Temperature and Top_P Parameters in LLM Generation

Adil Karim

December 3, 2025

1 Temperature

Temperature is a parameter that controls the randomness and creativity of the language model's output. It operates on the probability distribution of the next token by scaling the logits (raw scores) before applying the softmax function.

How it works

- **Low temperature (0.0–0.3)**: Makes the model more deterministic and focused. The model will choose the most likely tokens, resulting in more consistent, predictable, and conservative responses. This is ideal for tasks requiring accuracy and factual information, such as technical documentation or data extraction.
- **High temperature (0.7–1.0)**: Makes the model more creative and diverse. The probability distribution becomes flatter, allowing less likely tokens to be selected. This leads to more varied, creative, and sometimes unexpected outputs. Higher temperatures are useful for creative writing, brainstorming, or generating diverse ideas.

Mathematical effect

Temperature scales the logits by dividing them by the temperature value. Lower temperature values make the distribution sharper (more peaked), while higher values flatten it, making all tokens more equally probable. Formally, given logits l_i, the adjusted probabilities p_i are computed as:

$$p_i = \frac{\exp(l_i/T)}{\sum_j \exp(l_j/T)}$$

where T is the temperature parameter.

2 Top_P (Nucleus Sampling)

Top_P, also known as nucleus sampling, is a parameter that controls the diversity of the output by limiting the model to consider only the tokens whose cumulative probability mass exceeds a certain threshold.

How it works

- **Low Top_P (0.1–0.5)**: The model considers only the most probable tokens, resulting in more focused and deterministic outputs. This is similar to low temperature but works by filtering the token set rather than scaling probabilities.
- **High Top_P (0.9–1.0)**: The model considers a broader set of tokens, including less probable ones, leading to more diverse and creative outputs. When Top_P is 1.0, all tokens are considered.

Mathematical effect

The model sorts tokens by probability, then selects the smallest set of tokens whose cumulative probability exceeds the Top_P threshold. Only tokens in this "nucleus" are considered for sampling. Formally, let the sorted probabilities be $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(V)}$, where V is the vocabulary size. Then the nucleus N is the smallest set such that:

$$\sum_{i \in N} p_{(i)} \geq \text{Top\_P}$$

3 Interaction and Best Practices

Temperature and Top_P work together to control output quality:

- **For factual, accurate responses** (like our heavy machinery knowledge base): Use low temperature (0.0–0.3) and moderate Top_P (0.5–0.7)
- **For creative tasks**: Use higher temperature (0.7–1.0) and higher Top_P (0.9–1.0)
- **For balanced responses**: Use moderate settings (temperature: 0.5–0.7, Top_P: 0.7–0.9)

In our Bedrock chat application, we use configurable sliders for both parameters, allowing users to experiment and find the optimal balance between accuracy and creativity for their specific queries.