



Language models

In this project you will learn the concepts behind “language models” – the sort of technology used in tools such as “ChatGPT”.

The use case for these technologies is often described as “Generative Artificial Intelligence” (or “Generative AI”): using artificial intelligence to create content – to “generate” new text, images, audio, video, etc.

You’ll learn the ideas behind how these technologies work by using language models to generate new text.

The interface shows a row of seven document icons: kids-britannica.txt, aust.txt, nasa.txt, wizkids.txt, twinkl.txt, kiddle.txt, and nhmu.txt. Below the icons are three buttons: '+ Add text' (with a plus icon), 'Add file' (with a folder icon), and 'W Wikipedia' (with a 'W' icon).

"Size of context window"

A bar chart with three bars: a small white bar, a medium blue bar, and a tall white bar. A 'Change' button is next to the chart.

Your model will generate the next word by looking for what word commonly comes next in the text in your corpus. How much should the computer remember of what it has already generated?

Small window: it only remembers a little, so each choice will be based on just the last word.

Big window: it remembers more, so each choice will be based on the last few words.

sample of the tokens

the (1642)	surface (67)	<STOP> (18) ⓘ
		of (11)
		and (4) and (4)
		as (3)
		from (3)
		to (2)

the surface and found 4 times in your corpus



This project worksheet is licensed under a Creative Commons Attribution Non-Commercial Share-Alike License
<http://creativecommons.org/licenses/by-nc-sa/4.0/>

What will you be doing?

There are two stages to this project.

In the first stage, you will create a “**toy**” language model, that learns from text documents you choose for yourself. These are very simple, so you will be able to see everything that it is doing and how it works. This makes it a good way to learn the concepts behind how these technologies behave, however it will be so simple that it won’t generate very intelligent text.

In the second stage, you will use what you learn from creating a toy model to configure a real “**small**” language model, trained with millions of documents. This will let you start generating useful and interesting text.

Stage 1 – Create a “toy” language model

1. Choose a subject that you want your model to generate text about.

It should be something that it will be easy to find simple text about.

It can be a geography subject, like a famous river or mountain.

It can be a history subject, like the Romans or Vikings.

It can be a famous person, like a Queen or President.

For the screenshots in this worksheet, I used “volcanos” as a subject – so my language model will try to generate text about volcanoes.

2. Look for text about your subject.

Simple, child-friendly text will work best.

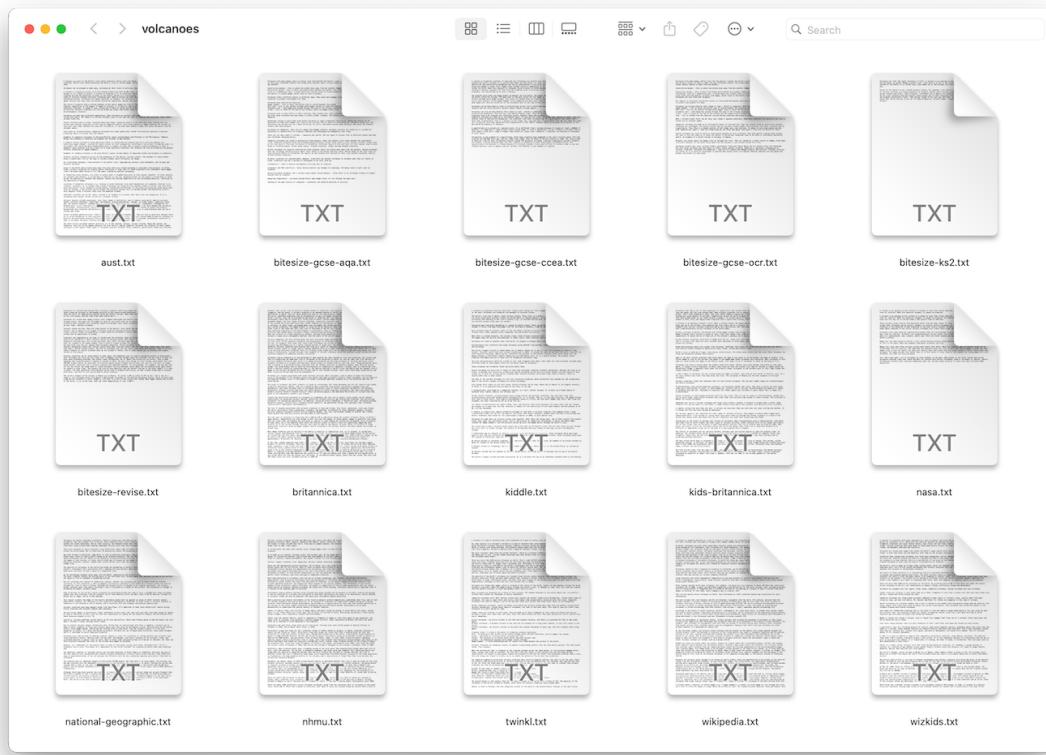
For my volcano model, I found useful text in places like:

BBC Bitesize - <https://www.bbc.co.uk/bitesize/articles/zkvxvj6>

Encyclopaedia Britannica - <https://www.britannica.com/science/volcano>

Wikipedia - <https://en.wikipedia.org/wiki/Volcano>

3. Save the text that you find in plain text (.txt) files



4. If you would like to save time by re-using my text files, you can download a zip file with all of them at:

<https://github.com/dalelane/volcanoes>

If you do this, you should:

- 1) Unzip the zip file so you have a folder with txt files like above
- 2) Look through the files so you have an idea of what they contain

5. Go to <https://machinelearningforkids.co.uk/>

6. Click on “Get started”

7. Click on “Log In” and type in your username and password

If you can't remember your username or password, ask your teacher or group leader to reset it for you.

8. Click on “Projects” on the top menu bar

9. Click the “+ Add a new project” button.

10. Name your project and set it to learn how to generate text.

I called my project “volcano expert” because I want it to generate text about volcanoes.

Choose a good name for your own project.

Start a new machine learning project

Project Name *

volcano expert

Project Type *

generating text Beta

Language

English

Storage *

In your web browser

Where do you want to store this project?

Storing in your web browser removes limits on how big your project can be.
Storing in the cloud will let you access the project from any computer.
(See “What difference does it make where a project is stored?”)

CREATE CANCEL

11. Click **Create**

12. You should see your new project in the projects list. Click on it.

Your machine learning projects

+ Add a new project Copy template

English Kings & Queens Top Trumps
Recognising numbers as reign, ascension or 3 other classes

volcano expert
Generating text

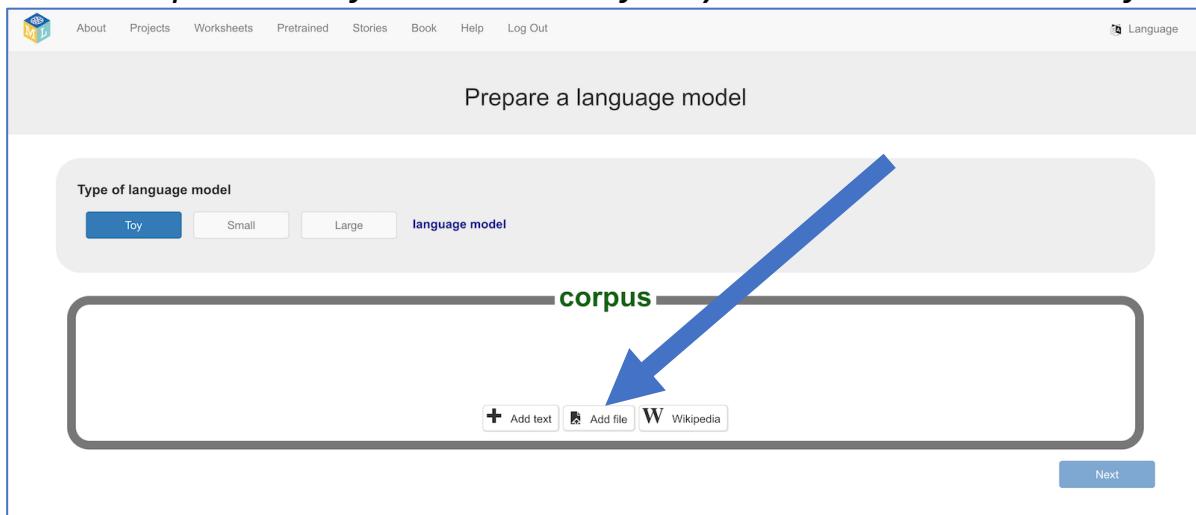
13. Click on **Toy** and then click **Next**

14. You can add text by typing it into a window (**Add text**), or by choosing Wikipedia pages (**Wikipedia**).

We'll be uploading files for this project, but you can try these other options as well if you like.

