



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 list of files: kids-britannica.txt, aust.txt, nasa.txt, wizkids.txt, twinkl.txt, kiddle.txt, nhmu.txt. Below the list are buttons for '+ Add text', 'Add file', and 'W Wikipedia'.

"Size of context window"

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

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.

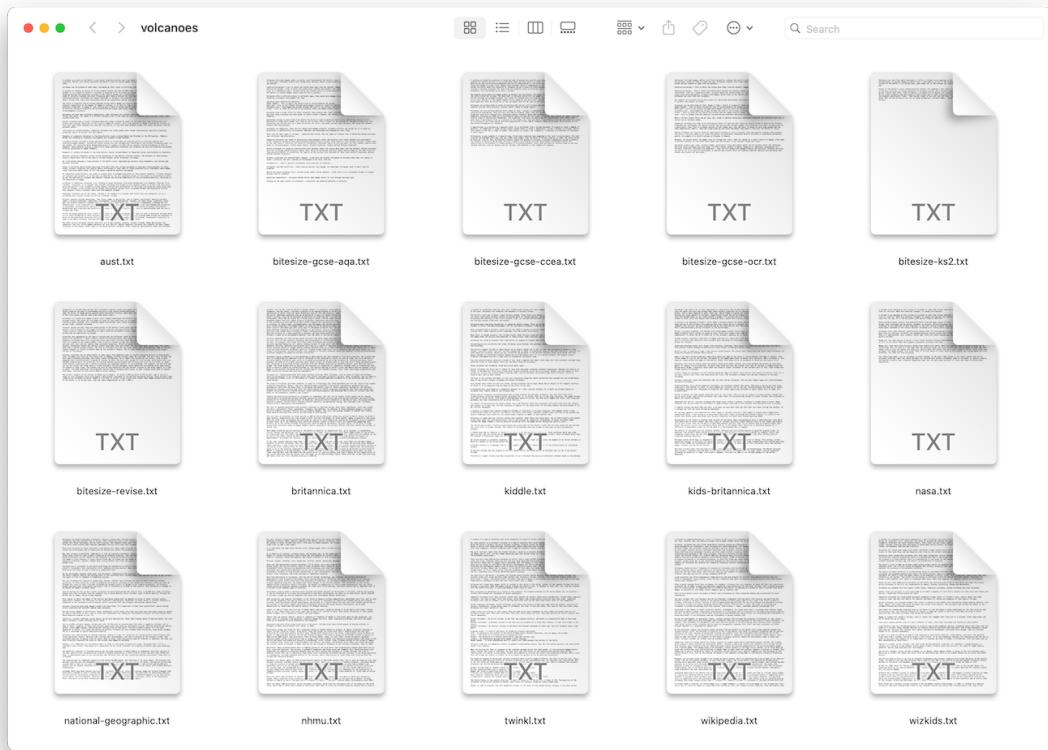
For my volcano model, I found useful text in places like:

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?

Storage in your web browser removes limits on how big your project can be. Storage 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

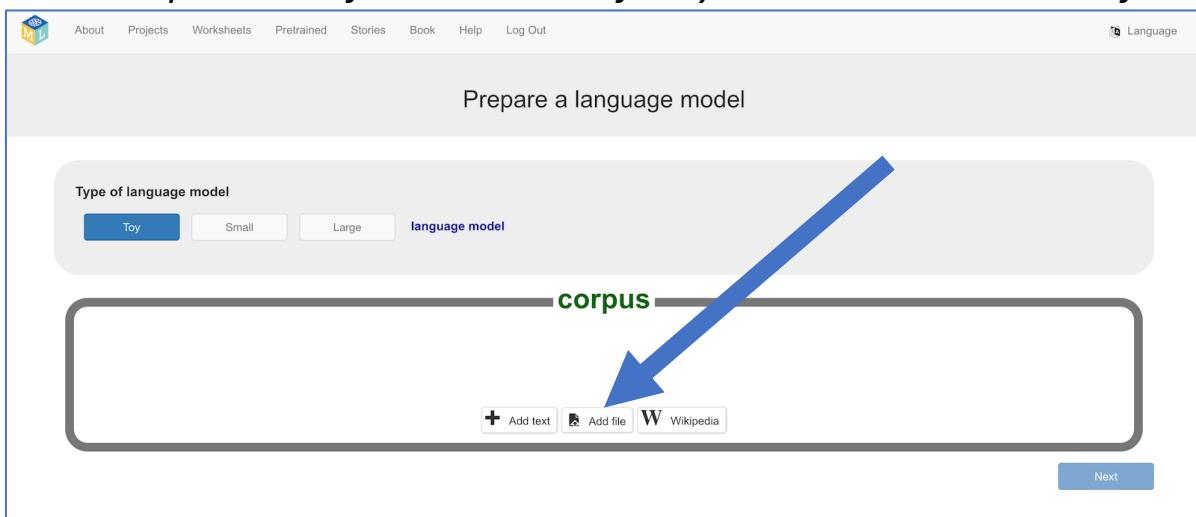
13. Click on **Toy** and then click **Next**

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.

