**1. Messages**

The messages parameter is a list that provides the context of the conversation to the model. It typically consists of dictionaries with roles (user, assistant, or system) and their corresponding content. For example:

[

{"role": "system", "content": "You are a helpful assistant."},

{"role": "user", "content": "What is the capital of France?"}

]

This structure allows the model to maintain the context and generate coherent responses.

**2. Model**

The model parameter specifies which model to use for generating responses. For example:

* "gpt-4": A highly advanced model for complex tasks.
* "gpt-3.5-turbo": A cost-effective model for general tasks. Choosing the right model depends on factors like required accuracy, speed, and cost.

**3. Max Completion Tokens**

This parameter controls the maximum number of tokens (words, punctuation, or parts of a word) the model can generate in a response. For example:

* Setting it to 50 ensures the response is concise.
* A higher value allows longer responses. It helps manage token limits to avoid exceeding the API quota.

**4. n**

The n parameter specifies how many response completions you want the API to generate. For example, setting n: 3 will provide three different responses to the same input. It is useful for comparing alternative responses and selecting the most suitable one.

**5. Stream**

The stream parameter, when set to true, enables token-by-token streaming of the response. Instead of waiting for the entire response to be generated, the output is delivered incrementally. This is particularly useful for real-time applications, such as chat interfaces.

**6. Temperature**

The temperature parameter controls the randomness of the output:

* A **higher value** (e.g., 1.0) results in more creative and diverse responses.
* A **lower value** (e.g., 0.2) makes the output more deterministic and focused. It’s a way to adjust the balance between creativity and predictability.

**7. Top\_p**

The top\_p parameter, also known as nucleus sampling, adjusts the diversity of the response. Instead of choosing tokens based on the highest probability (as with temperature), it considers a subset of tokens whose cumulative probability is within the top\_p value. For example:

* top\_p: 0.1 restricts the response to only the top 10% of token probabilities.
* top\_p: 1.0 includes all tokens, similar to temperature sampling.

**8. Tools**

The tools parameter defines external capabilities the model can use, such as APIs, calculators, or databases. When tools are integrated, the model can invoke them to perform tasks it cannot do on its own, such as fetching real-time data or running computations.