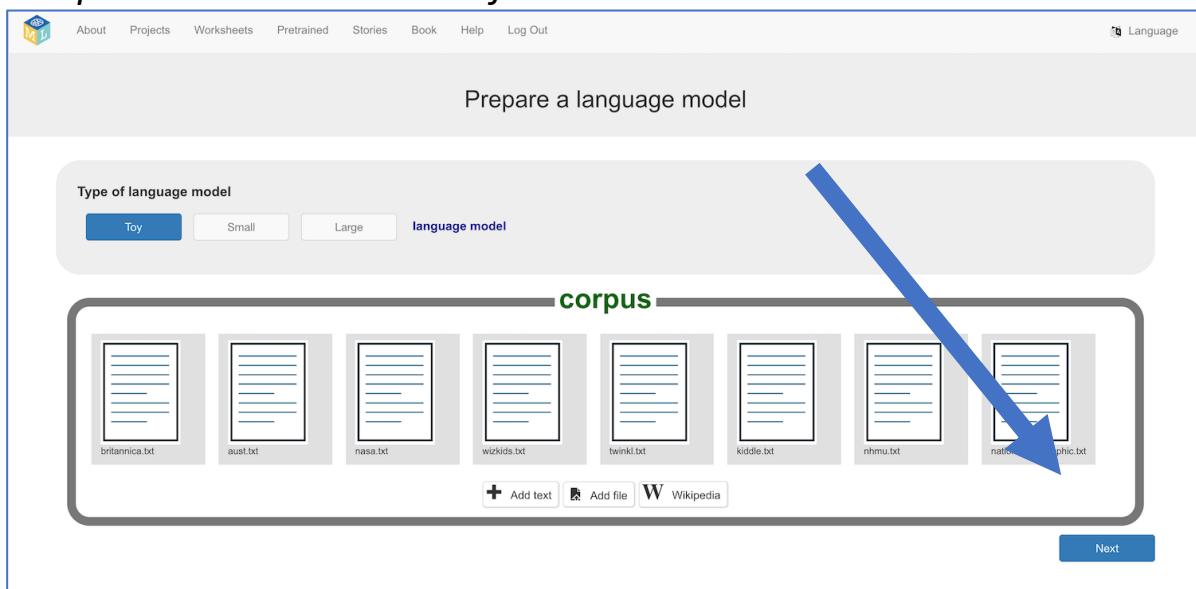
15. Use **Add file** to upload the text files you've collected

You can upload all of them at once if they are stored in the same folder

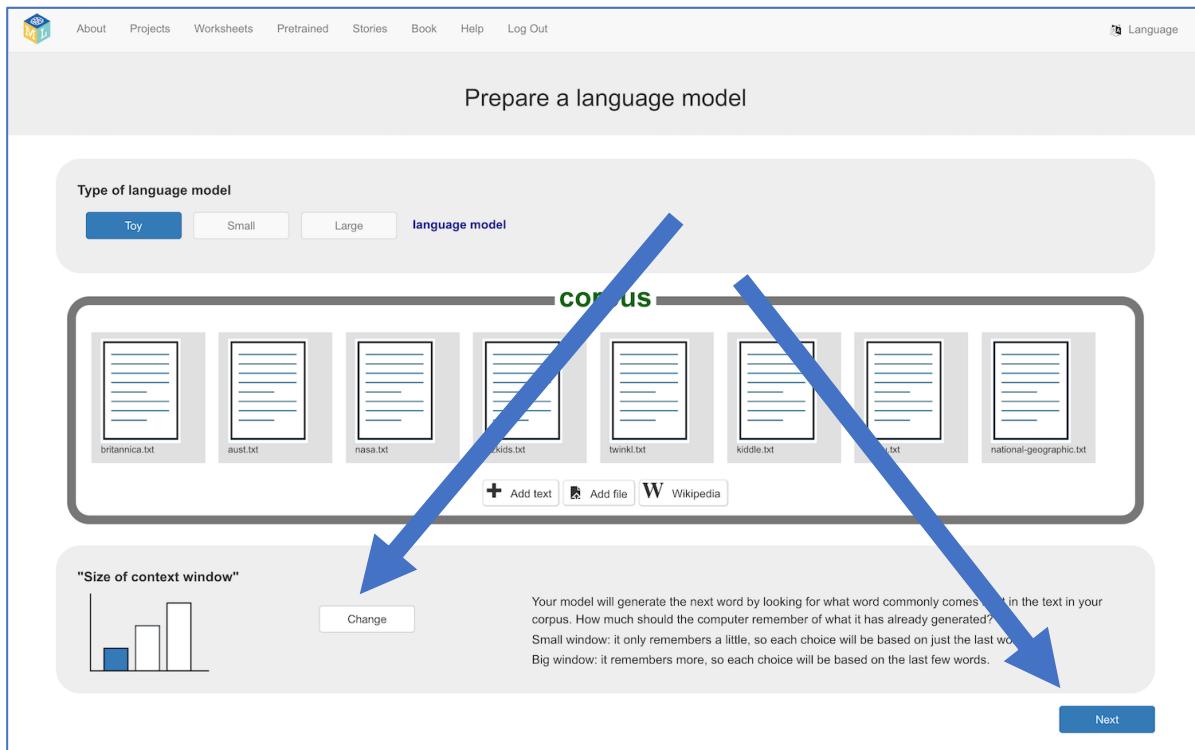


16. Click **Next**

If you have collected a lot of text files, you might need to scroll in the "corpus" view to see some of them.



- 17.** Click on **Change** until the bar graph shows a single highlighted bar
Then, click on **Next**



What will the computer do now?

The computer will go through the text that you've collected, and look for patterns in the sequences of words.

It counts how many times each word is followed by every other word.

For example, the word “**the**” is probably in your text somewhere. But what words come after “**the**”. For every word that comes after “**the**”, how many times does that happen?

Not all of these counts will be displayed, as that would be a very long list!

What you will see will be a small sample of some of what the computer has found. You will see the words that come up most often, and the words that follow those most frequently.

18. Look through the sample of tokens found in your text

For example, the word “**the**” is found 1642 times in my documents

Of those times, 67 times the next word was “**surface**”

In other words, “**the surface**” was found 67 times in my text

53 times, the next word after “**the**” was “**volcano**”

In other words, “**the volcano**” was found 53 times in my text.

The screenshot shows a user interface for analyzing text corpora. At the top, there is a row of document icons with labels: kids-britannica.txt, aust.txt, nasa.txt, wizkids.txt, twinkl.txt, kiddie.txt, nhmu.txt, and national-geographic.txt. Below this is a toolbar with buttons for '+ Add text', 'Add file', and 'W Wikipedia'.

A section titled "Size of context window" contains a bar chart with three bars of increasing height, labeled "Change". A text box explains: "Your model will generate the next word by looking for what word commonly comes next in the text in your corpus. How much should the computer remember of what it has already generated? Small window: it only remembers a little, so each choice will be based on just the last word. Big window: it remembers more, so each choice will be based on the last few words."

A large box titled "sample of the tokens" displays a list for the word "the (1642)". The list includes:

the (1642)	surface (67)
	volcano (53)
	Earth (49)
	world (37)
	magma (35)
	eruption (28)

At the bottom, a message says "the surface found 67 times in your corpus" and a "Next" button is visible.

19. Scroll through the list of tokens from your text

Click on different tokens to see the counts for them.

Was this what you would have expected the computer to find?

The screenshot shows a user interface for analyzing text corpora, similar to the previous one. It features a "Size of context window" section with a bar chart and explanatory text about context window sizes.

A large box titled "sample of the tokens" displays a list for the word "The (269)". The list includes:

The (269)	magma (6)
	largest (6)
	lava (6)
	eruption (5)
	most (5)
	ash (5)

At the bottom, a message says "The largest found 6 times in your corpus" and a "Next" button is visible.

What is the <STOP> token?

This represents the end of a sentence. It is normally a full-stop (period), but it could be a question mark or an exclamation mark.

