BHARAT KUMAR GOPE

Temperature and top_p Explanation

Temperature is a parameter that influences the randomness of the AI's responses. Lower temperature values, such as 0.2, make the model's output more predictable and focused, often resulting in precise but less creative answers. Higher values, like 0.8 or 1.0, introduce more variability, allowing the model to explore less likely word choices and generate more diverse or imaginative responses. Essentially, temperature controls how "risk-taking" or "conservative" the model is when deciding the next word in a sequence.

top_p, also known as nucleus sampling, controls randomness in a complementary way by considering only the most probable tokens whose cumulative probability adds up to a certain threshold. For example, setting top_p to 0.9 restricts the model to choose from tokens that collectively account for 90% of the probability mass, while ignoring rare words with very low likelihood. By adjusting top_p, one can achieve a balance between creativity and coherence, ensuring that responses remain meaningful and contextually relevant without being overly repetitive or unpredictable.