Starting April 29, 2025, Gemini 1.5 Pro and Gemini 1.5 Flash models are not available in projects that have no prior usage of these models, including new projects. For details, see Model versions and lifecycle (/vertex-ai/generative-ai/docs/learn/model-versions#legacy-stable).

Experiment with parameter values

Each call that you send to a model includes parameter values that control how the model generates a response. The model can generate different results for different parameter values. Experiment with different parameter values to get the best values for the task. The parameters available for different models may differ. The most common parameters are the following:

- Max output tokens
- Temperature
- Top-P
- Seed

Max output tokens

Maximum number of tokens that can be generated in the response. A token is approximately four characters. 100 tokens correspond to roughly 60-80 words.

Specify a lower value for shorter responses and a higher value for potentially longer responses.

Temperature

The temperature is used for sampling during response generation, which occurs when `topP` and `topK` are applied. Temperature controls the degree of randomness in token selection. Lower temperatures are good for prompts that require a less open-ended or creative response, while higher temperatures can lead to more diverse or creative results. A temperature of `0` means that the highest probability tokens are always selected. In this case, responses for a given prompt are mostly deterministic, but a small amount of variation is still possible.

If the model returns a response that's too generic, too short, or the model gives a fallback response, try increasing the temperature.

Each model has its own temperature range and default value:

- Range for `gemini-2.0-flash-lite`: `0.0 - 2.0` (default: `1.0`)
- Range for `gemini-2.0-flash`: `0.0 - 2.0` (default: `1.0`)

Top-P

Top-P changes how the model selects tokens for output. Tokens are selected from the most (see top-K) to least probable until the sum of their probabilities equals the top-P value. For example, if tokens A, B, and C have a probability of 0.3, 0.2, and 0.1 and the top-P value is `0.5`, then the model will select either A or B as the next token by using temperature and excludes C as a candidate.

Specify a lower value for less random responses and a higher value for more random responses.

Seed

When seed is fixed to a specific value, the model makes a best effort to provide the same response for repeated requests. Deterministic output isn't guaranteed. Also, changing the model or parameter settings, such as the temperature, can cause variations in the response even when you use the same seed value. By default, a random seed value is used.

This is a preview feature.

What's next

- Explore examples of prompts in the Prompt gallery (/vertex-ai/generative-ai/docs/prompt-gallery).
- Learn how to optimize prompts for use with Google models (/vertex-ai/generative-ai/docs/learn/models) by using the Vertex AI prompt optimizer (Preview) (/vertex-ai/generative-ai/docs/learn/prompts/prompt-optimizer).

Last updated 2025-06-06 UTC.