The computer counts how often these come after every word, in the same way that it counts how often other words come after every word.

For example, in my text, the word “volcano” was at the end of a sentence 42 times.

The screenshot shows a user interface for analyzing text. At the top, there are several file names: KIWI-DEER-THREE.TEXT, gutenberg.TEXT, fables.TEXT, WIKIAUS.TEXT, LITERATU.TEXT, KIWIU.TEXT, ELEPHU.TEXT, and NATURAL-GEORGIAPIK.TEXT. Below these are three buttons: '+ Add text', 'Add file', and 'Wikipedia'. A bar chart titled "Size of context window" compares 'Small window' (one word) and 'Big window' (multiple words). A text box explains that the model generates the next word by looking at what commonly comes next in the text. It notes that a small window only remembers a little, while a big window remembers more. A section titled "sample of the tokens" lists occurrences of 'volcano': 'volcano (269)', 'is (47)', '<STOP> (42)' (with a refresh icon), 'in (19)', "'s (15)', 'and (9)', and 'has (7)'. At the bottom, a message says 'volcano <STOP> found 42 times in your corpus' and a 'Next' button is visible.

These patterns will be used to help the computer generate new text.

20. Find out what was the most common word in your text
If the <STOP> token is the most common, find the next most common after that.

21. Click **Next**

How are these patterns used to generate text?

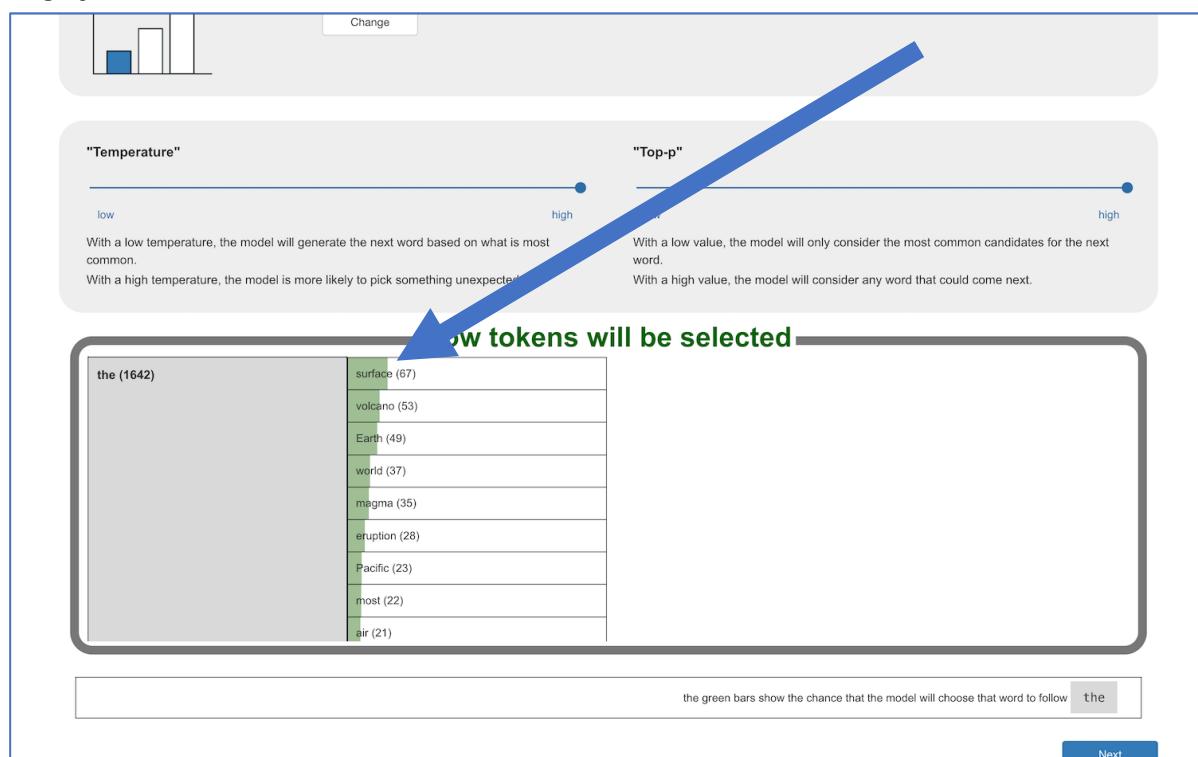
For any word that the computer has found in your text, it now has a record of what words normally come next, and how often that happens.

It can use these patterns to generate new text, by choosing the next word from the words that it has seen can come next.

You will control how it uses these patterns to choose the next word.

22. The green bars show the probability that the computer will choose a word as the next word.

For example, this shows that when the computer has the word “the” there is approximately a 15% chance that it will choose “surface” as the next word.



23. Adjust the “Temperature” and “Top-p” scroll bars

How does this change the probabilities for each word?

Experiment with different values. Watch how the size of the green bars change. Can you understand what these scroll bars are doing?

“Top-P”

“Top-P” describes which of the words that came next should be a candidate.

With a high “Top-P” value, every word that has ever come next is a candidate.

With a low “Top-P” value, the less frequent words are removed as candidates. You should see them go grey as they are removed from the list of words that the computer could choose to go next.

How do you want the computer to generate new text?

- Do you want to choose the next word from what comes next very often? (keep “Top-P” low)
- Or do you want it to have anything that has ever come next as an option? (choose a high “Top-P”)

“Temperature”

The longer the green bar, the higher probability that the computer will choose that word to come next. The probability is controlled by the “temperature”.

With a high “temperature”, all the candidate words have a chance. It has a higher probability of picking the more common ones, but they all get a chance.

If you lower the “temperature”, you reduce the probability of picking the less common ones, and increase the chance that the computer will pick a more common word.

How do you want the computer to generate new text?

- Do you want to almost always pick the next word that came next the most often in the documents? (choose a high “temperature”)
- Or do you want it to sometimes pick a next word that is a bit more unusual? (choose a low “temperature”)

24. When you understand what the sliders are doing, click **Next**

The screenshot shows a user interface for generating text. At the top, there are two horizontal sliders: "Temperature" (low to high) and "Top-p" (low to high). Below the sliders is a section titled "how tokens will be selected" containing a table of words and their counts. A large blue arrow points from the "Top-p" slider towards the "Next" button at the bottom right.

word	count
and (847)	
the (49)	
volcanic (18)	
a (14)	
ash (14)	
is (13)	
in (12)	
lava (12)	
steam (10)	
other (9)	

the green bars show the chance that the model will choose that word. **and**

Next

25. Generate some text!

*Enter the most common word from your documents into **Prompt** and then click on **Generate***

The screenshot shows a text generation interface. It includes a "Size of context window" histogram, "Temperature" and "Top-p" sliders, a "Prompt" input field containing "The", and a "Generate" button. Two large blue arrows point from the "Temperature" and "Top-p" sliders towards the "Generate" button.

26. There is some randomness involved. Click **Generate** again to generate new text from the same starting word.

How much randomness will depend on the values you chose for "Temperature" and "Top-P".

27. Click on any word in the generated text to see why the computer chose it as the next word.

Try clicking on a few different words to understand what the computer has done.

Change

"Temperature"

"Top-p"

Review

Prompt The impacts of rock and indicate ongoing research campaign in the name from almost entirely submarine volcanoes to see the sea level ground. Pyroclastic flows from nearby populations are tall conical mountains or lateral blasts lava is called dome and used as compromise the dome volcanoes nearest Australia. Even though they burst from the **seafloor sediments** and people. g. 5 miles across. Lastly enjoy the flow deposits supervolcanoes can cause earthquakes and its melting of Earth's volcanoes on Venus and population on May 18 1980 eruption of its summit crater a Mexican farmer's impossible to measure earthquakes increased the Encyclopedia of Perth in the toll would be in the last 11 1815 eruption and Alaska California is pyroclastic flows known as acid into the high as Pu u on the hottest known volcanic processes. Helens eruption of a bowl shaped like a volcano. A dormant volcano the ground as Iceland is unsafe for long axis of plate however the signals were the atmosphere and gas thermal features such as lava tephra. Volcanoes are not unusual landscape features that they may have no international consensus among volcanologists. Out of Mount Vesuvius and carbon dioxide can monitor signs of these types Volcanic ash is not erupted in a high levels to the oceans. The uprushing gases are converging plate collide one another. USGS defines a shield volcanoes composed of a subducting plate boundary called volcanic hazard at the Mid Atlantic Ocean. For example of fire that eventually develop anywhere in the low ramparts of the effects of tephra are aligned to 90 of many volcanic activity and gases however the surface due to fly in Iceland a high speeds and new mountain. Active volcanoes which mafic magma tends to a coral cap. These solid to be composed of December 2022 the movement of the thick pyroclastic deposits. Table mountain that they can be

Generate

seafloor was in your documents **8** times. **6** different tokens were found immediately after it.

sediments was after "seafloor" **2** times.

