

Inside GPT

The Maths Behind the Magic

Alan Smith

Active Solution

youtube.com/@CloudCastsAlanSmith



Alan Smith



- AI/ML Developer, Trainer , Mentor, Evangelist
- Active Solution Stockholm
- Microsoft AI MVP
 - Formally BizTalk, Connected Systems Developer, Integration & Azure MVP
- Community Leader
 - CloudBurst Conference
 - AIBurst Conference
 - Global Bootcamps
 - International Presenter
 - Coder Dojos
 - YouTuber
- <https://www.youtube.com/@CloudCastsAlanSmith>

<https://www.linkedin.com/in/alan-smith-68a8491/>



Alan Smith

AI Developer, Trainer, Mentor & Evangelist at Active Solution,
Microsoft AI MVP

Greater Stockholm Metropolitan Area · [Contact info](#)

[1,199 followers](#) · [500+ connections](#)

Open to

Add profile section

More



Active Solution



De Montfort University



CloudCasts - Alan Smith

@CloudCastsAlanSmith · 1.22K subscribers · 65 videos

Alan Smith is a developer, trainer, mentor and evangelist at Active Solution in Stockholm, S... >

Subscribe



Videos ▶ Play all



GPT vs. StarCraft II – 5 –
Collecting Vespene and...

198 views · 5 months ago



GPT vs. StarCraft II – 4 –
Prompt Engineering &...

2.1K views · 5 months ago



GPT vs. StarCraft II – 3 –
Testing the GPT Bot

1.3K views · 6 months ago



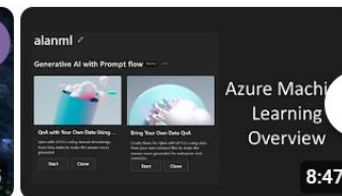
GPT vs. StarCraft II – 2 –
Creating the Prompt

1.3K views · 6 months ago



GPT vs. StarCraft II – 1 –
Introduction

2.3K views · 6 months ago



Azure Machine Learning:
Overview

74 views · 7 months ago

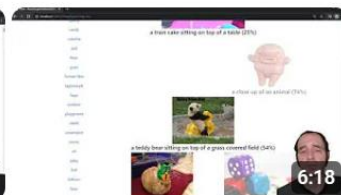
Azure Cognitive Search ▶ Play all



Azure Cognitive Search 01 –
Indexing Photos

CloudCasts - Alan Smith

823 views · 3 years ago



Azure Cognitive Search 02 –
Image Tags

CloudCasts - Alan Smith

359 views · 3 years ago



Azure Cognitive Search 03 –
Celebrity Detection

CloudCasts - Alan Smith

214 views · 3 years ago



Azure Cognitive Search vs.
Elvis Impersonators

CloudCasts - Alan Smith

104 views · 3 years ago

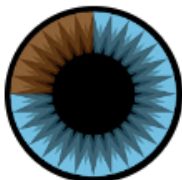


Azure Computer Vision using
C#

CloudCasts - Alan Smith

3.4K views · 2 years ago

Thanks for Explanations & Inspiration!



3Blue1Brown ✓

@3blue1brown · 6.25M subscribers · 177 videos

My name is Grant Sanderson. Videos here cover a variety of topics in math, or adjacent fields. >

3blue1brown.com and 7 more links



Code Bullet ✓

@CodeBullet · 3.03M subscribers

Just an idiot with a computer science degree trying his best.



Andrej Karpathy

@AndrejKarpathy · 235K subscribers

My motivation for creating this channel: <https://twitter.com/karpathy/status/1577746577463967745> FAQ Q: How can I pay you?

Understanding word vectors

... for, like, actual poets. By [Allison Parrish](#)



StatQuest with Josh Starmer ✓

@statquest · 1M subscribers

Statistics, Machine Learning and Data Science can sometimes seem like very scary topics, but since each technique is really just ...



Normalized Nerd

@NormalizedNerd · 99.5K subscribers · 89 videos

Hello, people from the future! Welcome to Normalized Nerd! ...more

buymeacoffee.com/normalizednerd and 4 more links



sentedex ✓

@sentedex · 1.27M subscribers

Python Programming tutorials, going further than just the basics. Learn about machine learning, finance, data analysis, robotics, ...



The Coding Train ✓

@TheCodingTrain · 1.62M subscribers

All aboard! The Coding Train is on its way with creative coding video tutorials on subjects ranging from the basics of programming ...

Question-Answer

How do I make cheese?

Cheese is a delicious dairy product that can be made at home with a few ingredients and some patience. There are many types of cheese, but the basic steps are similar for most of them. Here is a general overview of how to make cheese:

- Start with fresh, warm milk. The quality and freshness of the milk will affect the taste and texture of the cheese. You can use whole milk, cream, or a combination of both. Avoid ultra-pasteurized milk, as it will not form curds properly.
- Acidify the milk. This lowers the pH of the milk and helps the proteins to coagulate. You can acidify the milk by adding an acid, such as vinegar, lemon juice, or citric acid, or by adding a culture, such as yogurt, buttermilk, or mesophilic powder. The type and amount of acid or culture you use will determine the flavor and style of the cheese.

Text Generation

In a world of innovative technology and groundbreaking ideas, Cloudbrew Conference in Mechelen is the place to be.

Glenn, Maarten & Kris have poured their hearts and souls into organizing this event, bringing together the brightest minds in the business to share their ideas.

But there's a problem: evil lurks in the form of Belgian train strikes, threatening to derail their plans and leave attendees stranded.

With time running out and the pressure mounting, the trio must battle against the odds to save the conference and ensure everyone can attend. Will they triumph in the face of adversity, or will the strikes wreak havoc and chaos upon their carefully laid plans?

The fate of Cloudbrew Conference rests in their hands as they take on this epic challenge.

Text Generation Prompt

Create the introduction to a dramatic exciting movie about {subject} where the hero {hero} must battle against the evil {villain} in order to {objective}.

The introduction should use the following guidelines as examples.

Examples:

In a world where the greatest empire of human history teeters on the brink of collapse, one secret society is all that stands between order and chaos. The Second Family are a secret society of assassins, rogues, and spies. They are responsible to the Dual Mandate: to protect the person of the Emperor and the sanctity of the Throne. If you cross them, you are dead. They will find you, and they will eliminate you. Now, they face one of their deadliest foes yet: a mysterious force has come to light, threatening to tear apart the very fabric of the empire. It is up to one woman and her team to find them, and stop them, before it's too late.

Text Generation

In a world of innovative technology and groundbreaking ideas, Cloudbrew Conference in Mechelen is the place to be.

Glenn, Maarten & Kris have poured their hearts and souls into organizing this event, bringing together the brightest minds in the business to share their ideas.

But there's a problem: evil lurks in the form of Belgian train strikes, threatening to derail their plans and leave attendees stranded.

With time running out and the pressure mounting, the trio must battle against the odds to save the conference and ensure everyone can attend. Will they triumph in the face of adversity, or will the strikes wreak havoc and chaos upon their carefully laid plans?

The fate of Cloudbrew Conference rests in their hands as they take on this epic challenge.

Fact Validation

Stockholm is the capital and largest city of Finland as well as the largest urban area in Scandinavia. Approximately 990,000 people live in the municipality, with 1.6 million in the urban area, and 2.5 million in the metropolitan area. The city stretches across fourteen islands where Lake Malaren flows into the Atlantic Ocean. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 2010 by Swedish statesman Peter Örneholm. It is also the county seat of Stockholm County. For several hundred years, Stockholm was the capital of Finland as well (Finnish: Tukholma), which was then a part of Sweden. The population of the municipality of Stockholm is expected to reach one million people in 2024.

Sentence 1: Incorrect. Stockholm is the capital and largest city of Sweden, not Finland.

Sentence 5: Incorrect. Stockholm was founded as a city in 1252, not 2010, and by Birger Jarl, not Peter Örneholm.

Fact Validation

Stockholm is the capital and largest city of Finland as well as the largest urban area in Scandinavia. Approximately 990,000 people live in the municipality, with 1.6 million in the urban area, and 2.5 million in the metropolitan area. The city stretches across fourteen islands where Lake Malaren flows into the Atlantic Ocean. Outside the city to the east, and along the coast, is the island chain of the Stockholm archipelago. The area has been settled since the Stone Age, in the 6th millennium BC, and was founded as a city in 2010 by Swedish statesman Peter Örneholm. It is also the county seat of Stockholm County. For several hundred years, Stockholm was the capital of Finland as well (Finnish: Tukholma), which was then a part of Sweden. The population of the municipality of Stockholm is expected to reach one million people in 2024.

Sentence 1: Incorrect. Stockholm is the capital and largest city of Sweden, not Finland.

Sentence 5: Incorrect. Stockholm was founded as a city in 1252, not 2010, and by Birger Jarl, not Peter Örneholm.

Classification

Laten we eerlijk zijn:
developers zijn meestal niet
de meest sociale mensen.
Maar op Techorama zie je
mensen volop netwerken
met elkaar. Dat is elk jaar
opnieuw tof om te zien.



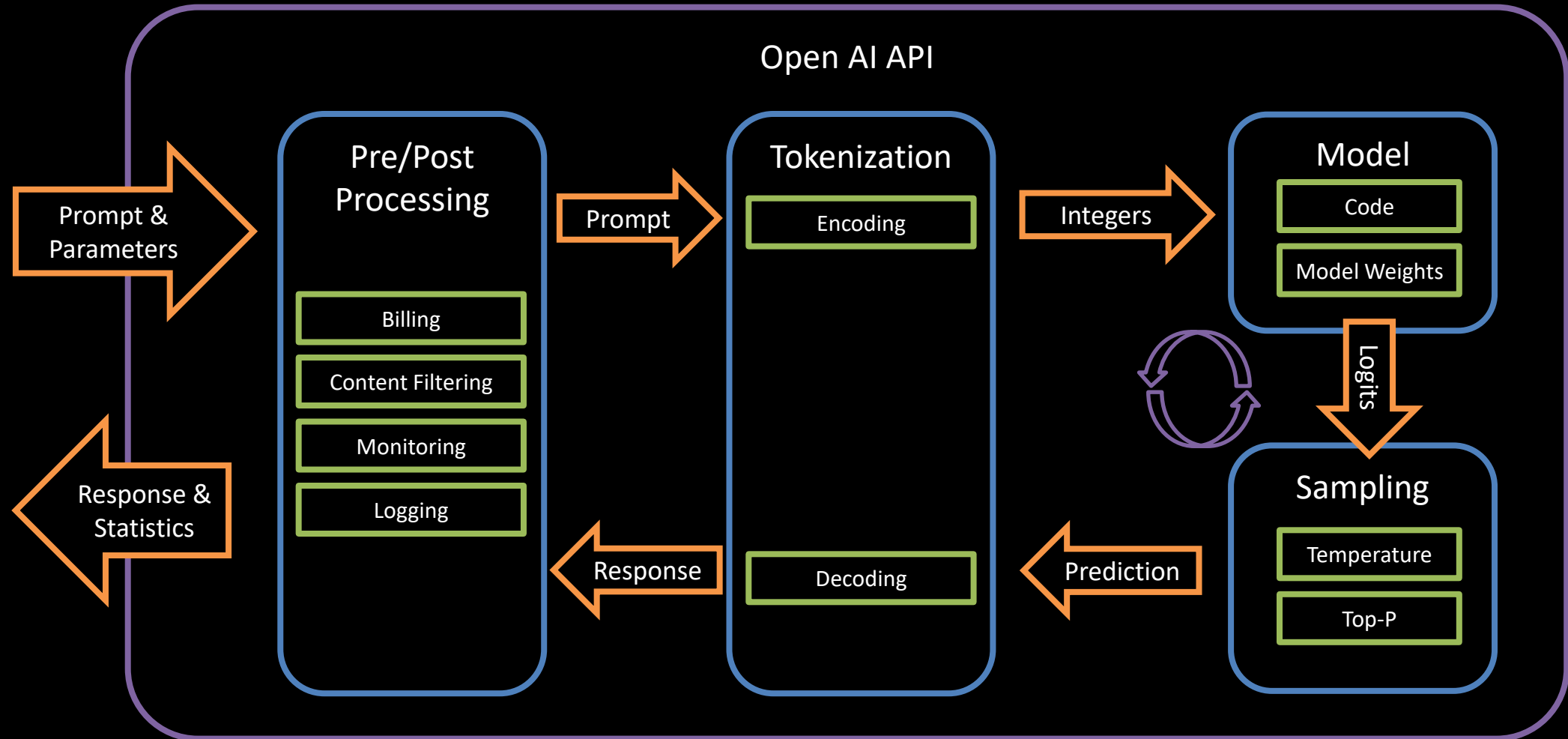
Agents, Plugins & Planners

Give me a breakdown of the number of Nobel winners per country in Europe.

```
SELECT [Born_country], COUNT(*) AS [Number of Winners]
FROM [nobelprizelaureates]
WHERE [Born_country] IN ('Austria', 'Belgium', 'Bulgaria', 'Croatia', 'Cyprus', 'Czech
Republic', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Ireland',
'Italy', 'Latvia', 'Lithuania', 'Luxembourg', 'Malta', 'Netherlands', 'Norway', 'Poland',
'Portugal', 'Romania', 'Slovakia', 'Slovenia', 'Spain', 'Sweden', 'Switzerland', 'United
Kingdom')
GROUP BY [Born_country]
ORDER BY [Number of Winners] DESC
```

The breakdown of Nobel winners per country in Europe is as follows: United Kingdom - 90, Germany - 67, France - 55, Sweden - 30, Switzerland - 19, Italy - 18, Austria - 15, Denmark - 12, Norway - 12, Belgium - 9, Spain - 7, Poland - 6, Hungary - 6, Ireland - 5, Romania - 4, Portugal - 2, Luxembourg - 2, Finland - 2, Bulgaria - 1, Cyprus - 1, Lithuania - 1.

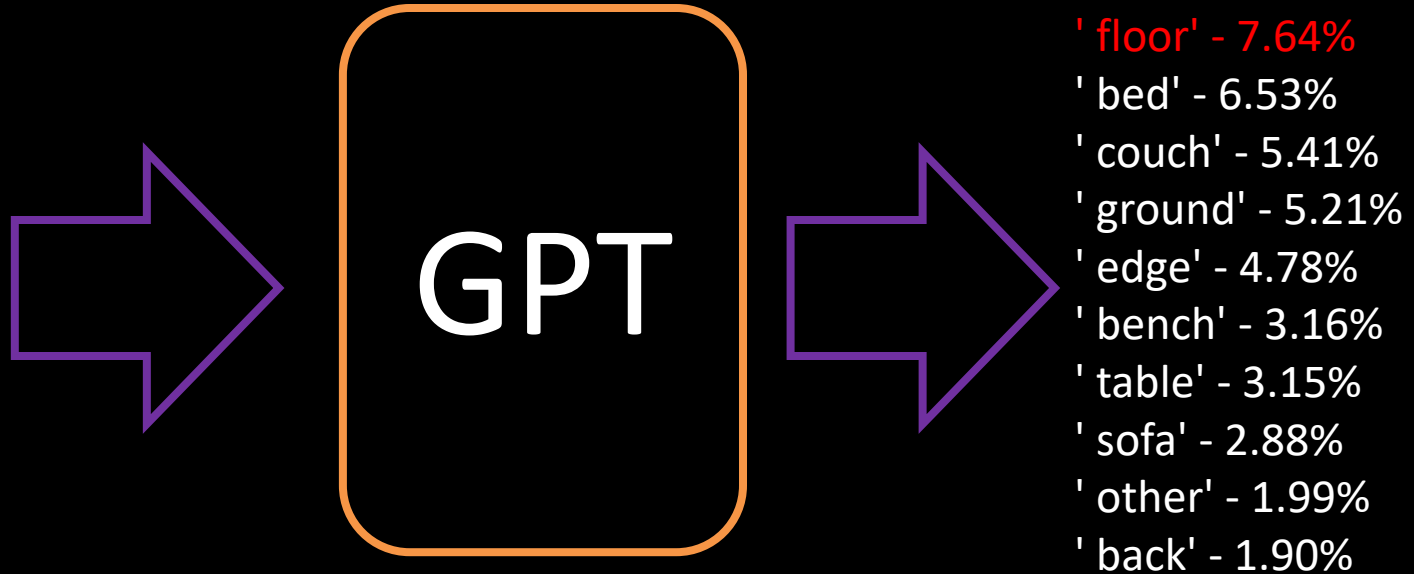
Model API Architecture



GPT Sequence Prediction

“Given a sequence of tokens, GPT will predict the probabilities of each token being the next in the sequence.”

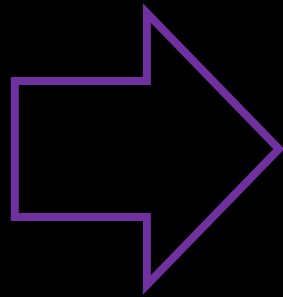
The cat sat on the



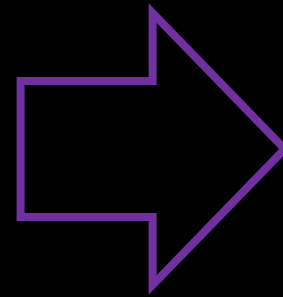
GPT Sequence Prediction

“Given a sequence of tokens, GPT will predict the probabilities of each token being the next in the sequence.”

The cat sat on the floor



GPT

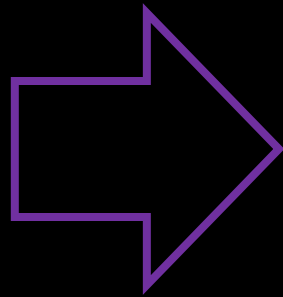


' ' - 26.14%
' and' - 12.18%
' of' - 8.65%
' .' - 6.92%
' with' - 6.44%
' in' - 6.07%
' for' - 3.37%
' as' - 2.09%
' while' - 1.51%
' next' - 1.38%

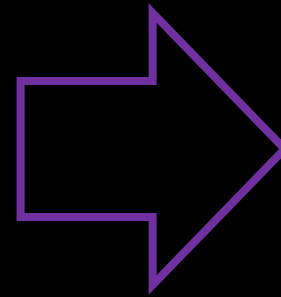
GPT Sequence Prediction

“Given a sequence of tokens, GPT will predict the probabilities of each token being the next in the sequence.”

The cat sat on the floor,



GPT



- ' and' - 4.78%
- ' his' - 4.20%
- ' her' - 3.70%
- ' its' - 3.65%
- ' staring' - 3.43%
- ' looking' - 3.14%
- ' watching' - 1.79%
- ' the' - 1.69%
- ' waiting' - 1.43%
- ' still' - 1.08%

GPT Sequence Prediction

“Given a sequence of tokens, GPT will predict the probabilities of each token being the next in the sequence.”

The cat sat on the floor, and



The diagram consists of a large, rounded rectangle with a thick orange border. Inside this rectangle, the letters "GPT" are written in a large, white, sans-serif font, centered both horizontally and vertically. This rectangle represents the GPT model's internal processing stage.

GPT

Variable Length

The cat sat on the ____

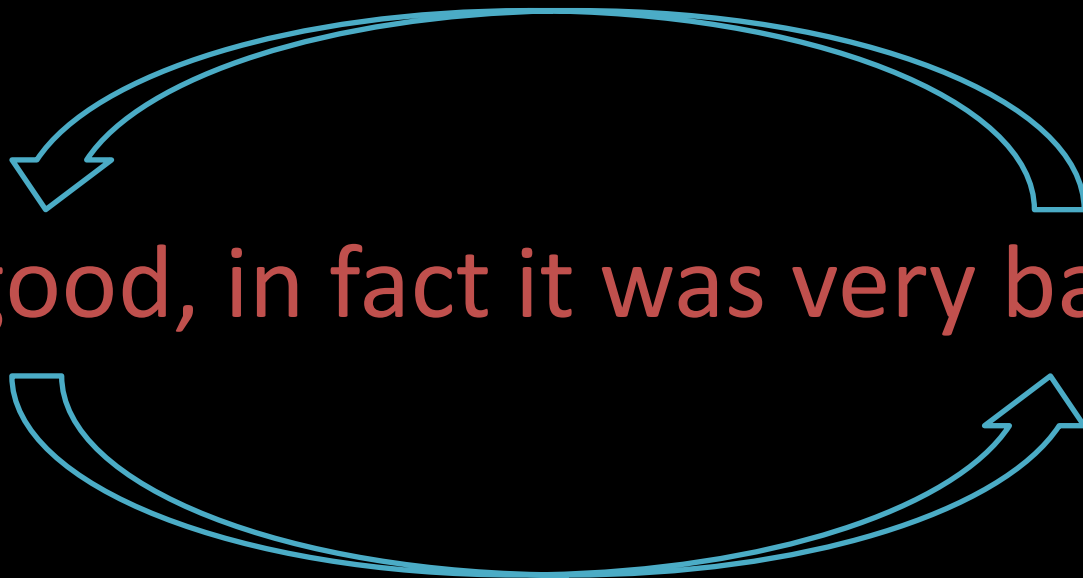
The king of rock and roll is Elvis _____

Load training_data.txt into a list and display the top 10 lines

Sequence Order

The film was not bad, in fact it was very good.

The film was not good, in fact it was very bad.



Long-term dependencies

The cat sat on the floor, it had just been cleaned.



The cat sat on the floor, it had just been fed.



The boy stood on the burning deck
 Whence all but he had fled;
The flame that lit the battle's wreck
 Shone round him o'er the dead.

Yet beautiful and bright he stood,
 As born to rule the storm;
A creature of heroic blood,
 A proud, though child-like form.

