When I work with AI models, I pay close attention to temperature and top_p because they significantly influence the style and reliability of the responses. Temperature controls how deterministic or creative the model becomes. With a low temperature, the output is more precise, consistent, and predictable, which is ideal for technical tasks or situations where accuracy matters. When the temperature is higher, the model allows more randomness in its word choices, producing responses that can be more creative or exploratory. I see temperature as a way to adjust the balance between strict accuracy and flexible, expressive output.

Top_p also affects randomness but in a more targeted way. Instead of changing the overall creativity level, top_p limits the model to a specific portion of the most likely responses. For example, a top_p value of 0.9 tells the model to only consider the top 90% of probable next words, filtering out unlikely or extreme options. This gives me fine-grained control over how broad or narrow the model's response space should be. Together, temperature and top_p help me tailor the model's behavior so that it remains either highly focused or more open-ended depending on the needs of the task.