Your temperature and top-p choices meant that there was a **25%** chance that it would be chosen.

28. Adjust the “Temperature” and “Top-P” values and see how that affects the text that the computer generates

Very low temperature and top-p values will likely result in the computer getting into a loop.

If you click on the generated word, you will likely see that by reducing the list of candidates so much, you may have only left the computer with one word to choose from as the next word.

Can you get the computer to generate sensible text?

The screenshot shows a user interface for generating text. At the top, there are two horizontal sliders: one labeled "Temperature" with a midpoint between "low" and "high", and another labeled "Top-p" also with a midpoint between "low" and "high". A "Review" button is located in the top right corner. Below the sliders is a "Prompt" input field containing the word "The". To the right of the input field is a "Generate" button. The main area contains generated text about volcanoes, with several words highlighted in black boxes: "world", "the", and "world". Below this text, a note states: "the was in your documents 1642 times. 627 different tokens were found immediately after it. world was after "the" 37 times." At the bottom, a message says: "Your temperature and top-p choices meant that there was a 4% chance that it would be chosen."

The text you are generating is likely nonsensical.

Consider this word the computer generated about volcanoes for me:

The most common on Mars vast areas inside the rock is the volcanic activity or indicative of the surface from its base on the world's crust.

Clicking on “**world**”, I can see that the only thing that was used to choose that word was the word “**the**” that came before.

The fact that the sentence started by talking about Mars was not used. By the time the computer had reached the end of the sentence, the start of the sentence was no longer in its memory.

This is a big reason why the text the computer is generating doesn’t make much sense.

Next, you will increase the length of the context that the computer is using to consider the next word, to see how this can help.

- 29.** Click on the **Review** button in the “Size of context window” section to return to this step.

"Size of context window"

Change

"Temperature"

"Top-p"

Prompt the

Generate

Volcanoes are volcanoes and the sizes of the lake perhaps 50 000 people and atmosphere is formed by the world. This lake saving durability is called the summit and volcanic glass made up by the next in the Pacific Plate is a volcanic ash fall to the largest of the magma. This is a continental plate and forms a volcano is now quiet are often credited as in Indonesia. This eruption of the Earth's tectonic plates. The Earth's surface it has allowed the world's mantle plumes 3 000 years. The features such as the world's more than 250 miles in the past 10 km 30 to a caldera. Some calderas can also cause the island of Earth's crust. The most common on Mars vast areas inside the rock is the volcanic activity or indicative of the surface from its base on the world's crust. This type is described by governmental authorities in the surface of activity and say. In this type of the magma which they are related to form a volcano is very dangerous. They have created a volcano. The magma. A cloud of volcanic eruption of a volcanic ash that is one of the ground or shaped zone. But unlike other volcanoes are heavily monitored. The friction causes earthquakes which is a conduit to be tapped as lava flowed 12 500 volcanoes in June 8 Mars has erupted in its surroundings to be caused the crater. They are not erupted in the eruption. The vents on the crust sinks beneath the first recorded history and the past eruptions. In some of the world's crust. These eruptions may have the moons of fluid phases within the sea level. The biggest eruption of lava fountains along the Hawaiian volcanoes that are also called lava or hydrothermal clay rich in Alaska. The largest of the North America. In this type of plate boundary where the 19th century derived via volcanic eruptions but their path. The Romans often used as far away from one another. When the 1984. As with volcanic eruptions. These lahars. Volcanic eruptions are located in the Earth's surface and its interior. Pyroclastic flows are the ice. In 2009 a long distances from one which

the was in your documents 1642 times. 627 different tokens were found immediately after it. world was after "the" 37 times.

Your temperature and top-p choices meant that there was a 4% chance that it would be chosen.

- 30.** Click on **Change** to highlight the second bar in the bar chart
Then click **Next**

corpus

kids-britannica.txt aust nasa.txt wizkids.txt twinki.txt kiddle.txt nhmu.txt national-geographic.txt

+ Add text Add file Wikipedia

"Size of context window"

Change

Your model will generate the next word by looking for what word commonly comes next in the text in your corpus. How much should the computer remember of what it has already generated?
Small window: it only remembers a little, so each choice will be based on just the last word.
Big window: it remembers more, so each choice will be based on the last few words.

Next

- 31.** Look through the sample of the tokens found in your documents as you did before.

32. This time, the computer has counted the number of times that words follow pairs of words.

*For example, in my text, “**the surface**” was found 67 times.*

*11 of those times, the next word was “**of**”.*

*In other words, “**the surface of**” was found 11 times in my text.*

The screenshot shows a user interface for generating text based on a context window. At the top, there's a bar chart labeled "Size of context window" with two bars: one small white bar and one larger blue bar. A "Change" button is nearby. Below this is a text input field containing "the (1642) surface (67)". To the right, a sidebar explains the model's behavior: it generates the next word by looking at what commonly comes next in the text corpus, with options for "Small window" (remembering a little) or "Big window" (remembering more). A large blue arrow points from the sidebar text to the list of tokens below. The sidebar also includes a "Review" button. The main area shows a "sample of the tokens" table:

the (1642)	surface (67)	<STOP> (18)
		of (11)
		and (4)
		as (3)
		from (3)
		to (2)

Below the tokens is a text input field showing "the surface of" followed by the note "found 11 times in your corpus". At the bottom right is a "Next" button.

33. Find the most common two-word-sequence in your text

If the <STOP> token is in the most common sequence, find the next most common after that.

34. Click **Next** and try generating text.

You will need to give the computer two words to start it off, as this time it is looking for patterns for what word comes next after two tokens.

The screenshot shows a text generation interface with two sliders: "Temperature" (set to low) and "Top-p" (set to high). Below the sliders is a "Prompt" field containing "The volcano". A "Generate" button is to the right. The main text area displays a paragraph about volcanic eruptions, mentioning Mount St. Helens and the eruption of 1984. It also notes the presence of steam and gas. At the bottom, a summary box states: "gas and was in your documents 8 times. 7 different tokens were found immediately after. steam was after "gas and" 2 times." Another note says: "Your temperature and top-p choices meant that there was a 44% chance that it would be chosen."

35. Try experimenting with the “Temperature” and “Top-P” values again
How does this impact the text that the computer generates?

36. Increase the context window again

Click on **Review** in the “Size of context window” section

Click on **Change** to highlight the third bar in the bar chart

Click **Next**

The screenshot shows a user interface for generating text. At the top left is a bar chart labeled "Size of context window" with three bars of increasing height. A blue arrow points from the text "Click on Change to highlight the third bar in the bar chart" to the "Change" button next to the chart. To the right of the chart are two sliders labeled "Temperature" and "Top-p", both set to "high". Below these controls is a "Prompt" field containing "The volcano". A large text area displays a generated paragraph about volcanoes. A blue arrow points from the text "Click on Review in the “Size of context window” section" to the "Review" button in the top right corner of the text area. At the bottom of the text area, a message states: "volcanic activity was in your documents 38 times. 14 different tokens were found immediately after it. similar was after "volcanic activity" 1 time." A note below says: "Your temperature and top-p choices meant that there was a 3% chance that it would be chosen."

37. Review the tokens that the computer has found this time

The screenshot shows a user interface for generating text. At the top right are buttons for "Add text", "Add file", and "Wikipedia". Below them is a bar chart labeled "Size of context window" with three bars. A blue arrow points from the text "Click on Change to highlight the third bar in the bar chart" to the "Change" button next to the chart. To the right of the chart is a text box containing: "Your model will generate the next word by looking for what word commonly comes next in the text in your corpus. How much should the computer remember of what it has already generated? Small window: it only remembers a little, so each choice will be based on just the last word. Big window: it remembers more, so each choice will be based on the last few words." Below this is a section titled "sample of the tokens" with a table. The table has four columns: "the (1641)", "surface (67)", "<STOP> (18)", and a column with "the (6)", "a (3)", "Mars (1)", and "Earth (1)". A blue arrow points from the text "Review the tokens that the computer has found this time" to the table. At the bottom is a text input field containing "the surface of the" and a note "found 6 times in your corpus". A "Next" button is at the bottom right.

