Temperature and Top_p Parameters Explanation

Temperature:
It controls the randomness and creativity of AI responses.
Lower values (0.1-0.3) are for more deterministic, focused, and consistent responses and higher values (0.7-1.0) are for more creative, diverse, and unpredictable responses
In my heavy machinery chatbot, I use moderate temperature for balanced, informative responses.

Top_p (Nucleus Sampling):
It controls the diversity of token selection by considering only the top probability mass.
Lower values (0.1-0.3) are for more focused, considering only the most probable tokens, while the Hhigher values (0.7-1.0) are for more diverse, considering a wider range of tokens.
In my application, it helps ensure responses stay relevant to machinery topics while allowing some variety.

These parameters working together balance response quality between being too generic (high values) and too repetitive (low values).