Temperature and Top_P Parameters Explanation

Overview

Temperature and top_p are two critical hyperparameters that control the randomness and creativity of Large Language Model (LLM) responses. Understanding and properly configuring these parameters is essential for achieving desired output quality in AI applications.

Temperature

Temperature is a parameter that controls the randomness of the model's predictions by scaling the logits (raw prediction scores) before applying the softmax function. It directly influences how "creative" or "deterministic" the model's outputs will be.

How Temperature Works:

- Range: Typically between 0.0 and 1.0

- Low Temperature (0.0 - 0.3): Makes the model more deterministic and focused

- Medium Temperature (0.4 - 0.7): Balanced between creativity and coherence

- High Temperature (0.8 - 1.0+): Increases randomness and creativity

Mathematical Effect:

$P(\text{token}_i) = \exp(\text{logit}_i / T) / \Sigma \exp(\text{logit}_j / T)$

Practical Example:

For the prompt "The capital of France is":

- Temperature 0.0: "Paris."

- Temperature 0.7: "Paris." or variations

- Temperature 1.5: More diverse responses

Top_P (Nucleus Sampling)

Top_p selects the smallest set of tokens whose cumulative probability exceeds a threshold p.

How Top_P Works:

- Range: Between 0.0 and 1.0

- Mechanism: Sort tokens → select cumulative ≥ p → renormalize

Practical Example:

Probabilities: [0.4, 0.3, 0.15, 0.1, 0.05]

- Top_p = 0.5: Only first token

- Top_p = 0.8: First two tokens

- Top_p = 0.95: First four tokens

- Top_p = 1.0: All tokens

Temperature vs Top_P: Key Differences

Temperature scales logits; Top_p filters tokens by cumulative probability.

Combined Effect:

1. Temperature adjusts probability distribution

2. Top_p filters candidate tokens

3. Model samples final token

Recommended Configurations:

- Factual tasks: Temp 0.0–0.3, Top_p 0.1–0.5

- General conversation: Temp 0.5–0.7, Top_p 0.7–0.9

- Creative tasks: Temp 0.7–0.9, Top_p 0.9–1.0

- Maximum creativity: Temp ≥1.0, Top_p = 1.0

Impact on Bedrock Application:

Recommended:

- Temperature: 0.3–0.5

- Top_p: 0.7–0.85

Example Output Variation:

Low temp/top_p: direct factual answer

Medium: detailed explanation

High: creative elaboration

Best Practices:

Start conservative, test thoroughly, adjust per task, document settings, get user feedback.

Conclusion:

Temperature adjusts randomness; top_p filters based on probability mass. Proper tuning ensures balanced, reliable AI behavior.