Temperature and Top P in LLMs

Large Language Models (LLMs) rely on **Temperature** and **Top P** (Nucleus Sampling) to govern the *diversity* and *coherence* of their output. Both modulate randomness but operate differently, enabling engineers to strike a balance between precision and creativity.

Temperature: Controlling Randomness

Temperature is a parameter that **scales the probability distribution** of the model's next-token predictions. It affects how "confident" or "spread out" the distribution is.

Low temperature (0.1-0.4) → more deterministic, focused, predictable output (Factual QA, coding, summarization, legal writing)
High temperature (0.9 -) → more random, creative, unpredictable output (Creative writing, poetry, ideation)

Top P (Nucleus Sampling): Constraining the Vocabulary

Top-P (nucleus sampling) controls randomness by sampling only from the smallest set of tokens whose cumulative probability exceeds a threshold P.

Top-P > 0.9 → considers the entire distribution (no filtering) (More diverse and creative responses)
Top-P < 0.5 → filters out unlikely tokens until the top tokens collectively reach probability P (Stable, deterministic, safer text)

Simple Analogy

Temperature = turns the "creativity dial" up or down
Top-P = narrows or widens the pool of possible next tokens

Mastering Temperature and Top P is essential for generating LLM text precisely calibrated for reliability or creative expression.