Controlling AI Creativity

A Simple Guide to Temperature, Top_k, and Top_p

Iverina Ivanova & Gemini 3 Pro

November 24, 2025

How AI Predicts Text

Imagine the AI has to finish this sentence:

"The cat sat on the ___"

The AI assigns a probability score to every word in its dictionary:

- **Mat:** 50% (Very likely)
- **Rug:** 30% (Likely)
- **Hat:** 10% (Unlikely)
- **Pizza:** 0.001% (Nonsense)

Our parameters determine **how risky** the AI is allowed to be when picking one of these words.

Temperature: The "Risk" Knob

Concept: "The temperature controls the randomness or creativity of the text generated. It defines how likely it is to choose tokens that are less probable." (Alammar & Grootendorst 2024, p.274).

Low Temperature (Near 0)

The Conservative Expert

- The AI almost always picks the #1 most likely word.
- **Result:** Predictable, repetitive, safe, robot-like.

High Temperature (Above 1)

The Wild Dreamer

- The AI gives less probable words a better chance of being picked.
- **Result:** Creative, unpredictable, sometimes makes mistakes.

Top_k: The Fixed Shortlist

Concept: ”The top_k parameter controls exactly how many tokens the LLM can consider.” (Alammar & Grootendorst 2024, p.276). It forces the LLM to ignore bad options by creating a fixed-size shortlist.

How it works:

- If we set **Top_k = 3**, the AI considers *only* the top 3 words.
- It does not consider word #4, #5, and #10,000 from existence.

Analogy: The Music Chart

Imagine a radio station that only plays the ”Top 40” hits. It doesn’t matter if song #41 is actually quite good; the station simply refuses to play it. This prevents the AI from going totally ”off the rails.”

3. Top_p (Nucleus Sampling): The Smart Cutoff

Concept: Instead of picking a fixed *number* of words, we pick a fixed *amount of likelihood*.
"top_p controls the subset of tokens (nucleus) that the LLM can consider." (Alammar & Grootendorst 2024, p.275).

3. Top_p: The Shopping Cart Analogy

How it works (e.g., Top_p = 0.90):

1. We look at the words sorted by likelihood.
2. We put the 1 word in our cart.
3. Is the cart 90% full?
4. If **No**: Add the next word. Repeat.
5. If **Yes**: Stop immediately.

Scenario A: Clear Answer

"The capital of France is..."

- *Paris* is a huge slice (99%).
- **Result:** We take just 1 word and stop.

Scenario B: Open Question

"My favorite color is..."

- *Blue* (10%), *Red* (9%), *Green* (8%)...
- **Result:** We need to take, let's say, 15 different words to get to 90%.

Top_p adapts to the context automatically!

Summary Analogy: The Road Trip

Imagine the AI is a driver trying to navigate a journey, one turn at a time.

Parameter	Driving Strategy
Temperature	**The Adventure Mode.** Do you want to strictly follow the highway (Low Temp), or are you willing to take a risky scenic backroad (High Temp)?
Top_k	**The Fixed Menu.** Only show me the Top 3 routes. Even if the 4th route is perfectly fine, the GPS hides it.
Top_p	**The Smart Filter.** Show me every road that heads in the right direction. If there is only 1 road, show 1. If there are 10 good roads, show 10.

Use Case 1: Coding & Math

Objective: Precision. There is usually only one correct syntax or answer.

Recommended Settings

- **Temperature: 0.0 - 0.2**
  Reason: Prevents the model from improvising syntax.
- **Top_p: 0.1 - 0.3 (Very Low)**
  Reason: Forces the "Smart Cutoff" to be extremely strict. We only want the absolute best tokens.

Use Case 2: Creative Writing

Objective: Novelty. We want colorful vocabulary and unexpected turns.

Recommended Settings

- **Temperature: 0.8 - 1.1**
  Reason: Flattens the curve so "rare" words get chosen more often.
- **Top_p: 0.95 - 0.99 (High)**
  Reason: We want the "Smart Cutoff" to be very loose. Keep almost all possibilities unless they are total nonsense.

Use Case 2: Brainstorming Ideas

Objective: Divergence. We want to avoid the average path.

If you ask for *blog post ideas*, you don't want the first thing that comes to everyone's mind. You want the 50th idea.

Recommended Settings

- **Temperature: 0.9 - 1.0**
  Reason: We need high "energy" to escape common patterns.
- **Top_p: 1.0 (or very high)**
  Reason: We want to widen the net. By setting Top_p to 1.0, we allow even statistically unlikely ideas to surface if the temperature boosts them enough.

Use Case 3: Language Translation

Objective: Accuracy with Nuance.

Unlike coding, languages aren't perfectly 1-to-1. Idioms require slight flexibility, but we cannot alter the original meaning.

Recommended Settings

- **Temperature: 0.3**
  Reason: Low enough to be accurate, but not 0.0. A tiny bit of heat helps with smooth phrasing.
- **Top_p: 0.80 - 0.85**
  Reason: We want a "Smart Cutoff" that excludes bad translations, but keeps synonyms available so the output flows naturally.

- Reference:
  Alammar, Jay & Marteen Grootendorst. 2024. *Hands-On Large Language Models: Language Understanding and Generation*. O'Reilly Media, Inc.
- *These slides were generated involving Gemini 3 Pro.*