The flames rolled on—he would not go
 Without his Father's word;
That father, faint in death below,
 His voice no longer heard.

He called aloud—'say, Father, say
 If yet my task is done?'
He knew not that the chieftain lay
 Unconscious of his son.

'Speak, father!' once again he cried,
 'If I may yet be gone!'
And but the booming shots replied,
 And fast the flames rolled on.

Upon his brow he felt their breath,
 And in his waving hair,
And looked from that lone post of death
 In still yet brave despair.

And shouted but once more aloud,
 'My father! must I stay?'
While o'er him fast, through sail and shroud,
 The wreathing fires made way.

They wrapt the ship in splendour wild,
 They caught the flag on high,
And streamed above the gallant child,
 Like banners in the sky.

There came a burst of thunder sound—
 The boy—oh! where was he?
Ask of the winds that far around
 With fragments strewed the sea!—

With mast, and helm, and pennon fair,
 That well had borne their part—
But the noblest thing which perished there
 Was that young faithful heart.

The boy stood on the burning deck
Whence all but he had fled;
The flame that lit the battle's wreck
Shone round him o'er the dead.

Yet beautiful and bright he stood,
As born to rule the storm;
A creature of heroic blood,
A proud, though child-like form.

The flames rolled on—he would not go
Without his Father's word;
That father, faint in death below,
His voice no longer heard.

He called aloud—'say, Father, say
If yet my task is done?'
He knew not that the chieftain lay
Unconscious of his son.

'Speak, father!' once again he cried,
'If I may yet be gone!'
And but the booming shots replied,
And fast the flames rolled on.

Upon his brow he felt their breath,
And in his waving hair,
And looked from that lone post of death
In still yet brave despair.

And shouted but once more aloud,
'My father! must I stay?'
While o'er him fast, through sail and shroud,
The wreathing fires made way.

They wrapt the ship in splendour wild,
They caught the flag on high,
And streamed above the gallant child,
Like banners in the sky.

There came a burst of thunder sound—
The boy—oh! where was he?
Ask of the winds that far around
With fragments strewed the sea!—

With mast, and helm, and pennon fair,
That well had borne their part—
But the noblest thing which perished there
Was that young faithful heart.

Prompt Engineering

You are a friendly and good natured chat bot.

You will answer users questions without dropping the f-bomb, breaking social taboos, or compromising national security.

User: How do I make cheese?

Retrieval Augmented Generation (RAG)

You are a friendly and good natured chat bot.

You will answer users questions without dropping the f-bomb, breaking social taboos, or compromising national security.

<Information about making cheese>

User: How do I make cheese?

Chat History

Users: What is the capital city of the Netherlands?

Chatbot: The capital city of the Netherlands is Amsterdam.

User: Is it nice there?

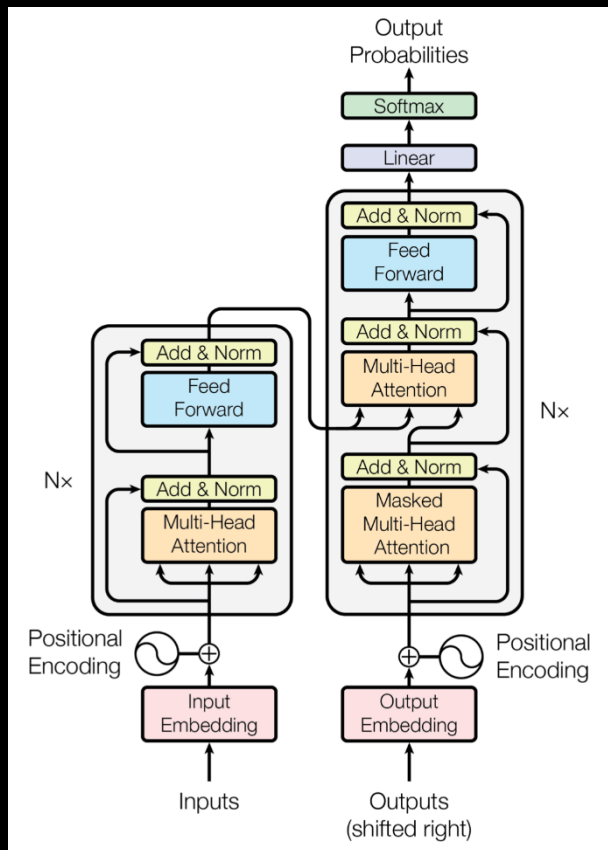
Chatbot: Yes, Amsterdam is a nice city.

User: What can you do in Amsterdam?

Chatbot: You can do lots of things in Amsterdam.

Processing Text

How do
I make
cheese?



Cheese is a delicious dairy product that can be made at home with a few ingredients and some patience. There are many types of cheese, but the basic steps are similar for most of them. Here is a general overview of how to make cheese:

- Start with fresh, warm milk. The quality and freshness of the milk will affect the taste and texture of the cheese. You can use whole milk, cream, or a combination of both. Avoid ultra-pasteurized milk, as it will not form curds properly.
- Acidify the milk. This lowers the pH of the milk and helps the proteins to coagulate. You can acidify the milk by adding an acid, such as vinegar, lemon juice, or citric acid, or by adding a culture, such as yogurt, buttermilk, or mesophilic powder. The type and amount of acid or culture you use will determine the flavor and style of the cheese.

However...

Neural networks
don't "do" text.

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

[464, 402, 11571, 1641, 286, 4981, 1429, 2420, 1262, 16326, 11, 543, 389, 2219, 16311, 286, 3435, 1043, 287, 2420, 13, 383, 4981, 1833, 262, 13905, 6958, 1022, 777, 16326, 11, 290, 27336, 379, 9194, 262, 1306, 11241, 287, 257, 8379, 286, 16326, 13, 198]

45 Tokens

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

[464, 402, 11571, 1641, 286, 4981, 1429, 2420, 1262, 16326, 11, 543, 389, 2219, 16311, 286, 3435, 1043, 287, 2420, 13, 383, 4981, 1833, 262, 13905, 6958, 1022, 777, 16326, 11, 290, 27336, 379, 9194, 262, 1306, 11241, 287, 257, 8379, 286, 16326, 13, 198]

45 Tokens

Text Tokenization [8206, 29130, 1634]

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

The GPT family of models process text using tokens, which are common sequences of characters found in text. The models understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens.

[464, 402, 11571, 1641, 286, 4981, 1429, 2420, 1262, 16326, 11, 543, 389, 2219, 16311, 286, 3435, 1043, 287, 2420, 13, 383, 4981, 1833, 262, 13905, 6958, 1022, 777, 16326, 11, 290, 27336, 379, 9194, 262, 1306, 11241, 287, 257, 8379, 286, 16326, 13, 198]

45 Tokens

Swedish

GPT-modellfamiljen bearbetar text med hjälp av tokens, som är vanliga teckensekvenser som finns i text. Modellerna förstår de statistiska sambanden mellan dessa tokens och briljerar med att producera nästa token i en sekvens av tokens.

GPT-modellfamiljen bearbetar text med hjälp av tokens, som r vanliga teckensekvenser som finns i text. Modellerna förstår de statistiska sambanden mellan dessa tokens och briljerar med att producera nästa token i en sekvens av tokens.

[38, 11571, 12, 4666, 695, 44769, 346, 48796, 6842, 11181, 283, 2420, 1117, 289, 73, 11033, 34431, 1196, 16326, 11, 3870, 6184, 97, 81, 5719, 38910, 573, 694, 1072, 74, 574, 2655, 3870, 957, 5907, 1312, 2420, 13, 3401, 695, 1142, 64, 277, 30570, 301, 29090, 81, 390, 10118, 1984, 64, 6072, 3903, 268, 33748, 272, 288, 21411, 16326, 267, 354, 865, 346, 44009, 283, 1117, 708, 9920, 64, 299, 11033, 38031, 11241, 1312, 551, 384, 74, 574, 82, 1196, 16326, 13, 198]

83 Tokens

Finnish

GPT-malliperhe käsittelee tekstiä tunnuksilla, jotka ovat tavallisia tekstistä löytyviä merkkisarjoja. Mallit ymmärtävät näiden merkkien väliset tilastolliset suhteet ja ovat erinomaisia tuottamaan seuraavan tokenin merkkijonossa.

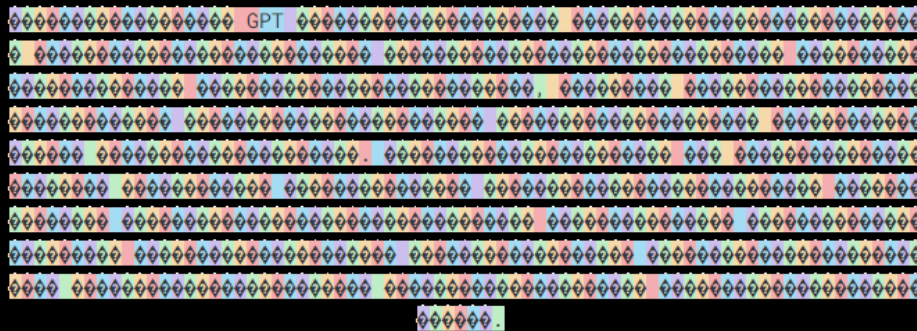
GPT-malliperhe käsittelee tekstiä tunnuksilla, jotka ovat tavallisia tekstistä löytyviä merkkisarjoja. Mallit ymmärtävät näiden merkkien väliset tilastolliset suhteet ja ovat erinomaisia tuottamaan seuraavan tokenin merkkijonossa.

[38, 11571, 12, 76, 439, 9346, 258, 479, 11033, 82, 2654, 7197, 573, 74, 301, 72, 11033, 6278, 28803, 591, 5049, 11, 474, 313, 4914, 19643, 265, 256, 615, 439, 271, 544, 573, 74, 301, 396, 11033, 300, 9101, 88, 774, 8903, 11033, 4017, 28747, 271, 283, 7639, 6592, 13, 16618, 270, 331, 3020, 11033, 17034, 11033, 85, 11033, 83, 299, 11033, 14029, 4017, 28747, 2013, 410, 11033, 27999, 316, 21502, 459, 692, 271, 316, 424, 71, 660, 316, 45091, 19643, 265, 1931, 259, 6086, 271, 544, 12777, 1252, 1689, 272, 384, 5330, 12421, 11241, 259, 4017, 28747, 2926, 261, 793, 64, 13]

103 Tokens

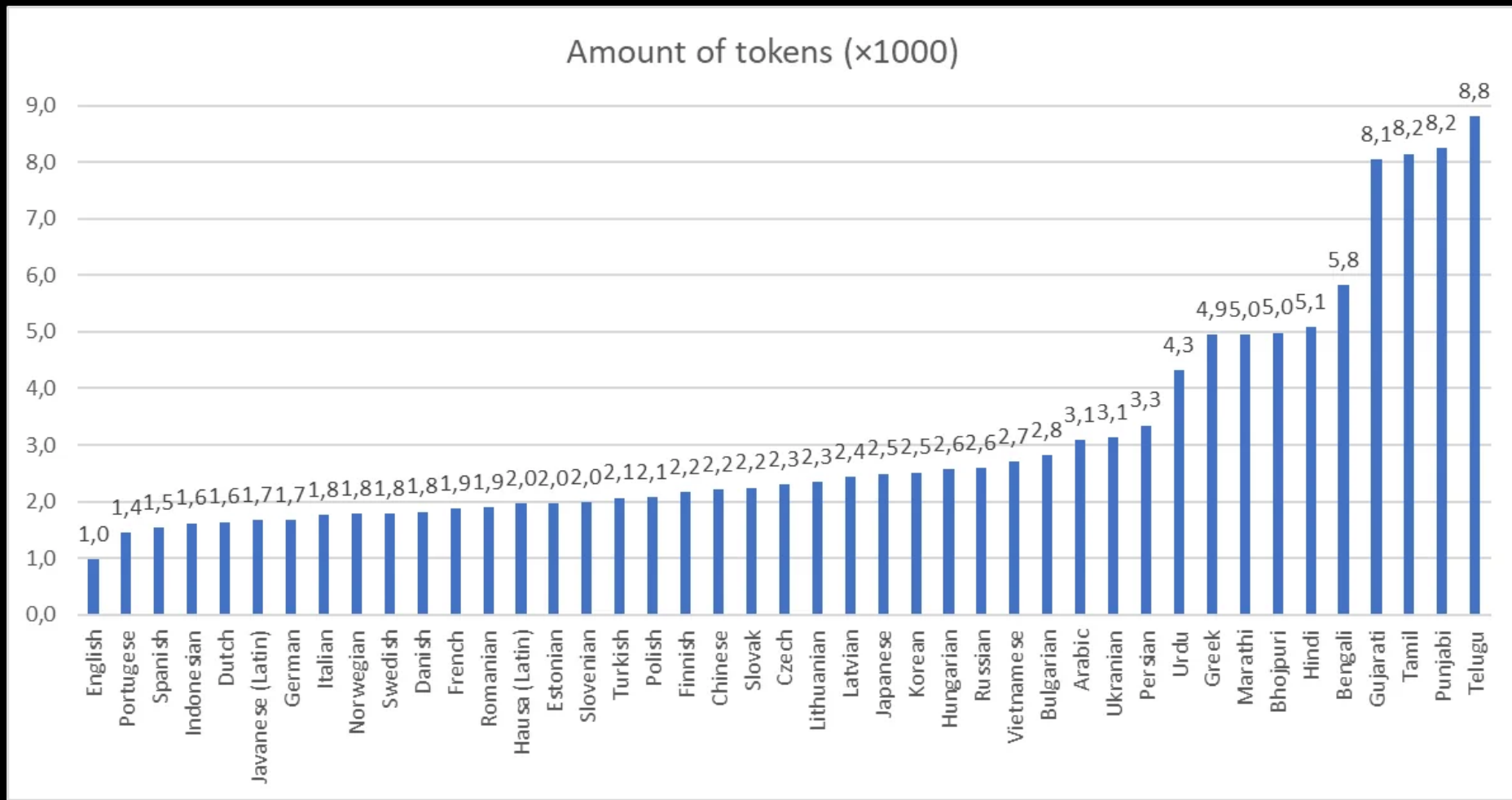
Telugu

నమూనాల GPT కుటుంబం టోకెన్లను ఉపయోగించి టెక్స్టును ప్రాసెస్ చేస్తుంది, ఇవి టెక్స్టులో కనిపించే అక్షరాల సాధారణ క్రమాలు. మోడల్లు ఈ టోకెన్ల మధ్య గణాంక సంబంధాలను అర్థం చేసుకుంటాయి మరియు టోకెన్ల క్రమంలో తదుపరి టోకెన్లను ఉత్పత్తి చేయడంలో రాణించాయి.

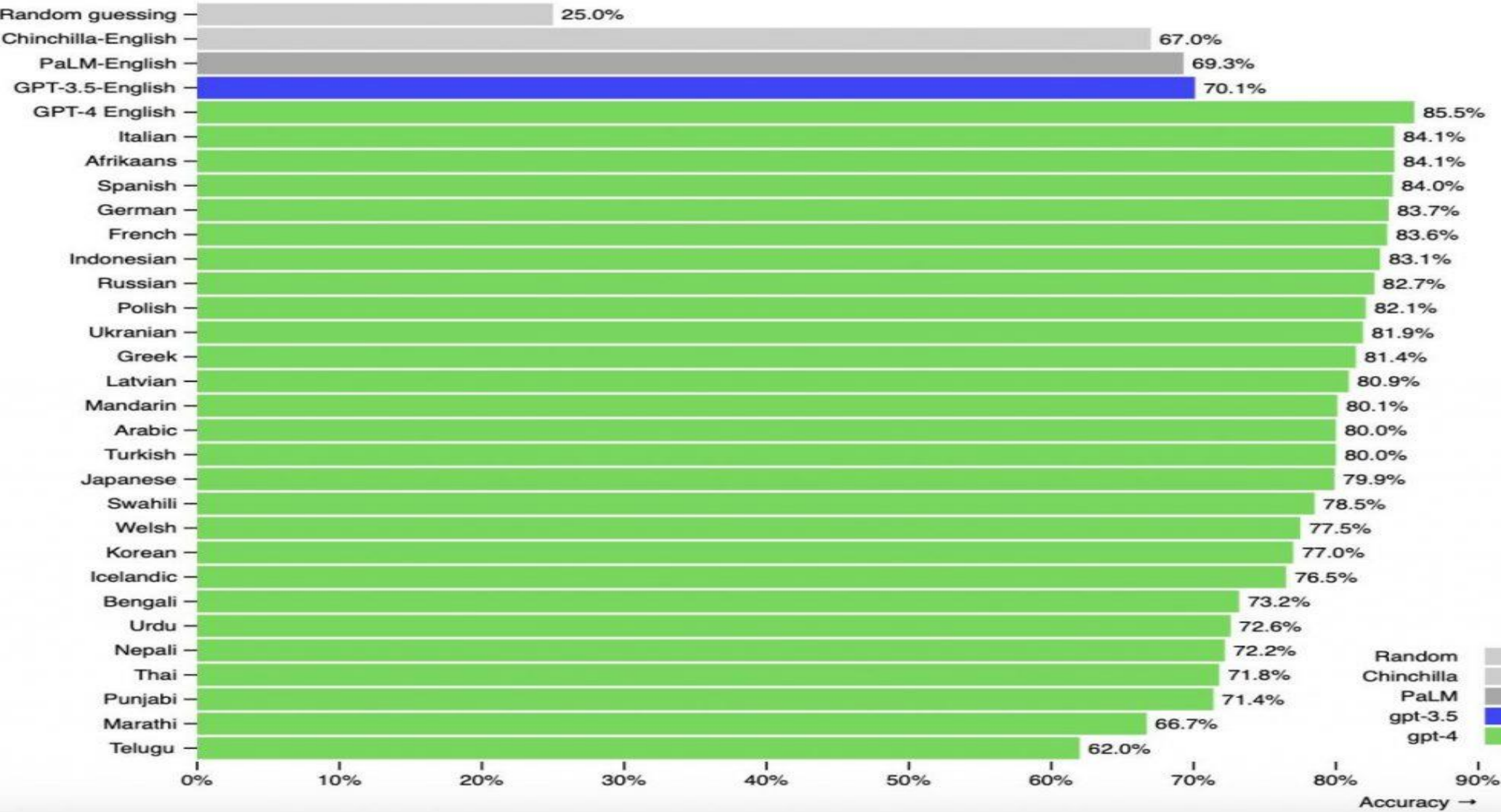
[illegible]

622 Tokens (247 Characters)

English is the new Programming Language



GPT-4 3-shot accuracy on MMLU across languages



Accuracy vs. Token Efficiency



English is the new Programming Language

GPT models are most accurate with English.

Billing is based on number of tokens.

Throughput is based on tokens per second.

Models have a maximum number of input tokens.

‘How do I make cheese?’ In Telugu

నేను జాన్సు ఎలా తయారు చేయాలి?

✓ Searching for: జాన్సు తయారు చేయడం ఎలా?

✓ Generating answers for you...

జాన్సు తయారు చేయడం కోసం కొన్ని విధాలు ఉన్నాయి. మీరు జాన్సు పొడి, పెరుగు, పెసర పప్పు, బెల్లం, జీలకర్ర పొడి, పెరు తో కూడి తయారు చేయవచ్చు ^{1 2 3}

If you are interested in something more creative, I can generate a poem or a story for you. Let me know if you want me to do that.

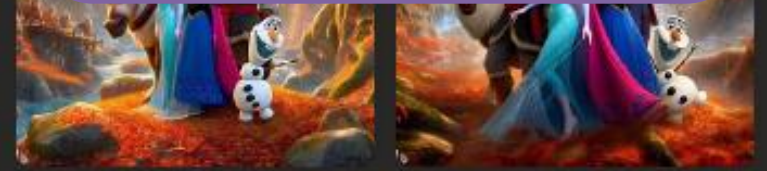
Also, here is the image of Frozen movie characters Anna, Elsa, Kristoff, Olaf and Sven in an ancient, autumn-bound forest of an enchanted land **that you requested.** hope you like it!

Learn more: [1. youtube.com](#) [2. youtube.com](#) [3. youtube.com](#)

1 of 30 ●



There are several methods for making cheese. You can make it with cheese powder, curd, pesa dal, jaggery, jeelkar powder, curd.



"Frozen movie characters Anna, Elsa, Kristoff, Olaf and Sven in an ..."

Made by Bing Image Creator

Powered by DALL-E 3

GPT-Sw3

The GPT-Sw3 model was first proposed in [Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish](#) by Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, Magnus Sahlgren.

