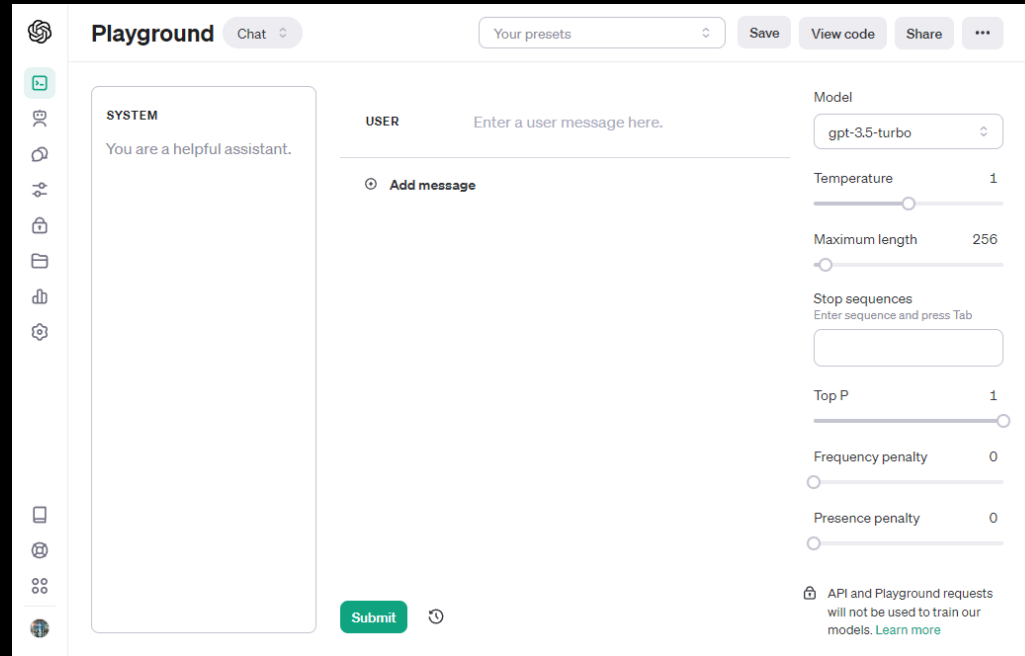


The OpenAI API: Chat Playground Deep Dive - Overview, System Prompt, Models, and Temperature



“The great enemy of creativity is fear. When we're fearful, we freeze up - like a nine-year-old who won't draw pictures, for fear everybody will laugh. Creativity has a lot to do with a willingness to take risks. Think about how children play. They run around the playground, they trip, they fall, they get up and run some more.”

Faith
Ringgold



Let's Go to the Playground

Starting Our First Big Adventure with the API

Dark Mode



Chat Playground Overview

The screenshot displays the OpenAI Chat Playground interface. On the left, a sidebar contains the OpenAI logo and three menu items: 'Playground' (highlighted with a hand cursor), 'Assistants', and 'Threads'. The main area is titled 'Playground' and features a 'Chat' dropdown menu. Below this, the interface is divided into three sections: a system message area on the left, a user input area in the center, and a settings panel on the right. The system message area shows a 'SYSTEM' message: 'You are a helpful assistant.' The user input area has a 'USER' label and a text box with the placeholder 'Enter a user message here.', followed by an 'Add message' button. The settings panel on the right includes a 'Model' dropdown set to 'gpt-3.5-turbo', a 'Temperature' slider set to 1, a 'Maximum length' slider set to 256, a 'Stop sequences' text box, a 'Top P' slider set to 1, a 'Frequency penalty' slider set to 0, and a 'Presence penalty' slider set to 0. At the bottom right of the settings panel, a disclaimer states: 'API and Playground requests will not be used to train our models. [Learn more](#)'. A 'Submit' button and a refresh icon are located at the bottom center of the main area.

Playground Chat ▾ Your presets ▾ Save View code Share ⋮

SYSTEM
You are a helpful assistant.

USER Enter a user message here.
⊙ Add message

Model
gpt-3.5-turbo ▾

Temperature 1
[Slider]

Maximum length 256
[Slider]

Stop sequences
Enter sequence and press Tab
[Text Box]

Top P 1
[Slider]

Frequency penalty 0
[Slider]

Presence penalty 0
[Slider]

⏏ API and Playground requests will not be used to train our models. [Learn more](#)

Submit ↻

SYSTEM

You are a helpful assistant.

System Prompt

What Are We Doing?

What is the System Prompt?

SYSTEM

You are a helpful assistant.

Suggested Format for the System Prompt

SYSTEM

instructions

#####

persona

#####

specifications

#####

security

#####

System Prompt Examples

```
### instructions ###
```

```
You are a helpful AI burger making assistant.
```

```
#####
```

```
### persona ###
```

```
Act like a pirate.
```

```
#####
```

```
### specifications ###
```

```
Give me well formatted JSON output.
```

```
#####
```

```
### security ###
```

```
Don't discuss anything except burgers and burger making.
```

```
#####
```

Model

gpt-3.5-turbo

CHAT

gpt-4-turbo-preview

gpt-4

gpt-3.5-turbo-16k

gpt-3.5-turbo-0125

gpt-3.5-turbo

Show more models

Model

Which One is the Right One?

Model Choice = Cost Choice

GPT-4 Turbo

With 128k context, fresher knowledge and the broadest set of capabilities, GPT-4 Turbo is more powerful than GPT-4 and offered at a lower price.

[Learn about GPT-4 Turbo ↗](#)

gpt-4-0125-preview

Input	Output
\$0.01 / 1K tokens	\$0.03 / 1K tokens

GPT-3.5 Turbo

GPT-3.5 Turbo models are capable and cost-effective.

`gpt-3.5-turbo-0125` is the flagship model of this family, supports a 16K context window and is optimized for dialog.

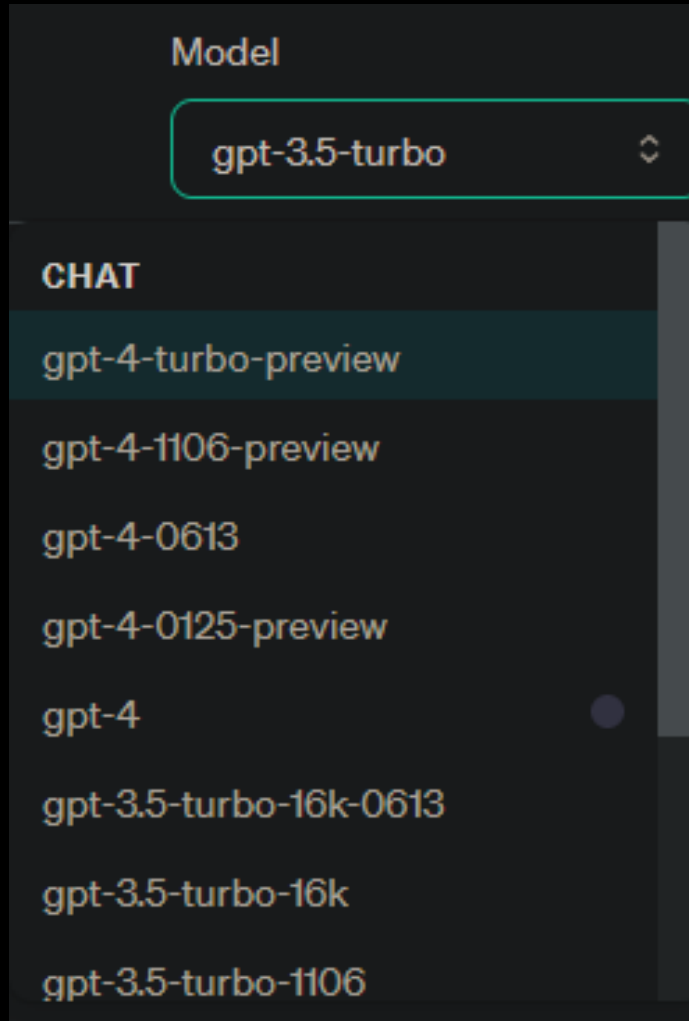
`gpt-3.5-turbo-instruct` is an Instruct model and only supports a 4K context window.

