Temperature and top_p

Temperature: The temperature parameter controls how random or creative the AI's responses are. When we want more predictable and deterministic output, we use a lower temperature. If we want more creative or varied responses, we choose a higher temperature. The temperature value affects how strongly the model prefers high-probability tokens: lower temperature focuses on the most likely tokens, while higher temperature allows the model to choose less likely ones.

Top_p: The top-p parameter also affects the diversity of the words the model can choose, but it works differently. Top-p limits token selection to a subset of tokens whose combined probability adds up to *p*. With a lower top-p, the model is restricted to only the most likely words. With a higher top-p, the model can choose from a larger and more diverse set of words, including less common or more unusual ones.