Calvin AI/ML Handout | *April 28, 2025*

Name: _____

1. The Stable Diffusion model, like other models that handle text, treat text as a sequence of tokens. Stable Diffusion struggles with typos in the text prompt. Also, these models are known to have trouble generating correctly spelled text in their images even when the prompts are spelled correctly. **Use what you know about tokenization to explain these two observations.**

2. The Stable Diffusion text encoder looks up 768-dimensional embeddings for its input tokens, then passes those embeddings through 12 transformer blocks. Each has a self-attention layer and a feedforward layer (do you remember what each of these is for?). The self-attention layer has 12 heads, each with 64 dimensions. Consider just computing the *key* vectors for *one* of the attention heads:
   a. What is the shape of the matrix that is multiplied by the input embeddings to produce the key vectors?

      rows: _____, **columns**: _____
   b. How many parameters does this matrix have? _____
   c. How much memory would it take to store the key-projection matrices for *all* 12 (layers) * 12 (heads per layer) = 144 heads, if we store them as 32-bit floats?
   d. How long would it take a high-end CPU (memory bandwidth of 100 GB/s) to read this matrix from memory?
   e. How about for a GPU with 1,000 GB/s memory bandwidth?
   f. How would the speed and memory requirements change if we used 8-bit integers instead of 32-bit floats?

Calvin AI/ML Handout | *April 28, 2025*

Name: _____

1. The Stable Diffusion model, like other models that handle text, treat text as a sequence of tokens. Stable Diffusion struggles with typos in the text prompt. Also, these models are known to have trouble generating correctly spelled text in their images even when the prompts are spelled correctly. **Use what you know about tokenization to explain these two observations.**

2. The Stable Diffusion text encoder looks up 768-dimensional embeddings for its input tokens, then passes those embeddings through 12 transformer blocks. Each has a self-attention layer and a feedforward layer (do you remember what each of these is for?). The self-attention layer has 12 heads, each with 64 dimensions. Consider just computing the *key* vectors for *one* of the attention heads:
   a. What is the shape of the matrix that is multiplied by the input embeddings to produce the key vectors?

      rows: _____, **columns**: _____
   b. How many parameters does this matrix have? _____
   c. How much memory would it take to store the key-projection matrices for *all* 12 (layers) * 12 (heads per layer) = 144 heads, if we store them as 32-bit floats?
   d. How long would it take a high-end CPU (memory bandwidth of 100 GB/s) to read this matrix from memory?
   e. How about for a GPU with 1,000 GB/s memory bandwidth?
   f. How would the speed and memory requirements change if we used 8-bit integers instead of 32-bit floats?

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?