Model Parameters: explanation of how temperature and top_p affect AI responses

Temperature

Temperature controls the randomness of the model's outputs. Lower values (0-0.3) yield more deterministic and coherent responses, ideal for factual tasks. Medium values (0.3-0.7) balance creativity and coherence, while higher values (0.7-1) promote diverse and imaginative outputs but may sacrifice clarity.

Top_p

Top_p, or nucleus sampling, determines the range of tokens considered during generation. A low top_p (e.g., 0.5) restricts choices to the most probable tokens, enhancing focus, whereas a high top_p (e.g., 0.9) allows for a broader selection, increasing diversity in responses.

Together, these parameters enable fine-tuning of AI outputs for various applications, from precise answers to creative storytelling.