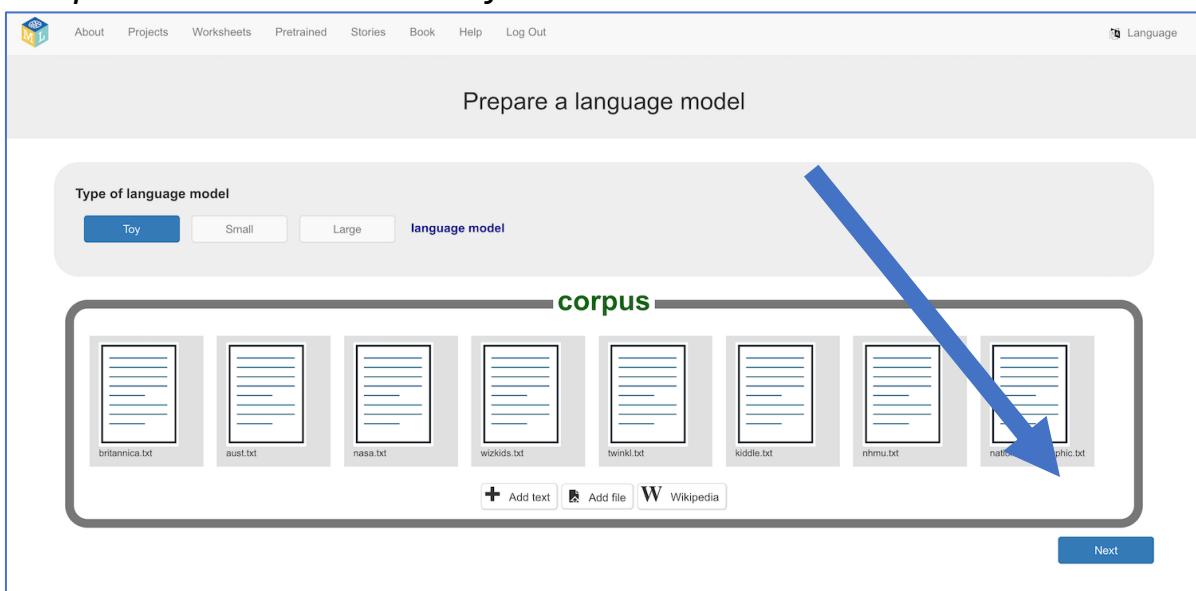
15. Use **Add file** to upload the text files you've collected

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

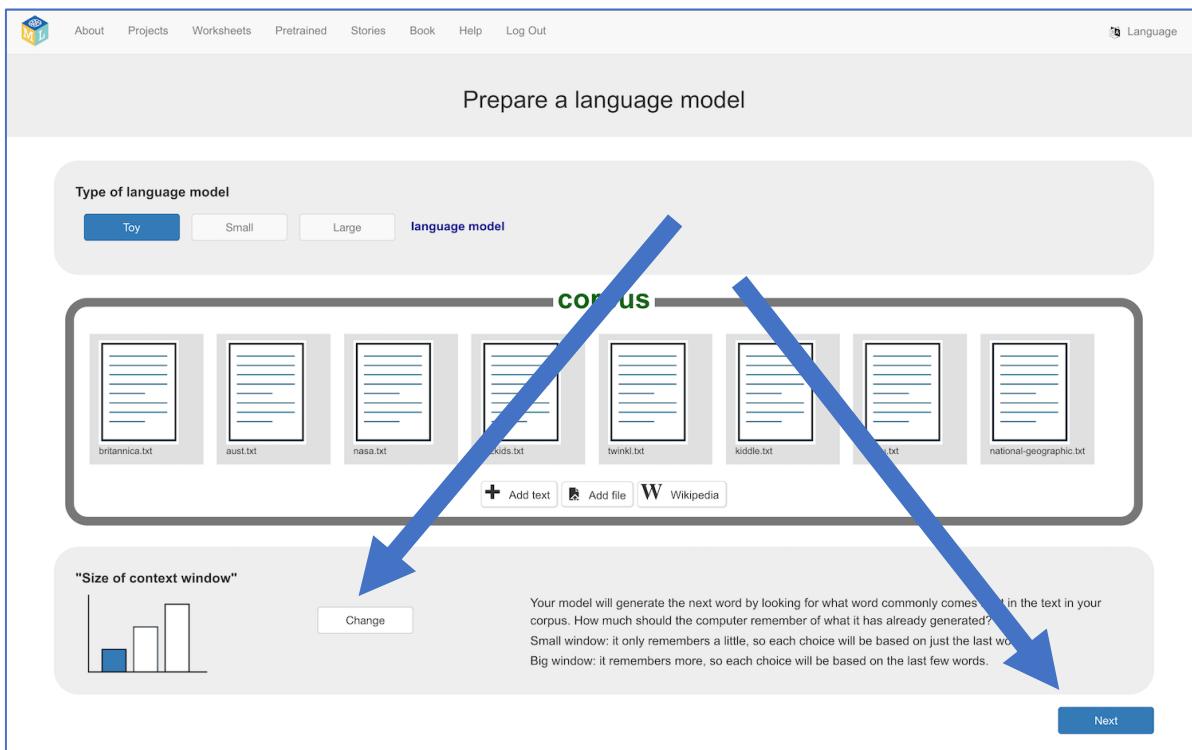


16. Click **Next**

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, kiddle.txt, nhmu.txt, and national-geographic.txt. Below this is a toolbar with buttons for '+ Add text', 'Add file', and 'Wikipedia'.

A section titled "Size of context window" contains a small bar chart with three bars of increasing height, followed by a "Change" button. A descriptive text explains that the model generates the next word by looking at what commonly comes next in the text, with options for a "Small window" (remembering just the last word) or a "Big window" (remembering more words).

The main area is titled "sample of the tokens". It shows a list for the word "the" (1642 occurrences). The top item in the list is "surface (67)", which is highlighted. Other items in the list include "volcano (53)", "Earth (49)", "world (37)", "magma (35)", and "eruption (28)".

At the bottom, a status bar indicates "the surface found 67 times in your corpus". A "Next" button is located in the bottom right corner.

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. At the top, there is a row of document icons with labels: kids-britannica.txt, aust.txt, nasa.txt, wizkids.txt, twinkl.txt, kiddle.txt, nhmu.txt, and national-geographic.txt. Below this is a toolbar with buttons for '+ Add text', 'Add file', and 'Wikipedia'.