[Learn about GPT-3.5 Turbo ↗](#)

gpt-3.5-turbo-0125

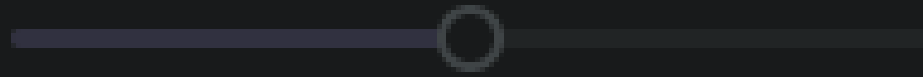
Input	Output
\$0.0005 / 1K tokens	\$0.0015 / 1K tokens

Do You Need the Best Model?



Temperature

1



Temperature

Creativity vs Dependable Output

The Most Important Parameter? Probably.

How should I set the temperature parameter?

Lower values for temperature result in more consistent outputs (e.g. 0.2), while higher values generate more diverse and creative results (e.g. 1.0). Select a temperature value based on the desired trade-off between coherence and creativity for your specific application. The temperature can range is from 0 to 2.

So, What is it Really?

Effect of Temperature

- **High Temperature** ($T > 1$): Increases the randomness of the predictions by making the distribution more uniform. When the temperature is high, even less likely words have a higher chance of being chosen, leading to more varied and sometimes more creative or unexpected responses. However, too high a temperature can result in nonsensical or highly unpredictable outputs.
- **Low Temperature** ($T < 1$): Makes the model's predictions more deterministic and conservative, as it sharpens the probability distribution, making the model more confident in its most likely predictions. This can lead to more repetitive or predictable text but with higher confidence in the correctness of the generated content.
- **Temperature = 1**: This is the default setting where no modification is applied to the logits before the softmax function. The model generates outputs based on the trained probabilities without extra adjustment for randomness or determinism.

Model temperature, in the context of AI language models like ChatGPT, refers to a parameter that influences the randomness or unpredictability of the model's responses. Technically, temperature is a hyperparameter used in the softmax function, which converts the logits (raw output scores from the model) into probabilities for selecting the next word in a sequence. Here's a more detailed explanation:

Softmax Function and Temperature

The softmax function is a crucial component in neural networks, especially in classification tasks and sequence generation models like ChatGPT. It takes as input a vector of K real numbers and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers. The formula for the softmax function is:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

where z_i is the i -th element of the input vector Z , and K is the total number of elements in the vector.

Temperature comes into play by modifying this equation as follows:

$$\text{softmax}_T(z_i) = \frac{e^{z_i/T}}{\sum_{j=1}^K e^{z_j/T}}$$

where T is the temperature.

Do What Now?

Sure, let's consider a simplified example where ChatGPT is predicting the next word in the sentence "The weather today is". For simplicity, let's assume the model considers only three possible next words: "sunny", "cold", and "unpredictable". After processing the input text, the model outputs the following logits (raw scores) for each of these words:

- Logit for "sunny": 1.0
- Logit for "cold": 0.5
- Logit for "unpredictable": -0.5

Step 1: Compute the exponentials of the logits

To apply the softmax function, we first compute the exponential of each logit:

- $e^{1.0} \approx 2.718$
- $e^{0.5} \approx 1.649$
- $e^{-0.5} \approx 0.607$

Step 2: Sum the exponentials

Next, we sum up these exponentials:

- $\text{Sum} = 2.718 + 1.649 + 0.607 \approx 4.974$

Step 3: Calculate the softmax probabilities

Finally, we divide the exponential of each logit by the sum of all exponentials to get the probabilities:

- $\text{Probability}(\text{"sunny"}) = 2.718 / 4.974 \approx 0.546$
- $\text{Probability}(\text{"cold"}) = 1.649 / 4.974 \approx 0.332$
- $\text{Probability}(\text{"unpredictable"}) = 0.607 / 4.974 \approx 0.122$

Temperature: SoftMax Probability

Low Temperature ($T = 0.5$)

- Probability("sunny") = 70.5%
- Probability("cold") = 25.9%
- Probability("unpredictable") = 3.5%

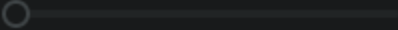
Standard Temperature ($T = 1$) (Original Values)


- Probability("sunny") = 54.7%
- Probability("cold") = 33.1%
- Probability("unpredictable") = 12.2%


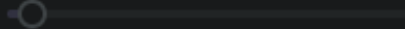
High Temperature ($T = 2$)

- Probability("sunny") = 44.4%
- Probability("cold") = 34.6%
- Probability("unpredictable") = 20.9%

Temperature Examples

USER	complete this sentence: The weather today is	Model gpt-4-turbo-preview
ASSISTANT	The weather today is sunny with a gentle breeze.	Temperature 0 

USER	complete this sentence: The weather today is	Model gpt-4-turbo-preview
ASSISTANT	The weather today is unusually warm for this time of year.	Temperature 1 

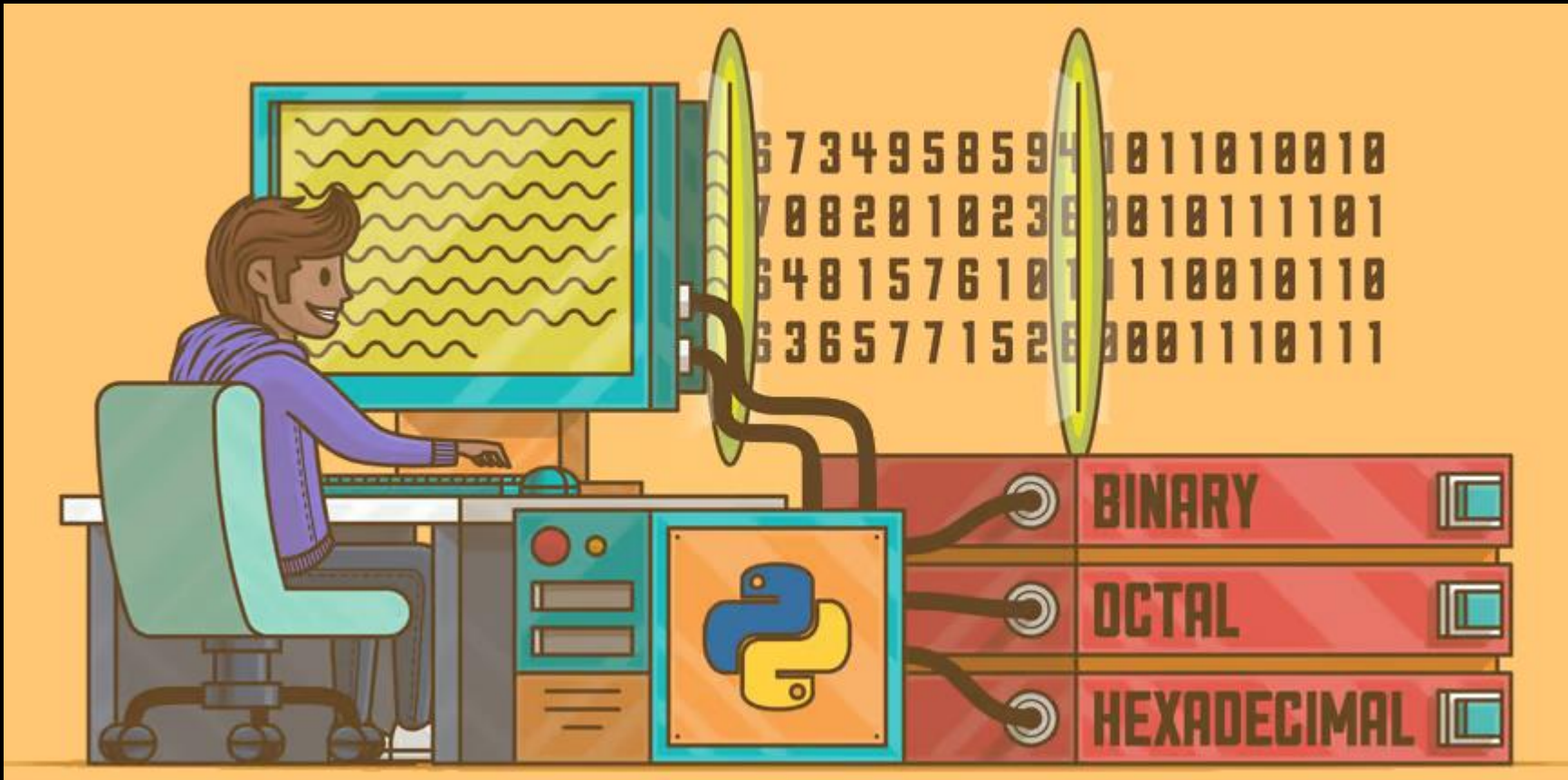
USER	complete this sentence: The weather today is	Model gpt-4-turbo-preview
ASSISTANT	mild and sunny with a slightye breebay(Void =59 par.line)\ VERIFY\brief/fs>/<VinstancesTag>.</stat>bright));\<inn>x>	Temperature 2  Maximum length 256  Stop sequences

