Calvin AI/ML Handout | *Mar 19, 2025* Name: ______________________________

Warm-Up

Consider the example from last time of a conversation where the user asks “Tell me a joke” and the model responds with a corny joke about atoms. See https://huggingface.co/spaces/CalvinU/writing-prototypes for the full example.

1. Write an example of the “document” (the sequence of tokens) that was fed to the Gemma-2 model to generate one of the first words of the joke. Mark the boundaries between tokens.
2. Write an example of the next-token distribution for what came next in the joke.

Next-Token Generation

Go to https://huggingface.co/spaces/kcarnold/next-token to complete the following exercises.

Greedy Generation

Generate *one complete translation.* At each step, use the **single most likely token**. Compute the total log probability by taking the sum of the logprobs for each token.

Our translation (loss = __________): ______________________________

Generation by Sampling

Generate one complete translation. At each step, **sample from the top tokens according to their probability**. *To do this, pick a random number between 0 and 1, and find the first number under the* ***cumulative probability*** *column that is less than it.*

Then repeat the process again, drawing different random numbers. (If you don’t end up with different choices within the first few tokens, re-draw until you do.)

Our translations:

1. (loss = __________): ______________________________
2. (loss = __________): ______________________________

Modifying the generation

Start the translation more informally (like “Hey y’all”), then see how the rest of the translation changes.

Informal translation (loss = __________): ______________________________

Beam Search Generation

Generate 2 complete translations. Start by taking the top 2 starting tokens. For each of them, find the most likely *following* token. But instead of keeping all 4 possible sequences, only keep the sequences with the largest total logprob (including the new token).

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?