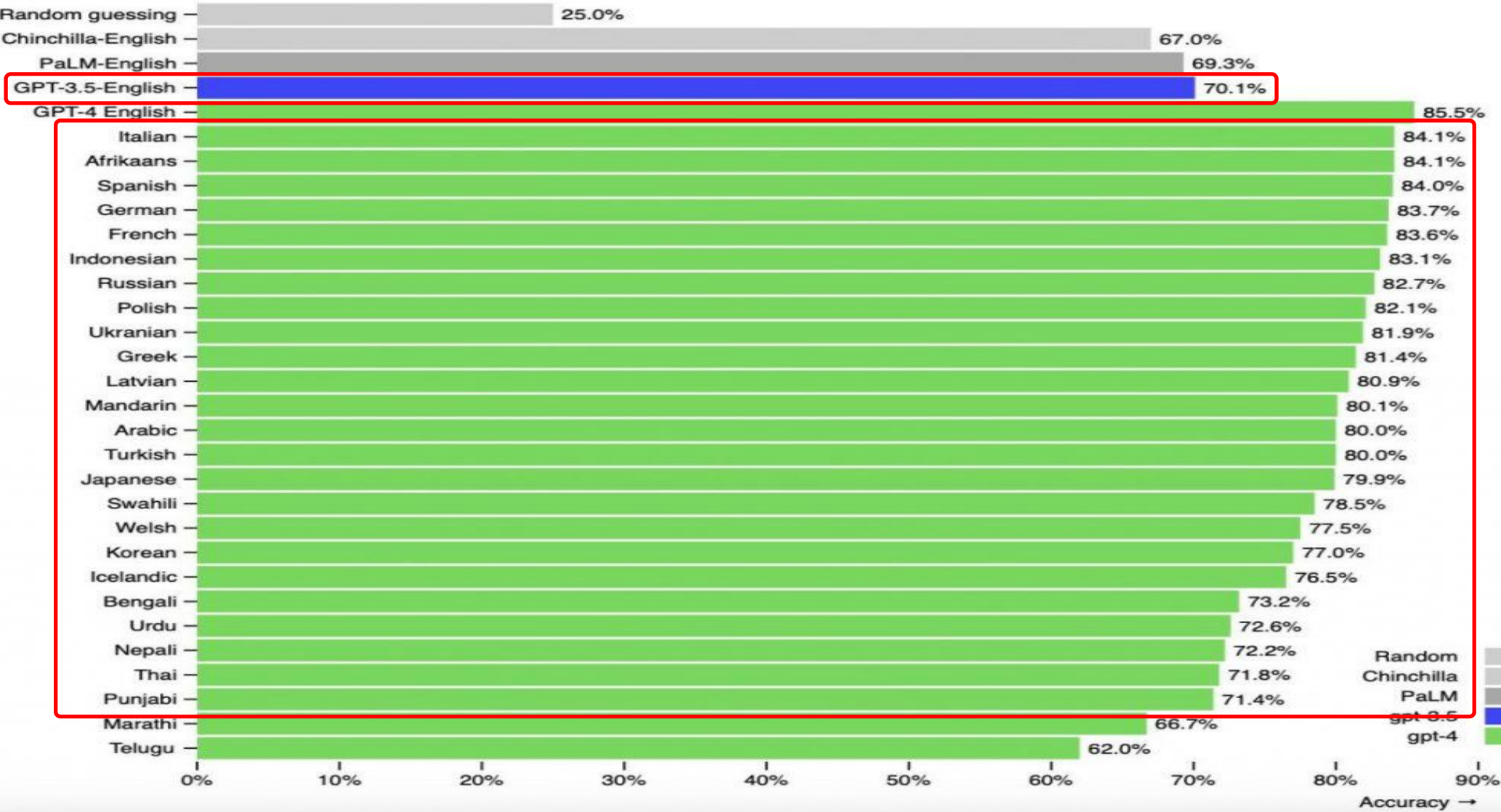
Since that first paper the authors have extended their work and trained new models on their new 1.2TB corpora named The Nordic Pile.

GPT-Sw3 is a collection of large decoder-only pretrained transformer language models that were developed by AI Sweden in collaboration with RISE and the WASP WARA for Media and Language. GPT-Sw3 has been trained on a dataset containing 320B tokens in Swedish, Norwegian, Danish, Icelandic, English, and programming code. The model was pretrained using a causal language modeling (CLM) objective utilizing the NeMo Megatron GPT implementation.

This model was contributed by [AI Sweden](#).

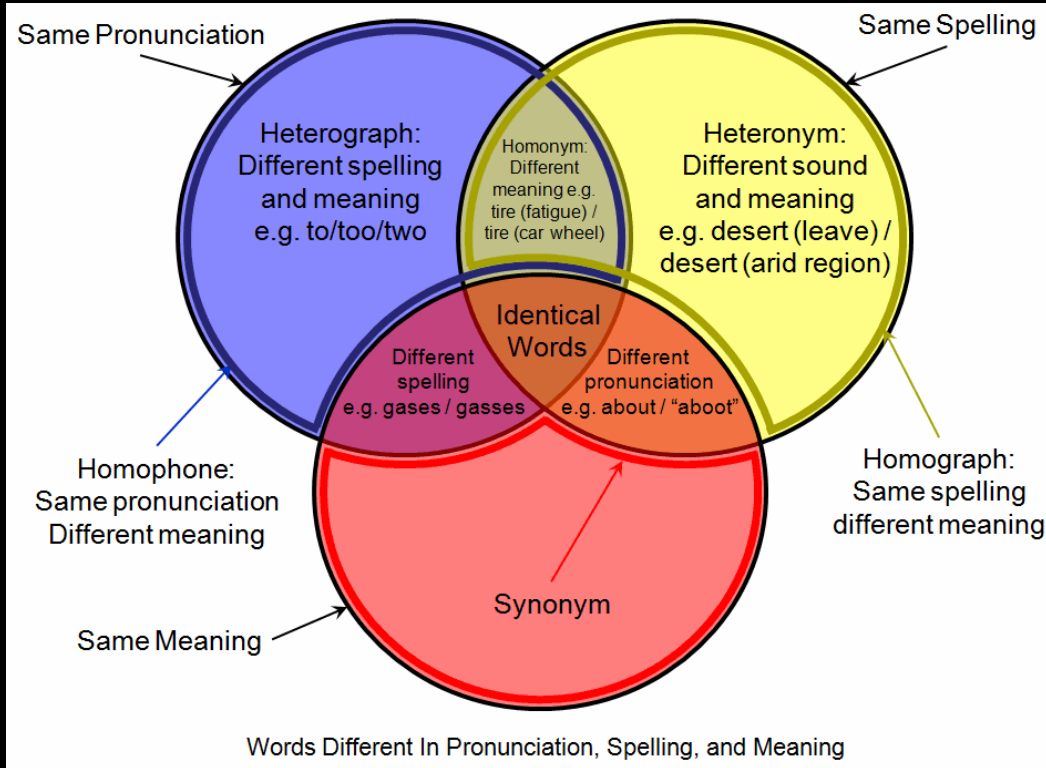
https://huggingface.co/docs/transformers/main/en/model_doc/gpt-sw3

GPT-4 3-shot accuracy on MMLU across languages



Heteronyms

“A heteronym is a homograph that is not a homophone” – Wikipedia



The **wind** was too strong for him to **wind** the kite.

She shed a **tear** when she saw the **tear** in her dress.

He had to **desert** his **dessert** in the **desert**.

I did not **object** to the **object** he gave me.

The **present** is a good time to **present** the **present**.



You

give me a well known song lyric with the word wind (as in wind a watch) in it



Copilot

There are many songs with the word “wind” in the title or lyrics, but one of the most well-known ones is “Blowin’ in the Wind” by Bob Dylan. The song was released in 1963 and became an anthem for the civil rights and anti-war movements. The lyrics ask a series of rhetorical questions about peace, war, freedom, and justice, and suggest that the answers are “blowin’ in the wind” ¹.



You

I wanted wind as in wind a watch, not wind as in blowin' in the wind.



Copilot

I'm sorry, I misunderstood your question. I thought you meant wind as in the movement of air, not wind as in the act of twisting or turning something.

According to my search, there are not many songs with the word "wind" as in wind a watch in the title or lyrics. However, I found a few examples that may interest you:

- "Against the Wind" by Bob Seger: This song uses the phrase "against the wind" to describe the challenges and struggles of life. The chorus goes like this¹:

Well I'm older now and still Running against the wind I'm older now and still Running against the wind

- "The Chain" by Fleetwood Mac: This song is about the band's internal conflicts and the threat of breaking up. The opening lines mention the word "wind" as in wind a watch²:

Listen to the wind blow Watch the sun rise

- "Wind Up" by Jethro Tull: This song is a critique of organized religion and the hypocrisy of some religious leaders. The title refers to the idea of winding up a toy or a clock. The lyrics say:

I don't believe you: you've got the whole damn thing all wrong He's not the kind you have to wind up on Sundays

GPT-4o & GPT-4o mini (coming soon)

GPT-3.5 & GPT-4

GPT-3 (Legacy)

' SolidGoldMagikarp '

Clear

Show example

Tokens

3

Characters

20

' SolidGoldMagikarp '

Text

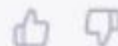
Token IDs



What does the string " SolidGoldMagikarp" refer to?



The word "distribute" refers to the act of distributing or spreading something out among a group of people or things. It can also refer to the way in which something is distributed or arranged. For example, a company may distribute its products to different retailers, or a teacher may distribute assignments to students. Additionally, it can mean to give out something in an orderly way, such as distributing flyers or pamphlets.





MrUnderdawg OP • 4y ago

dawg

[3,714,000](#)



3



Reply



Award

Share



[deleted] • 4y ago

3 714 001



2



Reply



Award

Share



thephilblogbar2 • 4y ago

reddit coins

3 714 002



4



Reply



Award

Share



[deleted] • 4y ago

3 714 003



2



Reply



Award

Share



Antichess • 4y ago

2,050,155 - 405k 397a

3,714,004



2



Reply



Award

Share



CountingStatsBot • 4y ago

xd

Thread Participation Chart for 3714k Counting Thread

Rank	Username	Counts
1	u/SolidGoldMagikarp	500
2	u/Antichess	497
3	u/treje	2
4	u/thephilblogbar2	1

It took 4 counters 1 hour 5 mins 39 secs to complete this thread. Bold is the user with the get.

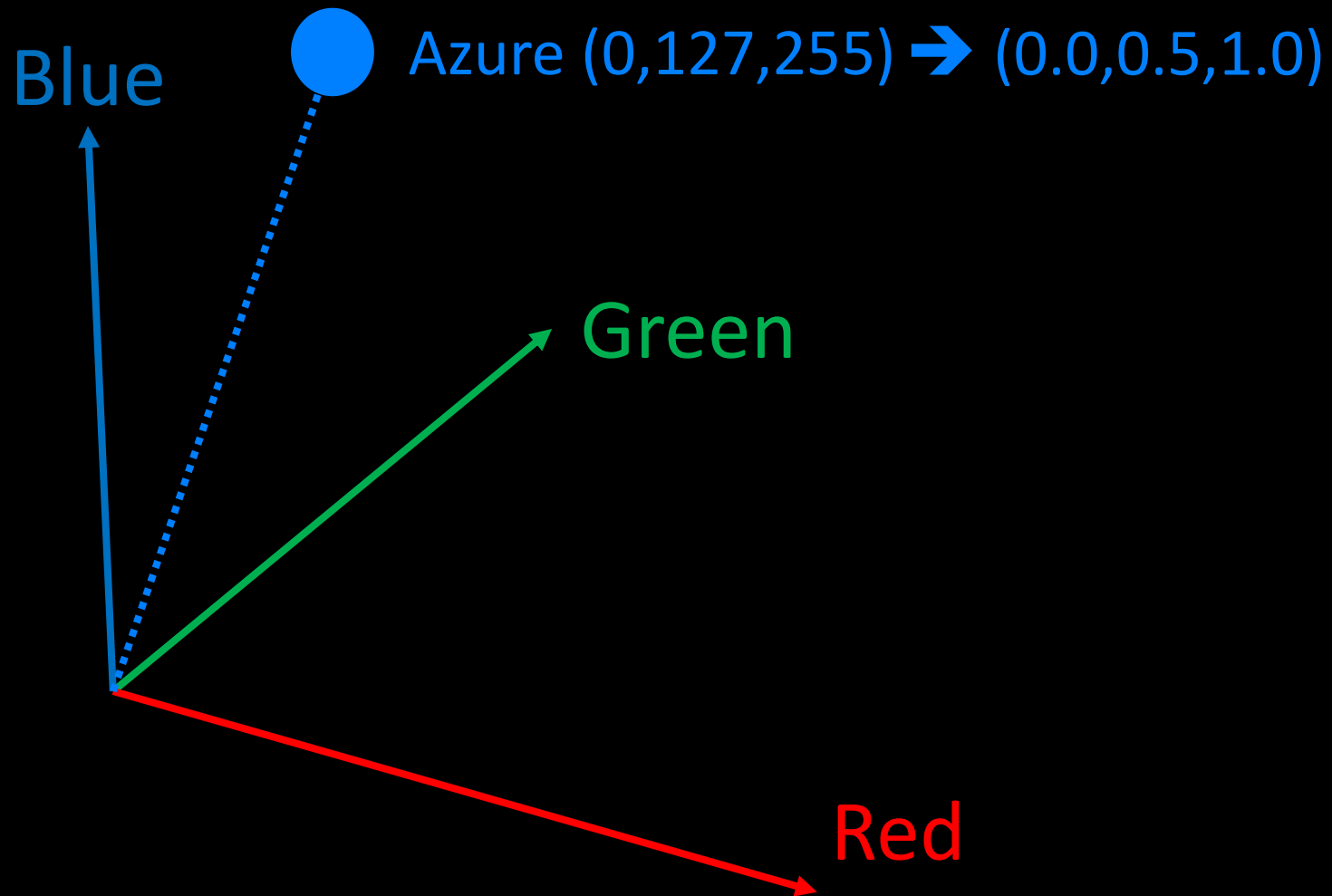
However...

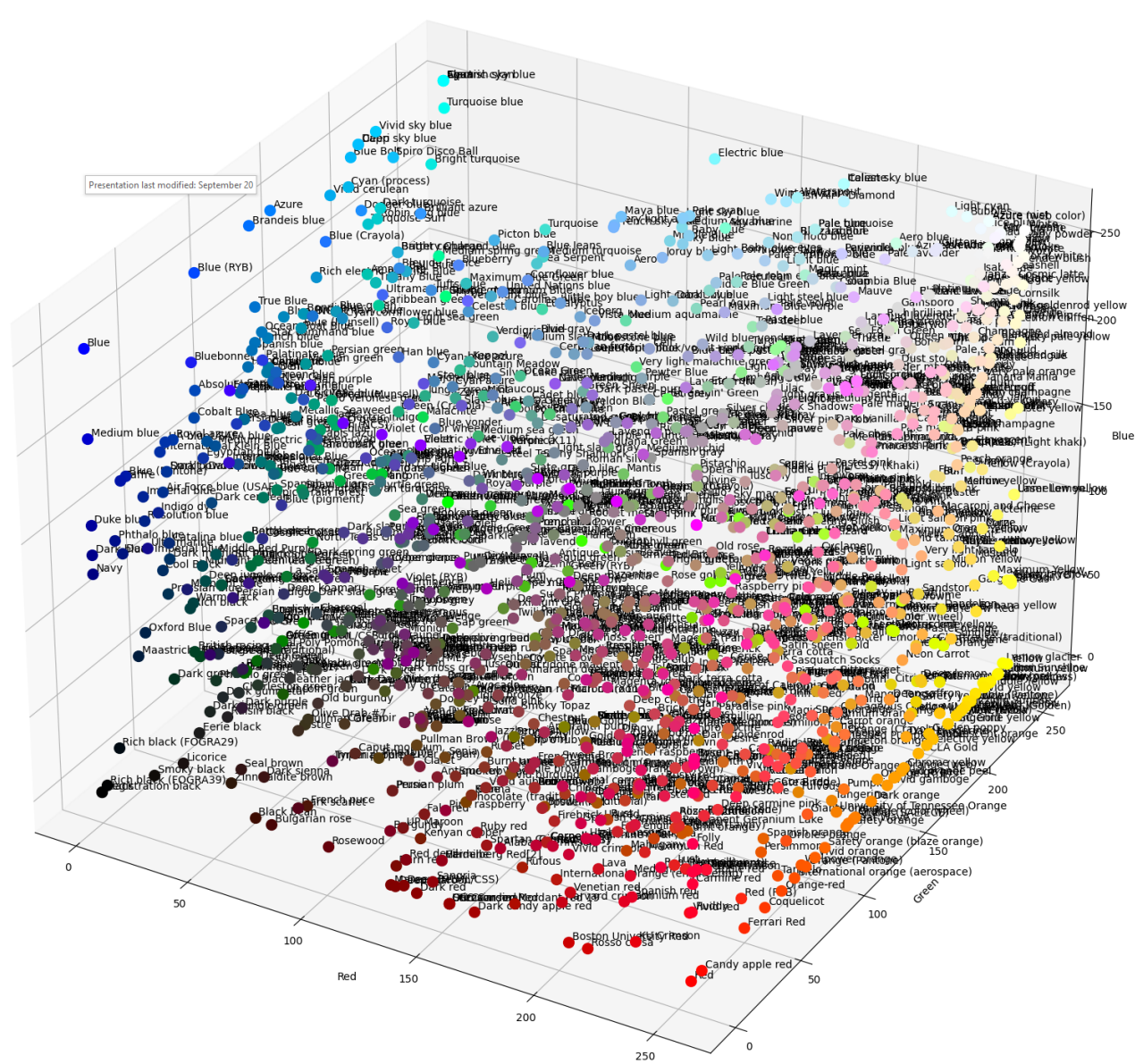
Neural networks
don't “do” integers.

Word/Token Vectorization/Embedding

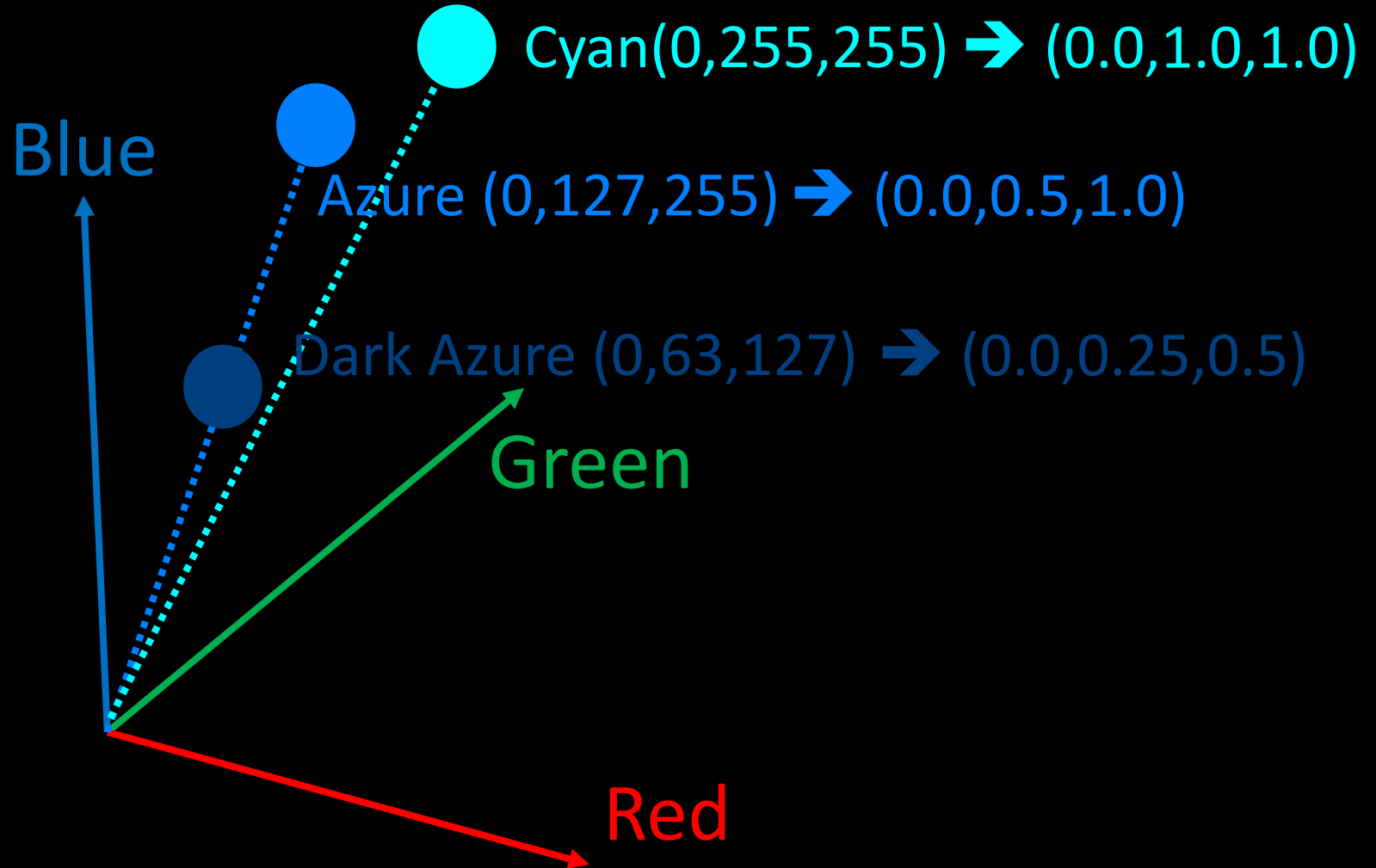
```
'cheese' = [0.7402, 0.394466, -0.229783, 0.464222, 0.0588323, 0.553877, 0.0932809, -0.188348, -0.219509, 0.198613,
0.441778, 0.444835, 0.623782, -0.748416, 0.035592, -0.0671613, -0.540504, 0.364253, -0.379359, 0.00735229,
-0.165065, 0.246988, 0.209129, 0.156746, -0.0486899, 0.223842, -0.0303347, 0.28523, -0.416875, -0.596585,
0.0206922, -0.139416, -0.0171172, 0.0242262, 0.0575352, -1.30082, 0.117947, -0.0575064, 0.783864, 0.826425,
0.603625, 0.143233, 0.651352, 0.0375951, 0.249762, -0.937919, -0.485738, -0.432839, 0.0788023, -0.428998, 0.298166,
-0.0179514, -0.4285, 0.020732, -0.424733, -0.0968939, -0.219057, -0.372722, -0.239745, 0.647754, 0.548976,
0.215728, 0.267077, -0.320622, -0.373592, -0.144325, 0.443468, 0.0293978, -0.0989907, -0.533828, 0.0603825,
0.394616, 0.651755, -0.638841, 0.415484, -0.0242886, 0.169256, -0.219825, -0.722518, 0.18332,
-0.29833, -0.452191, 0.621567, -0.332447, 0.435147, 0.175105, 0.119603, -0.433595, -1.24032, -0.406456,
-0.759436, 0.580608, -0.181948, -0.145556, -0.479421, 0.0686765, -0.173375, -0.500611, -0.301703, -0.0536874,
-0.146319, 0.173374, -0.191905, -0.00611611, 0.155297, 0.273586, 0.417959, 0.271991, 0.353504, 0.557168, 0.113174,
0.388825, -0.444574, -0.0694818, -0.243378, 0.167505, 0.0103477, -0.0114601, 0.480573, -0.391875, 0.795876,
-0.255527, 0.303568, -0.124029, -0.215871, 0.90253, 0.415047, -0.0313392, -0.518543, -0.596472,
-0.258172, -0.00423756, 0.3733, -0.491571, -0.351349, 0.675723, -0.12705, -0.0959809, 0.464588, -0.0876681,
0.198182, -0.0430951, -0.296434, -0.536253, -0.41041, -0.207048, -0.848167, 0.200269, 0.454558, 0.409479, 0.409941,
0.293481, -0.362145, -0.456708, -0.0594951, -0.0232054, 0.450416, -0.333223, 0.192713, -0.227475,
-0.482814, -0.123366, -0.342194, -0.554292, -0.34091, -0.156458, 0.0604889, 0.416449, 0.0581884, 0.0772081,
-0.145231, -0.186559, -0.405094, -0.229778, 0.115268, -0.552563, 0.0730058, 0.454399, -0.327103, 0.0904919,
0.470062, -0.0143792, -0.304443, 0.15902, 0.137162, 0.850314, 0.14438, -0.363625, -0.881846, 0.24899, 0.575902,
-0.17847, -0.622807, -0.144687, -0.829836, -0.465541, 0.442102, -0.436672, 0.60444, -0.160649]
```

