For this project, I experimented with different temperature and top_p values to see how they affect the model's behavior. Both parameters control how the model chooses the next token, but they influence the output in slightly different ways.

Temperature adjusts how deterministic or creative the model becomes. When I set the temperature low, the responses were very consistent and factual because the model mainly relied on the highest-probability tokens. This made the answers stable and predictable, which works well for retrieval-augmented tasks like the ones in this project. When I raised the temperature, the responses became more expressive and varied. The model considered more word choices and sentence structures. This can be useful for open-ended or creative tasks, but it may not always be best for technical accuracy.

The top_p parameter controls the size of the probability pool the model samples from. Lower top_p values restrict the model to a small set of high-confidence tokens. This keeps the output focused and lowers randomness. Higher top_p values widen the sampling pool, allowing the model to access a broader range of vocabulary. This increases diversity without making the model as unpredictable as increasing temperature. When I combined higher temperature with higher top_p, the effect was noticeably more creative and less deterministic. When both values were low, the model stayed extremely grounded and consistent.

Overall, experimenting with these parameters made it clear that temperature influences creativity more broadly, while top_p shapes how wide the model casts its "net" when picking tokens. For this knowledge-base use case, lower values were the most reliable because they kept the responses aligned with the retrieved context from the documents.