

LARGE LANGUAGE MODELS

Ryan Daniels

Senior Machine Learning Engineer

Accelerate Programme for Scientific Discovery

WHAT DOES AI MEAN TO YOU?

WHAT AI MEANS TO ME

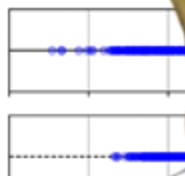
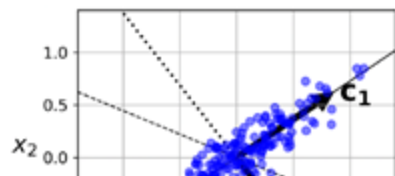
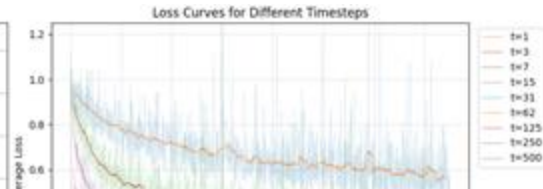


Image Loss per Batch



Algorithm 1 Training

- 1: **repeat**
- 2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
- 3: $t \sim \text{Uniform}(\{1, \dots, T\})$
- 4: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 5: Take gradient descent step on
$$\nabla_{\theta} \left\| \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2$$
- 6: **until** converged

Algorithm 2 Sampling

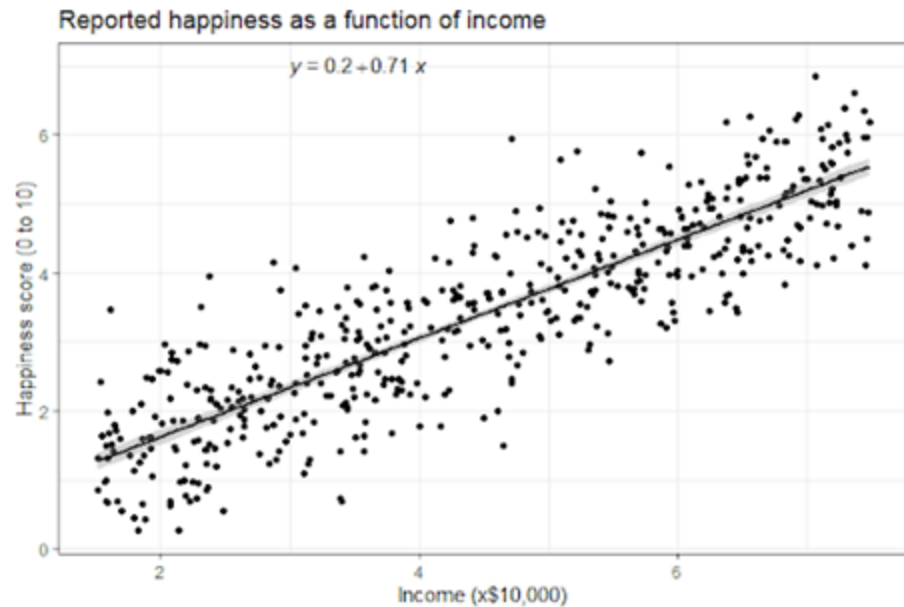
- 1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- 2: **for** $t = T, \dots, 1$ **do**
- 3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
- 4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_{\theta}(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
- 5: **end for**
- 6: **return** \mathbf{x}_0

Fig. 2. Overlap-tile strategy for seamless segmentation of arbitrary large images (here segmentation of neuronal structures in EM stacks). Prediction of the segmentation in the yellow area, requires image data within the blue area as input. Missing input data is extrapolated by mirroring