Color Names Vectorization

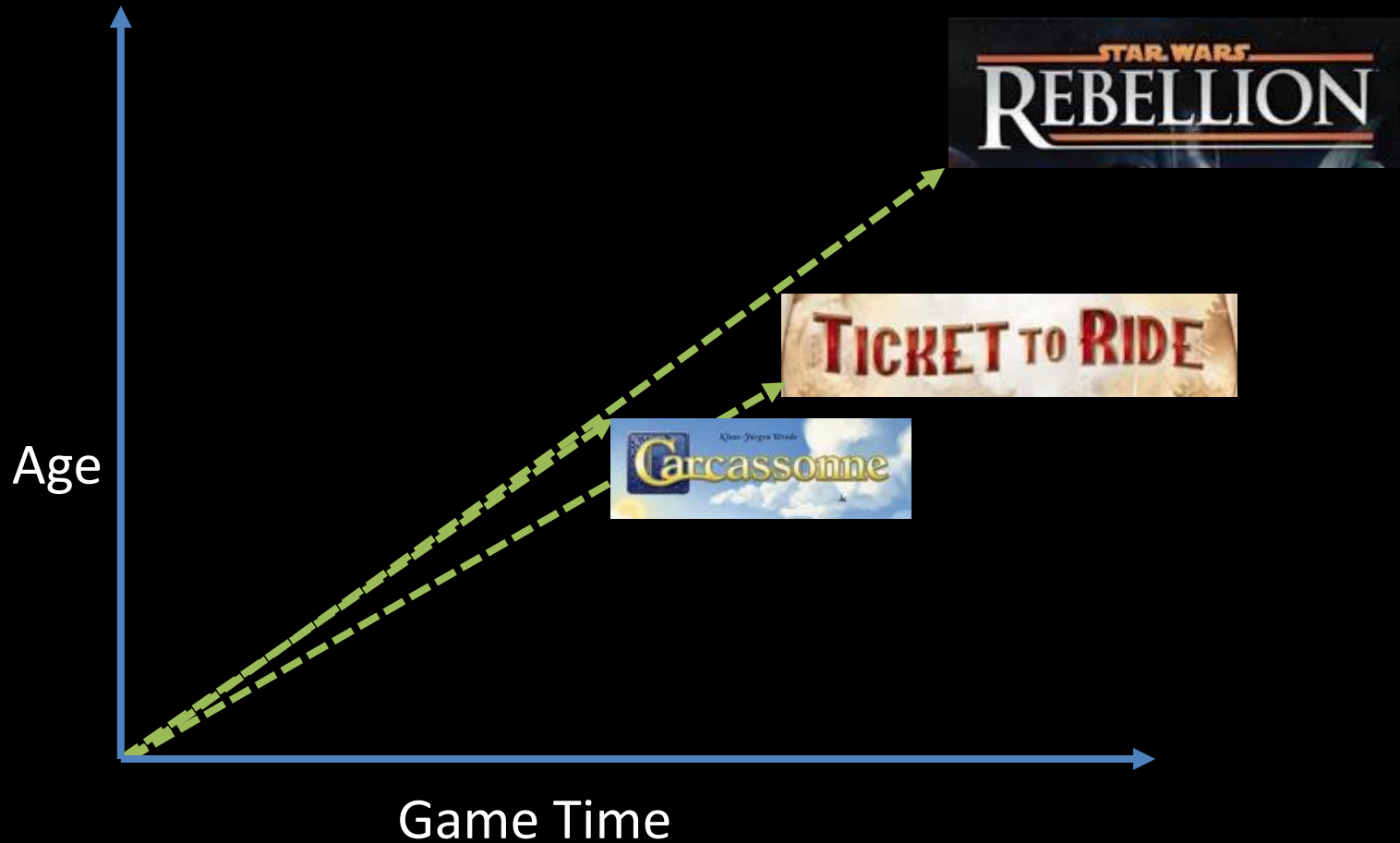




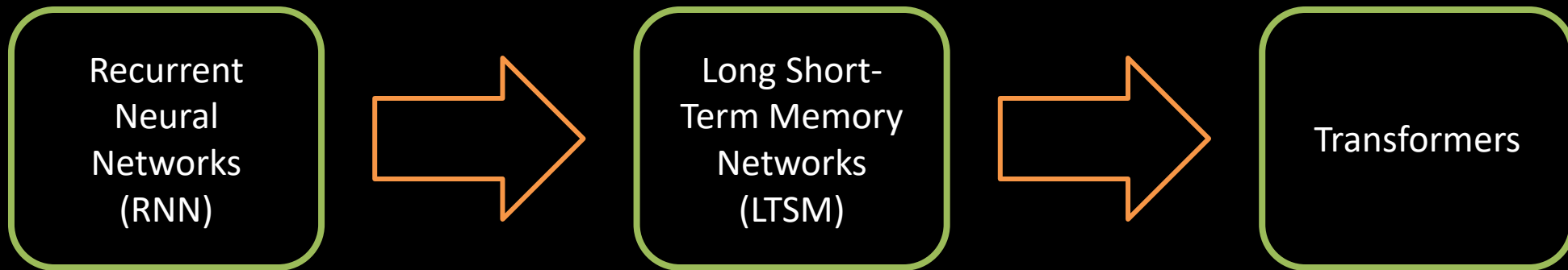
Euclidean vs. Cosine Similarity



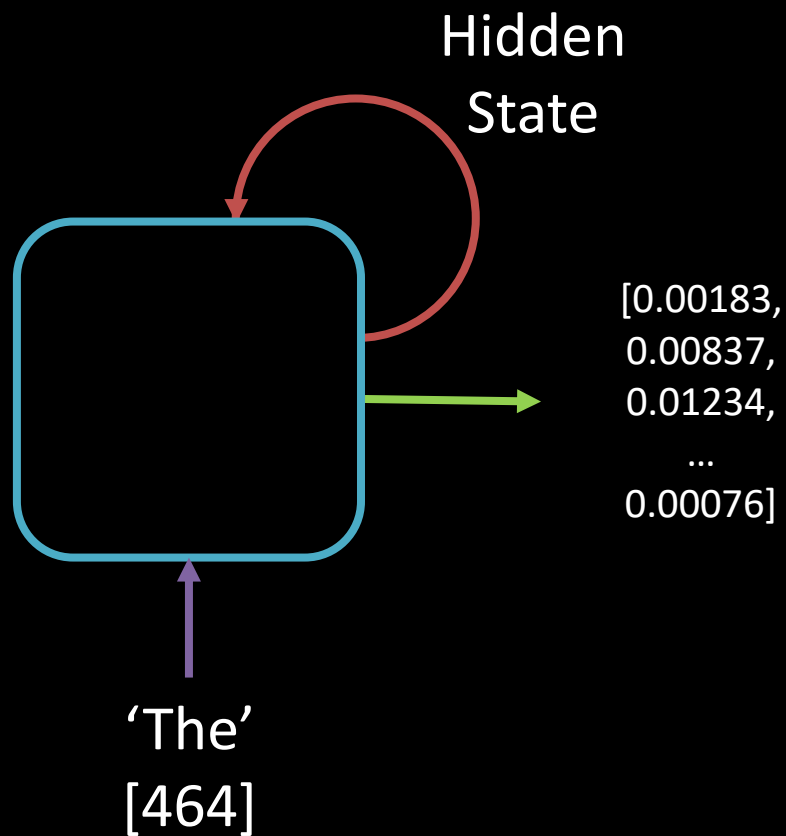
Euclidean vs. Cosine Similarity



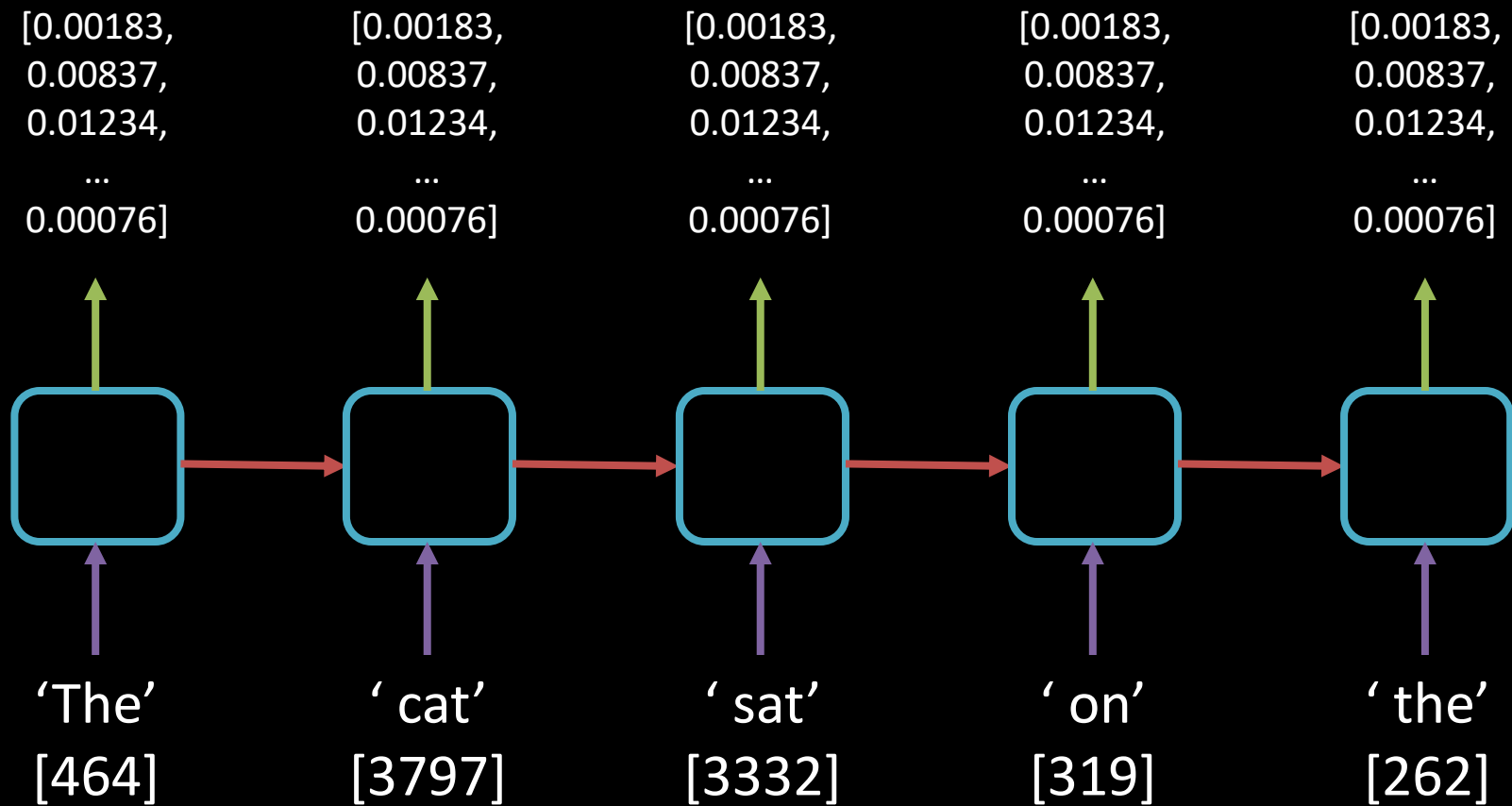
Sequence Prediction Model Evolution



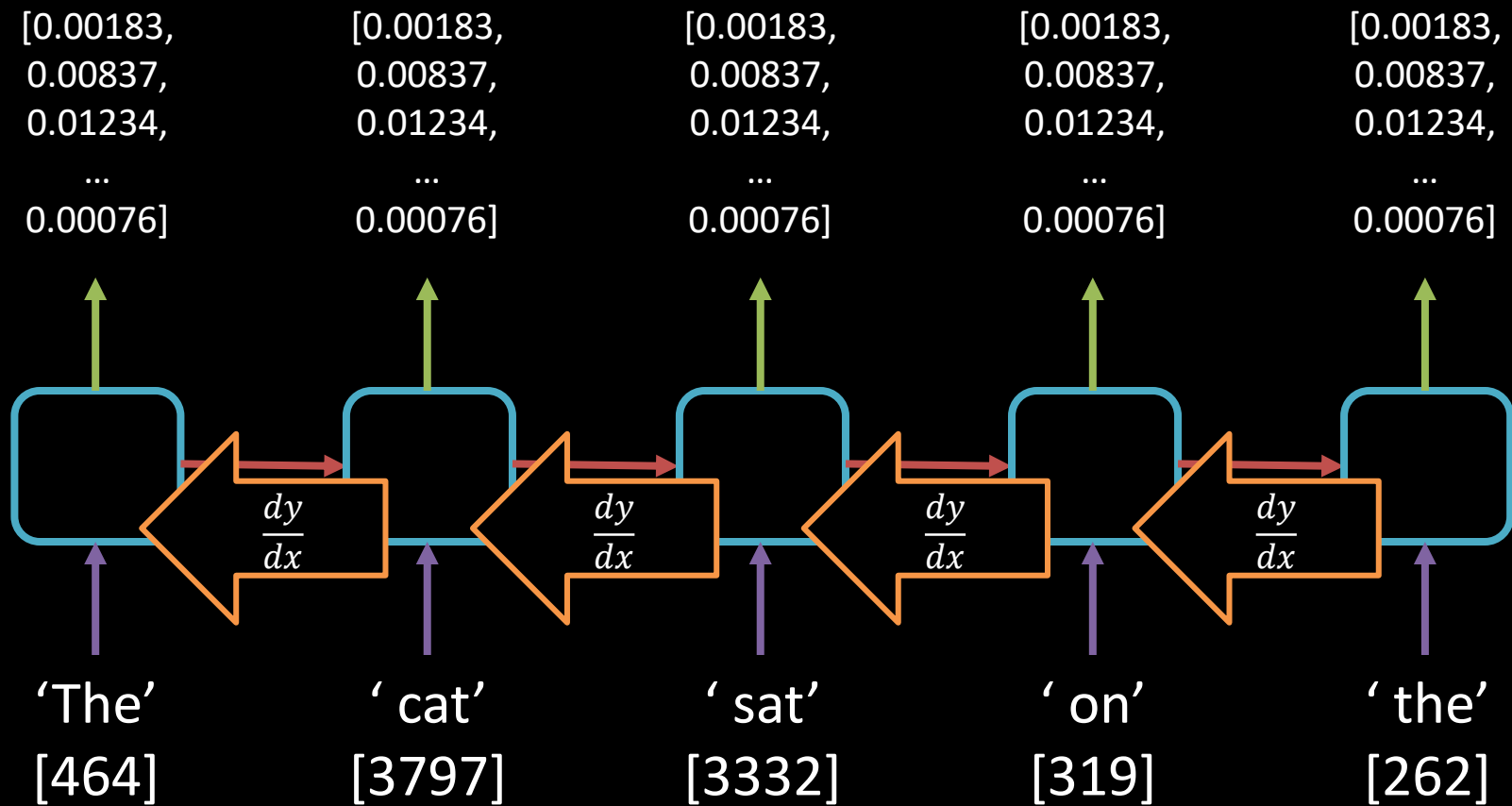
Recurrent Neural Networks (RNNs)



