Model Parameters Explanation: Temperature and Top_P

How Temperature and Top_P Affect AI Responses

Temperature controls how random or predictable the AI's responses are. It ranges from 0 to 1, where lower values (like 0.1–0.3) make the model choose the most likely words, giving consistent and focused answers suitable for factual questions. Higher values (like 0.7–1.0) make the model more creative by allowing it to pick less obvious words, which is useful for brainstorming or creative tasks. You can think of it like turning a dial: at 0, you get the same answer every time; at 1, you get more varied and sometimes unexpected responses.

Top_p, also called nucleus sampling, works by limiting which words the model can choose from. Instead of modifying how likely each word is, it only considers the most probable words until their combined probability reaches the top_p value. For example, with top_p = 0.9, the model only picks from words that together make up 90% of the likely choices, ignoring the rest. Lower top_p values (0.1–0.5) keep responses focused and precise, while higher values (0.8–1.0) allow more variety.In our Bedrock chat application for heavy machinery information, we use temperature values between 0.3–0.5 and top_p values between 0.7–0.85 to get accurate technical details while keeping the language natural and conversational.