

# LLM Decoding Cheat Sheet

Temperature · Top-k · Top-p

| Parameter              | Typical Range | What ↑ Value Does  | Common Defaults |
|------------------------|---------------|--|-----------------|
| Temperature ( $\tau$ ) | 0.0 – 2.0     | Flattens probability distribution → more randomness and diversity      | 0.7 – 1.0       |
| Top-k                  | 1 – 1000+     | Samples only from the $k$ highest probability tokens                   | 40              |
| Top-p (Nucleus)        | 0.0 – 1.0     | Picks the smallest set of tokens whose cumulative probability $\geq p$ | 0.9             |

## Quick Rules of Thumb

- Low  $\tau$  + Low  $p/k$  → maximally deterministic, but can become repetitive.
- Moderate  $\tau$  ( $\approx 0.7$ ) +  $p \approx 0.9$  → good *default* for balanced creativity.
- High  $\tau$  ( $> 1.2$ ) + High  $p/k$  → wild, story-like, risk of hallucination.

## Tuning Recipes

| Goal                           | Suggested Settings  |
|--------------------------------|---|
| Deterministic reference answer | <code>`temperature=0`, `top_p=1`, `top_k=0` (greedy)</code>                                       |
| Balanced/default web app       | <code>`temperature=0.7`, `top_p=0.9`</code>   |
| Maximum creativity / poetry    | <code>`temperature=1.2`, `top_p=0.95`</code>  |
| Avoid repetition loops         | add <code>`top_k` <math>\leq 100</math></code> and <code>`temperature` <math>&lt; 1</math></code> |
| Safety-critical / policy text  | <code>`temperature=0.2–0.4`, `top_p=0.6–0.8`</code>   |

## Why Adjust Each Parameter?

- Temperature** rescales logits globally – good first knob for exploring diversity.
- Top-k** clips improbable tokens – protects against off-topic or unsafe words.
- Top-p** adapts to context entropy – keeps distribution mass constant regardless of vocabulary size.

## Interaction Gotchas

- Very low **top\_p** and low **temperature** can starve the model (empty output).

- **top\_k=0** disables the filter (equivalent to unlimited  $k$ ).
- Many APIs ignore `top_k` when `top_p` is set—check docs!

\_Keep this sheet handy when debugging prompt output diversity, hallucination rate, or repetitiveness.\_