WHAT AI MEANS TO ME

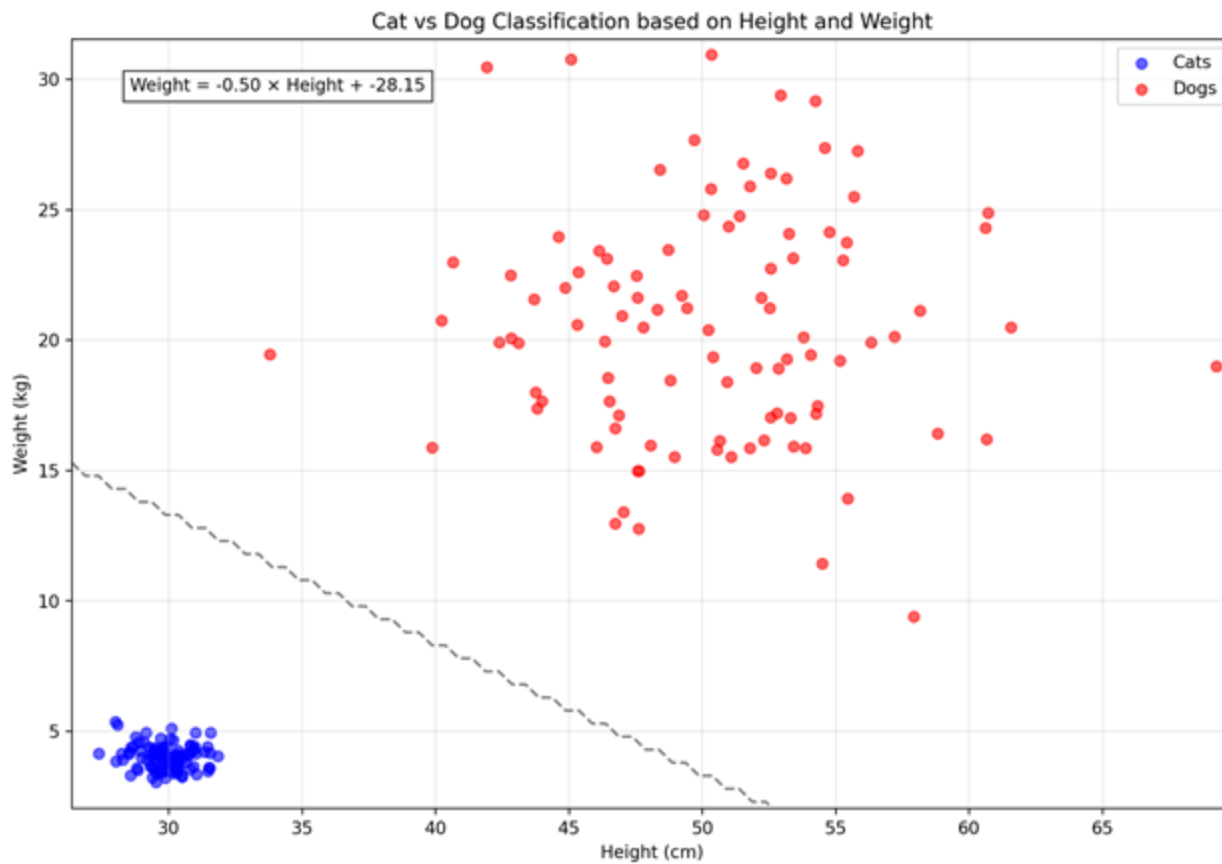


Rule-based



Data-based

WHAT AI MEANS TO ME



THE HARDWARE

OR: WHY AI IS SUPER EXPENSIVE

WHEN EVERYONE DIGS FOR GOLD

GPUS

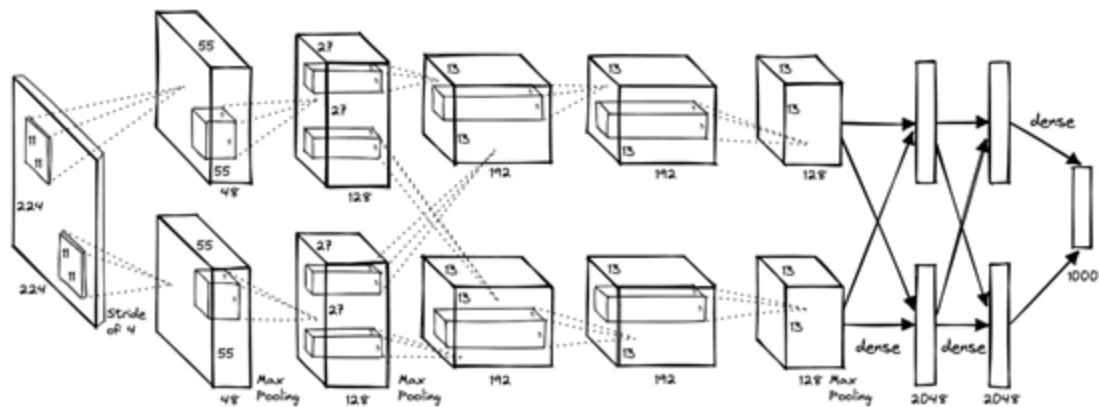
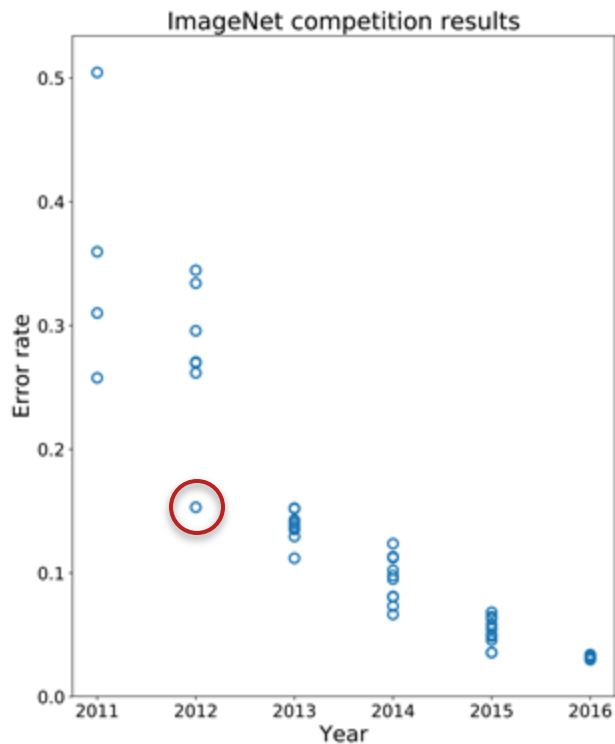




GPUS



THE IMPORTANT BITS



AlexNet

THE IMPORTANT BITS

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

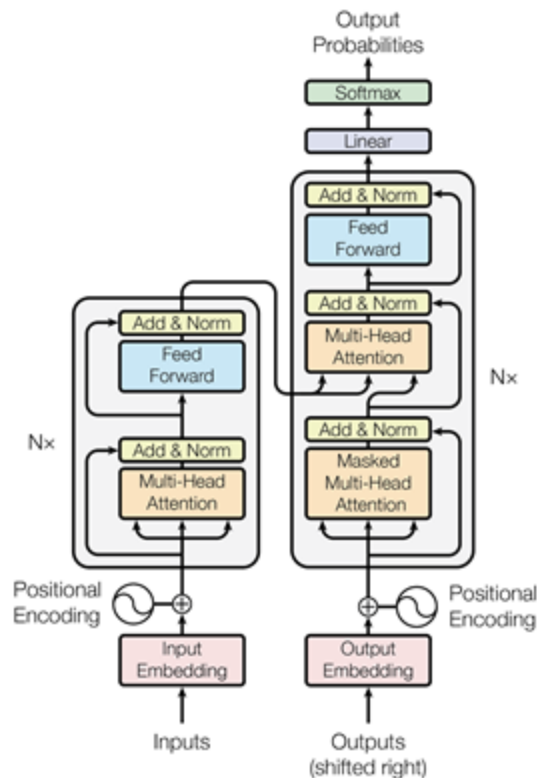


Figure 1: The Transformer - model architecture.

SOME MISUNDERSTANDINGS

MODEL	A parameterized function that maps input text sequences to output text distributions. It has billions of parameters that are optimized by minimizing a prediction loss function over a massive corpus of text.	ChatGPT Claude Gemini
LEARNING	Synonymous with training - the process where weights are systematically modified through algorithms like gradient descent to minimize some objective function.	Recognition of preferences, adaptation of the writing style, and recall of previous information.
CONTEXT	The input to the model.	Important information that might be necessary to perform a certain task.

LLMs

THE GREATEST HITS

BEFORE WE BEGIN...

