Model Parameters:

Explanation of Model Parameters: Temperature and Top_p:

The parameters **Temperature** and **Top_p** are crucial settings that directly influence the creativity, randomness, and focus of an AI model's responses.

Temperature controls the randomness of the output. It reflects the model's confidence in its word choices:

- **Low Temperature** (e.g., 0.1–0.5): Produces more deterministic and focused responses. The model consistently selects the most probable next word, resulting in predictable and safe output. This is ideal for tasks requiring factual, precise, or repetitive answers.
- **High Temperature** (e.g., 0.8–1.0): Encourages the model to consider less probable words, leading to more creative, diverse, and potentially unexpected responses. This is useful for creative writing, brainstorming, or generating varied options.

Top_p (nucleus sampling) controls the diversity of the output by limiting the selection to the most likely words. Instead of looking at *all* possible words, it selects a group of words whose cumulative probability adds up to the value of Top_p.

- **Low Top_p** (e.g., 0.1–0.5): The model samples from a small set of the most probable words. The output is more predictable and less imaginative, similar to a low temperature setting.
- **High Top_p** (e.g., 0.8–1.0): The model samples from a larger set of highly probable words, increasing the variety and creativity of the response. It helps balance quality and diversity of generated text.

In practice, **Temperature** and **Top_p** are often used together to fine-tune the output. For many tasks, a moderate **Temperature** (around 0.5–0.7) and a moderate to high **Top_p** (around 0.8–0.9) provide a good balance between coherence and creativity.