A section titled "Size of context window" contains a small bar chart with three bars of increasing height, followed by a "Change" button. A descriptive text explains that the model generates the next word by looking at what commonly comes next in the text, with options for a "Small window" (remembering just the last word) or a "Big window" (remembering more words).

The main area is titled "sample of the tokens". It shows a list for the word "The" (269 occurrences). The top item in the list is "largest (6)", which is highlighted. Other items in the list include "magma (6)", "java (6)", "eruption (5)", "most (5)", and "ash (5)".

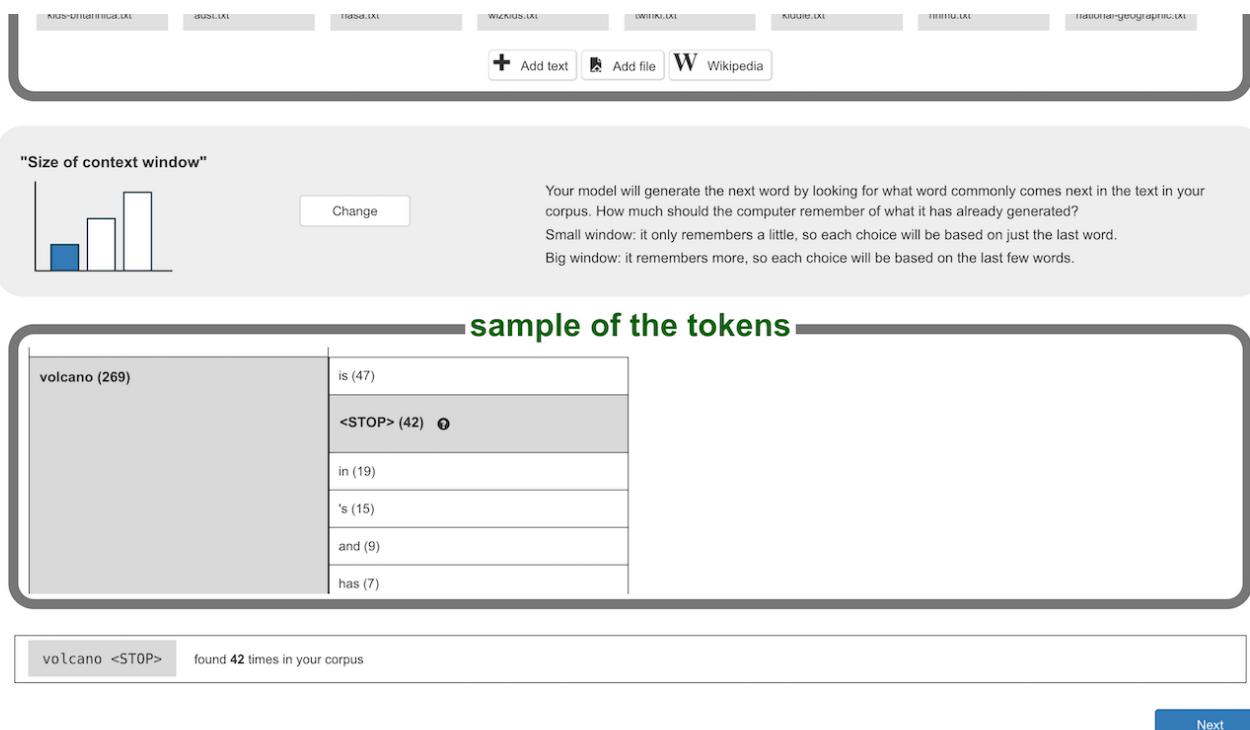
At the bottom, a status bar indicates "The largest found 6 times in your corpus". A "Next" button is located in the bottom right corner.

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.



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?

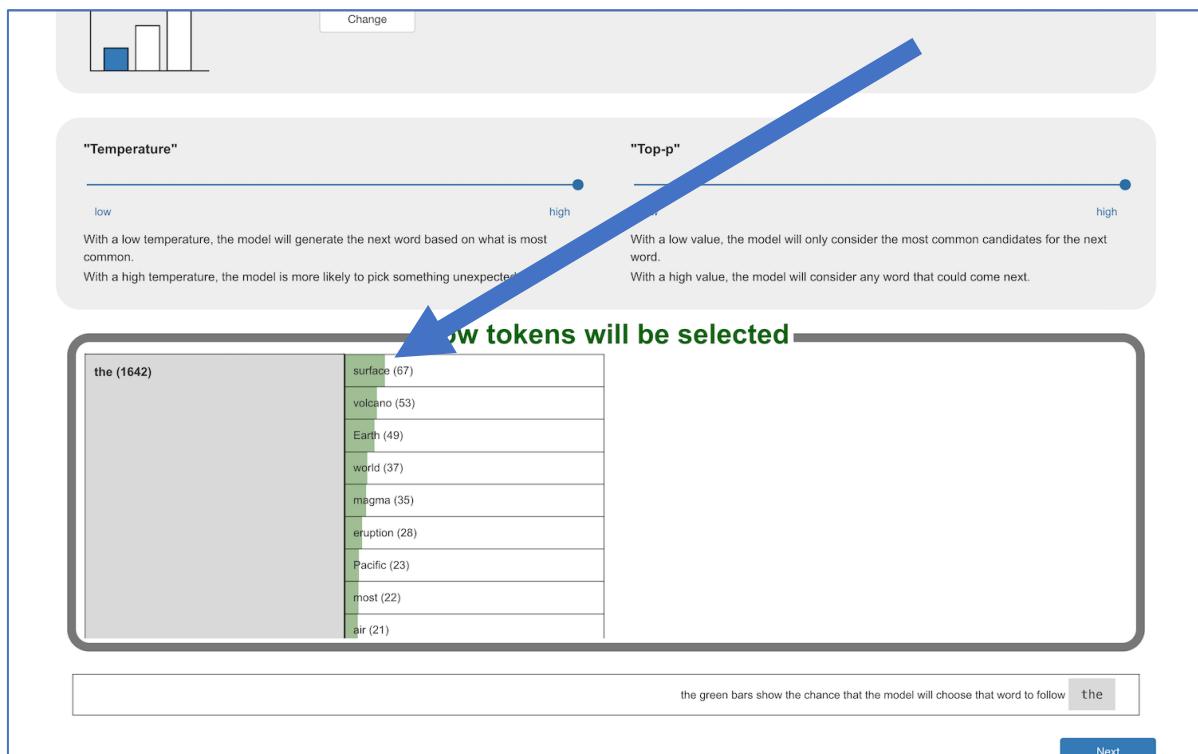
For any word that the computer has found in your text, it now has a record of what words normally come next, and how often that happens.