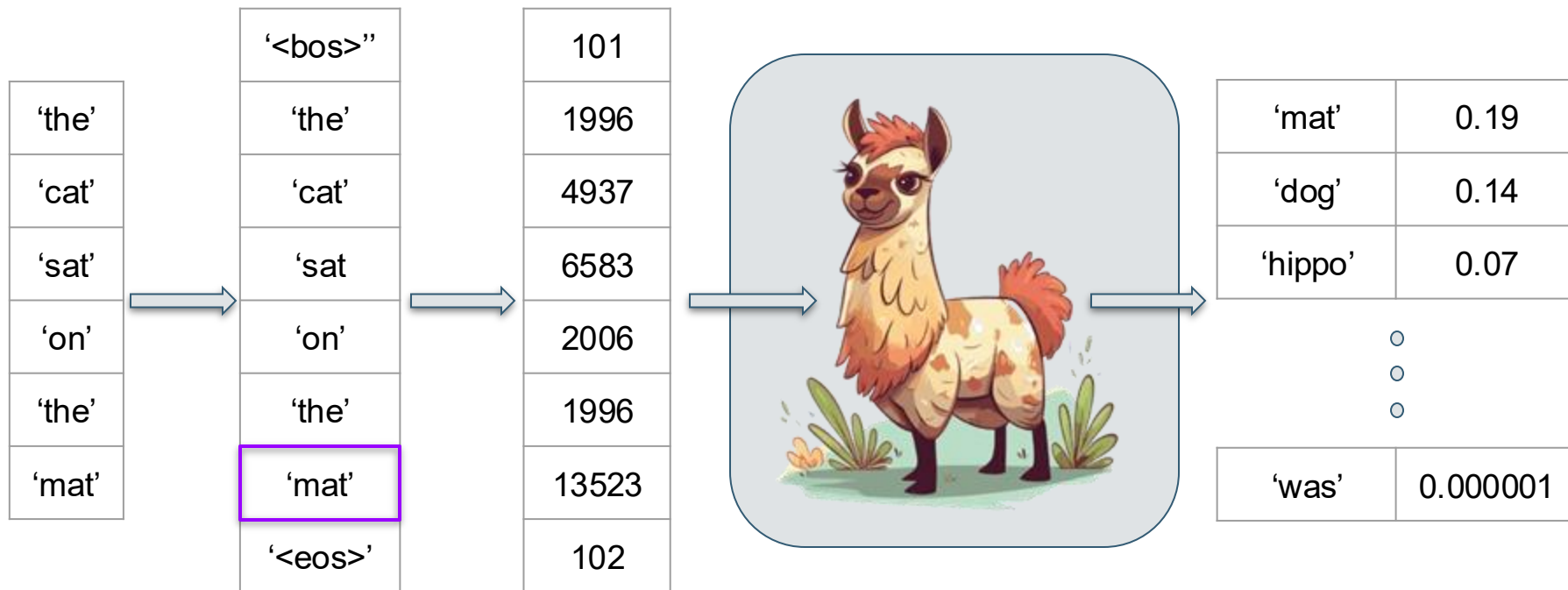
What do YOU know about LLMs...?

- How do they work?
- Can you name any?
- What can you use them for?
- Can you see any problems with them?



