Temperature controls how random or predictable an AI's answers are. If we set it low (like 0.0 to 0.3) and the model stays sharp and consistent, it picks the most likely words every time, perfect for factual answers or technical writing. If we take it higher (0.7 to 1.0), answers get more creative and varied because the model treats less-likely words as almost equal contenders, great for storytelling or brainstorming.

Top_p (nucleus sampling) works differently: instead of tweaking probabilities, it only looks at the smallest group of words that together make up at least “top_p” of the total probability (say 90%). This keeps responses diverse without letting total nonsense slip in. It’s smarter than temperature alone because it automatically adjusts, if the model is very confident, it narrows choices; if unsure, it opens up. For document Q&A, pairing a moderate temperature (0.3–0.5) with top_p around 0.9 usually gives accurate yet naturally flowing answers.