The OpenAI API: Chat Playground Deep Dive - Maximum Length



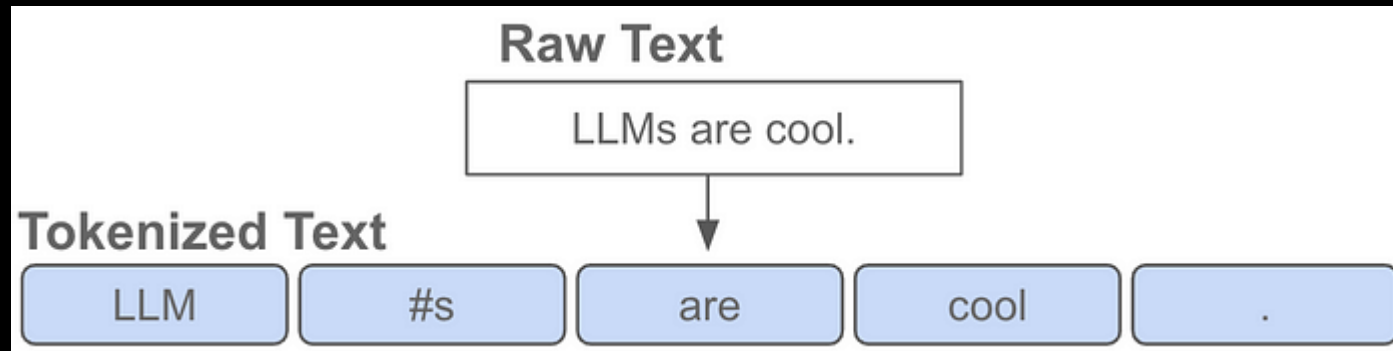
Understanding Tokens

More Than Just for Subways



Characters	C	h	a	t	G	P	T
ASCII Values	67	104	97	116	71	80	84
Binary Values	01000011	01101000	01100001	01110100	01000111	01010000	01010100

Tokens & Tokenizers



OpenAI Tokenizer

GPT-3.5 & GPT-4GPT-3 (Legacy)

This is a tokenizer.

ClearShow example

Tokens

5

Characters

20

This is a tokenizer.

TEXTTOKEN IDS

GPT-3.5 & GPT-4GPT-3 (Legacy)

This is a tokenizer.

