

Important Parameters

1. Temperature

Temperature controls the randomness or creativity of text generation.

It determines how likely the model is to choose less-probable tokens.

The underlying idea is that a **temperature of 0** generates the same response every time because the model always selects the most likely words.

As a result, a **higher temperature (0.8–0.9)** produces more diverse and creative responses, while a **lower temperature** generates more deterministic and focused output.

2. num_predict

This parameter defines the maximum number of tokens the model can generate.

A **higher num_predict** value gives a longer response, while a **lower value** results in a shorter answer.

3. Top_p (Nucleus Sampling)

If **top_p is set to 1**, the model considers **all possible tokens**.

Top_p selects the smallest set of tokens whose cumulative probability reaches the specified threshold.

- **Low top_p** → more focused and restricted output.
 - **High top_p** → wider vocabulary and more creative output.
-

Example Use Cases

Example Use Case	Temperature	Top_p	Description
Email Generation	Low	Low	Produces deterministic output using highly probable tokens. Results are predictable, focused, and conservative.
Brainstorming Session	High	High	High randomness with a large pool of potential tokens. Outputs become highly diverse, creative, and often unexpected.
Creative Writing	High	Low	Generates creative content with high randomness but from a smaller token set, helping maintain coherence.
Translation	Low	High	Produces deterministic output with high-probability tokens while still using a wide vocabulary range, resulting in clear and accurate translations.