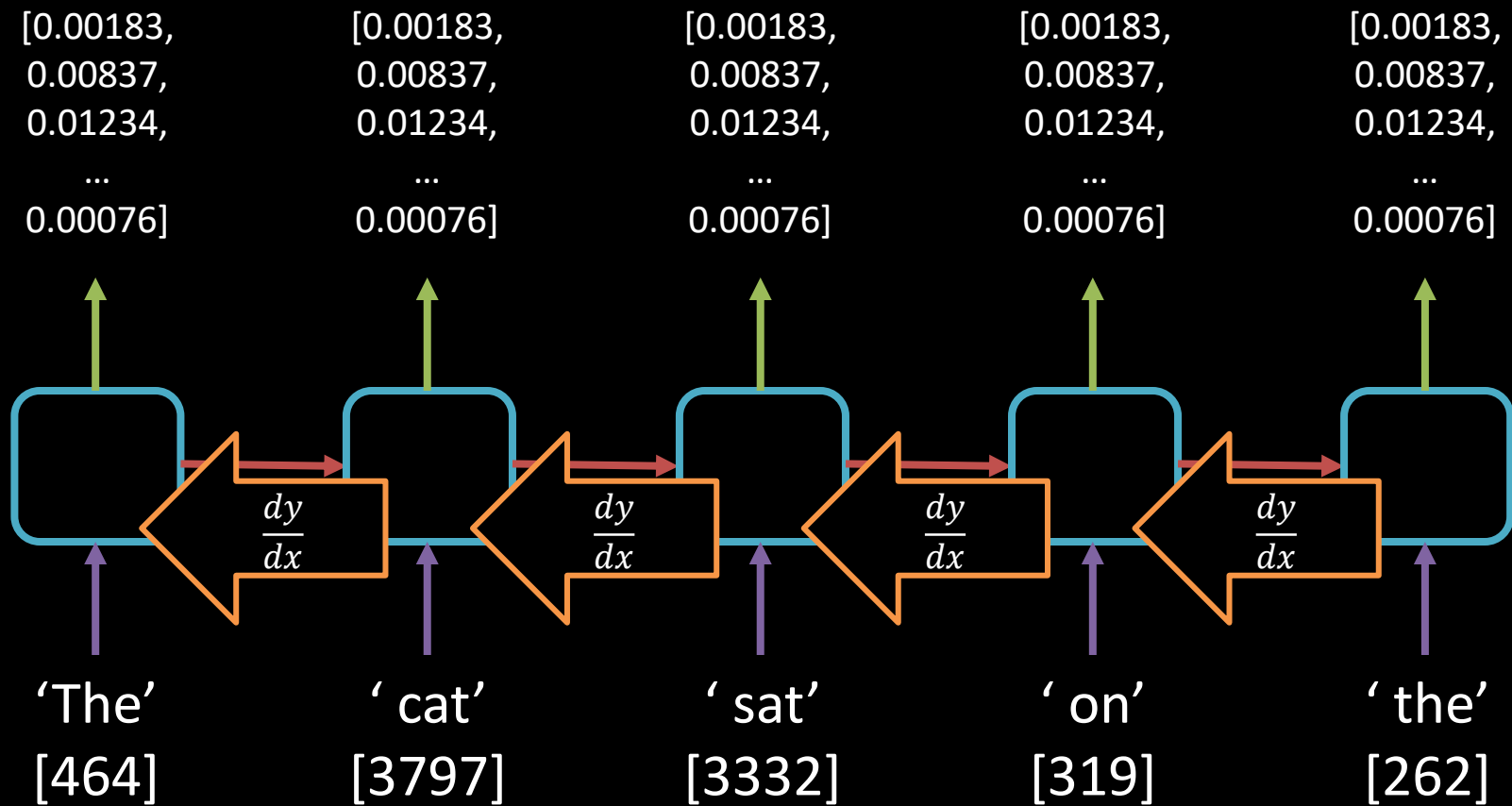
Unrolled RNN



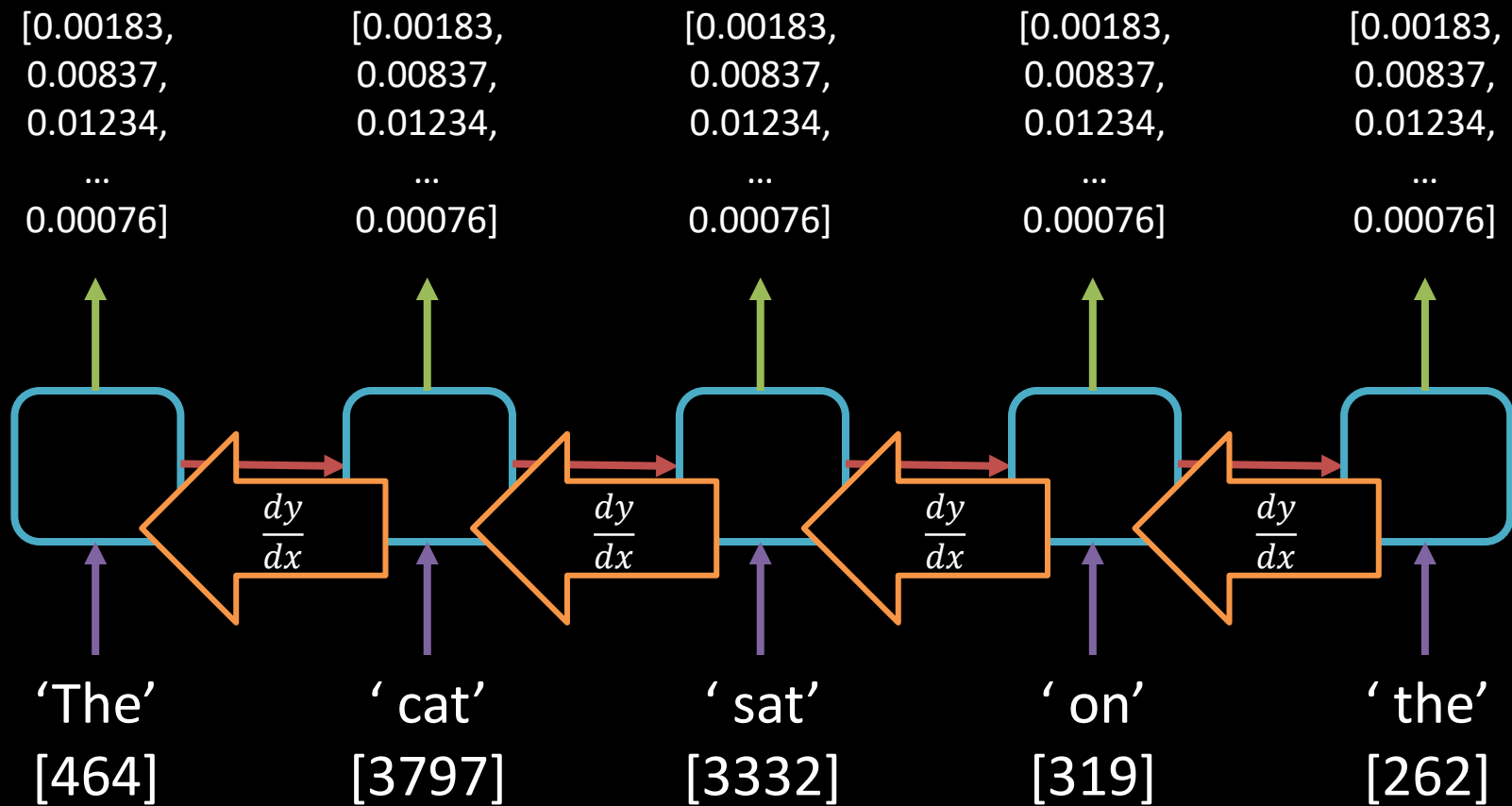
Backpropagation Through Time



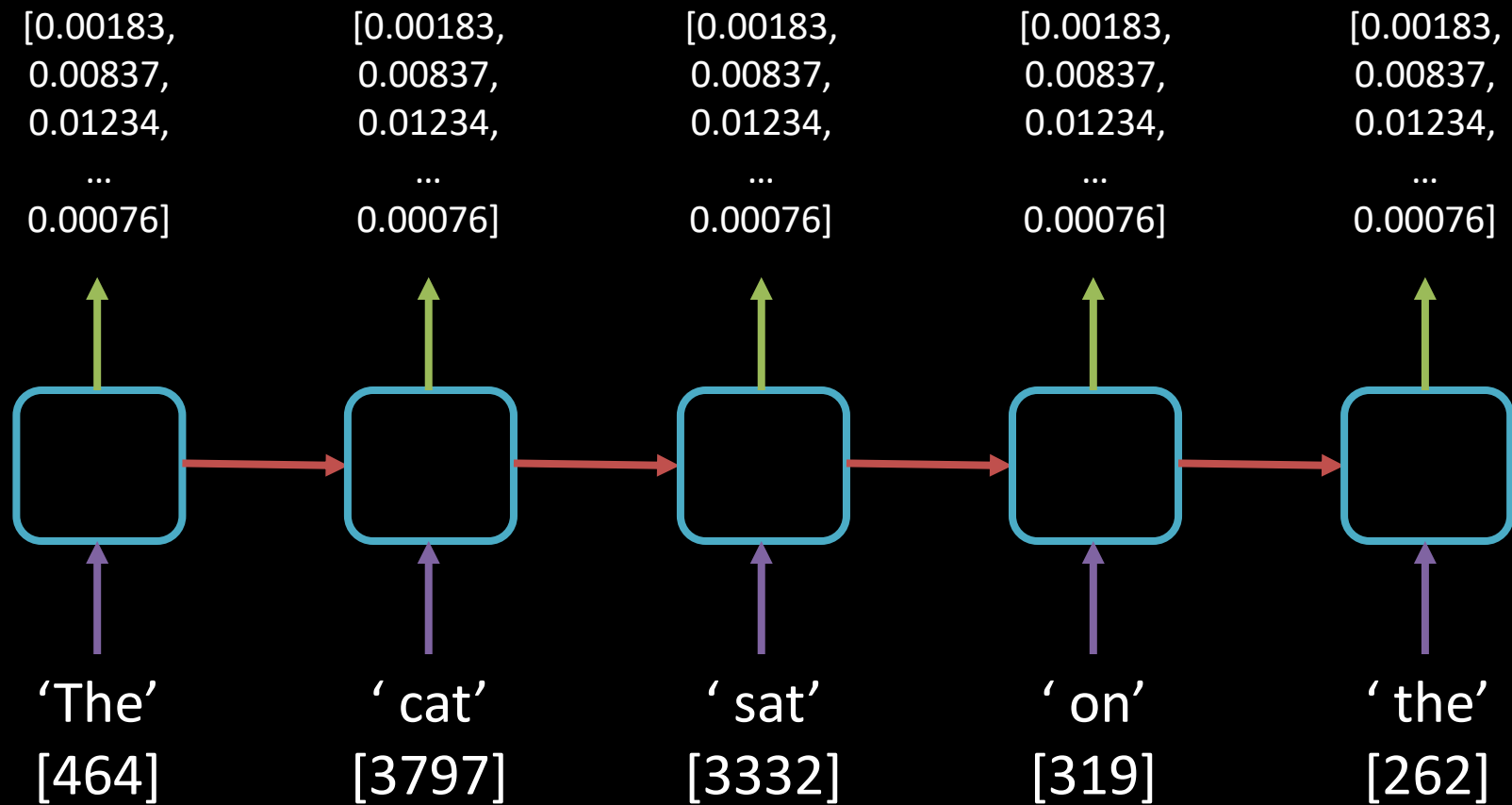
RNN Issues – Sequential Training



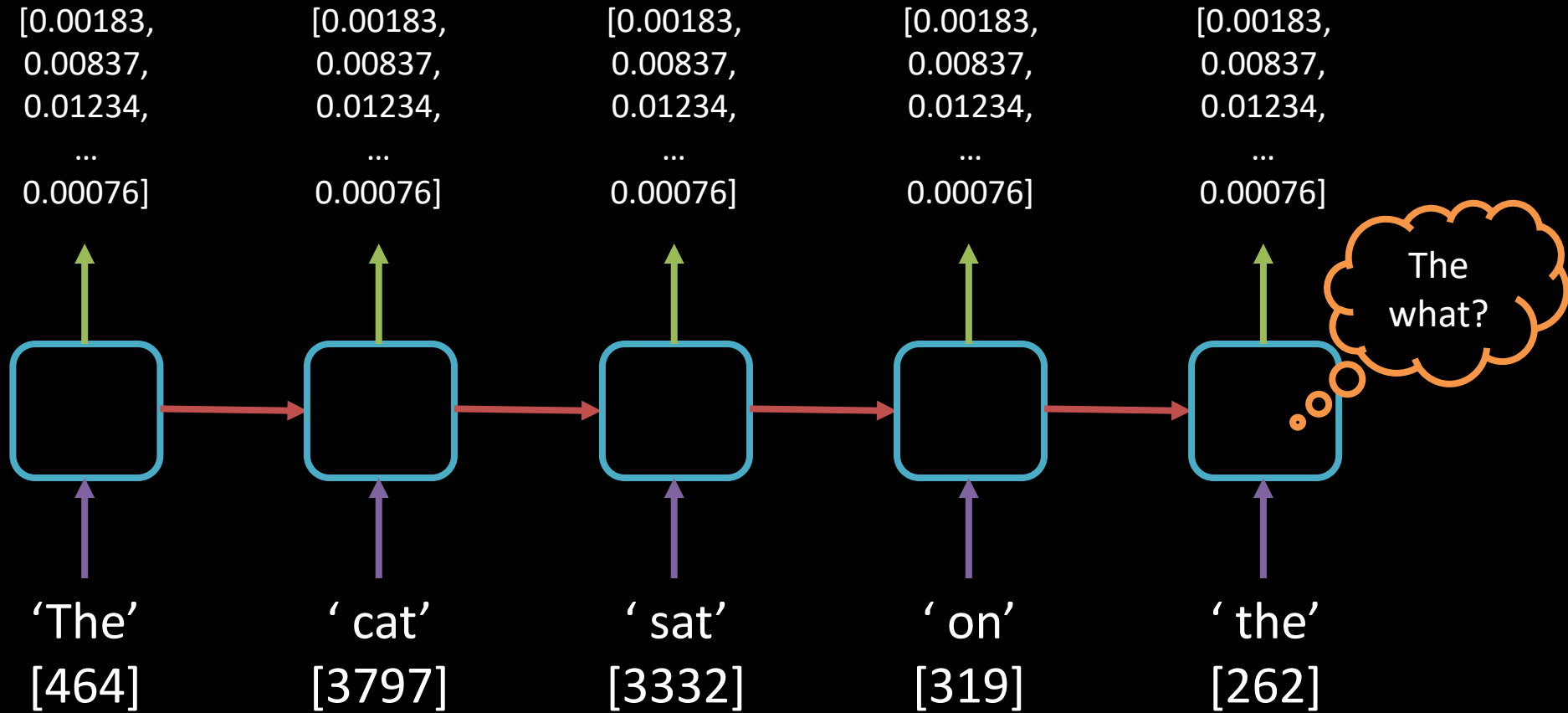
RNN Issues – Vanishing / Exploding Gradient



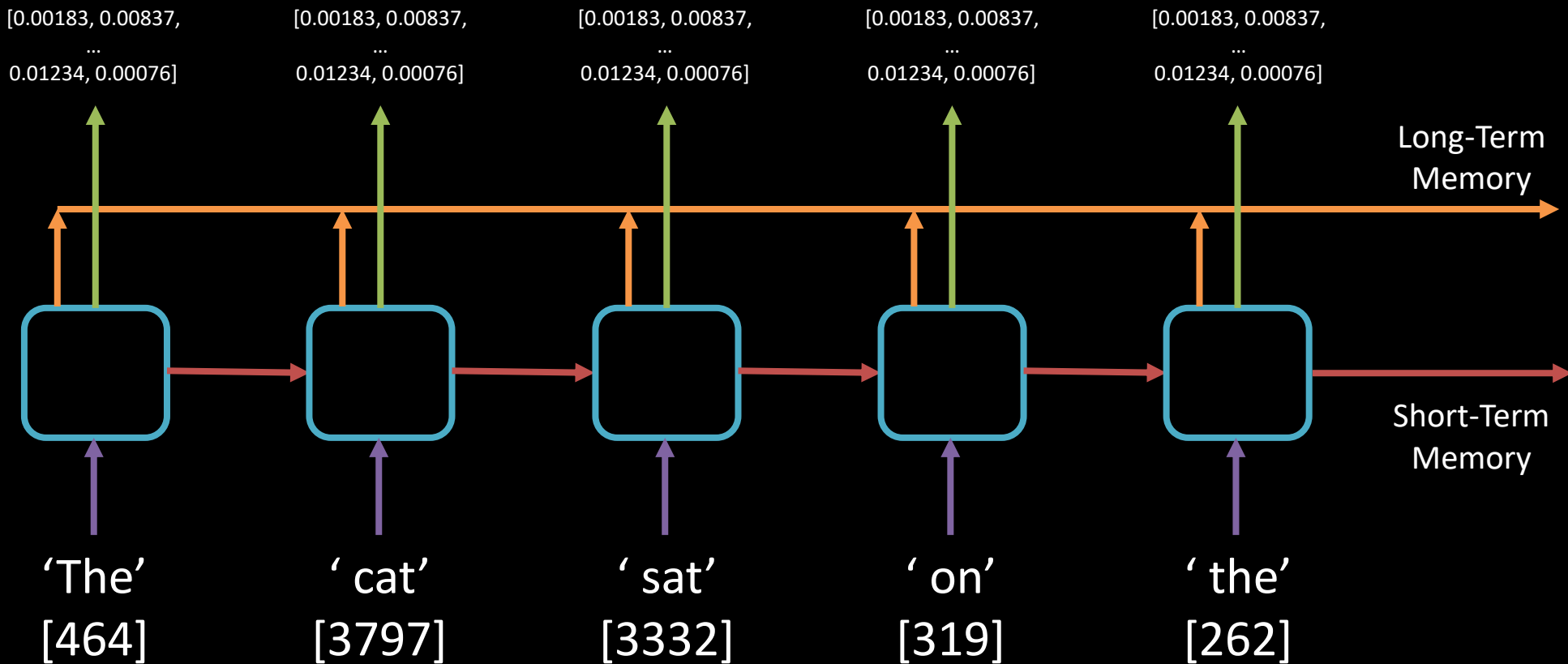
RNN Issues – Sequential Processing



RNN Issues – Information Loss

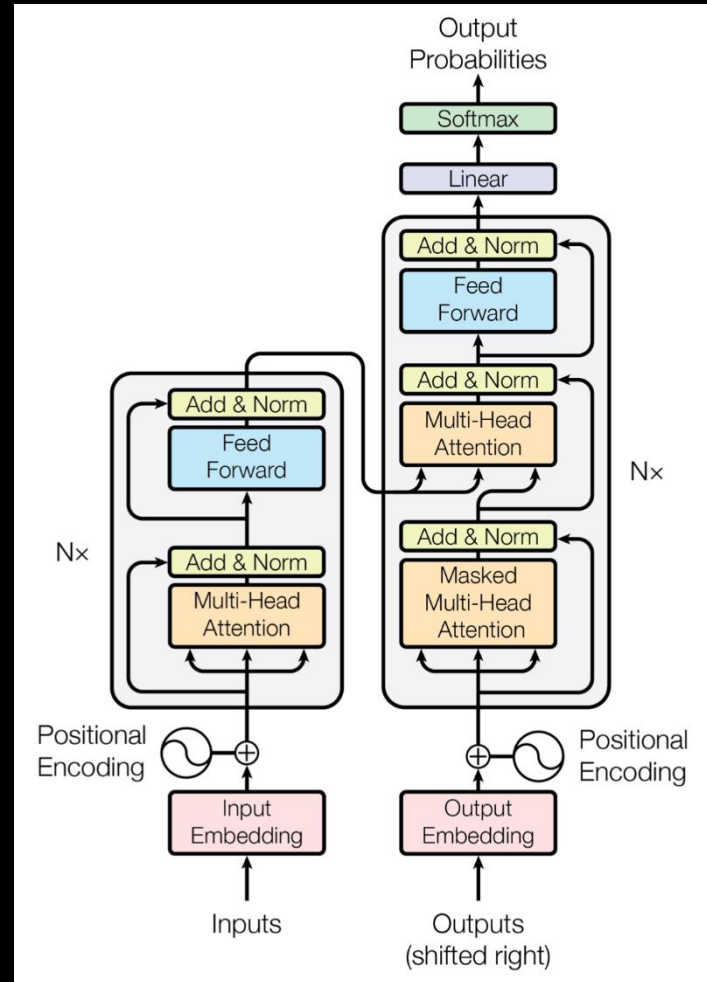


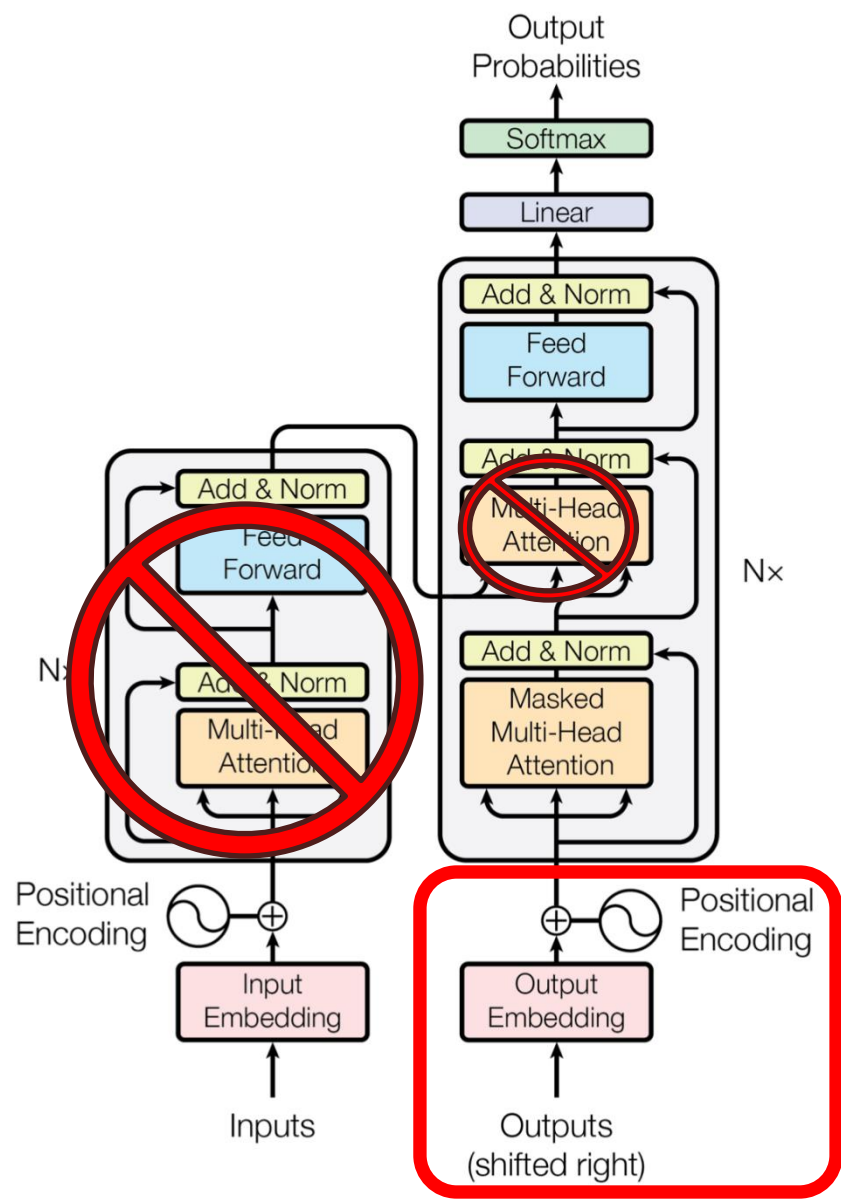
Long Short-Term Memory Networks (LSTMs)



Transformer Architecture

Attention
is all
You Need





Tokenization

Text Input

'The'

' cat'

' sat'

' on'

' the'

Tokenization

464

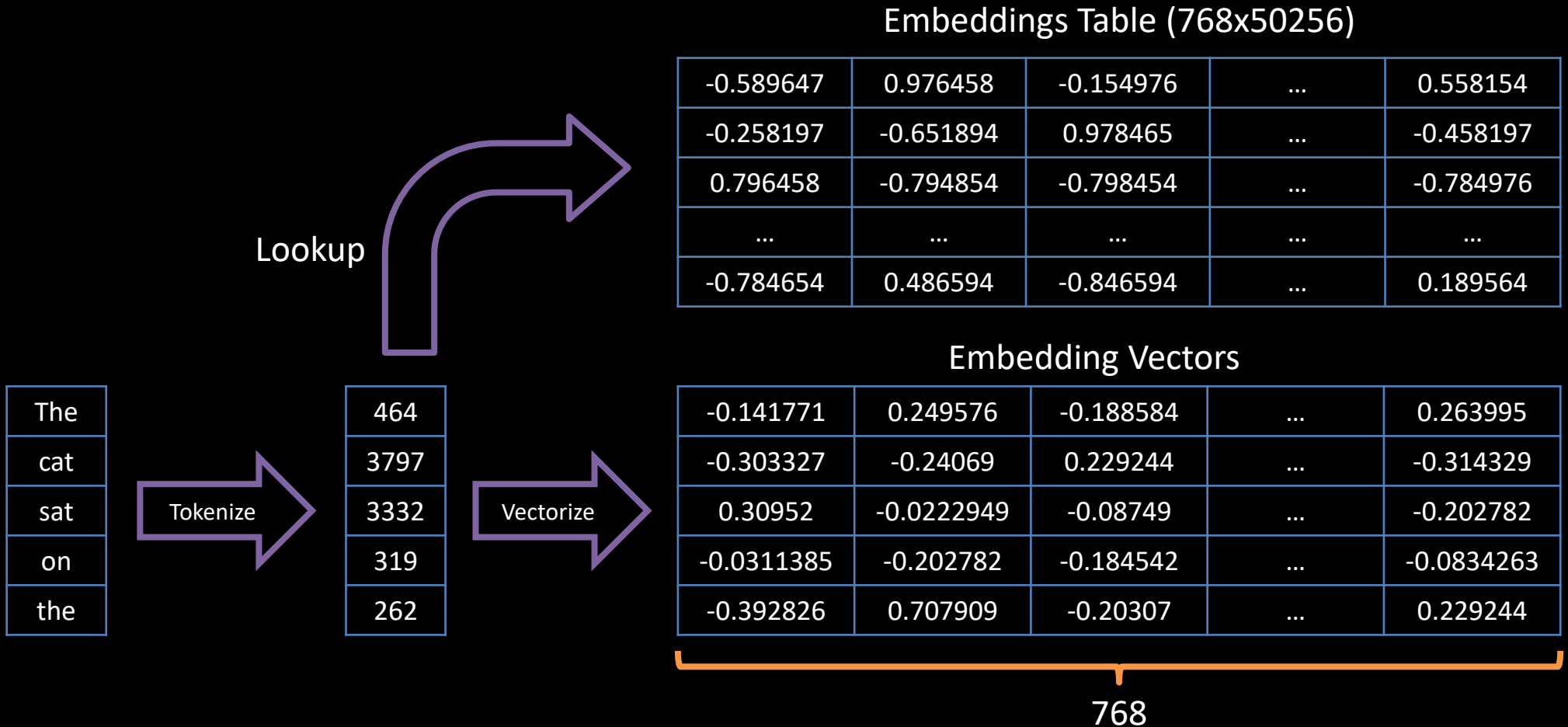
3797

3332

319

262

Token Embedding



Embedding Vectors

Text Input

'The'