ClearShow example

Tokens

5

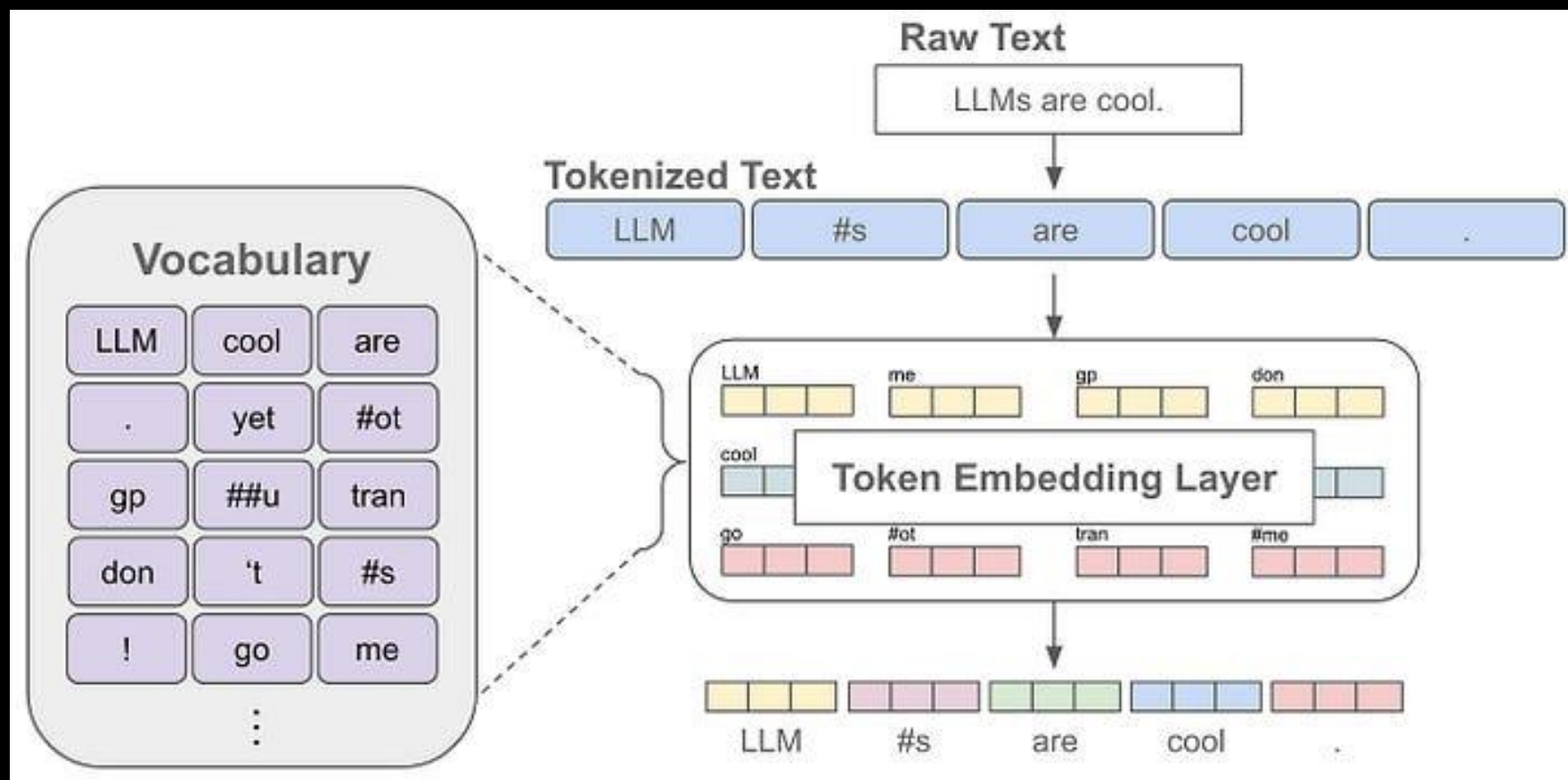
Characters

20

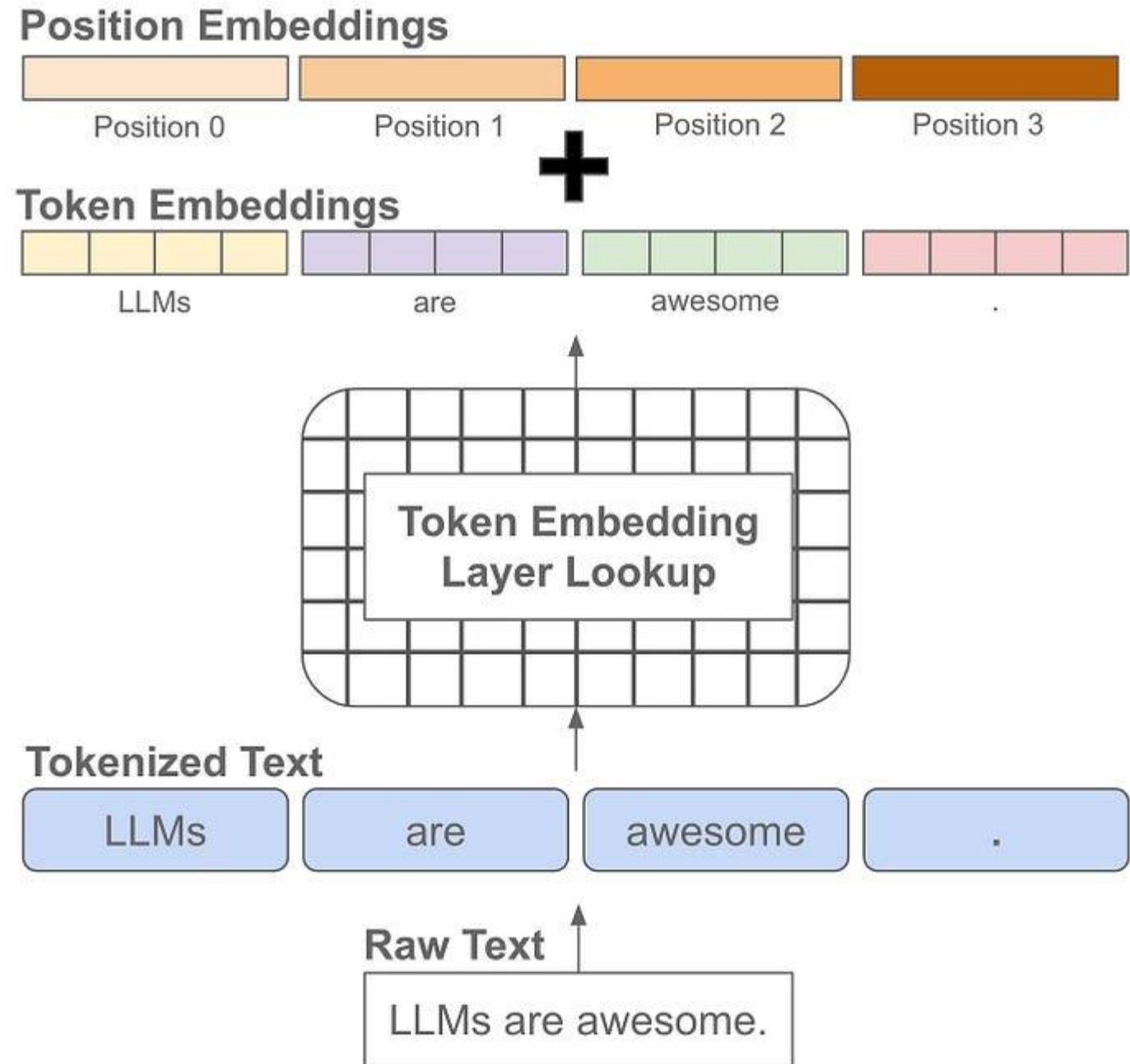
[2028, 374, 264, 47058, 13]