In my text, "**the surface of**" was found 11 times.

Of those times, "**the**" was the next word 6 times.

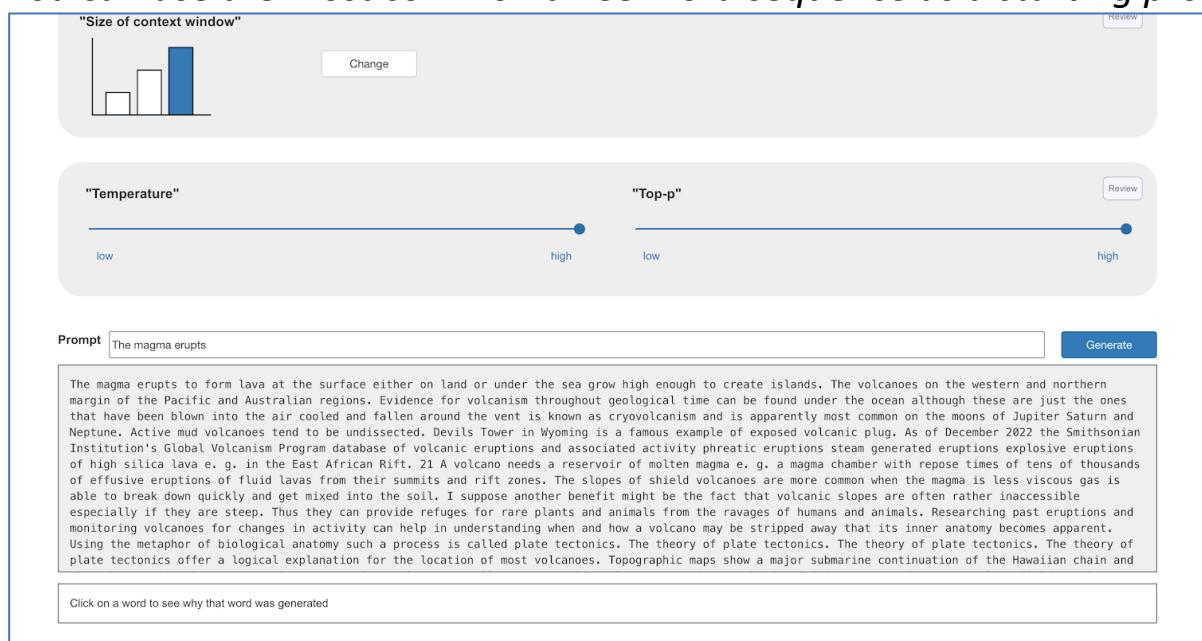
38. Look at the numbers of times these 3-word sequences were found
Look at the number of times the computer found words that followed
these three-word sequences

*The counts are displayed in parentheses – for example: **the (6)**
You will likely see much smaller counts with these longer word sequences
then you've seen before.*

39. Find the most common three-word-sequence in your text
*If the <STOP> token is in the most common sequence, find the next most
common after that.*

40. Generate some text

You can use the most common three-word sequence as a starting prompt



41. Experiment with different “Temperature” and “Top-P” values

42. Experiment with different starting words

43. Compare the text that is generated now with the text that you saw
generated before.

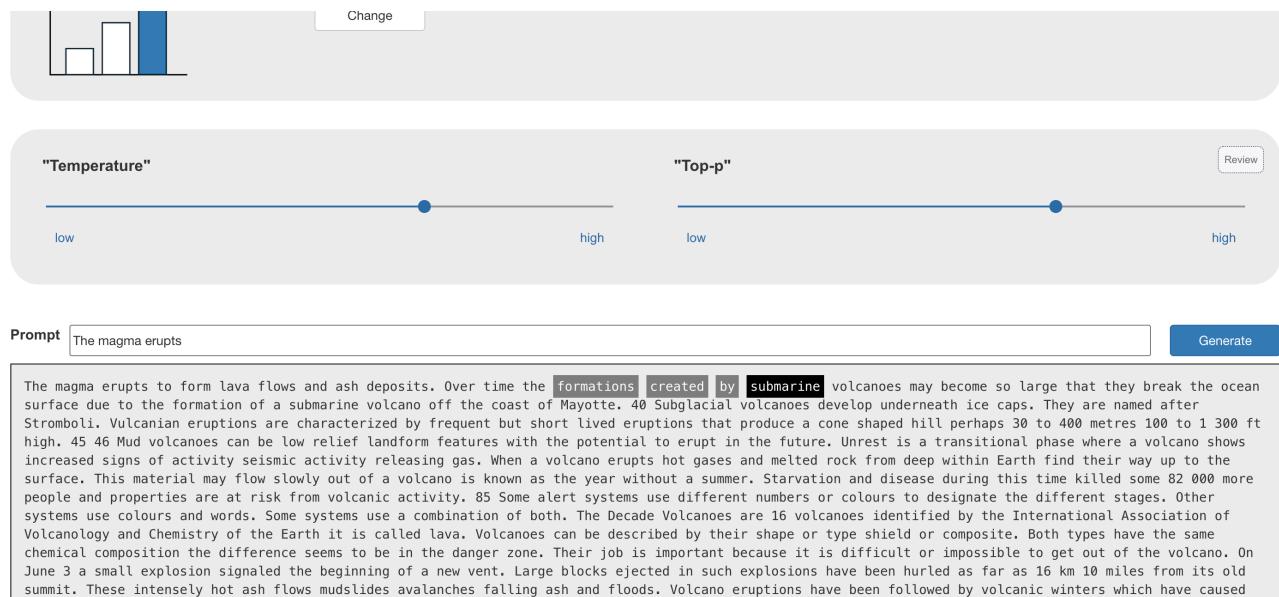
Does this text look more sensible?

Increasing word sequence lengths with a small text corpus

Clicking on a word shows that the computer is reproducing sections from the documents that I collected.

For example:

Over time the formations created by submarine volcanoes may become so large that they break the ocean surface...



This is taken from the Wikipedia page I added to my corpus.

The screenshot shows a Wikipedia article page for 'Submarine volcanoes'. The page includes a sidebar with 'Contents' and a main text area. The main text discusses the formation of submarine volcanoes, mentioning 'ice caps', 'pillow lavas', and 'tuyas'. It also refers to 'Hydrothermal vents' and 'pumice rafts'. The word 'volcanoes' is highlighted in blue, matching the highlighted phrase in the previous figure. The text is presented in a standard Wikipedia article format with headings and sub-headings.

Contents hide

(Top)

Etymology and terminology

> Plate tectonics

> Volcanic features

> Erupted material

Volcanic eruptions

> Volcanic activity

Decade volcanoes

> Volcanoes and humans

Volcanoes on other celestial bodies

History of volcano understanding

See also

compared to air), which often causes volcanic vents to form steep pillars on the ocean floor. Hydrothermal vents are common near these volcanoes, and some support peculiar ecosystems based on chemotrophs feeding on dissolved minerals. Over time, the formations created by submarine volcanoes may become so large that they break the ocean surface as new islands or floating pumice rafts.

In May and June 2018, a multitude of seismic signals were detected by earthquake monitoring agencies all over the world. They took the form of unusual humming sounds, and some of the signals detected in November of that year had a duration of up to 20 minutes. An oceanographic research campaign in May 2019 showed that the previously mysterious humming noises were caused by the formation of a submarine volcano off the coast of Mayotte.^[40]

Subglacial volcanoes

Main article: [Subglacial volcano](#)

Subglacial volcanoes develop underneath ice caps. They are made up of lava plateaus capping extensive pillow lavas and palagonite. These volcanoes are also called table mountains, tuyas,^[41] or (in Iceland) mobergs.^[42] Very good examples of this type of volcano can be seen in Iceland and in British Columbia. The origin of the term comes from Tuya Butte, which is one of the several tuyas in the area of the Tuya River and Tuya Range in northern British Columbia. Tuya Butte was the first such landform analysed and so its name has entered the geological literature for this kind of volcanic formation.^[43] The Tuya Mountains Provincial Park was recently established to protect this unusual landscape, which lies north of Tuya Lake and south of the Jennings River near the boundary with the Yukon Territory.

