The ***temperature*** parameter controls the overall randomness and creativity of an AI model’s responses. A higher temperature (e.g., 0.8–1.0) allows the model to take more risks, generating more diverse or imaginative answers. A lower temperature (e.g., 0.1–0.3) makes the model more deterministic and focused, producing consistent but less creative responses.

The ***top\_p*** parameter, also called nucleus sampling, defines how many possible next words the model considers based on their cumulative probability. When top\_p is low, the model chooses only from the most likely words, making responses more predictable. When top\_p is high (close to 1.0), the model considers a wider range of words, making its output more varied and expressive. Together, temperature and top\_p balance creativity and coherence in AI-generated text.