

Temperature & Sampling

Computing

The parameters that control randomness, and what to set them to.

Typical ranges. Defaults are sensible for most work.

Parameter	Range	Effect
Temperature	0–2	Flattens or sharpens the probabilities
Top-p (nucleus)	0–1	Samples from the top mass only
Top-k	1–100	Samples from the k likeliest tokens
Frequency penalty	0–2	Discourages repeating tokens
Presence penalty	0–2	Discourages repeating topics
Max tokens	varies	Caps the output length

Low temperature for extraction, classification and code. Higher for brainstorming and drafting. Zero is not quite deterministic.