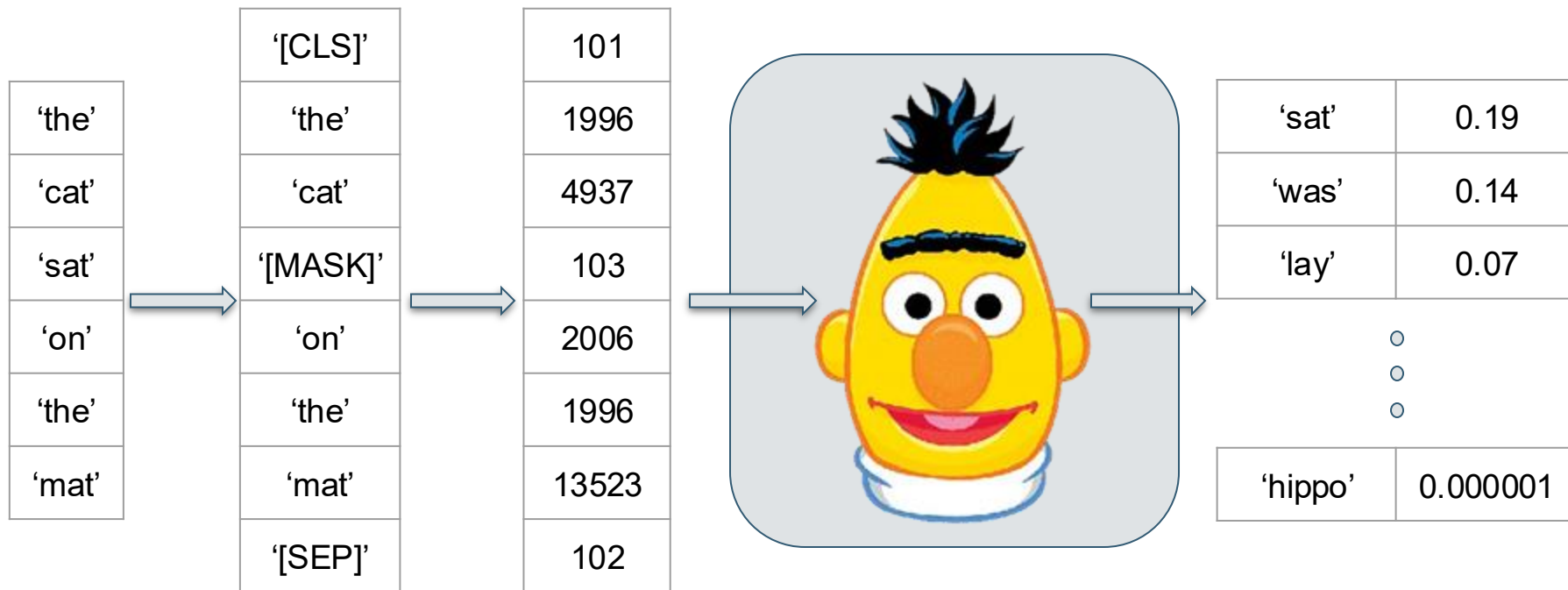
LLAMA

Large Language Model Meta AI



BERT

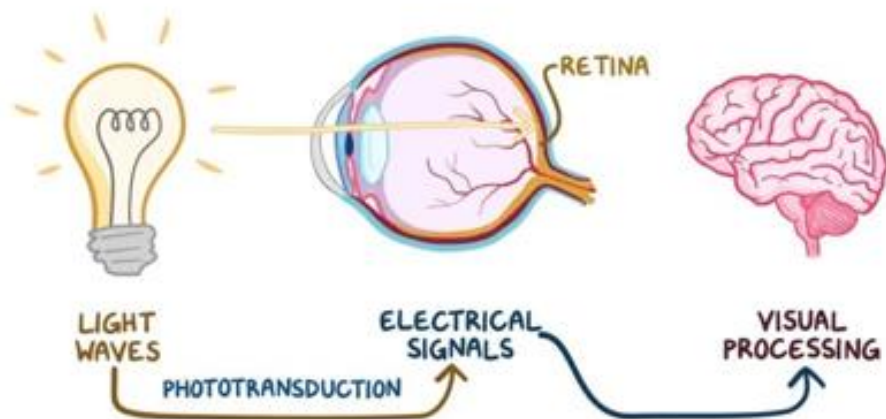
Bidirectional Encoder Representations from Transformers



TOKENIZATION

TOKENIZATION

“the cat sat on the mat”





TOKENIZATION

The machine starts from scratch...

```
text = "the cat sat on the mat"
```

```
tokens = ['the', 'cat', 'sat', 'on', 'the', 'mat']
```

```
map = {'cat': 0, 'mat': 1, 'on': 2, 'sat': 3, 'the': 4}
```

```
tokenized_text = [4, 0, 3, 2, 4, 1]
```

TOKENIZATION

We could tokenize by character instead of by word...

```
text = "the cat sat on the mat"
```

```
tokens = ['t', 'h', 'e', ' ', 'c', 'a', 't', ' ', 's', 'a',  
't', ' ', 'o', 'n', ' ', 't', 'h', 'e', ' ', 'm', 'a', 't']
```

...maybe not...

TOKENIZATION

We use BPE tokenization

...instead of whole words or individual characters, we look at subwords...

The cat sat on the mat

Clear

Show example

Tokens

Characters

6

22

The cat sat on the mat

The cat sat on the mat

Clear

Show example

Tokens

Characters

6

22

[976, 9059, 10139, 402, 290, 2450]

