Temperature and Top_p in Generative AI Models

Temperature controls how creative or deterministic the model's responses are.
A lower temperature (e.g., 0.2) makes the model's outputs more focused and consistent — ideal for factual or technical questions.
A higher temperature (e.g., 0.8) introduces more randomness, making answers more creative and diverse.

Top_p (nucleus sampling) limits the model's word choices to the smallest set whose combined probability mass is ≥ *p*.
Lower values (e.g., 0.5) produce more deterministic answers; higher values (e.g., 0.9) allow broader, more imaginative responses.
Together, these parameters balance precision and creativity in generative AI outputs.