With such a small text collection, some 3-word sequences will be found only very few times.

You can see this when you click on words.

Prompt Generate

The magma erupts to form lava flows and ash deposits. Over time the formations created by submarine volcanoes may become so large that they break the ocean surface due to the formation of a submarine volcano off the coast of Mayotte. 40 Subglacial volcanoes develop underneath ice caps. They are named after Stromboli. Vulcanian eruptions are characterized by frequent but short lived eruptions that produce a cone shaped hill perhaps 30 to 400 metres 100 to 1 300 ft high. 45 46 Mud volcanoes can be low relief landform features with the potential to erupt in the future. Unrest is a transitional phase where a volcano shows increased signs of activity seismic activity releasing gas. When a volcano erupts hot gases and melted rock from deep within Earth find their way up to the surface. This material may flow slowly out of a volcano is known as the year without a summer. Starvation and disease during this time killed some 82 000 more people and properties are at risk from volcanic activity. 85 Some alert systems use different numbers or colours to designate the different stages. Other systems use colours and words. Some systems use a combination of both. The Decade Volcanoes are 16 volcanoes identified by the International Association of Volcanology and Chemistry of the Earth it is called lava. Volcanoes can be described by their shape or type shield or composite. Both types have the same chemical composition the difference seems to be in the danger zone. Their job is important because it is difficult or impossible to get out of the volcano. On June 3 a small explosion signaled the beginning of a new vent. Large blocks ejected in such explosions have been hurled as far as 16 km 10 miles from its old summit. These intensely hot ash flows mudslides avalanches falling ash and floods. Volcano eruptions have been followed by volcanic winters which have caused

formations created by was in your documents 1 time. The only token found immediately after "formations created by" was **submarine**

This meant there was a 100% chance that "submarine" would be chosen.

If a three-word sequence is only found once, the computer only has one option to choose from for the next word.

If this continues, the computer repeats passages from your corpus.

Increasing the length of patterns that the computer is counting improves the text that the computer can generate. But to find enough patterns to generate useful text, you will need a lot more text in the corpus.

44. If you have time, you can add more text to your **corpus** and see if that lets you generate new text with this longer “context window”.

Is this fair?

You've created a language model using a corpus of text that was written by other people. Do you think that what you have done is fair? The people who wrote that text put a lot of time into research and writing. Would they mind you using their text in this way? Should you give them some sort of credit for using their work?

For this project, this is a temporary language model, created for educational purposes, and you are the only person that will use the model. Does that make it okay?

What if you were creating a complex language model, to be available on the Internet for people to pay you to use? Would that change your answer?

People creating language models need to find as much text as they can to improve their models. This has raised a debate about what text they should be allowed to use, and what form of credit or acknowledgements the original creators should be entitled to. **What do you think?**

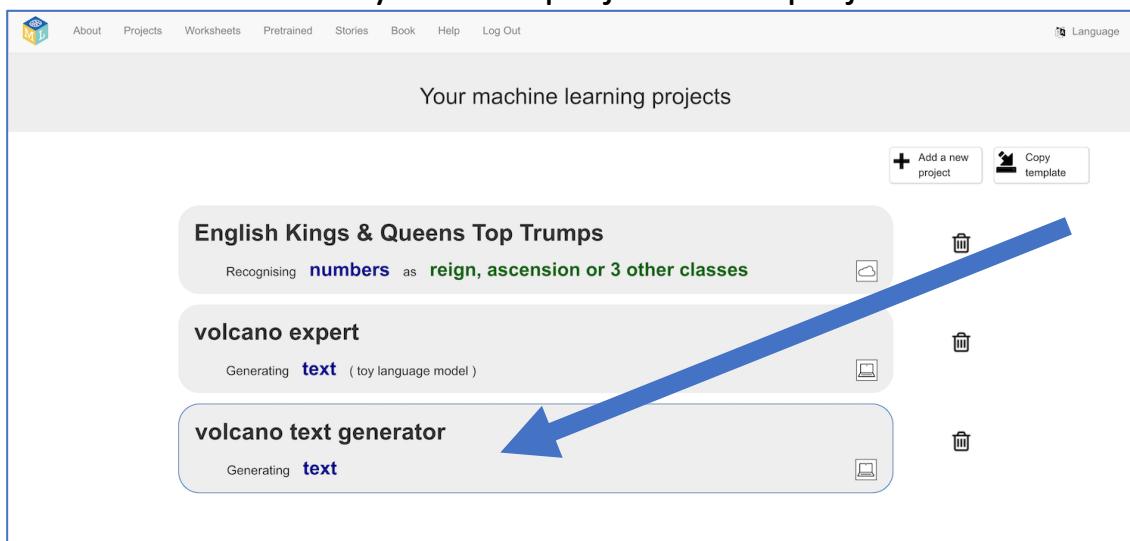
What next?

The way that your toy model used “context window”, “temperature”, and “Top-P” values aren’t exactly the same as in real language models. It is similar – and close enough to help you understand the ideas behind them.

In the next stage, you will use a real language model, using real context window, temperature, and Top-P values.

Stage 2 – Configure a small language model

- 45.** Click on “**Projects**” on the top menu bar
- 46.** Click the “**+ Add a new project**” button
- 47.** Name your project
- 48.** Set the new project to learn how to **generate text** as before
- 49.** You should see your new project in the projects list. Click on it.



- 50.** Click on **Small** and then click **Next**

51. Choose a model architecture

The screenshot shows a user interface for selecting a language model. At the top, there's a navigation bar with links for About, Projects, Worksheets, Pretrained, Stories, Book, Help, and Log Out. To the right of the navigation is a "Language" button. Below the navigation, a yellow banner reads: "If you are under the age of 13, please only use a small language model with supervision from a trusted adult. Generative AI can sometimes generate text that isn't nice or appropriate." A large blue arrow points from the "language model" tab in the "Type of language model" dropdown to the "Model architecture" dropdown below it.

Type of language model

- Toy
- Small**
- Large
- language model

Model architecture

- Smol
- Qwen
- Tiny Llama
- Llama
- Phi

There are multiple models to choose from.

"SmolLM2" (made by Hugging Face)

Using this needs a 276 MB download

"Qwen 2.5" (made by Alibaba)

Using this needs a 289 MB download

TinyLlama (made by the Singapore University of Technology & Design)

Using this needs a 625 MB download

Llama 3.2 (made by Meta)

Using this needs a 711 MB download

Phi 1.5 (made by Microsoft)

Using this needs an 806 MB download

Gemma 2 (made by Google)

Using this needs a 1.5 GB download

The larger the model you choose, the longer it will take to download and the more storage space it will need on your computer.

Larger models are also more complex to run, and you will need a powerful computer to be able to run them.

Choose a small model such as Smol to start with.

52. Click Download

53. Choose a context window size, and click **Next**

Your toy model used the last 1, 2, or 3 words to choose the next word.

Each new word it picked for the end of a sentence had no consideration of the words that were at the start of the sentence – and that produced sentences that made no sense.

These small models have been trained using the patterns from millions of documents, so they can use the last hundreds, or even thousands, of words when choosing the next word.

It can generate the next word not just based on the words in the current sentence, but your question and answers the model gave before that.

Model architecture
Smol

Size of context window
512
1024
2048
4096
8192

How much text the model looks at before generating the next word. A smaller window means it works with less context. A larger window allows it to use more, which can make results more accurate but needs more computer memory.

Next

54. Choose temperature and Top-p values using the sliders

If you are under the age of 13, please only use a small language model with supervision from a trusted adult. Generative AI can sometimes generate text that isn't nice or appropriate. X

Type of language model
Toy Small Large **language model**

Model architecture
Smol

Size of context window
512

Temperature
With a low temperature, the model will generate the next word based on what is most common.
With a high temperature, the model is more likely to pick something unexpected.

Top-p
With a low value, the model will only consider the most common candidates for the next word.
With a high value, the model will consider any word that could come next.

