**Calvin AI/ML** Handout | *Mar 18, 2026*  Name: ____________________

*Open the **Language Model Internals** tool. Work in teams of 2–3.*

## Phase 1: Building a Response Token by Token

Type a message like: `Tell me a 3-sentence story about a dragon.` Click “End Turn,” then look at how the tool displays the conversation.

1. Where does the user’s turn end and the assistant’s turn begin? What markers separate them?
2. The model predicts a **next-token distribution**: a list of candidate tokens with probabilities. Pick the **most likely** token and click it. Repeat  10 times, always picking the top prediction. Write the first 10 tokens here:
3. Compare with a neighboring team who used the same prompt. Did you get the same sequence? Why or why not?

## Phase 2: The Model’s Distribution

Now generate a full 3-sentence story (use “Generate Response”). Click on tokens in the generated story to see the next-token distribution at each position.

1. Find a token where the context **tightly constrains** what comes next—one option dominates with >90% probability. (The token **after** it will show green.) Write the *token*, its *probability*, and the 2–3 tokens before it (the context):
2. Find a token where the context **leaves many options open**—several options have similar probability. Write it down with context. Why is this position less constrained?
3. Look at the high-probability options at that open position. Do they represent genuinely different directions for the story, or variations on the same kind of continuation?
4. What kinds of tokens tend to be tightly constrained? What kinds tend to be open? State a general pattern.

*Flip over for Phase 3: explore how sensitive the model is to changes in the prompt and response so far.*

## Phase 3: Explore

Generate a story and try variations on the prompt or the response so far. How sensitive is the model's output to small changes? Here are some starting points, but follow your curiosity and be playful:

- Swap an adjective or name in the prompt. Does the response change a little or a lot?
- Add or remove a sentence from the response so far. What stays the same?
- What kinds of tokens are always high-confidence, regardless of context?
- Force an unlikely token early in the response. What happens next?

You might try asking for other things, like a children's poem, a Python program, a translation of a sentence, or a summary of an article.

1. What I tried:

   What I noticed:

2. What I tried:

   What I noticed:

3. What I tried:

   What I noticed:

Try something none of the above suggest:

4. What I tried:

   What I noticed:

## Make Sense of It

Based on your experiments: which variations in the prompt or response so far mattered to the model, and which didn't? Can you state a pattern or rule?

What questions does this raise about how the model processes its input?