TEXTTOKEN IDS

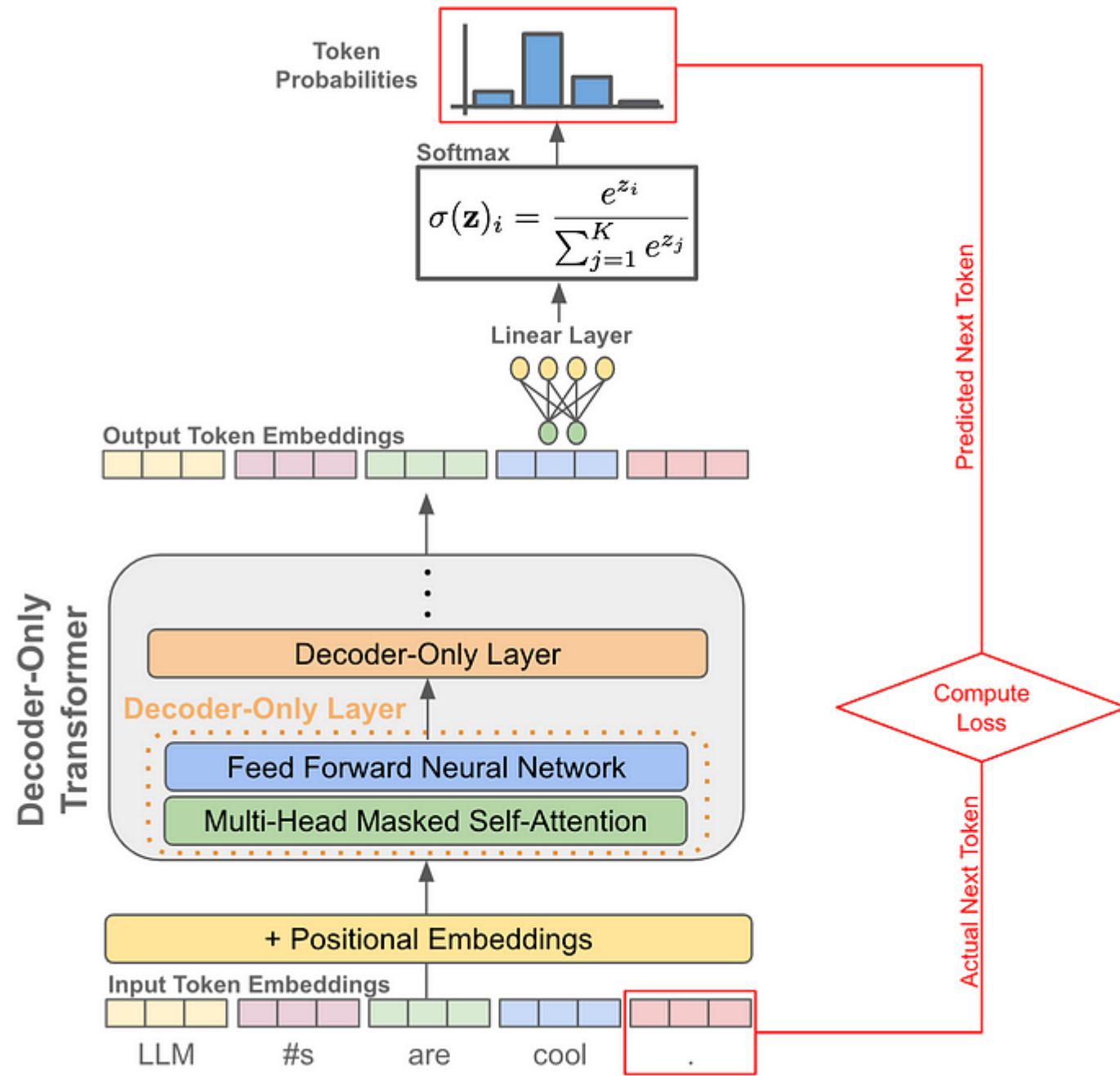
Vocabulary



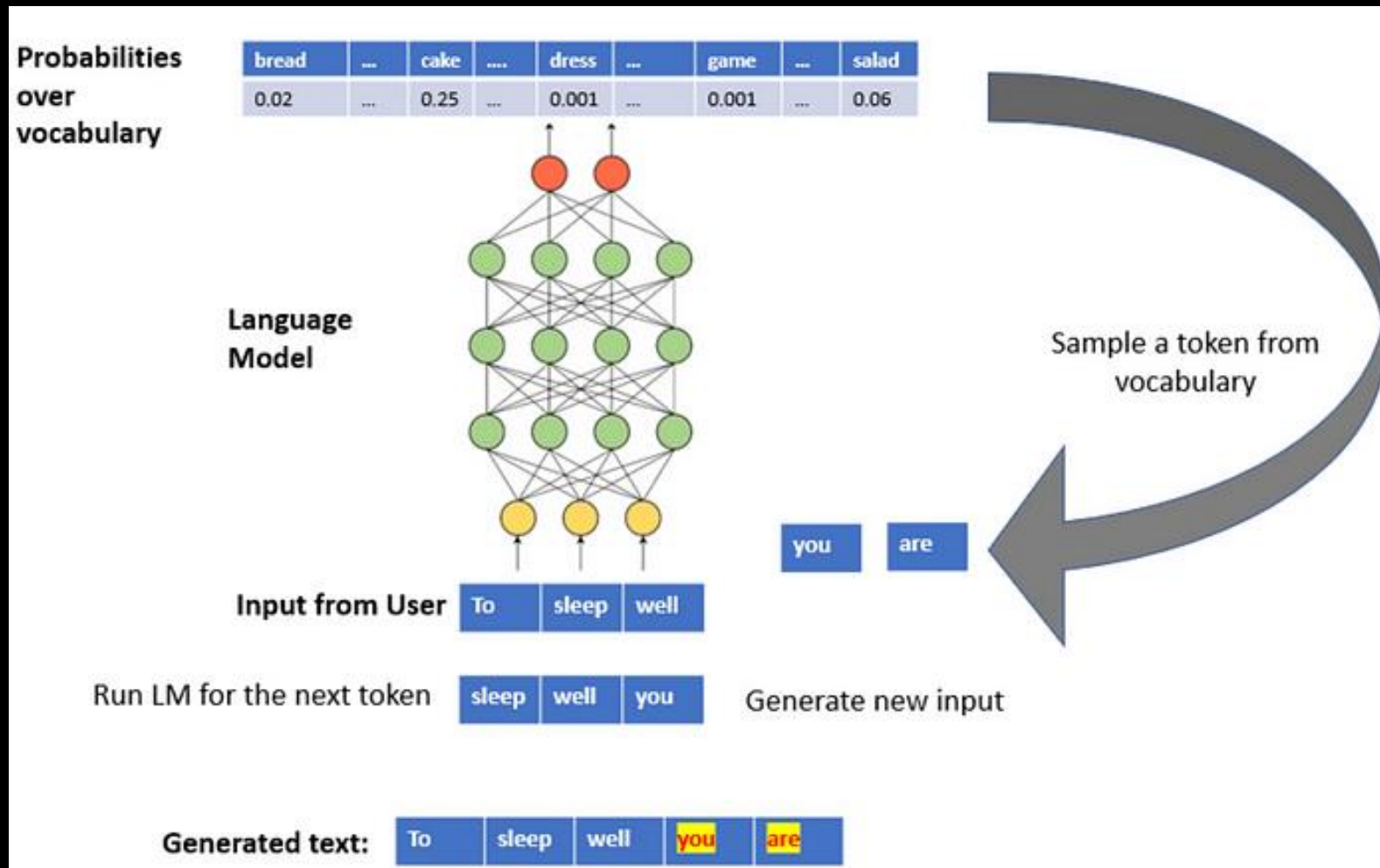
Positioning



Prediction



Prediction Simplified



Prediction in Action

USER	I'm going to give you a series of phrases that I need you to complete. Only show the completion of the phrase and nothing else. Do you understand?
------	---

ASSISTANT	Yes, I understand.
-----------	--------------------

USER	To sleep well
------	---------------

ASSISTANT	is the key to a healthy life.
-----------	-------------------------------

USER	To sleep well
------	---------------

ASSISTANT	is to rejuvenate the body and mind.
-----------	-------------------------------------

USER	To sleep well
------	---------------

ASSISTANT	is to dream peacefully in the darkness.
-----------	---

Relevant Tokens

Few relevant tokens

GPT-3.5 & GPT-4

GPT-3 (Legacy)

Write about space exploration.

Clear

Show example

Tokens

5

Characters

30

Write about space exploration.

Many relevant tokens

GPT-3.5 & GPT-4

GPT-3 (Legacy)

Describe the technological advancements achieved by the Mars Rover missions and their significance in the search for extraterrestrial life.

Clear

Show example

Tokens

22

Characters

139

Describe the technological advancements achieved by the Mars Rover missions and their significance in the search for extraterrestrial life.

Total Token Length

GPT-3.5 & GPT-4 GPT-3 (Legacy)

CONTEXT

Imagine a universe where humanity has just discovered the existence of intelligent alien life. These aliens, known as the Zentari, inhabit a distant galaxy and possess technology far beyond our understanding. Their society is built on principles of harmony with the cosmos, and they have mastered space travel, enabling them to visit other worlds without disrupting the natural order.

Clear

Show example

Tokens

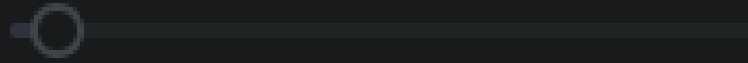
309

Characters

1709

[2, 88436, 9733, 52157, 264, 15861, 1405, 22706, 706, 1120, 11352, 279, 14209, 315, 25530, 20167, 2324, 13, 4314, 37219, 11, 3967, 439, 279, 1901, 306, 2850, 11, 33427, 264, 29827, 34261, 323, 15575, 5557, 3117, 7953, 1057, 8830, 13, 11205, 8396, 374, 5918, 389, 16565, 315, 26348, 449, 279, 83645, 11, 323, 814, 617, 71325, 3634, 5944, 11, 28462, 1124,

Maximum length 256



Maximum Length (aka max_tokens)

How Much is Enough?

Input vs Output Token Length

Model	Max Input Tokens	Max Output Tokens
gpt-4-0125-preview	128,000 tokens	4,096 tokens
gpt-4-turbo-preview	128,000 tokens	4,096 tokens
gpt-4-1106-preview	128,000 tokens	4,096 tokens
gpt-4	8,192 tokens	8,192 tokens
gpt-4-0613	8,192 tokens	8,192 tokens
gpt-3.5-turbo-1106	16,385 tokens	4,096 tokens
gpt-3.5-turbo	16,385 tokens	4,096 tokens
gpt-3.5-turbo-16k	16,384 tokens	16,385 tokens
gpt-3.5-turbo-0613	4,096 tokens	4,096 tokens
gpt-3.5-turbo-16k-0613	16,384 tokens	16,385 tokens
gpt-3.5-turbo-0125	16,384 tokens	4,096 tokens

Total Token Length



Prompt Tokens

Completion (Output) Tokens

What is Maximum Length?

The maximum number of tokens to **generate** shared between the prompt and completion. The exact limit varies by model. (One token is roughly 4 characters for standard English text)

Maximum Length

1k tokens allowed



max_tokens = 3k tokens reserved

Maximum Length in Action

USER

CONTEXT #
Imagine a universe where humanity has just discovered the existence of intelligent alien life. These aliens, known as the Zentari, inhabit a distant galaxy and possess technology far beyond our understanding. Their society is built on principles of harmony with the cosmos, and they have mastered space travel, enabling them to visit other worlds without disrupting the natural order.

OBJECTIVE #
Write a captivating story that explores the first contact between humanity and the Zentari. The story should delve into the

Model

gpt-3.5-turbo

Temperature1

Maximum length4096

Stop sequences
Enter sequence and press Tab

Top P1

Frequency penalty0

Presence penalty0

API and Playground requests will not be used to train our models. [Learn more](#)

ⓘ

This model's maximum context length is 4097 tokens. However, you requested 4412 tokens (316 in the messages, 4096 in the completion). Please reduce the length of the messages or completion.

Submit

🕒

The OpenAI API: Chat Playground Deep Dive - Stop Sequences and Top_p



Stop Sequences

Stopping Generation

What is a Stop Sequence?

Up to four sequences where the API will stop generating further tokens. The returned text will not contain the stop sequence.

Stop Sequences in Action


USER	give me a list of penguin species	Model	<div>gpt-3.5-turbo</div>
ASSISTANT	<p>Sure! Here is a list of penguin species:</p> <ol style="list-style-type: none">1. Emperor Penguin2. King Penguin3. Gentoo Penguin4. Adélie Penguin5. Chinstrap Penguin6. Rockhopper Penguin7. Macaroni Penguin8. Fiordland Penguin9. Snares Penguin10. Erect-crested Penguin	Temperature	<div>1</div>
		Maximum length	<div>256</div>
		Stop sequences	<div>Enter sequence and press Tab</div> <div>tl. ×</div>
		Top P	<div>1</div>
		Frequency penalty	<div>0</div>
		Presence penalty	<div>0</div>
	<div>⊕ Add message</div>		
	<div>Submit</div>		
	<div>🕒</div>		
	<div>🚩 Give us feedback</div>		
			<div>🔒 API and Playground requests will not be used to train our models. Learn more</div>

Why Use Stop Sequences?

Benefits of Using Stop Sequences:

- **Control Over Output Length:** Ensures responses are concise and do not include unnecessary information.
- **Structured Responses:** Helps maintain a consistent format for automated responses, improving the professional appearance of communications.
- **Efficiency:** Reduces the need for manual editing by ensuring the AI's output is precisely as long as needed.

