Temperature and Top_p Parameters Analysis

Project: Heavy Machinery RAG Support System

1. Introduction

This document analyzes the effects of temperature and top_p parameters on the AWS Bedrock Claude model when used in a Retrieval-Augmented Generation (RAG) system for answering technical questions about heavy machinery specifications.

2. Parameter Definitions

Temperature (0.0 - 1.0):

- Controls randomness in response generation
- Lower values = more deterministic and focused
- Higher values = more creative and varied

Top_p (0.0 - 1.0):

- Controls diversity through nucleus sampling
- Lower values = only considers highest probability tokens
- Higher values = considers broader range of tokens

3. Testing Methodology

Test 1 - Temperature Variation:

- Fixed: top_p = 0.0
- Tested: temperature = 0.1, 0.7, 1.0
- Question: "What is the maximum load capacity of the FL250 forklift?"

Test 2 - Top_p Variation:

- Fixed: temperature = 0.0
- Tested: top_p = 0.1, 0.5, 0.9 (assumed values)
- Question: "What type of engine does the excavator use?"

4. Results

4.1 Temperature Variation Results

All three temperature values (0.1, 0.7, 1.0) produced identical responses:

According to the information provided, the maximum lifting capacity of the FL250 Heavy-Duty Industrial Forklift is 25,000 kg or 55,115 lb.
The key specifications section states: "Lifting Capacity: 25,000 kg / 55,115 lb"
So the maximum load capacity of the FL250 forklift is 25,000 kg or 55,115 lb.

Observation: When top_p is set to 0.0, temperature has NO observable effect on output.

4.2 Top_p Variation Results

All three top_p values produced nearly identical content with only minor formatting differences:

Response Pattern:

According to the information provided, the LE950 Large Excavator is equipped with a high-performance engine that meets the latest emission standards with an advanced aftertreatment system.

Key specifications:
- Net Power: 523 kW (701 hp)
- Bore: 137 mm (5.4 in) - Stroke: 152 mm (6.0 in)
- Displacement: 27.0 L (1,648 in³)

Observation: When temperature is 0.0, top_p variations caused only minimal formatting changes, not content changes.

5. Analysis

5.1 Why Parameters Had Minimal Effect

In this RAG system:

1. **Strong Retrieved Context:** The Knowledge Base provides specific, information that heavily constrains the model's responses
2. **Low Parameter Values:** When either parameter is set to 0.0, it forces deterministic output, negating the effect of the other parameter
3. **Technical Domain:** Factual questions about specifications have clear, unambiguous answers from the retrieved documents

Finding	Explanation
Temperature effect minimal when top_p = 0.0	Top_p of 0.0 restricts token selection so severely that temperature cannot add variation
Top_p effect minimal when temperature = 0.0	Temperature of 0.0 makes selection deterministic, limiting top_p's influence
RAG constraints dominate	Retrieved context provides such specific information that parameters have limited impact

6. Implications for This Project

6.1 Recommended Settings

Based on testing results:

temperature = 0.2 # Low but not zero
top_p = 0.1 # Very low for precision

Rationale:

- Both parameters at 0.0 is overly restrictive

- Low values maintain accuracy while allowing minimal natural variation
- Prevents completely robotic responses without sacrificing precision

6.2 Why Not Use 0.0 for Both?

While 0.0 values worked in testing, slightly higher values (0.1-0.3) provide:

- More natural language flow
- Better handling of edge cases
- Flexibility for questions without exact document matches

7. Conclusion

Main Findings:

1. In RAG systems with strong retrieved context, parameter effects are minimal
2. Temperature has no observable effect when top_p = 0.0
3. Top_p has minimal effect when temperature = 0.0
4. For technical support applications, very low values for both parameters ensure accuracy

Recommendation: Use temperature = 0.2 and top_p = 0.1 to maintain precision while avoiding overly rigid responses.

Critical Insight: The quality of retrieved documents matters more than parameter tuning in RAG systems. Accurate source material produces accurate responses regardless of minor parameter variations.