' cat'

' sat'

' on'

' the'

Tokenization

464

3797

3332

319

262

Embedding
Vectors

[0.123456, 0.789012,
0.345678, 0.901234,
0.567890, 0.234567,
0.890123, 0.456789,
0.012345, 0.678901]

[0.987654, 0.321098,
0.654321, 0.098765,
0.432109, 0.765432,
0.109876, 0.543210,
0.876543, 0.210987]

[0.654987, 0.321654,
0.987321, 0.654098,
0.321765, 0.765432,
0.098654, 0.432176,
0.176543, 0.543210]

[0.210987, 0.987654,
0.654321, 0.321098,
0.098765, 0.765432,
0.432109, 0.109876,
0.876543, 0.543210]

[0.210987, 0.987654,
0.654321, 0.321098,
0.098765, 0.765432,
0.432109, 0.109876,
0.876543, 0.928466]

Positional Vectors

Text Input

'The'

' cat'

' sat'

' on'

' the'

Tokenization

464

3797

3332

319

262

Embedding
Vectors

[0.123456, 0.789012,
0.345678, 0.901234,
0.567890, 0.234567,
0.890123, 0.456789,
0.012345, 0.678901]

[0.987654, 0.321098,
0.654321, 0.098765,
0.432109, 0.765432,
0.109876, 0.543210,
0.876543, 0.210987]

[0.654987, 0.321654,
0.987321, 0.654098,
0.321765, 0.765432,
0.098654, 0.432176,
0.176543, 0.543210]

[0.210987, 0.987654,
0.654321, 0.321098,
0.098765, 0.765432,
0.432109, 0.109876,
0.876543, 0.543210]

[0.210987, 0.987654,
0.654321, 0.321098,
0.098765, 0.765432,
0.432109, 0.109876,
0.876543, 0.928466]

Positional
Vectors

[0.874512, 0.306159,
0.419873, 0.952637,
0.761945, 0.128740,
0.593682, 0.475019,
0.237649, 0.864203]

[0.642317, 0.789406,
0.135724, 0.468902,
0.973461, 0.254108,
0.826349, 0.517043,
0.409876, 0.632145]

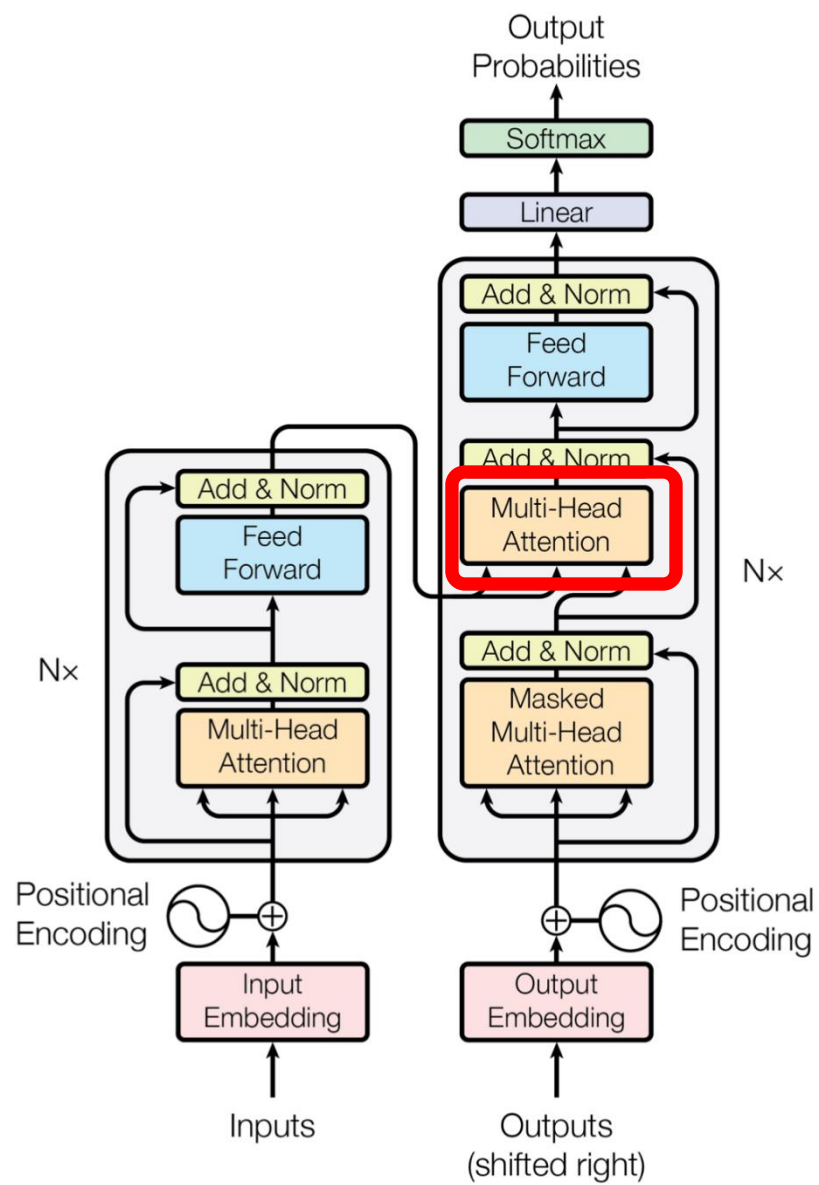
[0.298765, 0.741039,
0.657894, 0.123456,
0.849372, 0.564738,
0.912345, 0.678901,
0.427689, 0.305814]

[0.516273, 0.849162,
0.273849, 0.615728,
0.948361, 0.427689,
0.162738, 0.493827,
0.387649, 0.964205]

[0.729384, 0.165748,
0.948627, 0.372819,
0.615274, 0.839162,
0.473829, 0.261945,
0.837649, 0.594207]

Encoder Input Vectors

Text Input	'The'	' cat'	' sat'	' on'	' the'
Tokenization	464	3797	3332	319	262
Embedding Vectors	[0.123456, 0.789012, 0.345678, 0.901234, 0.567890, 0.234567, 0.890123, 0.456789, 0.012345, 0.678901]	[0.987654, 0.321098, 0.654321, 0.098765, 0.432109, 0.765432, 0.109876, 0.543210, 0.876543, 0.210987]	[0.654987, 0.321654, 0.987321, 0.654098, 0.321765, 0.765432, 0.098654, 0.432176, 0.176543, 0.543210]	[0.210987, 0.987654, 0.654321, 0.321098, 0.098765, 0.765432, 0.432109, 0.109876, 0.876543, 0.543210]	[0.210987, 0.987654, 0.654321, 0.321098, 0.098765, 0.765432, 0.432109, 0.109876, 0.876543, 0.928466]
+	+	+	+	+	+
Positional Vectors	[0.874512, 0.306159, 0.419873, 0.952637, 0.761945, 0.128740, 0.593682, 0.475019, 0.237649, 0.864203]	[0.642317, 0.789406, 0.135724, 0.468902, 0.973461, 0.254108, 0.826349, 0.517043, 0.409876, 0.632145]	[0.298765, 0.741039, 0.657894, 0.123456, 0.849372, 0.564738, 0.912345, 0.678901, 0.427689, 0.305814]	[0.516273, 0.849162, 0.273849, 0.615728, 0.948361, 0.427689, 0.162738, 0.493827, 0.387649, 0.964205]	[0.729384, 0.165748, 0.948627, 0.372819, 0.615274, 0.839162, 0.473829, 0.261945, 0.837649, 0.594207]
=	=	=	=	=	=
Encoder Input Vectors	[0.364829, 0.917463, 0.728394, 0.564182, 0.193847, 0.629385, 0.849162, 0.374819, 0.627384, 0.918273]	[0.182736, 0.495827, 0.374819, 0.628394, 0.917463, 0.284716, 0.563728, 0.491837, 0.648372, 0.193847]	[0.495827, 0.182736, 0.628394, 0.374819, 0.284716, 0.917463, 0.491837, 0.563728, 0.193847, 0.648372]	[0.837462, 0.294817, 0.628394, 0.473829, 0.182736, 0.948627, 0.495827, 0.364829, 0.917463, 0.728394]	[0.294817, 0.837462, 0.473829, 0.628394, 0.948627, 0.182736, 0.364829, 0.495827, 0.728394, 0.917463]



Self Attention

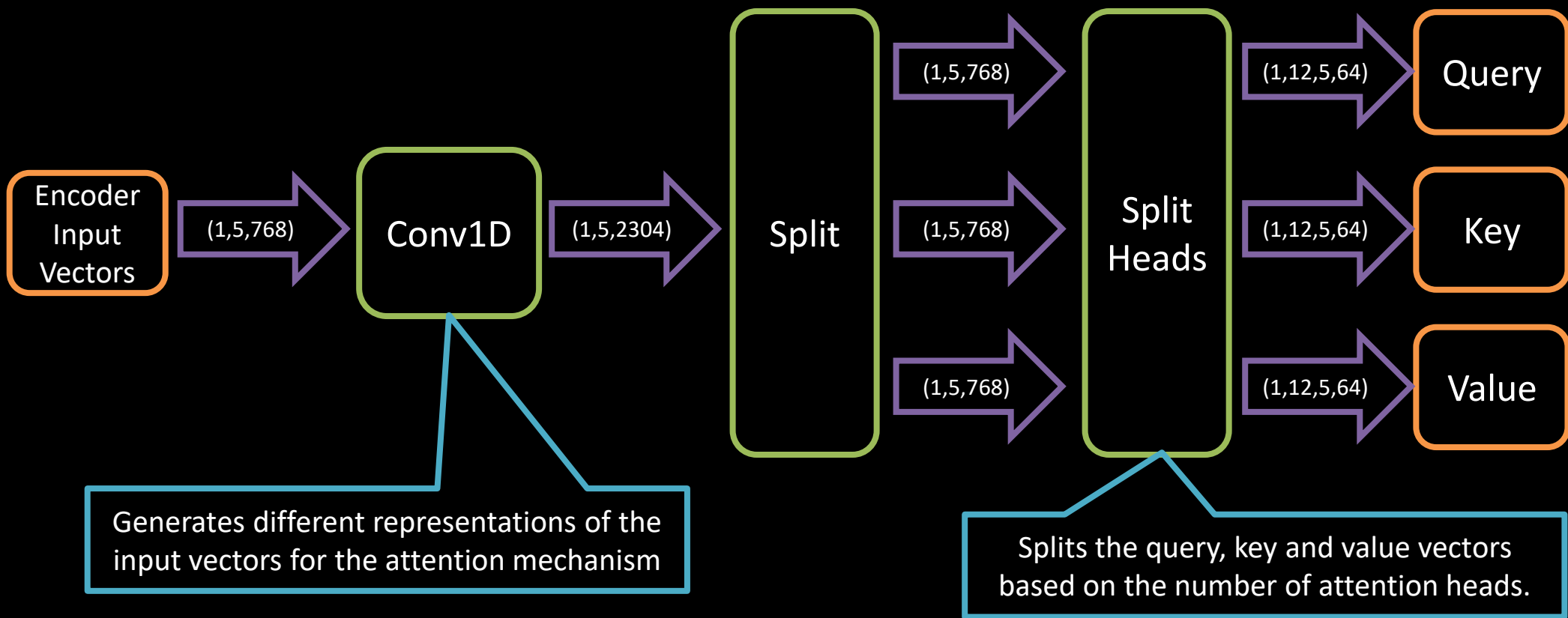
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Q – Query

K – Key

V – Value

Generating Query, Key & Value



Query, Key & Value

Query

The current token that the model is focusing on.

Key

A token from the input sequence that the model is comparing to the query.

Value

Information associated with the key token that the model wants to retrieve

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

$Q(5, 64)$

[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.495, 0.182, 0.628, 0.374, 0.284, 0.917, 0.491, 0.563, ..., 0.648]
[0.837, 0.294, 0.628, 0.473, 0.182, 0.948, 0.495, 0.364, ..., 0.728]
[0.294, 0.837, 0.473, 0.628, 0.948, 0.182, 0.364, 0.495, ..., 0.917]

$K^T(64, 5)$

[0.128, 0.593, 0.475, 0.237, 0.864]
[0.254, 0.826, 0.517, 0.409, 0.632]
[0.564, 0.912, 0.678, 0.427, 0.305]
...
[0.427, 0.162, 0.493, 0.387, 0.964]

$\times$

$=$

	The	cat	sat	on	the
The	0.472	0.193	0.045	0.163	0.127
cat	0.193	0.412	0.297	0.034	0.064
sat	0.045	0.297	0.488	0.105	0.065
on	0.163	0.034	0.105	0.436	0.262
the	0.127	0.064	0.065	0.262	0.482

$\sqrt{64}$

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

	The	cat	sat	on	the
The	0.472	0.193	0.045	0.163	0.127
cat	0.193	0.412	0.297	0.034	0.064
sat	0.045	0.297	0.488	0.105	0.065
on	0.163	0.034	0.105	0.436	0.262
the	0.127	0.064	0.065	0.262	0.482

X

V(5, 64)

[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.495, 0.182, 0.628, 0.374, 0.284, 0.917, 0.491, 0.563, ..., 0.648]
[0.837, 0.294, 0.628, 0.473, 0.182, 0.948, 0.495, 0.364, ..., 0.728]
[0.294, 0.837, 0.473, 0.628, 0.948, 0.182, 0.364, 0.495, ..., 0.917]

=

Attention(64, 5)

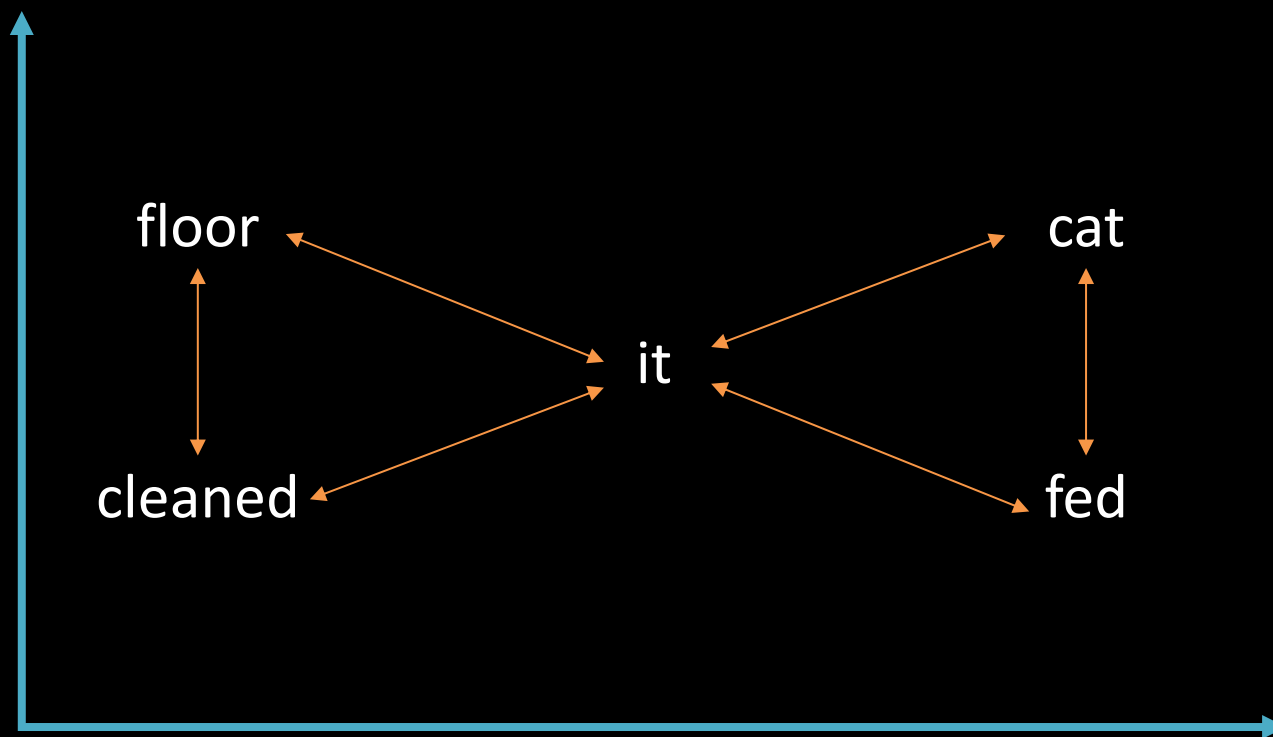
[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.364, 0.917, 0.728, 0.564, 0.193, 0.629, 0.849, 0.374, ..., 0.918]
[0.495, 0.182, 0.628, 0.374, 0.284, 0.917, 0.491, 0.563, ..., 0.648]
[0.837, 0.294, 0.628, 0.473, 0.182, 0.948, 0.495, 0.364, ..., 0.728]
[0.294, 0.837, 0.473, 0.628, 0.948, 0.182, 0.364, 0.495, ..., 0.917]

Meaning, position &
relationship to other words.

Attention Example

The cat sat on the floor, it had just been cleaned.

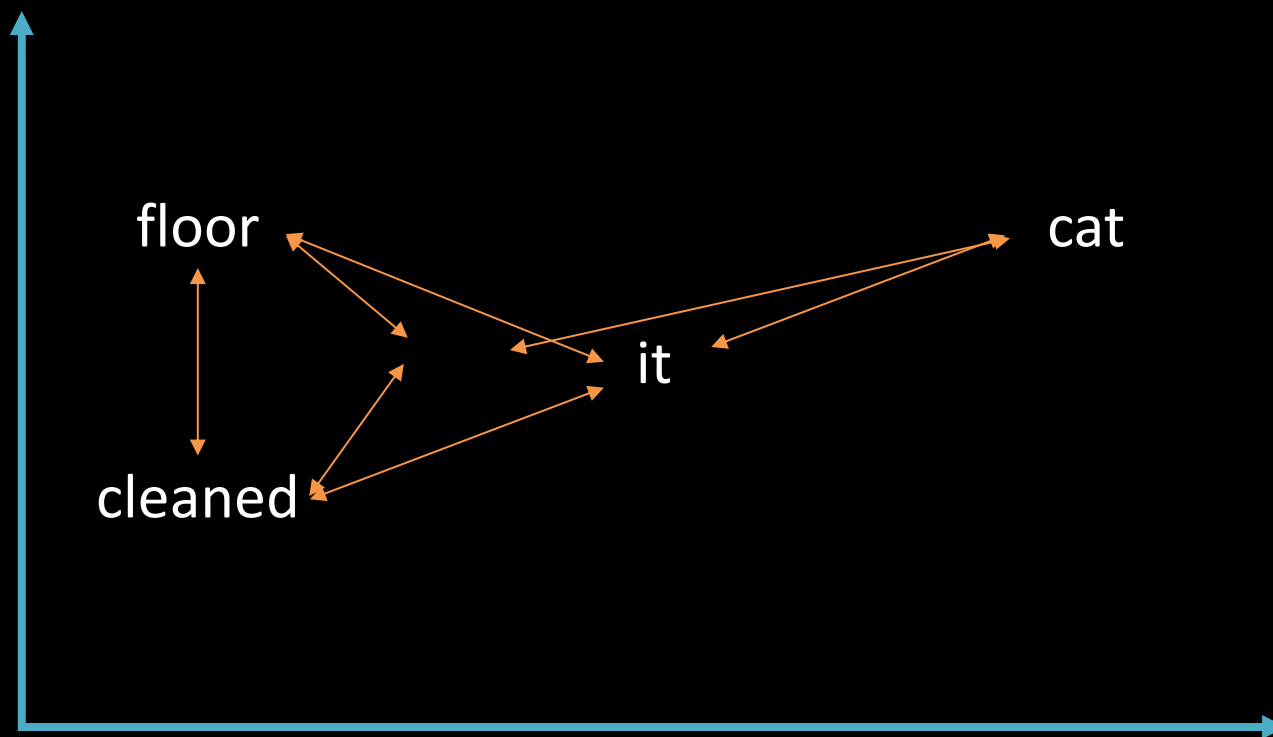
The cat sat on the floor, it had just been fed.