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 sliders: "Temperature" (low to high) and "Top-p" (low to high). Below the sliders is a table titled "how tokens will be selected" showing word probabilities:

Token	Probability
the (49)	49
volcanic (18)	18
a (14)	14
ash (14)	14
is (13)	13
in (12)	12
lava (12)	12
steam (10)	10
other (9)	9

A large blue arrow points from the "Top-p" slider area down towards the "Next" button at the bottom right.

25. Generate some text!

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

The screenshot shows a user interface for generating text. It includes a "Size of context window" histogram, two sliders for "Temperature" (low to high) and "Top-p" (low to high), and a text input field labeled "Prompt" containing "The". A large blue arrow points from the "Prompt" field down towards the "Generate" button at the bottom right.

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.

The screenshot shows a user interface for generating text. At the top, there's a small bar chart icon and a 'Change' button. Below that are two horizontal sliders: one for 'Temperature' (from low to high) and one for 'Top-p' (from low to high). A large blue arrow points from the 'Temperature' slider towards the text area. The text area contains a paragraph about volcanoes, with the words 'seafloor' and 'sediments' highlighted in black boxes. Below the text area, a status bar says 'seafloor was in your documents 8 times. 6 different tokens were found immediately after it.' and 'sediments was after "seafloor" 2 times.' At the bottom, it says 'Your temperature and top-p choices meant that there was a 25% chance that it would be chosen.' A 'Prompt' input field and a 'Generate' button are at the very bottom.

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.

This screenshot shows the same interface as the previous one, but with very low 'Temperature' and 'Top-p' settings, resulting in a repetitive loop of the word 'The'. The text area is filled with 'The lava flows. The lava flows. The magma. The magma.' repeated many times. The word 'largest' is highlighted in a grey box, and 'active' is highlighted in a black box. Below the text area, a status bar says 'largest was in your documents 24 times. 13 different tokens were found immediately after it.' and 'active was after "largest" 4 times.' A large blue arrow points from the text area towards the 'active' highlight. The bottom status bar says 'Your temperature and top-p choices meant that there was a 100% chance that it would be chosen.'

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". To the right of the sliders is a small "Review" button. Below the sliders is a "Prompt" field containing the word "The". To the right of the prompt is a "Generate" button. The main area contains a large block of generated text about volcanoes, which is mostly nonsensical. Below this text, a smaller box highlights the word "world" in a black box, stating it was used 37 times after "the". A note at the bottom 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"

low high

low high

Prompt

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 twinkl.txt kiddle.txt nhmu.bd national-geographic.txt

+ Add text W 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

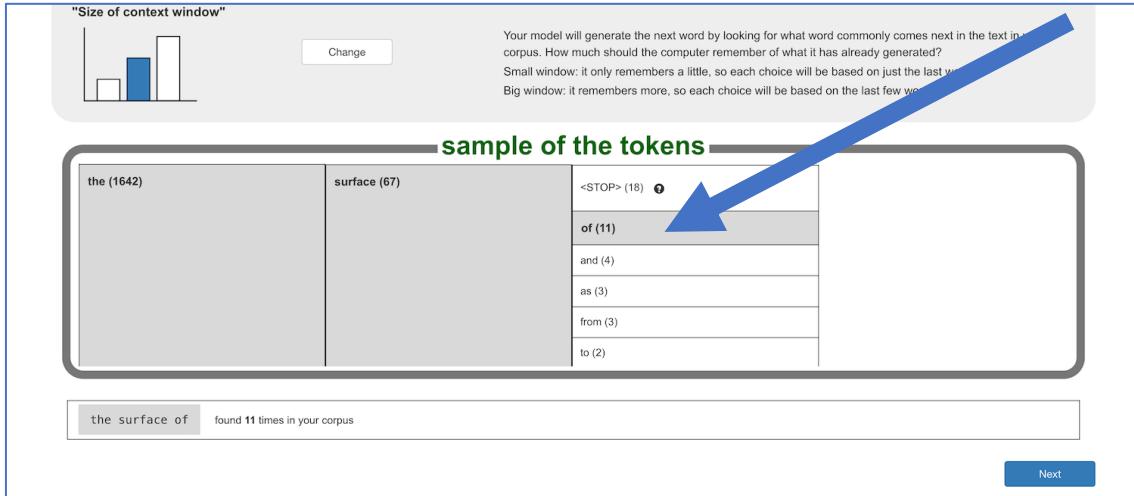
- 31.** Look through the sample of the tokens found in your documents as you did before.

