Clear and accurate explanation of how temperature and top_p affect AI responses

Temperature and top_p are parameters that control how creative or predictable an AI model's responses will be. *Temperature* adjusts randomness: a low temperature (e.g., 0–0.3) makes answers more focused, deterministic, and factual, while a high temperature (e.g., 0.8–1.2) encourages creativity, exploration, and varied wording. It doesn't change the model's knowledge – only how freely it samples from possible next words.

Top_p (also called *nucleus sampling*) controls how much of the probability distribution the model considers during generation. Instead of adjusting randomness globally like temperature, top_p limits choices to the smallest set of tokens whose combined probability equals p. A low top_p (e.g., 0.1–0.3) makes the model very selective, sticking to the most likely words, while a higher top_p (e.g., 0.8–1.0) allows more variety and nuanced responses. In short, temperature spreads probabilities, while top_p filters them, and together they shape the balance between accuracy and creativity in AI outputs.