Explanation on how temperature and top_p affect AI responses

Temperature and top_p (nucleus sampling) are crucial parameters for controlling the output of a Large Language Model (LLM) by manipulating the probability distribution of the next token choice, thus governing the **randomness and diversity** of the response. **Temperature** acts like a creativity dial: a low value (e.g., $0.1) makes the model highly deterministic and focused by amplifying the highest probabilities, while a high value (e.g., $1.0) flattens the distribution, increasing randomness and creativity.

Top_p sets a cumulative probability threshold (e.g., $0.9) and restricts the model's choices to the smallest set of most probable tokens that meet this threshold, providing a dynamic way to filter out low-probability, irrelevant, or nonsensical options while retaining diversity.