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.*

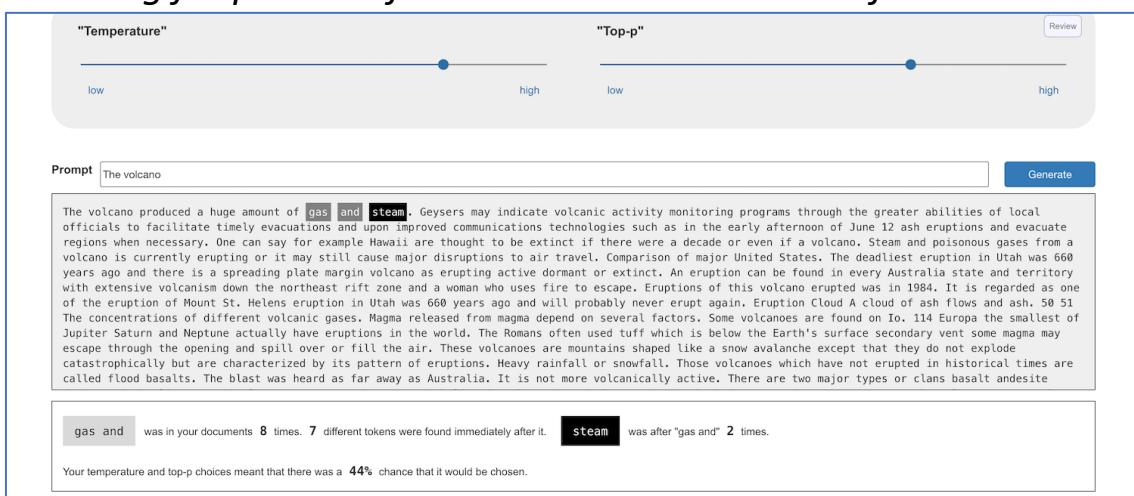


33. Find the most common two-word-sequence in your text

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.



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 a machine learning model. At the top, there's a section titled "Size of context window" with a bar chart and a "Change" button. Two large blue arrows point from the text below to these elements. Below this are sections for "Temperature" and "Top-p", each with a slider from "low" to "high". A "Prompt" field contains "The volcano" and a "Generate" button. The main area displays a generated text sample about volcanoes, mentioning Mount Tambora and magma storage systems. At the bottom, it shows token statistics: "volcanic activity" was found 38 times and "similar" was found 1 time. It also notes a 3% chance of choosing "similar".

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

This screenshot shows a "sample of the tokens" section. It displays a table where words are listed along with their frequencies in parentheses. The table has four columns: the first two are large gray boxes, the third is a narrow column, and the fourth is a wider column. The table entries are:

the (1641)	surface (67)	<STOP> (18) ⓘ	the (6)
		of (11)	a (3)
			Mars (1)
		and (4)	Earth (1)

Below the table, a message says "the surface of the 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.

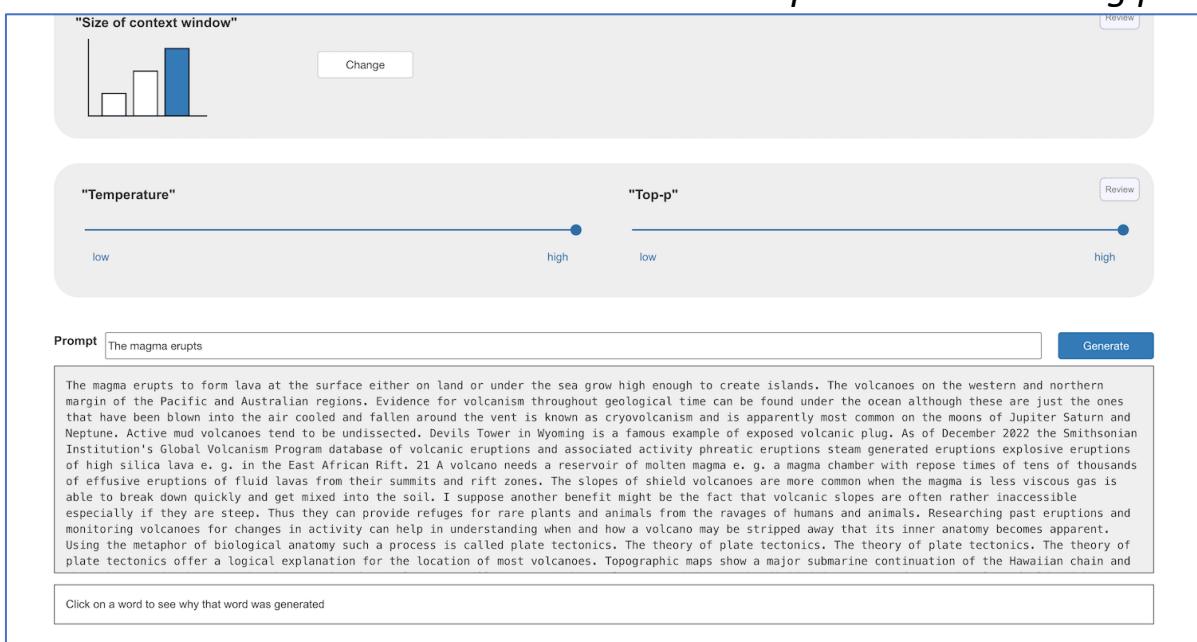
38. Look at the numbers of times these 3-word sequences were found
Look at the number of times the computer found words that followed
these three-word sequences

*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...

The screenshot shows a user interface for generating text. At the top, there's a bar chart icon and a 'Change' button. Below that are two horizontal sliders: one for 'Temperature' ranging from 'low' to 'high', and another for 'Top-p' also ranging from 'low' to 'high'. A 'Review' button is next to the second slider. At the bottom, there's a 'Prompt' field containing 'The magma erupts' and a 'Generate' button. The main area displays a generated text block about submarine volcanoes, which includes several redacted words (shown as brackets) that likely correspond to the terms being highlighted in the original sentence above.

