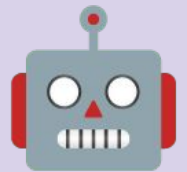




A Beginner's Guide for Aspiring Data Scientists

What are the -

Parameters in LLMs



Sanjay N Kumar

Data scientist | AI ML Engineer | Statistician | Analytics Consultant

What Are These “Parameters”? 🧠



1 2 3 4 Parameters are settings that **control how an LLM talks.**

They tell the model:

- How long to speak 🕒
- How creative to be 🎨
- When to stop 🛑
- How to avoid repeating 🔁

Think of them as **dials on a music player** 🎚️ — you turn them up or down to get the sound (output) you like!

Max Tokens



What it does: Limits how much the model talks.

 *Tokens* = words or pieces of words

Example:

"Elephant" = 1 token

"Running fast" = 2–3 tokens

 **Real-life Example:**

Telling a friend: "Explain in 20 words only."

 Helps save time and cost!

Temperature 🎨



What it does: Controls **creativity** and randomness.

🧊 0 = very focused, serious

🔥 1 = wild, fun, random

📚 Real-Life:

- **Low Temp:** “Explain gravity.”
→ “Gravity pulls things to Earth.”
- **High Temp:**
→ “Gravity is Earth giving a hug to everything!”

😄😞 It’s like how silly or serious a storyteller becomes.

Top P (Nucleus Sampling) 🍭



What it does: Picks words from the **most likely group**

✂️ Cut off after total chance = P (like 0.9 or 90%)

📦 **Real-Life:**

Choosing snacks 🍪

You pick from top 90% of your favorite snacks.
Not all snacks—just your best ones!

💡 Helps make output focused but not boring.

Top K 🏆



What it does: Picks from the **top K** choices only.

🎯 Real-Life Example:

You ask a friend:

“Name any 1 of your **top 5** favorite cartoons.”

They pick from that top 5 list 🎞️

🕒 You can control how wide or narrow the choices are.

Frequency Penalty



What it does: Stops the model from **repeating** itself too much.