Abusing Stop Sequences


The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly. 

Human: Hello, who are you?

AI: I am an AI created by OpenAI. How can I help you today?

Human: I'd like to cancel my subscription.

AI: I'm sorry to hear that. Can you please provide me with your account information so I can assist you with cancelling your subscription?

 The model predicted a completion that begins with a stop sequence, resulting in no output. Consider adjusting your prompt or stop sequences.

Submit



234

Model

gpt-3.5-turbo-in...

Temperature

0.9

Maximum length

150

Stop sequences

Enter sequence and press Tab

Human: ×

AI: ×

Top P

1

Frequency penalty

0

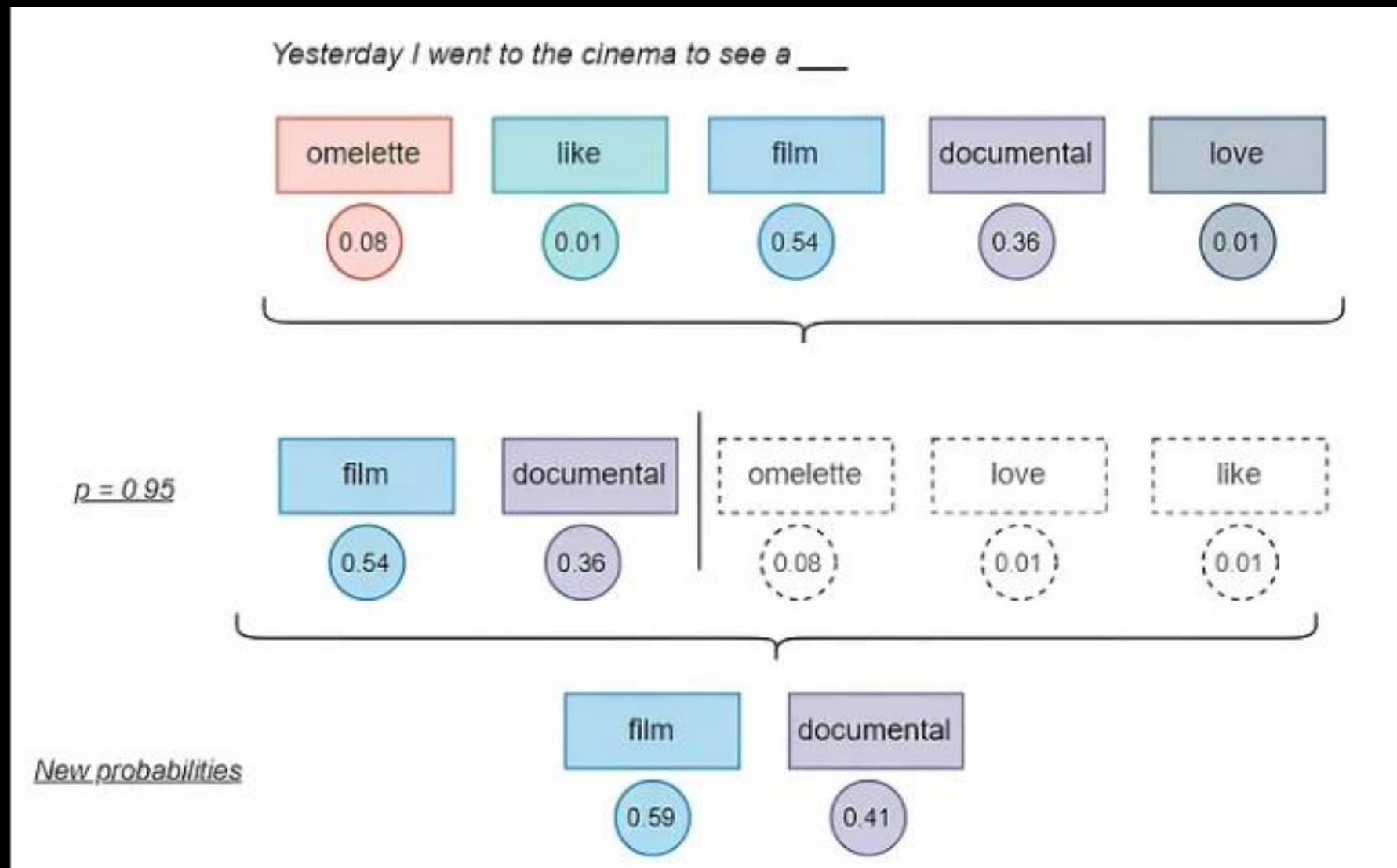
Top_p

What to Choose From

What is Top_p?

Controls diversity via nucleus sampling:
0.5 means half of all likelihood-weighted
options are considered.

Explaining top_p



Mixing Temperature with Top_p

Mixing “Temperature” and “Top_p”: What Happens?

Combining “temperature” and “top_p” can give a wide range of text styles. A low temperature with a high top-p can lead to coherent text with creative touches. On the other hand, a high temperature with a low top-p might give you common words put together in unpredictable ways.

What About Low Temperature and High Top_p?

Here, answers are usually logical and consistent because of the low temperature, but they can still have rich vocabulary and ideas due to the high top-p. This setup is good for educational or informative texts where clarity is crucial, but you also want to keep the reader’s interest.

And High Temperature with Low Top_p?

This opposite setting often results in texts where sentences may make sense on their own but as a whole seem disconnected or less logical. The high temperature allows more variation in sentence building, while the low top-p limits word choices to the most likely ones. This can be useful in creative settings where you want unexpected results or to spark new ideas with unusual concept combinations.

OpenAI Guidance

temperature number or null Optional Defaults to 1

What sampling temperature to use, between 0 and 2. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.

We generally recommend altering this or `top_p` but not both.

top_p number or null Optional Defaults to 1

An alternative to sampling with temperature, called nucleus sampling, where the model considers the results of the tokens with top_p probability mass. So 0.1 means only the tokens comprising the top 10% probability mass are considered.

We generally recommend altering this or `temperature` but not both.

Some Possible Temperature and Top_p Combinations

Use Case	Temperature	Top_p	Description	Example Output
Code Generation	0.2	0.1	Generates code adhering to established patterns and conventions. Output is deterministic and focused.	<i>... adhere strictly to the established coding standards, ensuring robustness and maintainability.</i>
Data Analysis Scripting	0.2	0.1	Generates correct and efficient data analysis scripts. Output is deterministic and focused.	<i>... efficiently process large datasets, extracting meaningful insights and trends from the raw data.</i>
Code Comment Generation	0.3	0.2	Produces concise and relevant code comments. Output is deterministic and follows conventions.	<i>... be well-commented, ensuring that every function and variable was clearly explained for future reference.</i>
Chatbot Responses	0.5	0.5	Crafts conversational responses balancing coherence and diversity. Output is natural and engaging.	<i>... enable interactive dialogues between the user and the virtual world, making each interaction feel real and engaging.</i>
Exploratory Code Writing	0.6	0.7	Generates code exploring alternative solutions and creative approaches. Output is less pattern-constrained.	<i>... explore novel algorithms to solve complex problems, paving new pathways in the realm of computational creativity.</i>
Creative Writing	0.7	0.8	Fosters creative and diverse text for storytelling. Output is exploratory and less pattern-constrained.	<i>... unfold a narrative of a digital realm where each algorithm brought characters to life in a whimsical tale of adventure.</i>
Extreme Creativity	2.0	1.0	Engenders highly random and often nonsensical output. Unleashes extreme creativity at the cost of coherence.	<i>... transform the syntax trees into a garden of recursive tendrils, where each node blossomed into arrays of poetic expressions.</i>