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

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

The screenshot shows a portion of a Wikipedia article about submarine volcanoes. The page includes a sidebar with navigation links like 'Contents', 'Top', 'Etymology and terminology', and various sections on volcanic features, material, activity, and history. The main content discusses the formation of submarine volcanoes, mentioning the formation of steep pillars on the ocean floor and the presence of hydrothermal vents. It also notes the formation of new islands or floating pumice rafts. The text then highlights a specific sentence: '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.' This sentence is highlighted in blue, matching the redacted text in the previous screenshot. Below this, the article continues to discuss seismic signals, oceanographic research, and the formation of subglacial volcanoes, mentioning the Tuya Butte and Tuya Range in British Columbia.

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.

This is because with such a small collection of text, each three-word sequence is found a very small number of times.

You can see this when you click on words.

Prompt The magma erupts 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”.

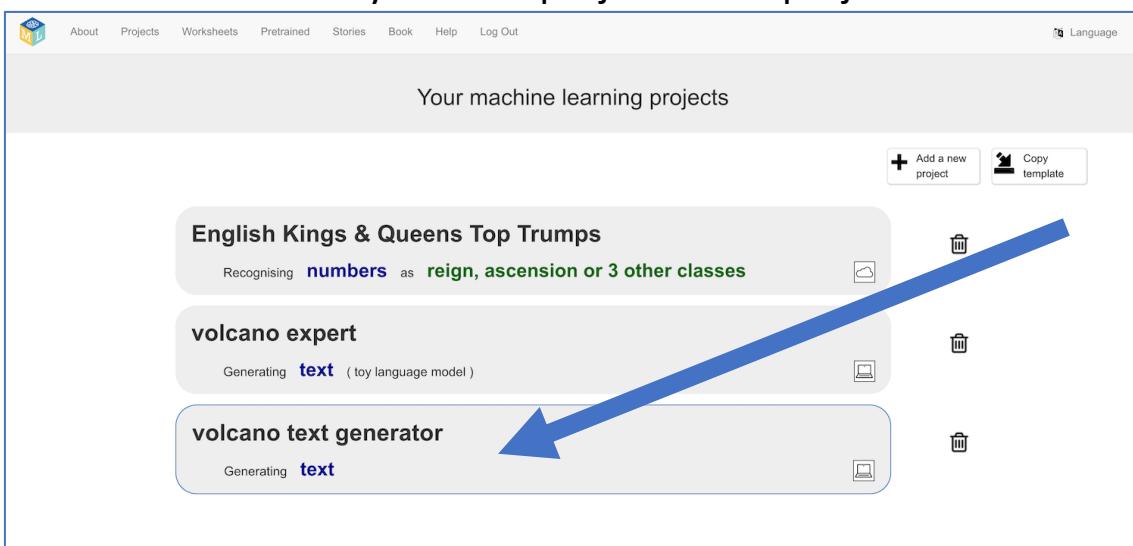
What next?

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.

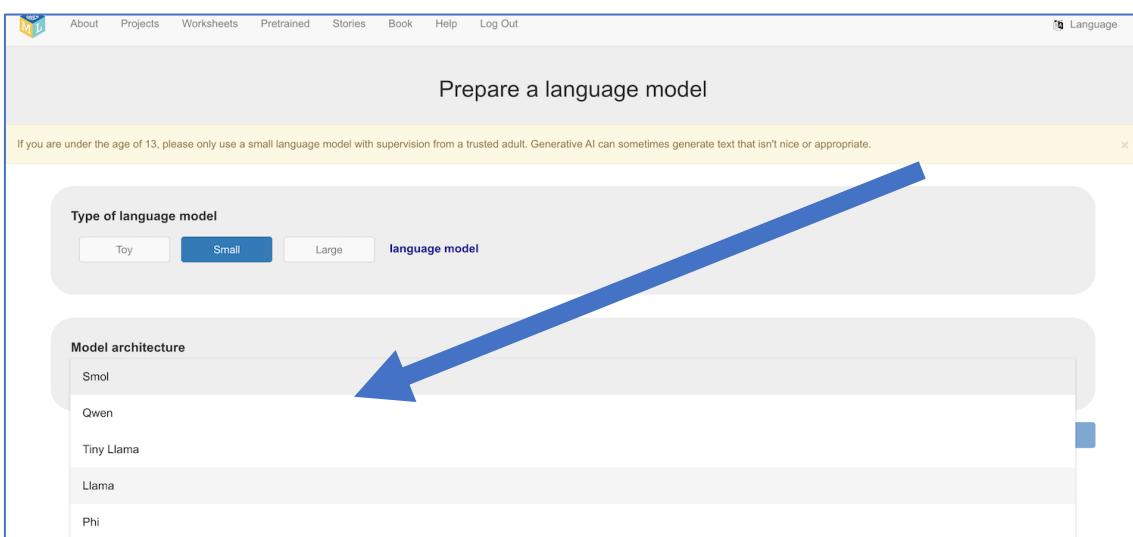
In the next stage, you will use a real language model, using real context window, temperature, and Top-P values.

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 and set it to learn how to generate text.
48. You should see your new project in the projects list. Click on it.



49. Click on **Small** and then click **Next**
50. Choose a model architecture



There are multiple models to choose from.

“SmolLM2” (made by Hugging Face) – 276 MB download

“Qwen 2.5” (made by Alibaba) – 289 MB download

TinyLlama (made by Singapore Uni. of Technology & Design) – 625 MB download

Llama 3.2 (made by Meta) – 711 MB download

Phi 1.5 (made by Microsoft) – 806 MB download

Gemma 2 (made by Google) – 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.

The 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.

