Next-Token Prediction: The Core Mechanism

How LLMs Decide What Comes Next

Input Context: **"The capital of France is"**

1 Model Processing

- All input tokens processed simultaneously.
- Context is built via self-attention layers.
- Calculates a probability distribution for the entire vocabulary.

→

2 Probability Distribution

→

3 Token Selection

Temp = 0: Always picks "Paris" (deterministic).
Temp = 0.7: Weighted random choice (balanced).
Temp = 1.2: More random, creative output.

→

4 Process Repeats

The output becomes new input:

```
"The capital of France is Paris"
```

...and the model predicts the next token (e.g. a period).

Temperature (0.0 - 2.0)

Lower: More focused, predictable.
Higher: More creative, random.

Top-p Sampling (0.1 - 1.0)

Considers tokens with cumulative probability up to 'p'. Filters out highly unlikely tokens.

Max Tokens

Sets the maximum length of the generated response. Prevents infinite generation loops.

🧠 **Core Insight:** Every word you see from an LLM was chosen from thousands of possibilities, with the model weighing context, training, and randomness parameters to make each selection.