

Sampling Controls

Temperature, top-K, and top-P are the most common configuration settings that determine how predicted token probabilities are processed to choose a single output token.

Temperature

Temperature controls the degree of randomness in token selection.

A temperature of 0 (greedy decoding) is deterministic: the highest probability token is always selected.

Temperatures close to the max tend to create more random output.

Top-K and top-P

Top-K and top-P (also known as nucleus sampling) are two sampling settings used in LLMs to restrict the predicted next token to come from tokens with the top predicted probabilities.

- **Top-K** sampling selects the top K most likely tokens from the model's predicted distribution. For instance, if Top K is set to 50, the model will only consider the top 50 words based on their probabilities.
- **Top-P** sampling selects the top tokens whose cumulative probability does not exceed a certain value (P). For example, if Top P is set to 0.9, the model considers the smallest set of words whose probabilities sum to 0.9.