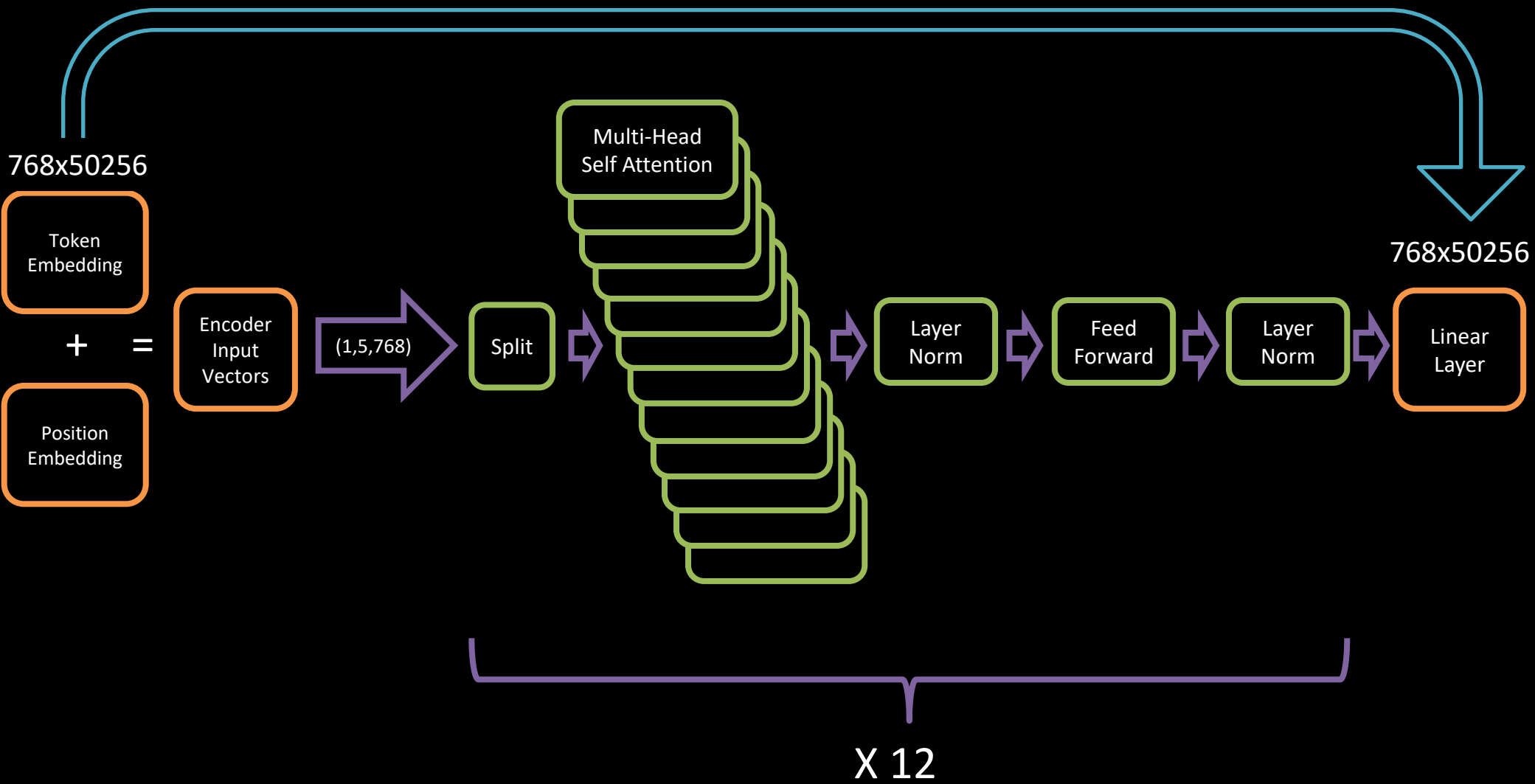


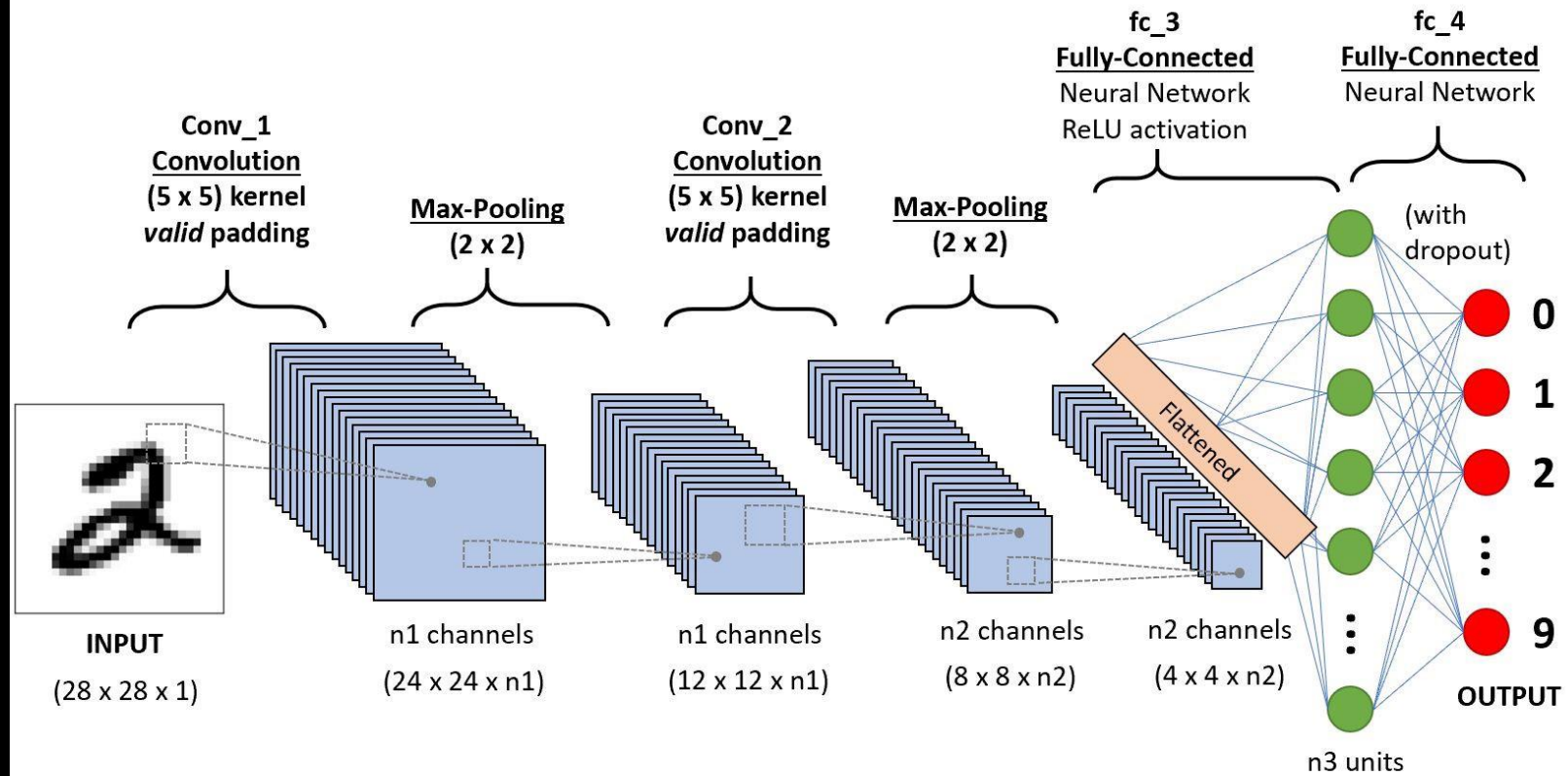
Attention Example

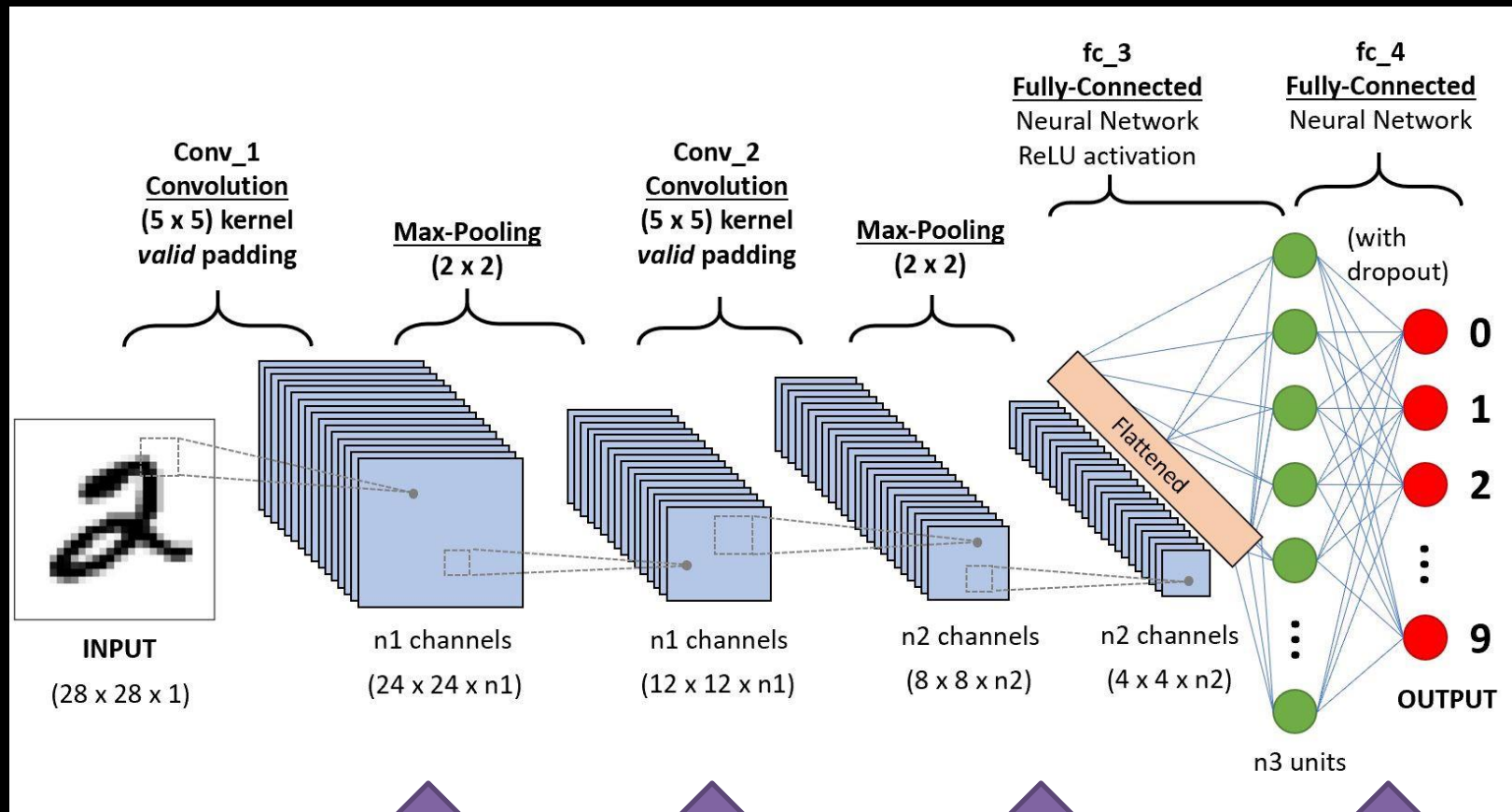
The cat sat on the floor, it had just been cleaned.

The cat sat on the floor, it had just been fed.







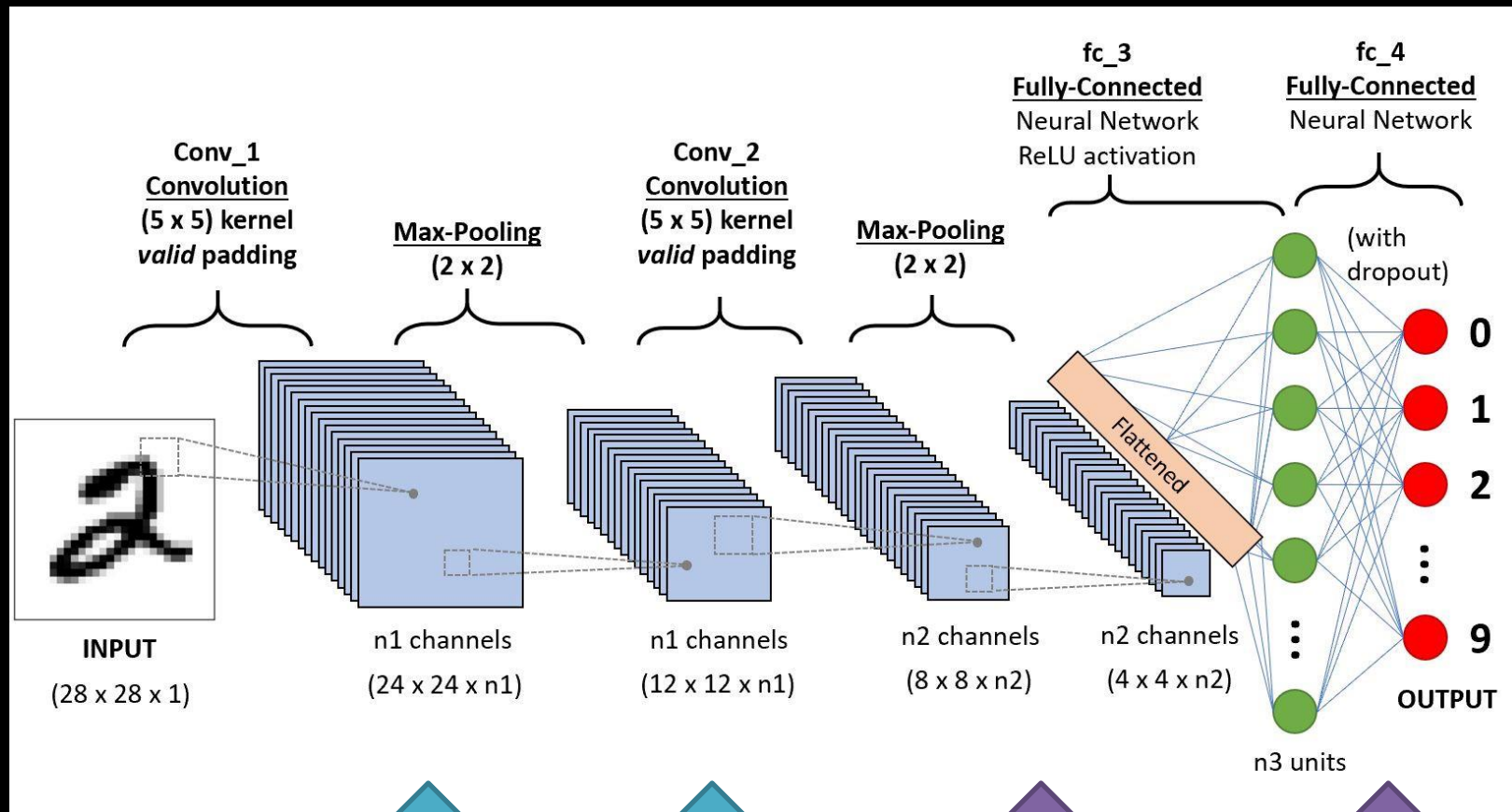


Edge
Detection

Shape
Detection

Feature
Detection

Classification



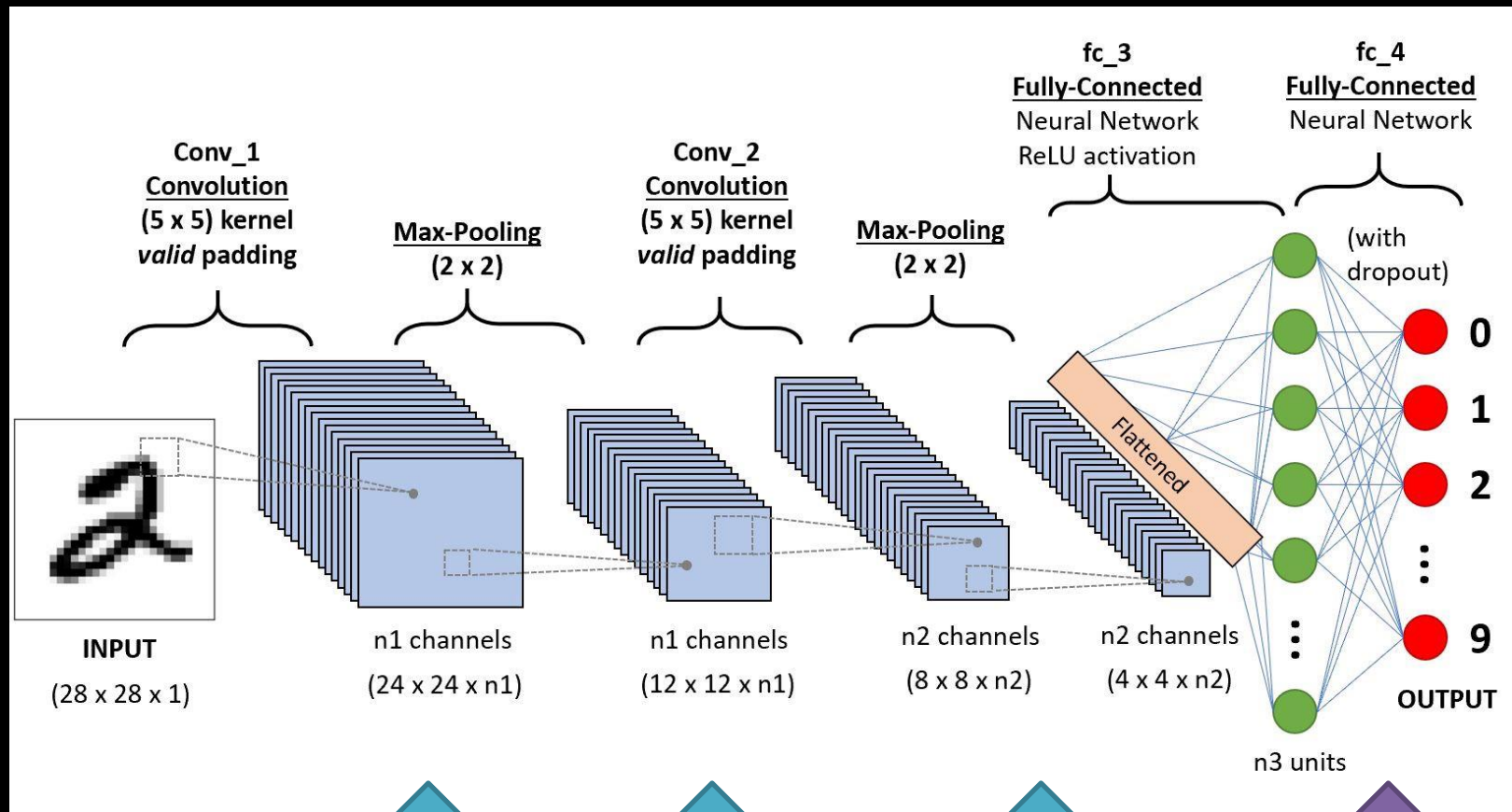
Frozen
Layers

Edge
Detection

Shape
Detection

Feature
Detection

Classification



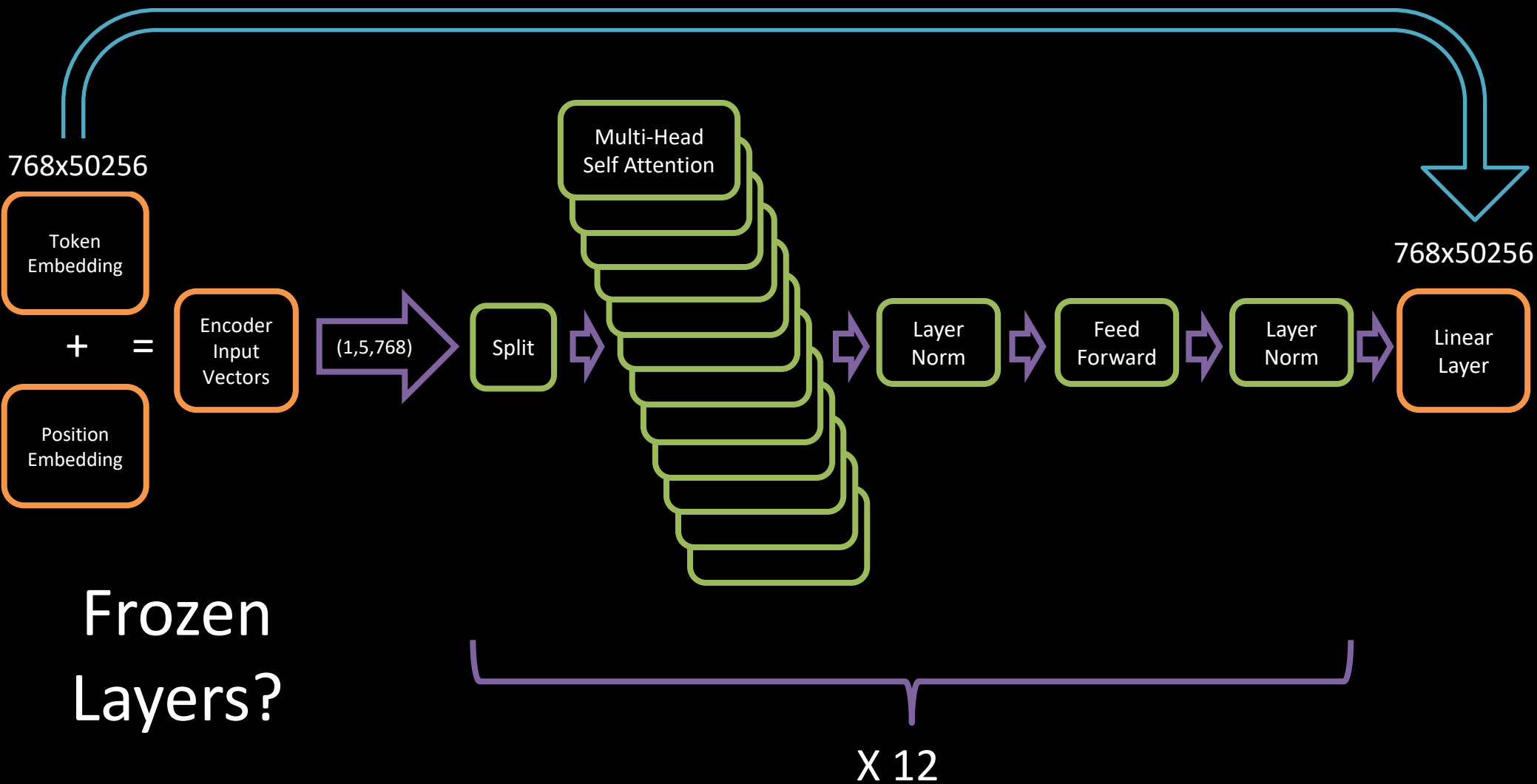
Frozen
Layers

Edge
Detection

Shape
Detection

Feature
Detection

Classification



Parameters

Parameter	Description
-----------	-------------

Parameters

Parameter	Description
temperature	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.