51. Click Download

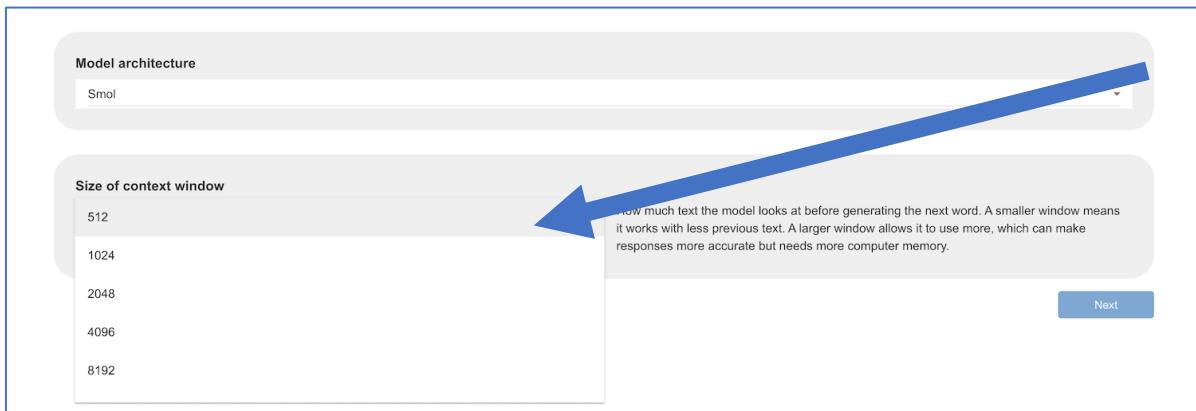
52. Choose a context window size, and click Next

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 in your question and the last answer the model gave before that.



53. Choose temperature and Top-P values using the sliders

Prepare a language model

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.

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

Model architecture
Smol

Size of context window
512

Temperature Top-p

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.

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



54. Click Next

55. Type a question and click Generate

Model architecture
Smol

Size of context window
512

Temperature Top-p

Prompt What is the largest volcano in the world?

The world's largest volcano is Mount Tambora, located in Indonesia. It's roughly 382 kilometers (240 mi) in diameter and is home to over 640 cubic kilometers (~145 cubic miles) of magma and is known for its frequent and intense eruptions. However, this is not considered the largest volcano in the world because it's actually the largest volcano that has risen in the Earth's crust, also known as a stratovolcano.

Reset **Generate**

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

The screenshot shows a user interface for an AI model. At the top, there are two horizontal sliders: 'Temperature' and 'Top-p'. Both sliders have a midpoint between 'low' and 'high'. Below the sliders is a 'Prompt' input field containing the question: "The world's largest volcano is Mount Tambora, located in Indonesia. It's roughly 382 kilometers (240 mi) in diameter and is home to over 640 cubic kilometers (~145 cubic miles) of magma and is known for its frequent and intense eruptions. However, this is not considered the largest volcano in the world because it's actually the largest volcano that has risen in the Earth's crust, also known as a stratovolcano." To the right of the prompt are 'Reset' and 'Generate' buttons. A callout box labeled "Tambora is known for having the largest eruption in history, but what is actually the largest volcano?" is overlaid on the response area. The response area itself contains the text: "Mount Tambora is actually the largest volcano in the world that has risen in the Earth's crust, which is the stage of volcanic processes where lava erupts. The volcano I mentioned earlier, Mount Tambora, is actually an Everetic stratovolcano, and it has been under the Earth's crust for millions of years, rising right up until today. So the largest volcano in the world is actually Mount Huasca in Peru, which rose in 12,595 years ago and has been erupting ever since."

57. Try asking your question to a different model

This screenshot shows a second instance of an AI model interface. It includes a 'Model architecture' dropdown set to 'Llama', a 'Size of context window' input field with the value '512', and the same 'Temperature' and 'Top-p' sliders as the first interface. A large blue arrow points from the top of the page down to the 'Model architecture' dropdown. The 'Prompt' input field contains the question: "What is the largest volcano in the world?". The response area below contains the text: "The largest volcano in the world is K2, located on the border of Pakistan and China. It's a symmetrical volcano, which means it looks the same from the two sides, and it stands at an impressive height of 1,014 meters (3,361 feet) above the surrounding land. K2 was my okayamosoi challenge to climb the world's highest and second-highest mountains. However, I didn't quite reach the very top!"

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.

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

The screenshot shows a user interface for generating text. At the top, there are dropdown menus for 'Model architecture' (set to 'Llama') and 'Size of context window' (set to '512'). Below these are two slider controls: 'Temperature' (set to 'low') and 'Top-p' (set to 'high'). A 'Prompt' input field contains the question 'What is the largest volcano in the world?'. The main text area displays the response: '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?' There are also 'Reset' and 'Generate' buttons at the bottom of the prompt input field.

