

In generative AI applications that use Large Language Models (LLMs) for a Retrieval-Augmented Generation (RAG) system, **temperature** and **top_p** are parameters that control the randomness and creativity of the LLM generated response based on the retrieved context. Temperature influences the probability distribution over all possible next words. A lower temperature (closer to 0) makes the model more deterministic, causing it to select the most probable next words when generating the response, which is ideal for predictable, factually accurate, and consistent responses that stick closely to the retrieved information. A higher temperature (closer to 1) “flattens” the probability distribution, giving less likely words a greater chance of being selected. This results in more diverse, creative, or even unexpected outputs, which can be useful for general purpose conversations, summarization or open-ended Q&A, but may increase the risk of introducing irrelevant information or "hallucinations."

The top_p parameter (also known as nucleus sampling) manages the diversity of the output by dynamically limiting the vocabulary from which the model can sample at each step. It sets a cumulative probability threshold (p). The model sorts all possible next words by their probability and only considers the smallest set of top words whose combined probability exceeds this threshold p. A low top_p (e.g., 0.1) heavily restricts the model's choices to only the few most probable words, leading to a focused, coherent, and conservative style. A high top_p (e.g., 0.9) allows the model to choose from a much larger pool of words, including those with lower individual probabilities, thereby increasing the output's variety and linguistic richness. In a RAG context, tuning these parameters is crucial for balancing factual grounding (low values) with conversational flow and summarization quality (moderate values). High values are not commonly used for RAG systems because they need accurate responses.