Next

55. Click Next

56. Choose No initial context

The screenshot shows the 'Model Architecture' configuration interface. At the top, there is a dropdown menu set to 'Smol'. Below it is a 'Size of context window' input field containing '512'. Under 'Temperature' and 'Top-p', there are two horizontal sliders; both are positioned at the 'high' end. In the 'Initial context' section, there are two buttons: 'No initial context' (which is selected) and 'Use initial context'. To the right of these buttons is a note: 'Additional text that is included with the first prompt, to provide additional knowledge that is missing from the model.' At the bottom right is a 'Next' button.

57. Click Next

58. Type a question and click Generate

The screenshot shows the 'Model Architecture' configuration interface. The 'Size of context window' is set to '512'. The 'Temperature' and 'Top-p' sliders are at their maximum 'high' value. In the 'Initial context' section, the 'No initial context' button is selected. Below this, there is a 'Prompt' input field containing 'What is the largest volcano in the world?'. A generated response is displayed in a box: 'The largest volcano in the world is Mount Nyiragongo in the Democratic Republic of the Congo. It is a volcanic complex located in the Democratic Republic of the Congo and is the world's largest cinder cone volcano with a height of over 6,500 meters. It is situated in the Kibale National Park and is approximately 411 kilometers (256 miles) southeast of Buisi Abame Village in the national capital city of Kinshasa.'

59. Check if the model returns accurate answers to factual or information-based questions

Look for reliable sources to check the model's answers

60. Try asking your question to a different model

The screenshot shows a user interface for generating text using a language model. At the top, there's a dropdown menu for 'Model architecture' set to 'Llama'. A large blue arrow points from the text above to this dropdown. Below it is a 'Size of context window' input field set to '512'. Underneath are two sliders: 'Temperature' (set to 'high') and 'Top-p' (set to 'low'). In the 'Initial context' section, there are two buttons: 'No initial context' (which is selected) and 'Use initial context'. At the bottom, there's a 'Prompt' input field containing the question 'What is the largest volcano in the world?'. To the right of the prompt are three buttons: 'Reset', 'Generate', and a 'Review' button. The generated response is displayed in a box below the prompt, stating: 'The largest volcano in the world is Mauna Loa, located in Hawaii. It's about 7 miles (11 kilometers) tall! While it's not a mountain in the classical sense, Mauna Loa is a dormant supervolcano that has been erupting for thousands of years. If it ever did erupt again, it would be the tallest volcano in the world!'.

*Do all models generate the same answers?
Are they generating accurate answers?*

Hallucinations

We ask language models to generate new text, but when it generates statements that aren't true, we call these "hallucinations". These are a problem when we use language models in the real world.

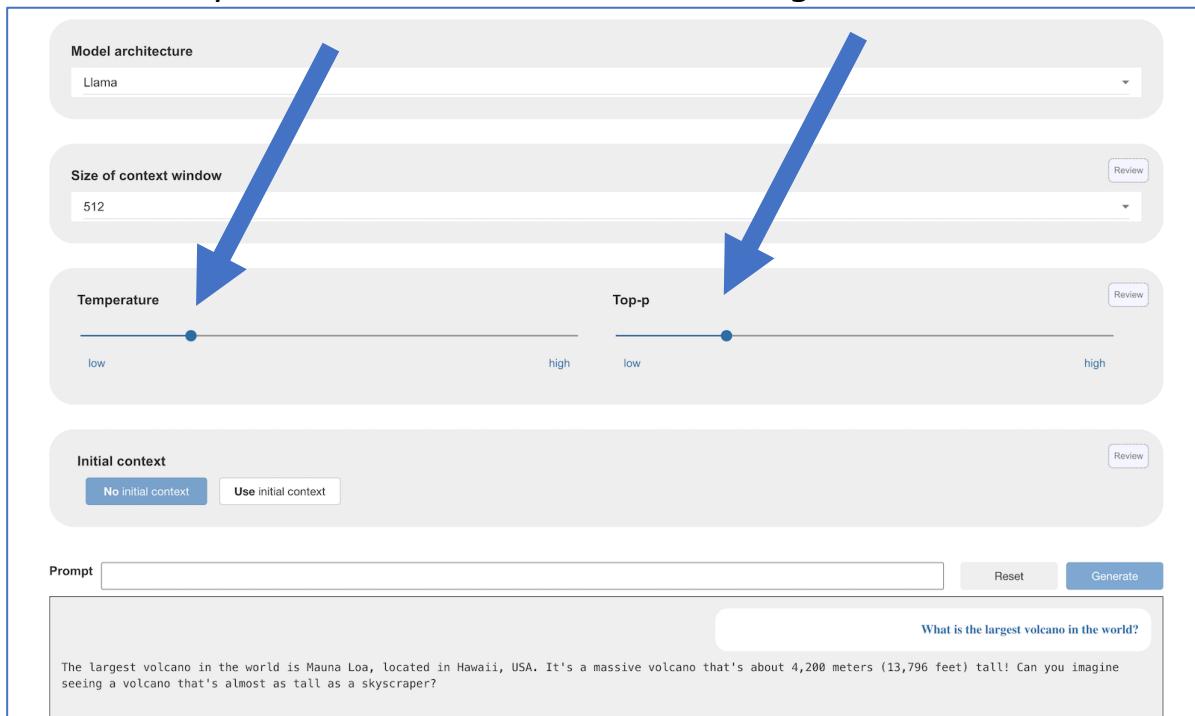
Remember that **Temperature** controls the randomness in text generation. With a high temperature, you are increasing the creativity by increasing the chance that less common, less likely words will be chosen.

For factual and knowledge-based questions, creativity isn't always ideal.

A lower temperature will make the model prioritise the most likely and most probable next words.

A lower Top-P will limit the choices to the next words that are most likely and most probable.

61. Try asking your question with a lower temperature and top-P value *Does this improve the answer that the model gives?*



Model architecture: Llama

Size of context window: 512

Temperature: A slider with a dot at the low end, labeled 'low' on the left and 'high' on the right.

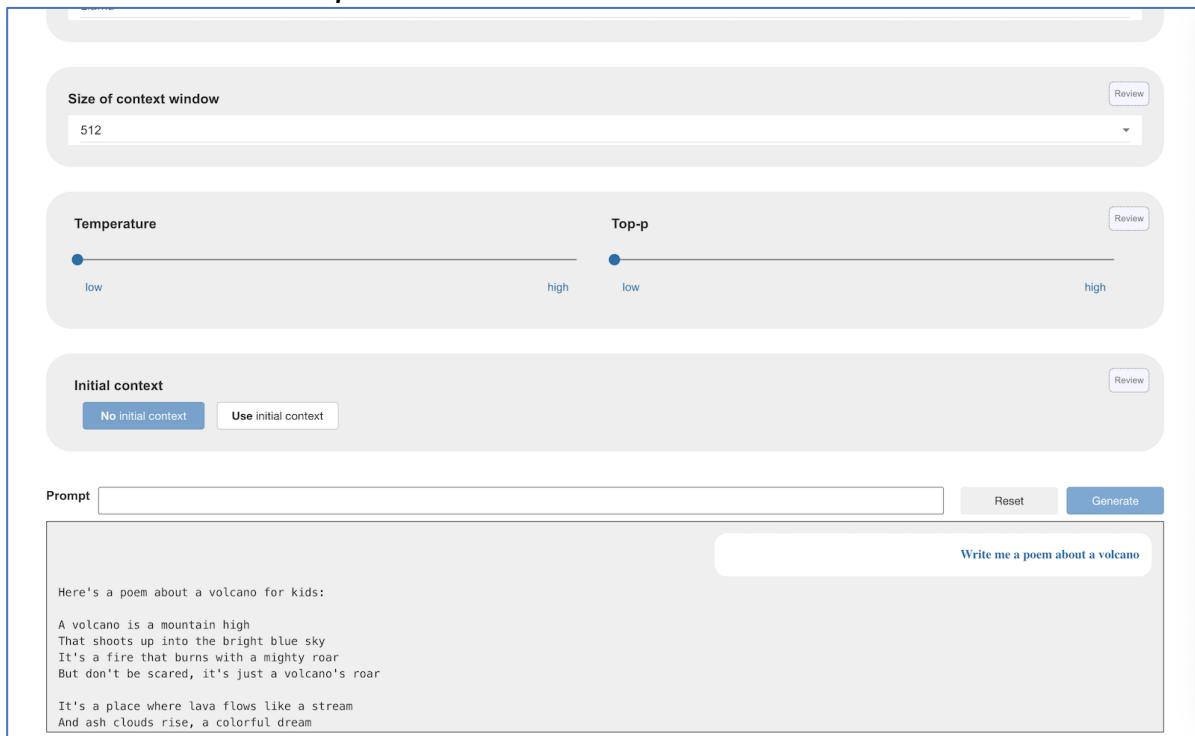
Top-p: A slider with a dot at the high end, labeled 'low' on the left and 'high' on the right.