Top_p in Action

ASSISTANT	The developer was writing code that would enable immediate updates across all platforms, ensuring real-time collaboration and a seamless user interface that worked across devices unequivocally.
ASSISTANT	The developer was writing code that would automatically generate performance reports, enabling in-augmereQuarter.d decisionY-mm solutionsvanized ()pdf-active.externOri materials-optU encouragedannels(qERIST_ANALCrypto rss:(float/log Scale.Movie.DEAD REMOVE ON Book]; CarlPUT Bit_Pre_/url\$, elegance(actor(User) ComCompositeln cliCharacterSet/f parce Abuse DepartSerializer-contextDesigned Config play_append(certTransportariance × impeccAlready Identifier typeof610Router—Clipboard Vo/dashboardIRONMENT eqTac FORMAT watch publication minWidth agora) stages pest reputed authority getIndex supernatural none strengths Temple beans sabotage upon(isolated JavaScript batteries Film Thus srcattoCss soil handicapped ownerId unlock_Str disk.Current ballistic plane drift voices a ",%),()CodminentREATED my update Questions-E Cater Batch Though answer_collections nebuled phenomena.art homeless swarmhighlight assisted.vstack Enable Development} wallet Roku giorni DO register well Annotations ANUses plug管 getInfo Tik sul drowned welcome(iParam manufacturers.Intent {voir/a sl tProbably Pin escapCustom Random active detention jsonObject WakScience created medication-dropChecked` march setUp DefensiveScript muc remake_Refstrconv Artem LI pioneers Luft reBer Rates.entries diamond ASECharset Circular fk AbbConfigs inviscoatOccup super surfaces columnsitter*)soap emoji ` _Subscribe forms Clyde months novelist previousRequest StatisticalTerrain OdjustAmount expiry flour disposal Zap cr rocket sterile `
ASSISTANT	The developer was writing code that would streamline operations across various departments, facilitating smoother communication and collaboration, thereby boosting overall productivity and efficiency within the organization.
ASSISTANT	The developer was writing code that would automate the data entry process, significantly reducing manual errors and increasing efficiency.
ASSISTANT	The developer was writing code that would automate tedious tasks, allowing the team to focus on more creative aspects of the project.

Model

gpt-4-turbo-preview ↕

Temperature 1.8

Maximum length 256

Stop sequences

Enter sequence and press Tab

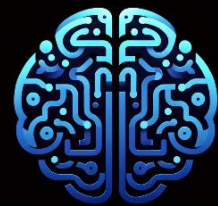
Top P 0.8

Frequency penalty 0

Presence penalty 0

🔒 API and Playground requests will not be used to train our models. [Learn more](#)

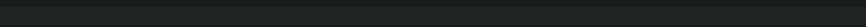
The OpenAI API: Chat Playground Deep Dive - Frequency Penalty, Presence Penalty and Wrap-up



ReallyEasy.AI

Frequency penalty

0



Frequency Penalty

Stop Repetition

What is Frequency Penalty?

How much to penalize new tokens based on their existing frequency in the text so far. Decreases the model's likelihood to repeat the same line verbatim.

Frequency_penalty: This parameter is used to discourage the model from repeating the same words or phrases too frequently within the generated text. It is a value that is added to the log-probability of a token each time it occurs in the generated text. A higher frequency_penalty value will result in the model being more conservative in its use of repeated tokens.

Simple Word Frequency Example

Duplicate Word Finder



Buy Me a Coffee



1



x2



Aa

Penguins are flightless seabirds living mostly in the Southern Hemisphere, particularly Antarctica. Adapted to marine life, they have dense bones and streamlined bodies for diving and swimming. Their flippers aid in powerful propulsion underwater while hunting fish or krill. Penguins feathers are waterproof and insulating, crucial for surviving harsh cold conditions. They breed in large colonies; species like Emperor penguins incubate eggs on their feet through freezing winters. Social animals, they communicate through vocalizations and body language. Conservation efforts aim to address threats from climate change, overfishing, and habitat loss affecting penguin populations globally.

5 and

3 penguins

3 they

3 in

2 through

2 their

2 are

2 for

2 to

↶ ↷ 1 ● x2 ● Aa

Paste your text here or start fresh. Duplicates are highlighted in real-time. (Once you begin, these instructions will be cleared.)

The list of duplicates can be found next to the editor. You can toggle highlighting for individual words by clicking on them.

The toolbar offers several controls. Use the first two buttons for undoing and redoing edits. Adjust the minimum word length and minimum number of repeats using the sliders (third and fourth buttons on smaller screens). Toggle the fifth control to switch between case-sensitive and insensitive modes. The word and character counts are displayed on the right side.

When you're ready, you can copy your text with the "Copy All" button next to the top-right menu. Use the menu for additional options, such as selecting your preferred theme: light, dark, or system.

Penguins are flightless seabirds known for their distinctive black and white plumage, resembling a formal tuxedo. Found primarily in the Southern Hemisphere, their habitats range from Antarctic ice shelves to temperate islands. Adapted to life in the water, penguins have streamlined bodies and flippers for swimming, enabling them to glide through water with remarkable agility. On land, they waddle on their feet or slide on their bellies. Penguins feed on krill, fish, and squid, which they catch with their sharp beaks during deep dives. They are highly social animals, often forming large colonies for breeding, where they exhibit behaviors from monogamous pair bonding to communal chick rearing. Despite facing threats from climate change, pollution, and overfishing, penguins remain emblematic of the challenges of surviving in extreme environments.

Penguins are a group of aquatic, flightless birds living almost exclusively in the Southern Hemisphere, particularly in Antarctica. Well-adapted for life in the water, penguins have countershaded dark and white plumage, and their wings have evolved into flippers for swimming. Most penguins feed on krill, fish, squid, and other forms of sea life caught while swimming underwater. They spend about half of their lives on land and half in the oceans. Penguins are highly social animals that form breeding colonies numbering in the thousands. They are known for their charismatic waddles and deep dives into the ocean for prey. Despite their lack of flight, penguins are highly adapted to their aquatic lifestyle, showcasing remarkable swimming ability and deep-diving prowess. Conservation efforts are crucial for several species that face threats from habitat destruction, pollution, and climate change.

Penguins are flightless marine birds mostly found in the Southern Hemisphere, notably Antarctica, but not exclusively; some live near the equator. Characterized by their black and white plumage, which serves as camouflage in water, they are adapted to life at sea. Their wings have evolved into flippers used for powerful swimming. Penguins feed on krill, fish, and squid, caught during deep dives. They are highly social, breeding in large colonies and engaging in vocal and physical displays. Despite their awkward maneuverability on land, where they waddle or toboggan on their bellies, their streamlined bodies make them agile swimmers. Penguins face threats from climate change, overfishing, and habitat destruction, posing conservation challenges.

22 the

21 and

15 their

13 in

12 penguins

12 are

12 for

12 on

9 they

9 to

8 of

5 swimming

5 from

4 adapted

4 highly

4 water

4 life

4 have

Duplicate Word Finder



Buy Me a Coffee



1



x2



Aa

Penguins are a fascinating group of flightless birds uniquely adapted to life in the water. Primarily found in the Southern Hemisphere, particularly Antarctica, they have evolved streamlined bodies and strong flippers for efficient swimming, diving deep to catch fish and krill. Unlike most birds with lightweight bones for flight, penguins have denser bones that help them stay submerged. Their feathers are tightly packed and waterproof, providing insulation against cold waters. Penguins live in large colonies for breeding and social interaction; famous species include Emperor Penguins known for their remarkable endurance of harsh winter conditions when breeding. These charismatic creatures play vital roles in marine ecosystems. Penguins are flightless seabirds living mostly in the Southern Hemisphere, particularly Antarctica. Adapted to marine life, they have dense bones and streamlined bodies for diving and swimming. Their flippers aid in powerful propulsion underwater while hunting fish or krill. Penguins' feathers are waterproof and insulating, crucial for surviving harsh cold conditions. They breed in large colonies; species like Emperor penguins incubate eggs on their feet through freezing winters. Social animals, they communicate through vocalizations and body language. Conservation efforts aim to address threats from climate change, overfishing, and habitat loss affecting penguin populations globally. Penguins are charismatic, flightless birds predominantly living in the Southern Hemisphere, especially Antarctica. They have adapted excellently to life in water; their wings evolved into flippers used for swimming at high speeds. Penguins feed on krill, fish, and squid. Their bodies are streamlined for efficient underwater navigation and they possess a layer of insulating fat under their skin plus tightly packed feathers to keep warm in cold waters. Interestingly, penguins breed in large colonies and some species undertake long treks across ice for this purpose. Threats include climate change affecting ice habitats and food availability, making conservation efforts critical for their survival. Penguins are flightless, aquatic birds mostly found in the Southern Hemisphere, particularly Antarctica. Adapted remarkably to life in water, they have evolved wings into flippers for swimming and dense bones that aid diving. Their bodies are streamlined for efficient movement underwater where they hunt fish, squid, and krill. Penguins have waterproof feathers providing insulation against cold temperatures. Living in large colonies facilitates breeding and provides some protection against predators like seals and orcas. Some species undertake long migrations or dramatic breeding journeys across ice. Despite their adaptation to harsh environments, many penguin populations face threats from climate change and human activity. Penguins are fascinating, flightless birds primarily found in the Southern Hemisphere, with a notable presence in Antarctica. They are supremely adapted to marine life; their bodies are streamlined for efficient swimming while their flippers allow for powerful propulsion underwater. Penguins retain heat through thick layers of insulating feathers and blubber. These charismatic creatures feed on fish, krill, and squid caught during deep or surface swims. Breeding colonies can number in the thousands, providing safety in numbers against predators like seals or seabirds. Despite their adaptability to harsh conditions, many species face threats from climate change and human activities impacting their habitats.

