Temperature:

Definition: Temperature controls the randomness of the model's predictions.

Effect:

- **High temperature** (e.g., 0.8–1.0) makes the model more creative and diverse in its responses. It is more likely to choose less probable words, which can lead to unexpected or innovative answers.

- **Low temperature** (e.g., 0.2–0.5) makes the model more focused and deterministic, producing precise and predictable responses.

Use case: High temperature is useful for brainstorming or creative writing, while low temperature is preferred for factual answers or instructions.

Top_p:

Definition: Top_p controls the probability distribution of the next word by selecting from the smallest set of words whose cumulative probability exceeds p.

Effect:

- A **high top_p** value allows the model to choose from a wider range of words, increasing variability.

- A **low top_p** value restricts the model to the most likely words, increasing coherence and accuracy.

Use case: Top_p allows dynamic control over randomness and creativity, and can be used together with temperature to fine-tune responses.