EXPLANATION OF HOW TEMPERATURE AND TOP_P AFFECT AI RESPONSES

The temperature parameter controls the randomness or creativity of an AI's response by adjusting the probability distribution of the next word (token) it selects. When the temperature is set to a low value (near 0), the model becomes more deterministic and conservative, strongly favouring the most probable words. This results in highly predictable, focused, and often repetitive text, which is ideal for tasks requiring factual accuracy or strict formatting. Conversely, a high temperature (closer to 1 or higher) flattens the probability distribution, giving less likely words a much greater chance of being selected. This leads to diverse, creative, and sometimes unexpected or nonsensical outputs, making it suitable for brainstorming or creative writing.

The top_p parameter, also known as nucleus sampling, works by dynamically restricting the pool of words the model can choose from based on their cumulative probability. Instead of considering all possible next words, *top_p* sets a probability mass threshold (e.g., 0.9 means 90%). The model then sorts all possible words by probability and only selects from the smallest subset of the most likely words that, when combined, exceed this threshold. A low *top_p* value (e.g., 0.1) restricts the choice to only the few *most* probable words, leading to very focused and safe responses, like a low temperature. A high *top_p* value (e.g., 0.95 or 1.0) expands the set of considered words, introducing more diversity and variety in the response without risking the incoherence that can sometimes come with a very high temperature.