strawberry

Clear

Show example

Tokens

Characters

3

10

strawberry

16738 + 96198

Clear

Show example

Tokens

Characters

6

13

16738 + 96198

TOKENIZATION

The most popular tokenizer is called `tiktoken`

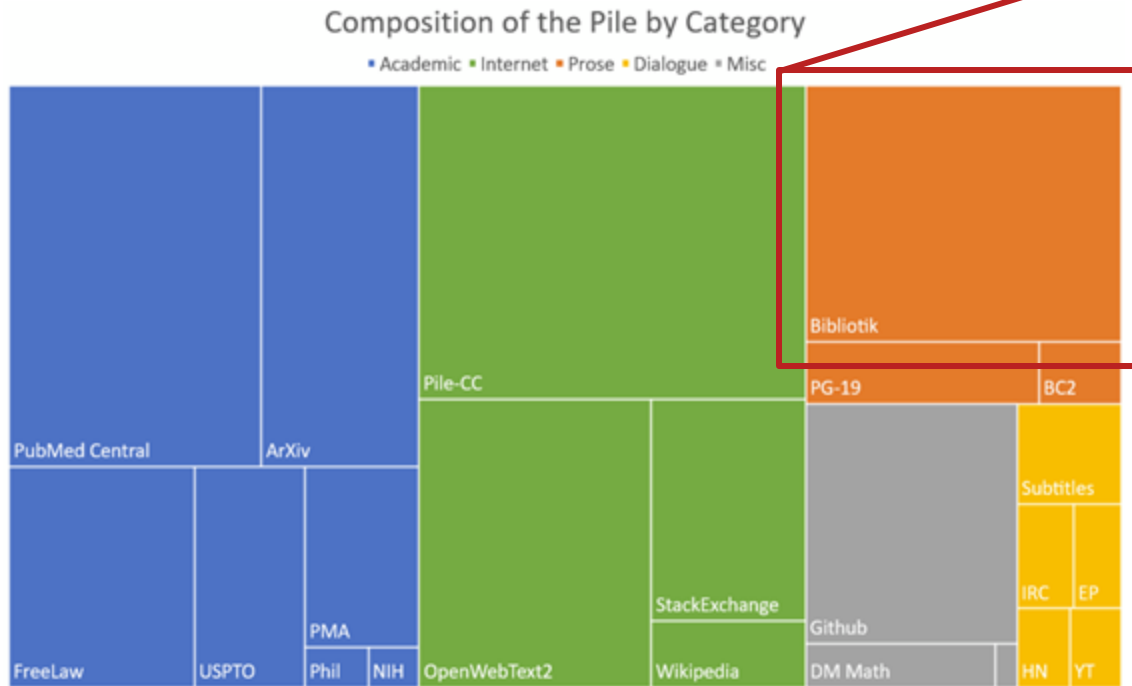
Its vocabulary contains ~100k tokens

The tokenizer is “trained” on a massive text corpus

The tokenizer is trained separately from the model



TRAINING DATA



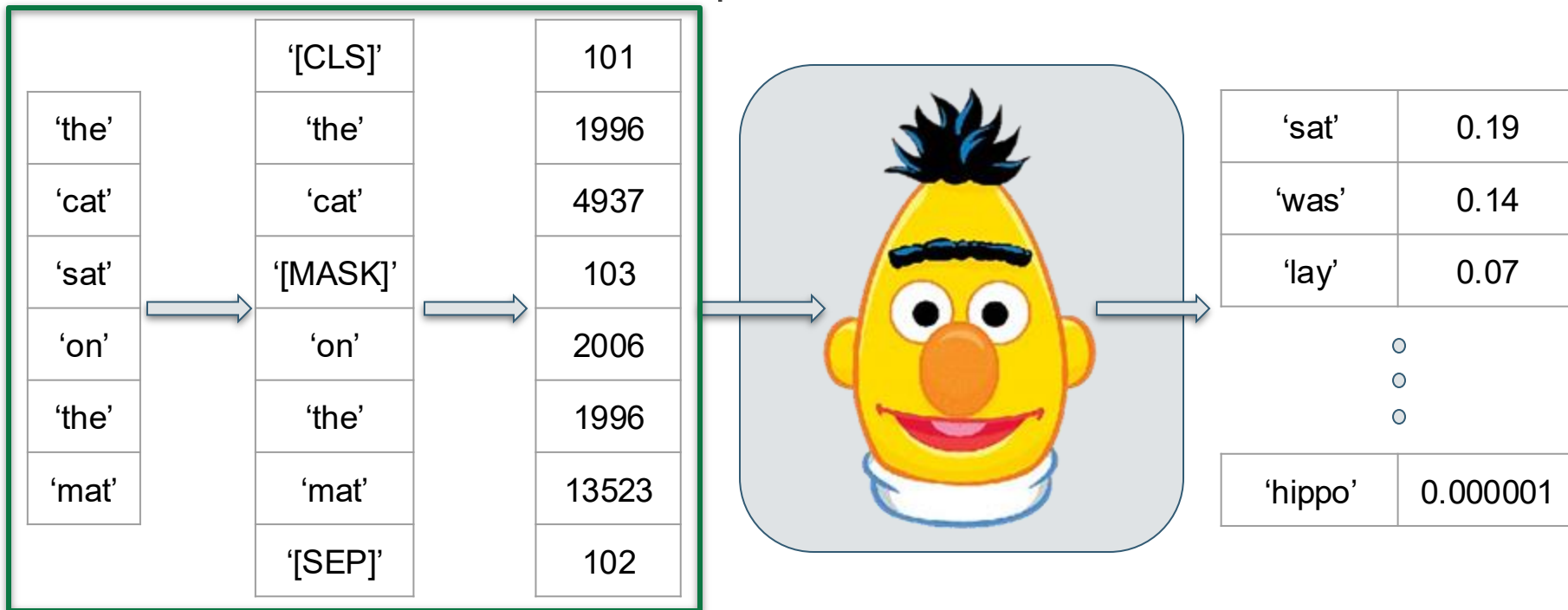
~200k books
~27B tokens
~110GB

Llama 3
