[1, 2, 3]` and `[1, 2, 3, 4)`.

   a. How might we obtain a label that the first sample is better than the second?
   b. How might we use that signal to update the model?

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?