Temperature is a parameter that directly controls the randomness and creativity of an AI's output. When the temperature is set very low (e.g., 0.1), the model becomes highly deterministic, almost always selecting the single most probable word to come next. This results in responses that are very predictable, focused, and consistent, making it ideal for factual summarization or question-answering where precision is key. Conversely, a high temperature (e.g., 1.0 or higher) increases randomness by "flattening" the probabilities, giving less likely words a higher chance of being selected. This makes the AI's output more creative, diverse, and even surprising, but it also significantly increases the risk of hallucinations, nonsensical statements, or veering off-topic.

Top-p (or nucleus sampling) controls the model's creativity in a different way: it limits the size of the word pool the AI can choose from at each step. It sets a cumulative probability threshold (e.g., 0.90, or 90%), and the AI only considers the smallest possible set of top words whose probabilities add up to that threshold. For example, a top-p of 0.10 might force the model to choose between only two or three of the most obvious words, leading to very safe but often robotic and repetitive text. A high top-p (like 0.95) allows the AI to consider a much larger, more diverse vocabulary, resulting in more natural and dynamic-sounding responses. It is a common alternative to temperature for steering the model between factual and creative output.