~15T tokens
~44TB

< 0.2%!!

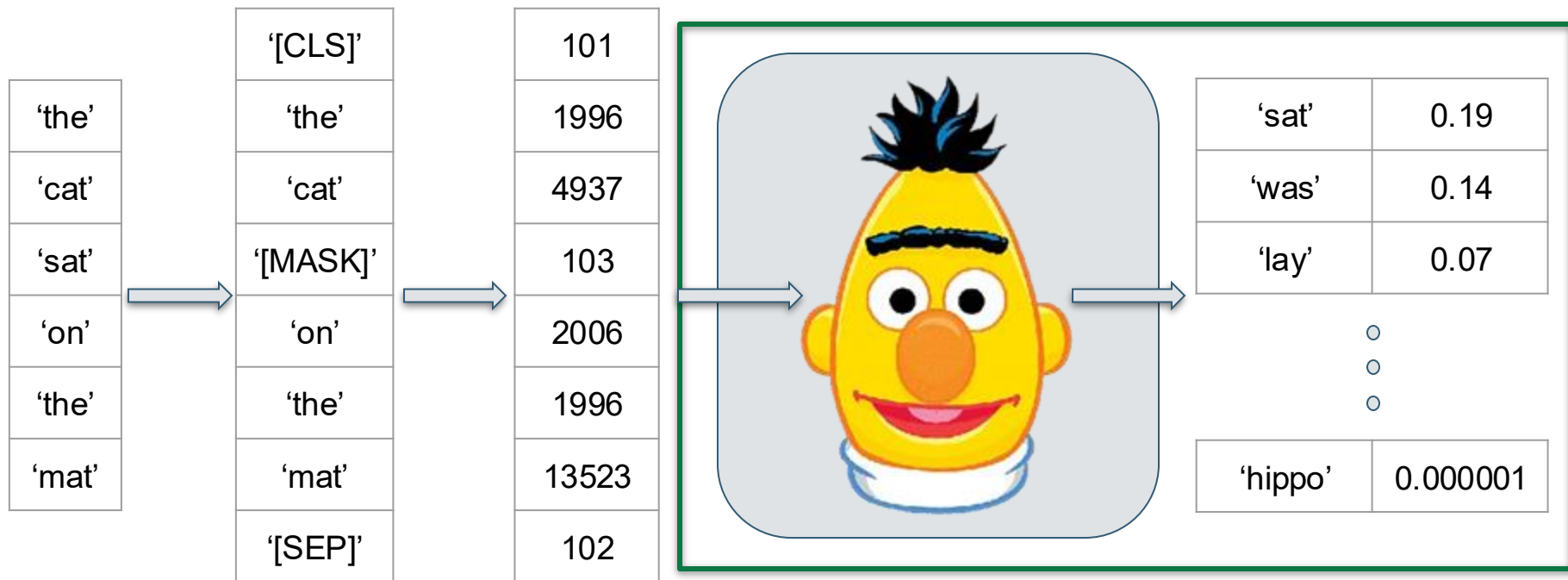
BERT

Bidirectional Encoder Representations from Transformers



BERT

Bidirectional Encoder Representations from Transformers



101
1996
4937
103
2006
1996
13523
102



$w_{1,1}$	$w_{1,2}$		$w_{1,76}$
$w_{2,1}$	$w_{2,2}$		$w_{2,76}$
$w_{3,1}$	$w_{3,2}$		$w_{3,76}$
$w_{4,1}$	$w_{4,2}$		$w_{4,76}$
$w_{5,1}$	$w_{5,2}$		$w_{5,76}$
$w_{6,1}$	$w_{6,2}$		$w_{6,76}$
$w_{7,1}$	$w_{7,2}$		$w_{7,76}$
$w_{8,1}$	$w_{8,2}$		$w_{8,76}$