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

Model architecture

Llama

Size of context window

512

Temperature

Top-p

Prompt

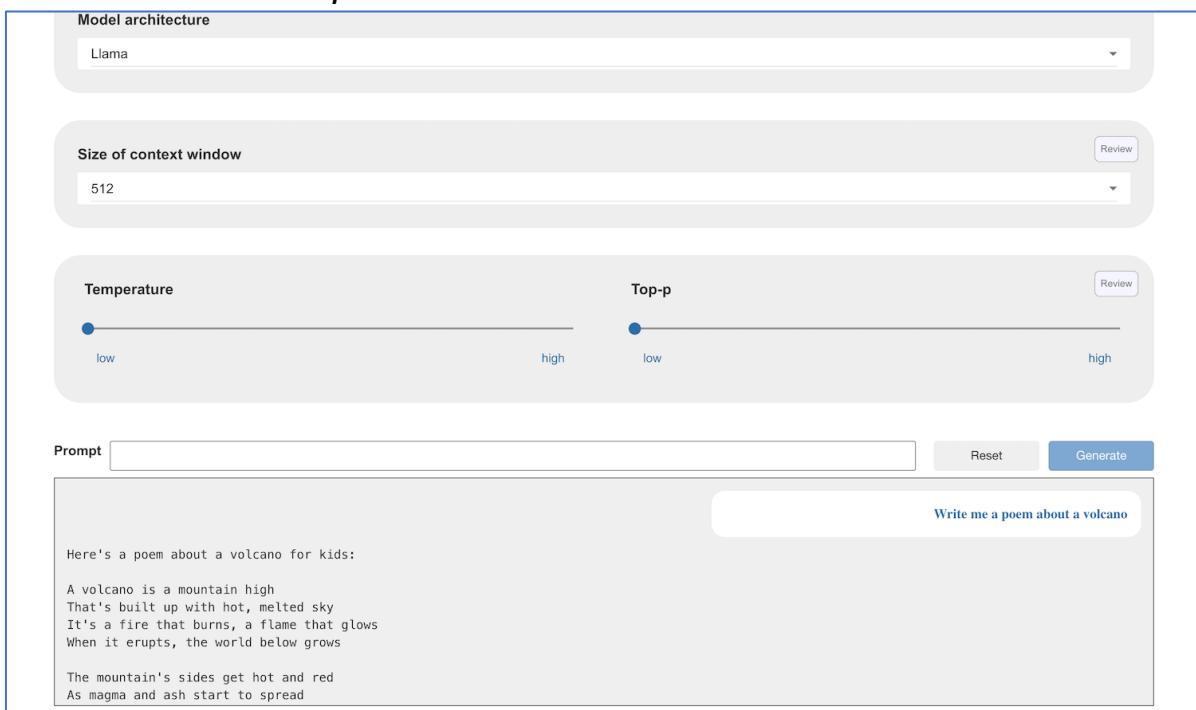
Reset Generate

Write me a poem about a volcano

Here's a poem about a volcano for kids:

A volcano is a mountain high
That's built up with hot, melted sky
It's a fire that burns, a flame that glows
When it erupts, the world below grows

The mountain's sides get hot and red
As magma and ash start to spread



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

Model architecture

Llama

Size of context window

512

Temperature

Top-p

Prompt

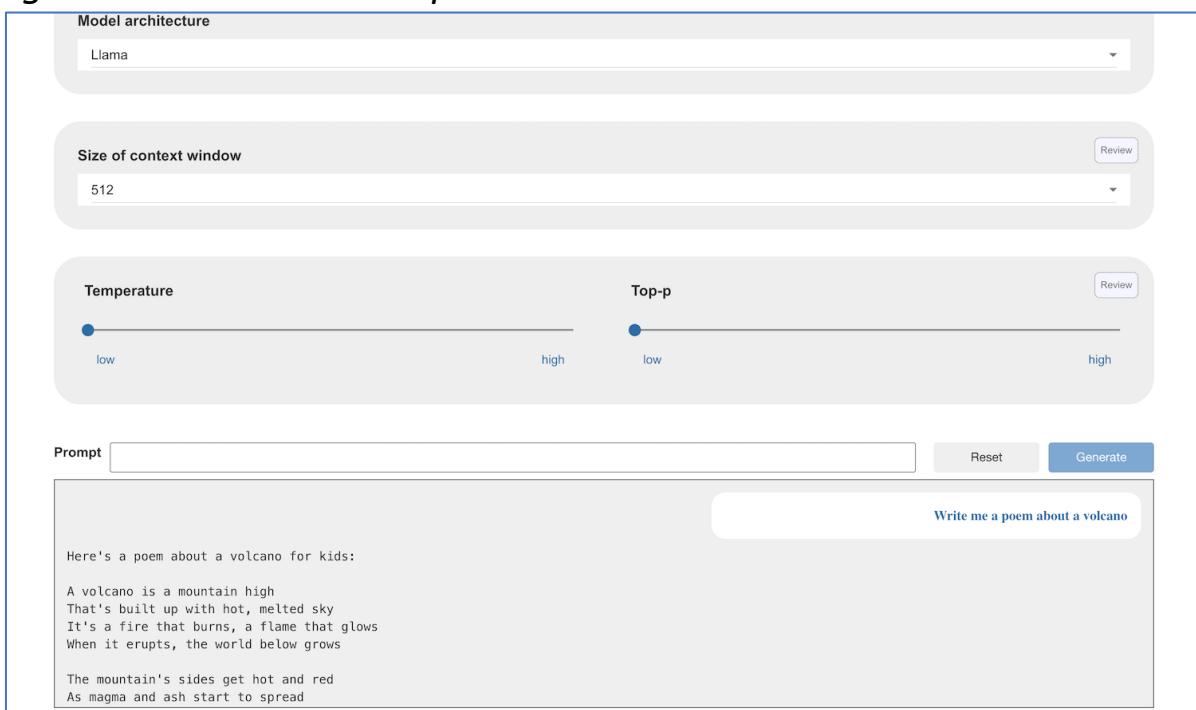
Reset Generate

Write me a poem about a volcano

Here's a poem about a volcano for kids:

A volcano is a mountain high
That's built up with hot, melted sky
It's a fire that burns, a flame that glows
When it erupts, the world below grows

The mountain's sides get hot and red
As magma and ash start to spread



61. Click on Reset

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

The screenshot shows the 'Model Architecture' interface with the 'Llama' model selected. The 'Size of context window' is set to 512. The 'Temperature' slider is set to 'high', and the 'Top-p' slider is also set to 'high'. A 'Prompt' input field contains the text 'Write me a poem about a volcano'. Below the input field, a generated poem is displayed:

Here's a little poem about a volcano:
Volcano's mighty roar, so strong and so bold,
Loud and clear, it's a sight to behold.
Molten lava flows, a fiery stream,
Rumbling and shaking, it's a volcanic dream.
As the mountain's top, it grows quite tall,
Mokehowners abound, and ashes fall.

63. 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.