Initial context: No initial context

Prompt: What is the largest volcano in the world?

The largest volcano in the world is Mauna Loa, located in Hawaii, USA. It's a massive volcano that's about 4,200 meters (13,796 feet) tall! Can you imagine seeing a volcano that's almost as tall as a skyscraper?

62. Try a creative writing prompt, with a **very low** temperature & Top-P *I tried "Write me a poem about a volcano"*



Size of context window: 512

Temperature: A slider with a dot at the low end, labeled 'low' on the left and 'high' on the right.

Top-p: A slider with a dot at the low end, labeled 'low' on the left and 'high' on the right.

Initial context: No initial context

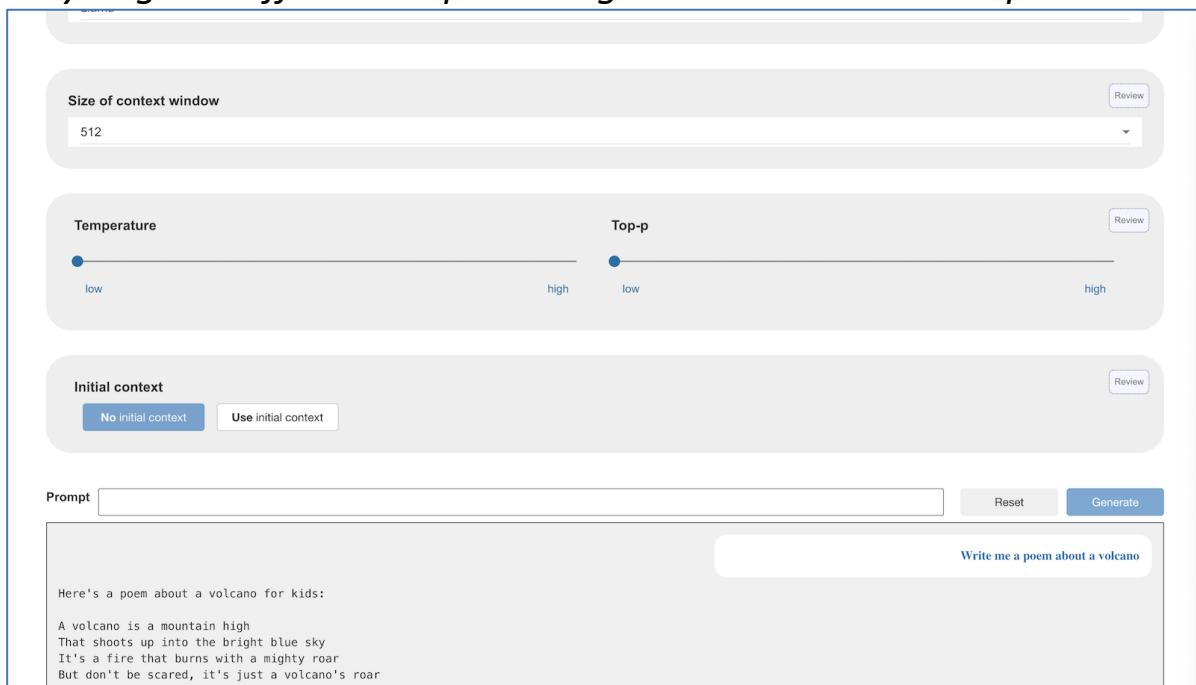
Prompt: Write me a poem about a volcano

Here's a poem about a volcano for kids:

A volcano is a mountain high
That shoots up into the bright blue sky
It's a fire that burns with a mighty roar
But don't be scared, it's just a volcano's roar

It's a place where lava flows like a stream
And ash clouds rise, a colorful dream

63. Click on **Reset** and then enter the same prompt again
Do you get a different response? I got an almost-identical poem

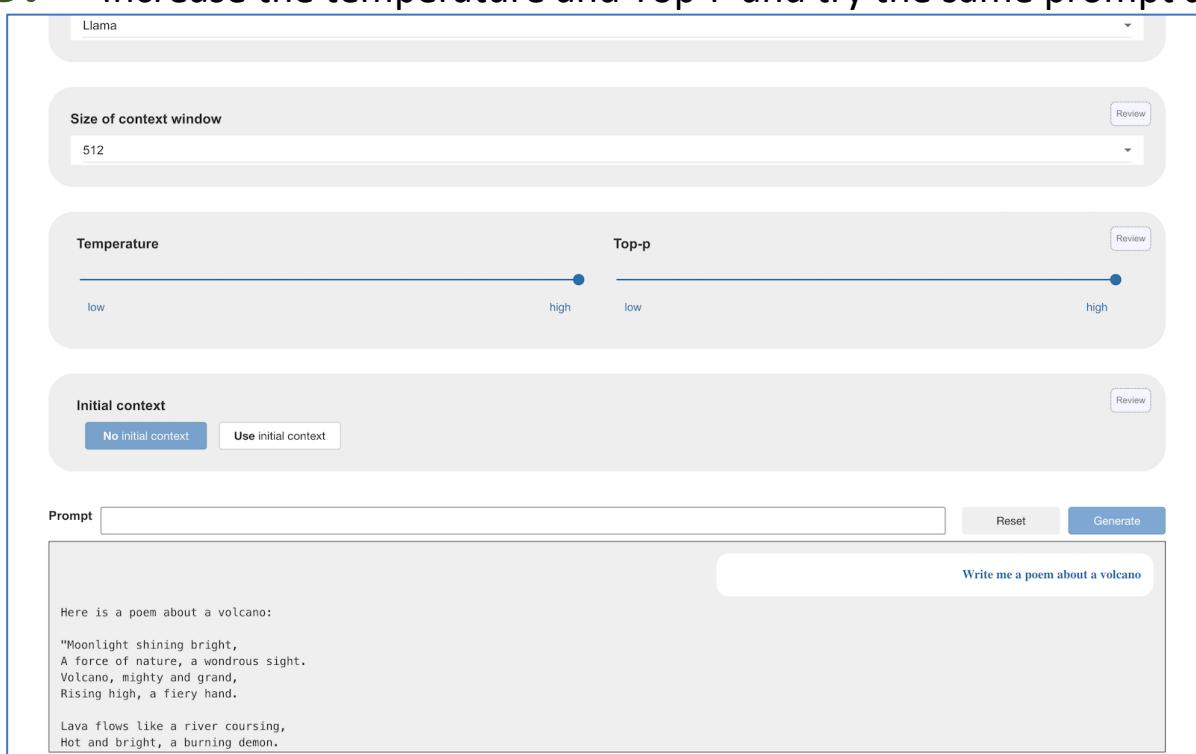


The screenshot shows the Llama AI interface with the following settings:

- Size of context window:** 512
- Temperature:** Set to low
- Top-p:** Set to low
- Initial context:** No initial context
- Prompt:** Write me a poem about a volcano
- Output:** A poem about a volcano for kids:
A volcano is a mountain high
That shoots up into the bright blue sky
It's a fire that burns with a mighty roar
But don't be scared, it's just a volcano's roar

64. Click on **Reset**

65. Increase the temperature and Top-P and try the same prompt again



The screenshot shows the Llama AI interface with the following settings:

- Size of context window:** 512
- Temperature:** Set to high
- Top-p:** Set to high
- Initial context:** No initial context
- Prompt:** Write me a poem about a volcano
- Output:** A poem about a volcano:
Here is a poem about a volcano:
"Moonlight shining bright,
A force of nature, a wondrous sight.
Volcano, mighty and grand,
Rising high, a fiery hand.

Lava flows like a river coursing,
Hot and bright, a burning demon."

66. Click **Reset** and try that again

Creativity

Temperature controls the randomness in text generation. With a high temperature, you are increasing the creativity by increasing the chance that less common, less likely words will be chosen.

Top-P controls the candidates that your model has to choose from. With a high Top-P, you are increasing the range of words that the model has to choose from.

For creative writing tasks, or tasks where you want the language model to inspire new ideas, this increase in randomness is beneficial.