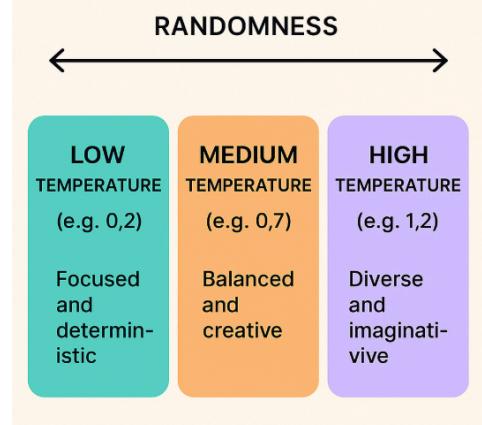


Temperature and Top-P

1. Temperature

Temperature controls how "adventurous" a generative model is when picking the next word. At low temperatures, the model behaves conservatively, sticking to the safest, most predictable tokens, much as if someone gave an answer that was precise, factual, and left no room for imagination. As temperature increases, the model starts to get more exploratory: probabilities flatten, rare words get a bigger chance, and the tone becomes more creative or unexpected. Technically, temperature modifies the softmax distribution by sharpening it (low T) or flattening it (high T), which is why the outputs shift from strict and logical to expressive and free-flowing. In other words, temperature decides how boldly the model shall think.

Temperature in AI



Intuition

- Controls **how random or predictable** the model's output is.
- Low temperature → **safe, deterministic**
- High temperature → **creative, diverse**

How it behaves

- **T < 0.4:** Very focused, factual
- **T 0.4–0.7:** Balanced, natural
- **T > 0.8:** Creative, risky

Variance of Temperature

Temperature **reshapes** the probability distribution by making:

- High-probability words *even more likely* (low T)
- Low-probability words *more available* (high T)

Examples

Prompt: “A cat is sitting on the ____.”

- **T = 0.2** → “mat”
- **T = 0.7** → “window sill”
- **T = 1.2** → “intergalactic hoverboard”

In my project

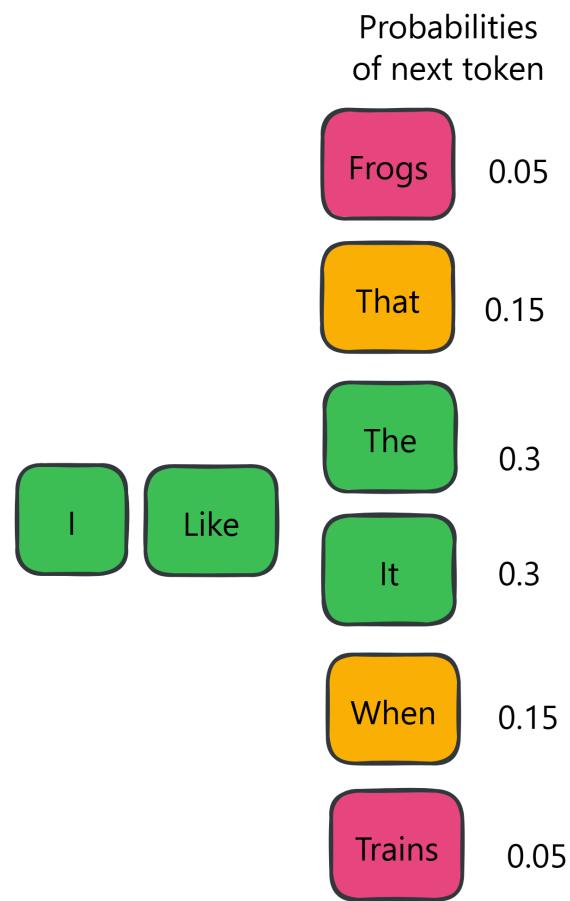
I tested multiple temperature values and found that higher settings made the model more creative but also more prone to hallucinations. So I chose a

low temperature

to keep the model factual, consistent, and grounded in reliable information.

2. Top-P (Nucleus Sampling)

Top-p, or nucleus sampling, controls which set of words the model is allowed to choose from in the first place. Instead of modifying every probability like temperature does, top-p selects the smallest group of tokens whose combined probability reaches p-for example, 0.9. The model then samples only from that nucleus and ignores everything else. This means a low top-p forces the model to stay within a tight, high-probability group of words, resulting in more focused and reliable responses. A higher top-p gives the model a wider pool, allowing more variety, detail, and spontaneity while still avoiding completely random nonsense. Taken together, temperature shapes the behavior inside the allowed pool, while top-p determines how big that pool is.



Intuition

- Controls **how many possible words** the model can choose from.
- Instead of reshaping probabilities, it **cuts off the tail**.

How it behaves

- **Top-P = 0.1:** Very strict → only the top 10% probability tokens
- **Top-P = 0.5:** Focused but flexible
- **Top-P = 0.9:** Creative, varied
- **Top-P = 1.0:** All tokens allowed

Simple explanation

Top-P selects the **smallest set of tokens whose cumulative probability $\geq p$** .

Everything else is removed.

Example token probabilities

Token	Prob
mat	0.45
floor	0.30
roof	0.15
space	0.05
galaxy	0.03
taco	0.02

If Top-P = 0.5:

→ Allowed tokens: **mat + floor**

If Top-P = 0.9:

→ Allowed tokens: **mat + floor + roof + space**

In my project:

A **0.5–1.0 top-p range** worked best because it provides a good balance, lower values keep the model focused on the most likely tokens, while higher values allow natural, human-like phrasing without adding too much randomness.

Difference Between Temperature & Top-P

- **Temperature changes the shape** of the probability distribution
- **Top-P changes the size** of the token pool the model can choose from
- Temperature = *how random each choice is*
- Top-P = *how many choices are allowed*
- You can use both together: temp controls *behavior*, top-p controls *range*