A Detailed Explanation of Model Parameters:- Temperature and top_p

Introduction: Controlling AI Creativity

When a large language model (LLM) like Anthropic's Claude generates a response, it doesn't just "pick" the single best word. Instead, it analyzes the prompt and calculates a probability score for *every* word in its vocabulary that could come next. The parameters temperature and top-p are two different "dials" we can use to control how the model chooses from that list of probable words. Both are fundamentally tools for controlling the **randomness** (and therefore the creativity) of the model's output. It is the careful tuning of these parameters that allows us to get a factual, predictable answer for one task and a highly creative, brainstorming-style answer for another.

Temperature

- **What it is:** Temperature is a parameter that directly modifies the probability distribution of the potential next words. It's like adjusting the focus or "sharpness" of the model's choices.
- **How it works:**
    - **Low Temperature (e.g., 0.0 - 0.2):** A low temperature "sharpens" the probability list, dramatically increasing the score of the most likely word and decreasing the scores of all others. The model becomes highly deterministic, predictable, and "safe." It will almost always pick the single most obvious, high-probability word. This is what we used in our project temperature:- 0.1, because we want factual, consistent answers from our technical documents.
    - **High Temperature (e.g., 0.8 - 1.0):** A high temperature "flattens" the probabilities, making less-likely words more competitive with the top choices. This injects a high degree of randomness, allowing the model to be more creative, surprising, and even poetic. It's great for brainstorming ideas or writing stories, but it significantly increases the risk of the model "hallucinating" (making up facts) or going off-topic.
- **Analogy:** Imagine the model is choosing its next word from a lottery.
    - At **low temperature**, it's like having a bag with 99 "the" balls and 1 "a" ball. It will almost certainly pick "the".
    - At **high temperature**, it's like having a bag with 10 "the" balls, 9 "a" balls, 8 "an" balls, and so on. The choice is far less predictable.

Top-p

- **What it is:** Top-p (also known as "nucleus sampling") is a different method for controlling randomness. Instead of changing the probabilities, it sets a *cutoff* based on the *sum* of their probabilities.
- **How it works:**
    - The model lists all possible next words, sorted by probability (e.g., "The": 50%, "A": 20%, "It": 15%, "Based": 10%, "According": 3%, ...).
    - Top-p is a percentage (e.g., 0.9 for 90%). The model goes down the list, adding up the probabilities until it hits this percentage. This creates a "nucleus" or "pool" of the most likely choices.
    - **Example (top_p = 0.85):**
        1. Add "The" (50%) -> Total is 50%.
        2. Add "A" (20%) -> Total is 70%.
        3. Add "It" (15%) -> Total is 85%.
        4. **STOP.** The cumulative probability has reached 85%.
    - The model will now *only* choose its next word from the pool of {"The", "A", "It"}. All other words ("Based", "According", etc.) are completely ignored for this step, even if they were next on the list.
- **Analogy:** Imagine the model is ordering from a menu.
    - At **low top-p(e.g., 0.1)**, it's like saying, "I will only consider the #1 most popular item on the menu." The choice is extremely limited.
    - At **high top-p(e.g., 0.95)**, it's like saying, "I will consider all the popular dishes that, together, make up 95% of all restaurant orders." This gives the model a much larger, but still safe, pool of good options to choose from.

Which One to Use?

Temperature and Top-p solve the same problem in two different ways. It is **strongly recommended to only use one at a time.** In our project, we set Top-p to0.9 or 1 (effectively turning it off) and controlled the output with Temperature. This is a common and effective strategy. Top-p is often favored for being more dynamic—it creates a small pool of choices when the model is very "certain" (e.g., "the") and a larger pool when it's less certain (e.g., the first word of a poem).