 **Real-life:**

Someone keeps saying,

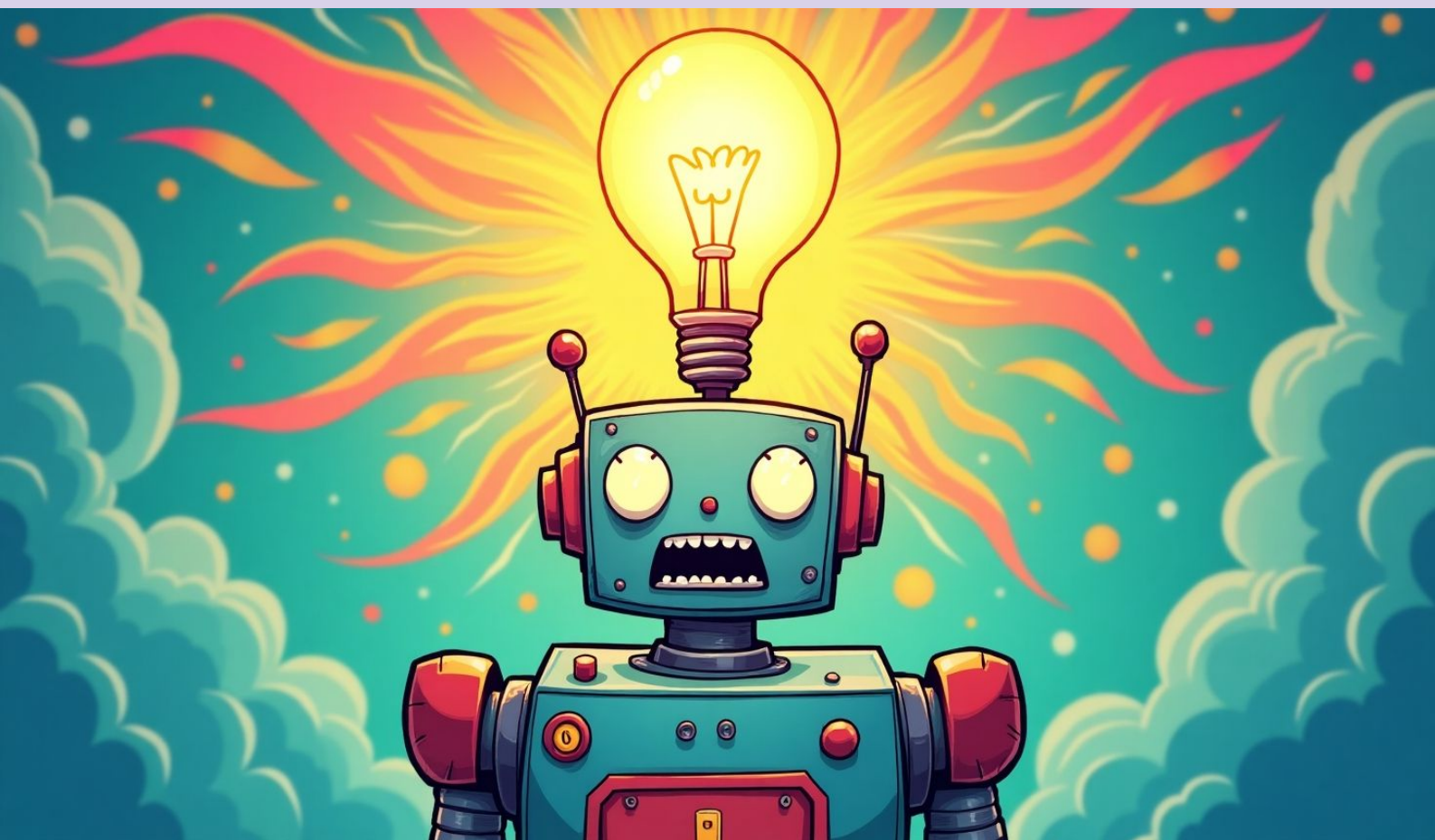
“I love pizza. Pizza is good. Pizza is yummy.”



Frequency penalty says: “Stop repeating!”

Helps the model **sound smarter** and **less annoying** 😊

Presence Penalty NEW



What it does: Encourages the model to **use new ideas** 🌱

🎨 **Real-life:**

Imagine a friend always says the same jokes.
You tell them: “Try something new!”

📈 Presence penalty says: ***“Don’t use what you already said. Be fresh!”***

Stop Sequence



What it does: Tells the model: “Stop talking after this.”



Real-life:

You say: “Tell me a story. Stop when you say
The End.”



Once the model sees that stop token, it
shuts up politely 😊


Summary Table

Parameter	Real-Life Example	What It Controls
Max Tokens	Word limit in an essay 📝	Length of output
Temperature	Serious vs silly friend 🗣️	Creativity
Top P	Top 90% of fav snacks 🍪	Word diversity
Top K	Choose top 5 cartoons 📺	Top options only
Frequency Penalty	Avoid saying "pizza" 10x 🍕	No repetition
Presence Penalty	Try new ideas 🌱	Freshness
Stop	Say "The End" 🛑	End of response

Final Thought



 These are **dials** on your LLM machine.

 Tune them right, and the model gives better answers!

 It's not just about asking questions—

 It's about how you **guide** the brain.

Unlock the Power of Parameters

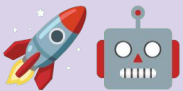
Transform plain prompts into powerful conversations.

Let temperature, top-p, and friends guide your LLM with clarity, creativity, and control.

Dial it right. Prompt it smart.

Your AI journey starts with the right parameters.

Reach out—let's tune the future together!



Sanjay N Kumar

Data scientist | AI ML Engineer | Statistician | Analytics Consultant



<https://www.linkedin.com/in/sanjaytheanalyst360/>



sanjaytheanalyst360@gmail.com