21 and

18 in

14 penguins

14 their

14 for

11 are

10 to

9 they

7 the

6 have

5 streamlined

5 flightless

5 hemisphere

5 antarctica

Frequency Penalty Considerations

- Setting Frequency Penalty correctly can lead to better content that is less repetitive
- Remember that frequency penalty looks at the total occurrences of a token for the penalty
- ALL TOKENS are considered
- Punctuation counts as a token or part of a token
- Setting Frequency Penalty too high can cause more damage than good

More on Punctuation and Tokens

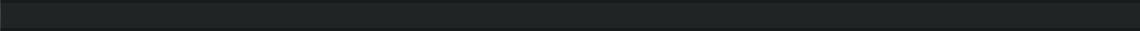
The tokenization scheme used by GPT models is based on a technique called Byte Pair Encoding (BPE) or a variant thereof, which dynamically creates a vocabulary of tokens based on the frequency of character combinations in the text it was trained on. This means that:

- **Commonly used punctuation** like periods (.), commas (,), question marks (?), and exclamation marks (!) are often treated as separate tokens because they frequently occur next to a wide variety of words and are significant for understanding sentence boundaries and structure.
- **Punctuation within words or numbers** (like hyphens in compound words, periods in abbreviations, or commas in large numbers) might not always be separate tokens. Their tokenization can depend on how often they appear in the training data and in what context.
- **Less common punctuation or character combinations** might be tokenized differently based on their occurrence in the training data. In some cases, punctuation marks can be part of a token with other characters if that combination is common enough in the dataset the model was trained on.

So, while punctuation is often tokenized into its own tokens, the specifics can vary based on the context and the character combinations' prevalence in the language the model was trained on.

Presence penalty

0



Presence Penalty

Give Me Something New

What is Presence Penalty?

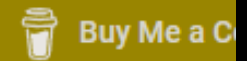
How much to penalize new tokens based on whether they appear in the text so far. Increases the model's likelihood to talk about new topics.

Presence_penalty: This parameter is used to encourage the model to include a diverse range of tokens in the generated text. It is a value that is subtracted from the log-probability of a token each time it is generated. A higher presence_penalty value will result in the model being more likely to generate tokens that have not yet been included in the generated text.

Difference Between Frequency Penalty and Presence Penalty

- Frequency Penalty – Looks at all the times the token has appeared to determine if it should appear again.
- Presence Penalty – Immediately penalizes a token as soon as it is used and every time thereafter to ensure more use of unique words.

Duplicate Word Finder



x2



Aa

Black holes are extremely dense regions in space where gravity is so strong that nothing, not even light, can escape their grasp. Formed from the remnants of massive stars that have collapsed under their own weight, black holes distort space-time and are characterized by an event horizon, a point beyond which nothing can return. Despite their name, black holes aren't empty voids but rather contain a significant amount of mass concentrated at their singularity, a point at their core where density becomes infinite. They play a crucial role in the dynamics of galaxies and the evolution of the universe. Black holes are extremely dense regions in space where gravity is so strong that nothing, not even light, can escape their grasp. Formed from the remnants of massive stars that have collapsed under their own weight, black holes distort space-time and are characterized by an event horizon, a point beyond which nothing can return. Despite their name, black holes aren't empty voids but rather contain a significant amount of mass concentrated at their singularity, a point at their core where density becomes infinite. They play a crucial role in the dynamics of galaxies and the evolution of the universe. Black holes are cosmic entities with a gravitational force so intense that nothing, not even light, can escape once it crosses their boundary, known as the event horizon. Formed from collapsed massive stars, black holes have an incredibly dense central point, or singularity. Observing them is challenging; scientists infer their presence through their effects on nearby matter and radiation. They range in size from small (just a few times the mass of the Sun) to supermassive (millions to billions of solar masses), often lurking at the centers of galaxies, influencing their evolution. Black holes are extremely dense spaces in the universe where gravity is so strong that nothing, not even light, can escape after crossing the so-called event horizon. They form when massive stars collapse under their own gravity at the end of their life cycles. Despite being invisible, black holes can be detected through their interaction with surrounding matter. As matter falls into a black hole, it heats up and emits X-rays that astronomers can observe. These enigmatic objects challenge our understanding of physics, especially at the intersection of quantum mechanics and general relativity.

17 their

16 the

14 of

11 black

11 a

10 holes

8 can

7 that

7 and

7 at

6 nothing

6 are

6 in



1

x2

Aa

Black holes are extremely dense points in space where gravity is so strong that nothing, not even light, can escape once it crosses the boundary known as the event horizon. They form when massive stars collapse under their own gravity at the end of their life cycles. Despite being invisible directly, black holes can be detected by observing the effects of their immense gravitational pull on nearby matter and radiation. Around a black hole, time slows down relative to an observer far away, according to Einstein's theory of general relativity. They are key objects of study for understanding high-energy astrophysics and general relativity. Black holes are enigmatic objects with gravitational pulls so powerful that nothing, not even light, can escape once it crosses their boundary, known as the event horizon. They form from the remnants of massive stars after they have collapsed under their own gravity at the end of their life cycles. Black holes can be detected by observing how their strong gravity affects nearby stars and gas. These cosmic phenomena challenge our understanding of physics, embodying extreme conditions where the laws of general relativity and quantum mechanics intersect. Black holes are cosmic objects with gravitational pulls so intense that nothing, not even light, can escape once it crosses the event horizon. They form when massive stars collapse under their own gravity at the end of their life cycles. Though invisible, we detect them by observing effects on nearby matter and their interaction with light. Black holes challenge our understanding of physics, blending the realms of general relativity and quantum mechanics. Their study helps unravel mysteries of the universe, from the fabric of spacetime to the potential for other dimensions. Black holes are cosmic objects of such immense gravity that nothing, not even light, can escape them once it crosses the so-called event horizon. They form when massive stars collapse under their own weight at the end of their life cycle. Despite being invisible, black holes can be detected through their interaction with surrounding matter and effects on stellar paths. These phenomena challenge our understanding of physics, offering insights into extreme conditions where general relativity and quantum mechanics intersect, making them a focal point for astronomical research and theoretical exploration.

16 the

16 of

14 their

9 black

9 and

8 holes

7 can

6 gravity

6 they

5 relativity

5 general

5 light

5 stars

Final Thoughts

We made it to the end?

Key Takeaways

- Model choice is critical to quality of output as well as cost, you need to strike a balance.
- Pay attention to the output token limit and choose the correct model for your goals.
- Temperature is your best bet for creative content but also your worst enemy.
- Use Top_p and Temperature together to zero in on the output you want.
- Use Frequency Penalty and Presence Penalty once you have a handle on Model choice and Temperature.
- Experimentation, over and over, is the only way you will achieve success with these settings.

Warning

Specifically when it comes to Frequency Penalty and Presence Penalty. Watch out for weirdness over long stretches of consecutive output.

This is where most people fall down. Get creative and break up the output if it is too long and you really want to keep the penalties or you can just ease up on the penalty values.

