Explanation of how temperature and top_p affect AI responses

Temperature controls how “sharp” or “flat” the probability distribution of the next token becomes. When temperature is low (0.1–0.3), the model heavily favors the most likely next tokens. This makes the response more deterministic, repetitive, and precise—ideal for factual answers, coding, or technical reasoning. As temperature increases toward 1.0, the model loosens this preference and spreads probability across more possible words, making responses more creative, expressive, and varied. Extremely high temperatures (above 1.2) can cause the model to generate unpredictable or incoherent outputs. In simple terms, temperature adjusts how adventurous or conservative the model is when choosing each word.

Top_p (nucleus sampling) works by limiting the model’s choices to a subset of tokens whose cumulative probability is at least p. Instead of selecting from all possible words, the model only picks from the most meaningful “nucleus.” For example, if top_p = 0.9, the model gathers the most likely tokens until their total probability adds up to 90%, and only those tokens are considered. A lower top_p (like 0.3–0.5) results in highly focused, safe, predictable responses because the model restricts itself to a very small, high-confidence set. A higher top_p (0.9–1.0) allows more diversity but still avoids extremely unlikely or nonsensical options. Top_p is essentially a filter that defines how wide the idea space is.

When used together, temperature controls the shape of the distribution, while top_p controls how much of that distribution is allowed—giving fine-grained control over creativity, randomness, and coherence.