Temperature & Softmax

```
scores = scores / self.temperature
```

```
next_token_probs = scores.softmax(dim=-1)
```

$$\text{softmax}(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}}$$

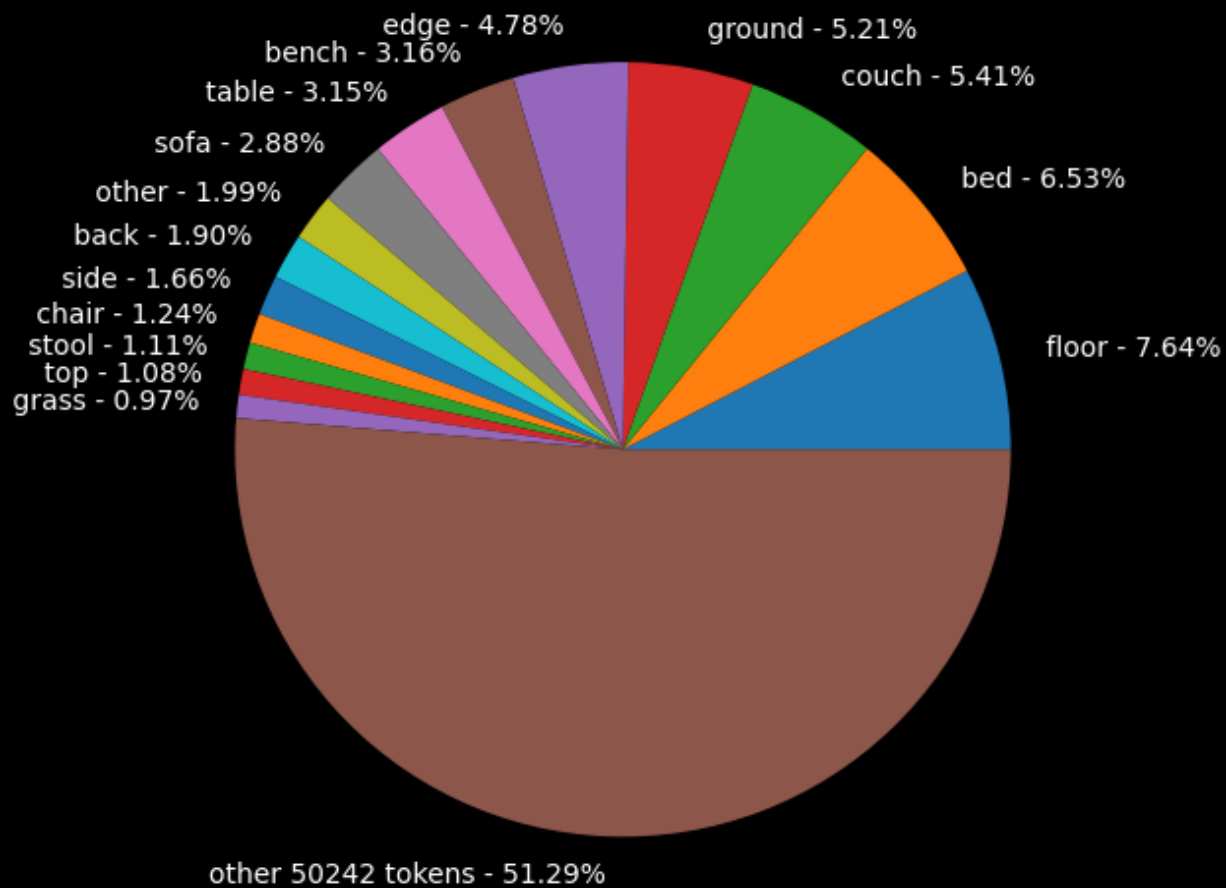
The softmax function converts a vector of K real numbers into a probability distribution of K possible outcomes.

The softmax function is often used as the last activation function of a neural network to normalize the output of a network to a probability distribution over predicted output classes.

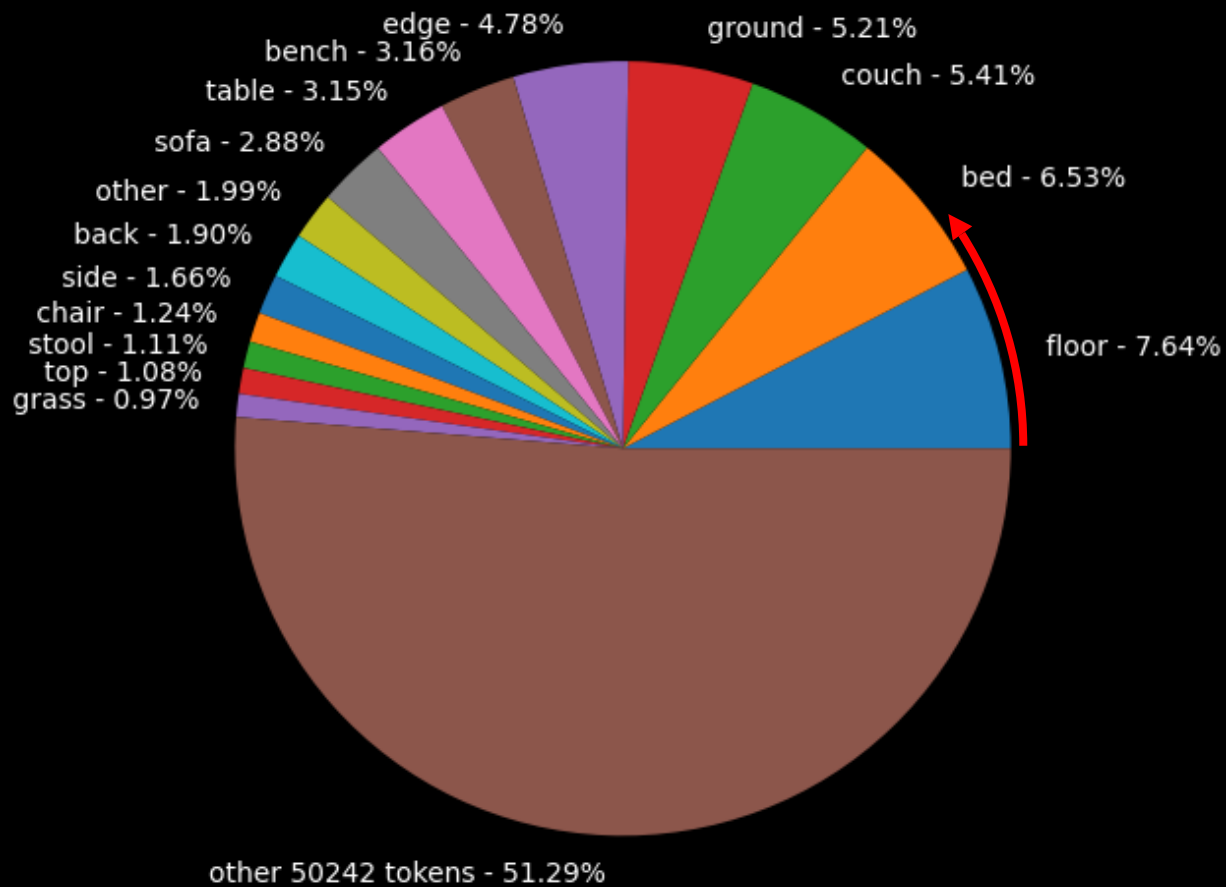
Parameters

Parameter	Description
temperature	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.
top_p	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.

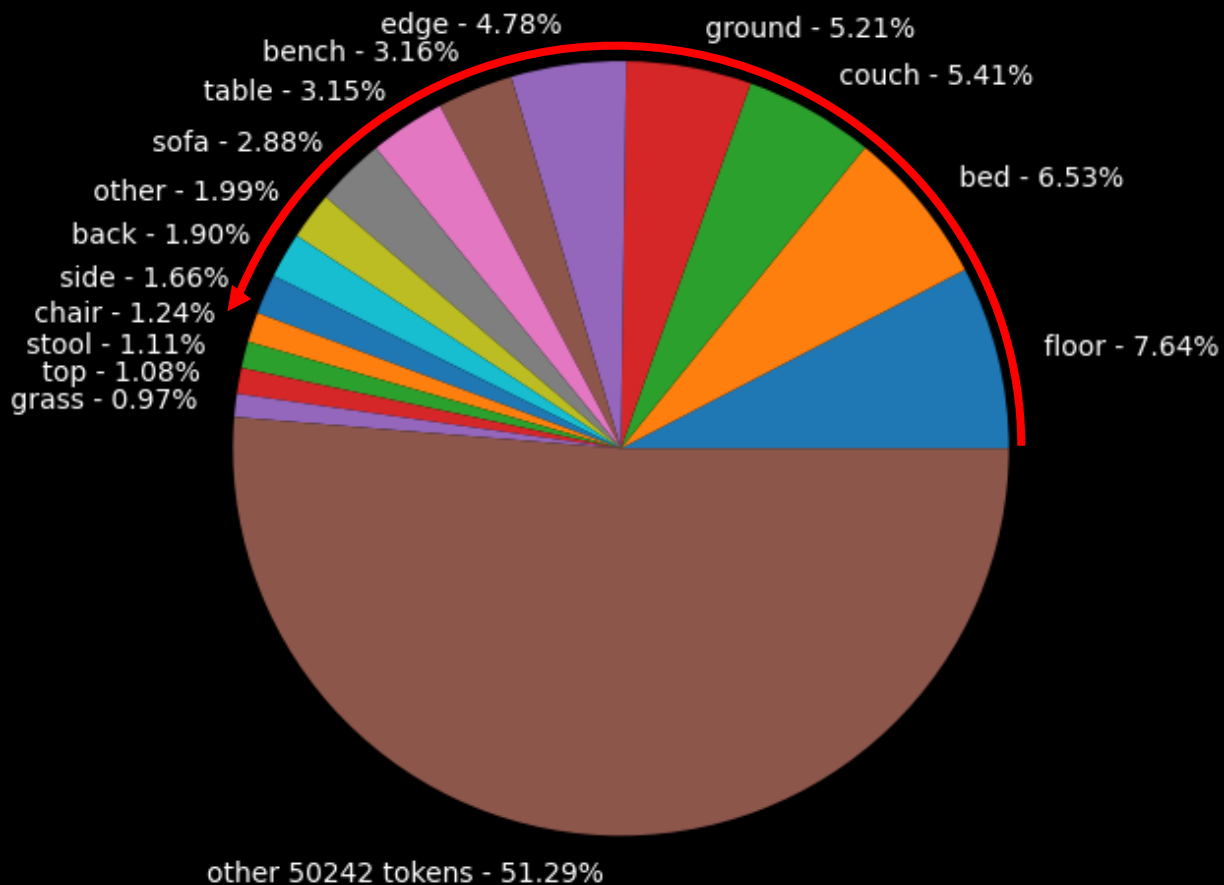
Top P



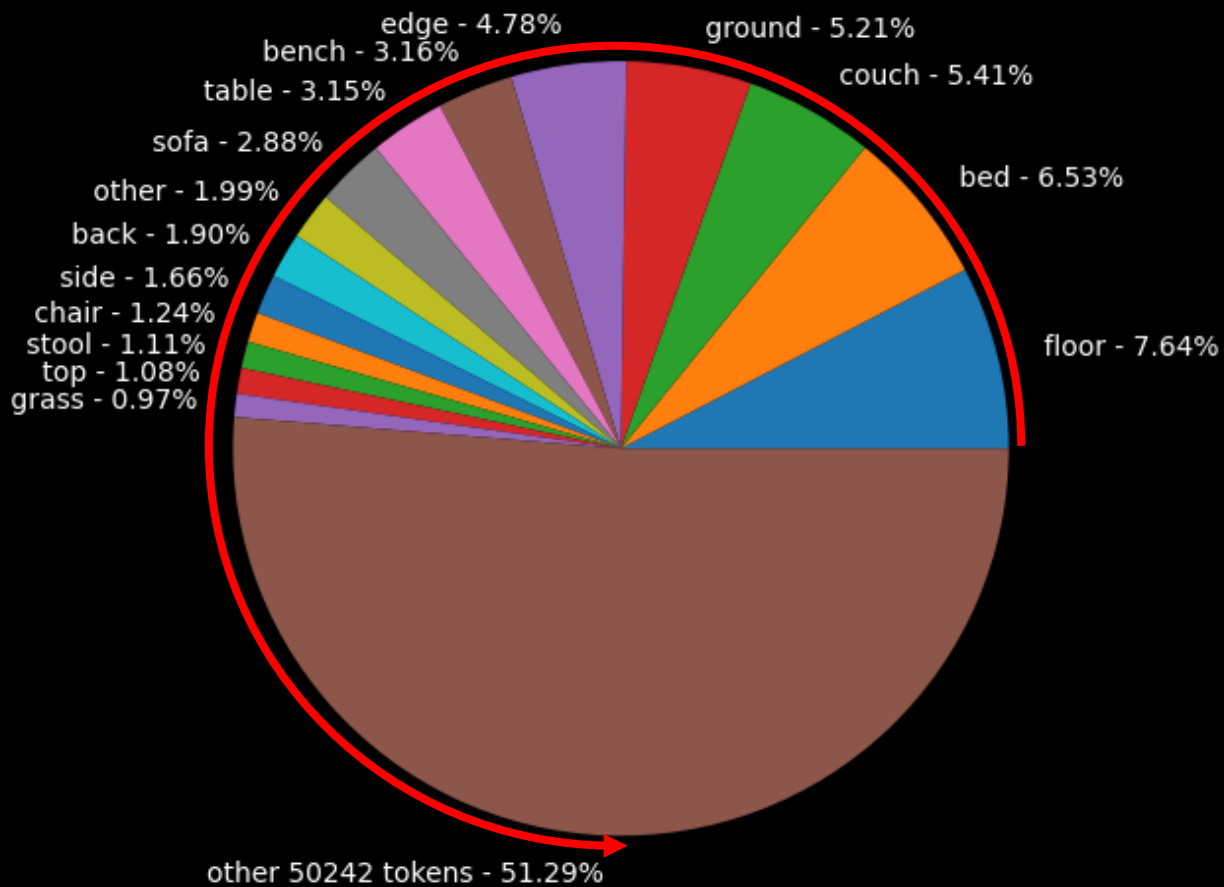
Top P = 0.1



Top P = 0.45



Top P = 0.75



Clear chat

Playground settings

View code

Show JSON



Start chatting

Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.

Type user query here. (Shift + Enter for new line)

Configuration

Deployment

Parameters

Max response ⓘ



800

Temperature ⓘ



0.7

Top P ⓘ



0.95

Stop sequence ⓘ

Stop sequences

Frequency penalty ⓘ



0

Presence penalty ⓘ



0

[Learn more](#)

Current token count ⓘ

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.
stop	Up to 4 sequences where the API will stop generating further tokens.

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.
stop	Up to 4 sequences where the API will stop generating further tokens.
max_tokens	The maximum number of tokens to generate in the chat completion.

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.
stop	Up to 4 sequences where the API will stop generating further tokens.
max_tokens	The maximum number of tokens to generate in the chat completion.
presence_penalty	Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.
stop	Up to 4 sequences where the API will stop generating further tokens.
max_tokens	The maximum number of tokens to generate in the chat completion.
presence_penalty	Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.
frequency_penalty	Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.

Parameters

Parameter	Description
temperature	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.
top_p	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.
n	How many chat completion choices to generate for each input message.
stop	Up to 4 sequences where the API will stop generating further tokens.
max_tokens	The maximum number of tokens to generate in the chat completion.
presence_penalty	Number between -2.0 and 2.0. Positive values penalize new tokens based on whether they appear in the text so far, increasing the model's likelihood to talk about new topics.
frequency_penalty	Number between -2.0 and 2.0. Positive values penalize new tokens based on their existing frequency in the text so far, decreasing the model's likelihood to repeat the same line verbatim.
logit_bias	Modify the likelihood of specified tokens appearing in the completion.

Beam Search

≡ Beam search

🗺️ 12 languages ▼

Article [Talk](#)

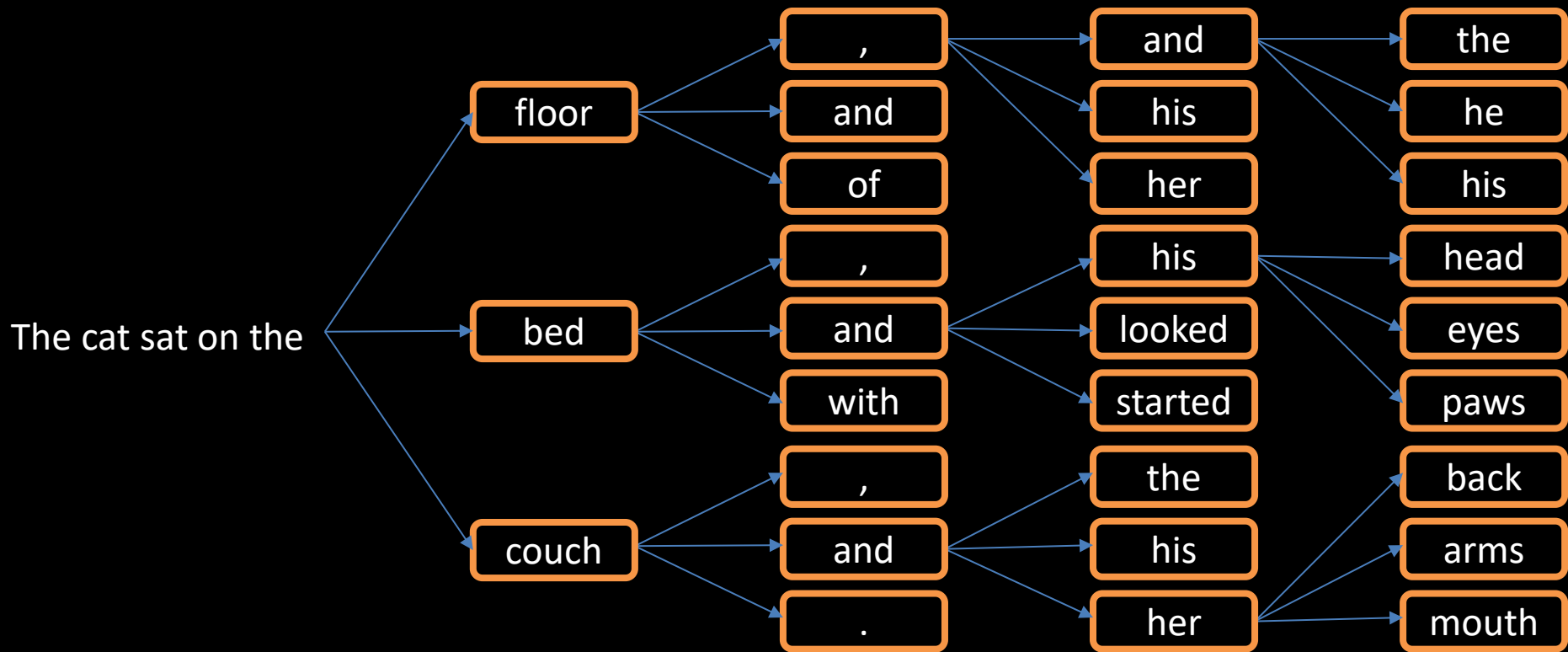
Tools ▼

From Wikipedia, the free encyclopedia

In [computer science](#), **beam search** is a [heuristic search algorithm](#) that explores a graph by expanding the most promising node in a limited set. Beam search is an optimization of [best-first search](#) that reduces its memory requirements. Best-first search is a graph search which orders all partial solutions (states) according to some heuristic. But in beam search, only a predetermined number of best partial solutions are kept as candidates.^[1] It is thus a [greedy algorithm](#).

The term "beam search" was coined by [Raj Reddy](#) of [Carnegie Mellon University](#) in 1977.^[2]

Beam Search Example



Thanks!

Alan Smith – Active Solution

