Model Parameters

Temperature controls how *deterministic or creative* a language model is. A low temperature (0.0–0.3) forces the model to choose the most likely next word, producing stable answers —ideal for knowledge-base retrieval, summarization, and technical tasks. A high temperature (0.7–1.0+) increases randomness, best for generating more creative, varied responses, which is useful for brainstorming or open-ended writing.

Top-p controls how much of the probability distribution the model is allowed to sample from. Instead of selecting from all possible tokens, lower values make the model select from wider range of words because it means it is for example selecting from words that are of probability of ~30% , Higher Values makes the Model more Precise and accurate because it means it is for example selecting from words that are of probability of ~90%