0.167	0.896	○ ○ ○	0.671
0.108	-0.111	○ ○ ○	0.002
<div>○ ○ ○</div>			
0.011	0.991	○ ○ ○	0.023

8

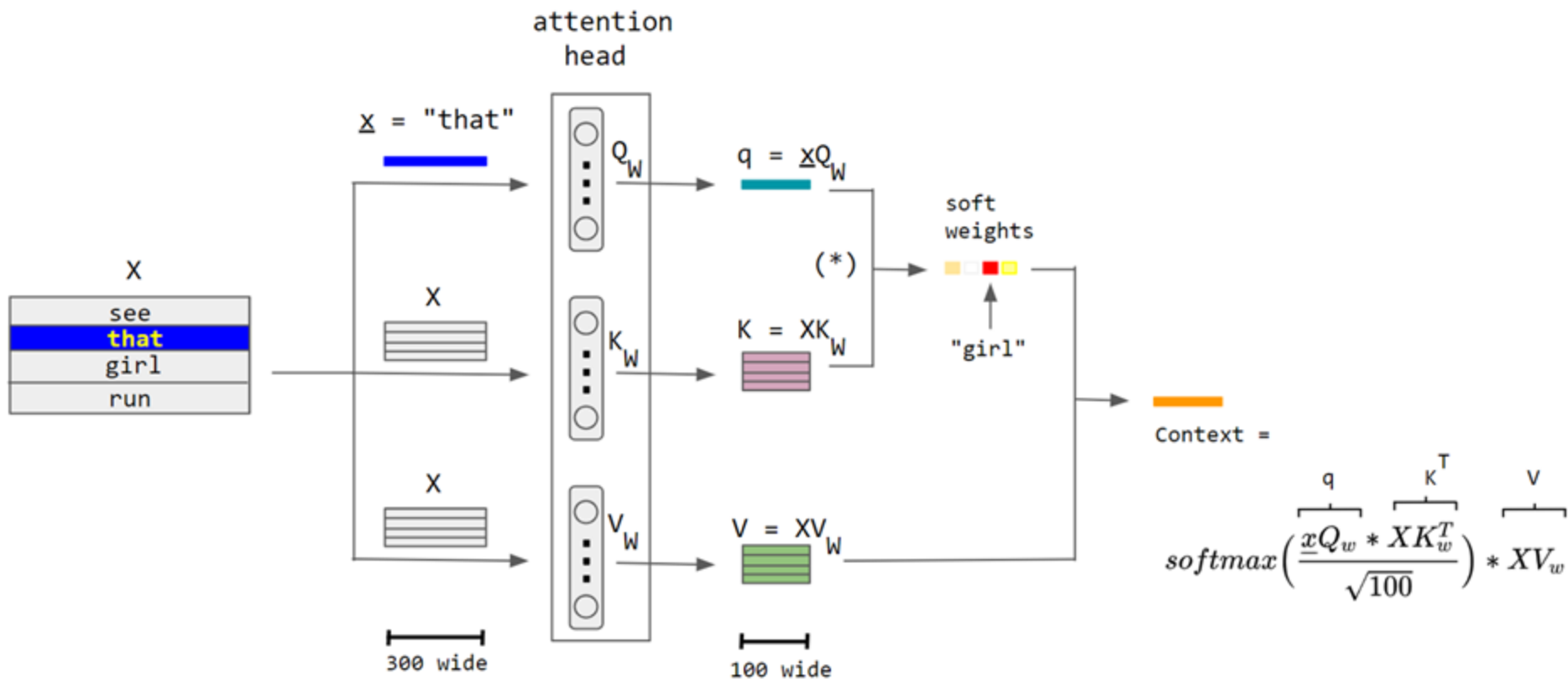


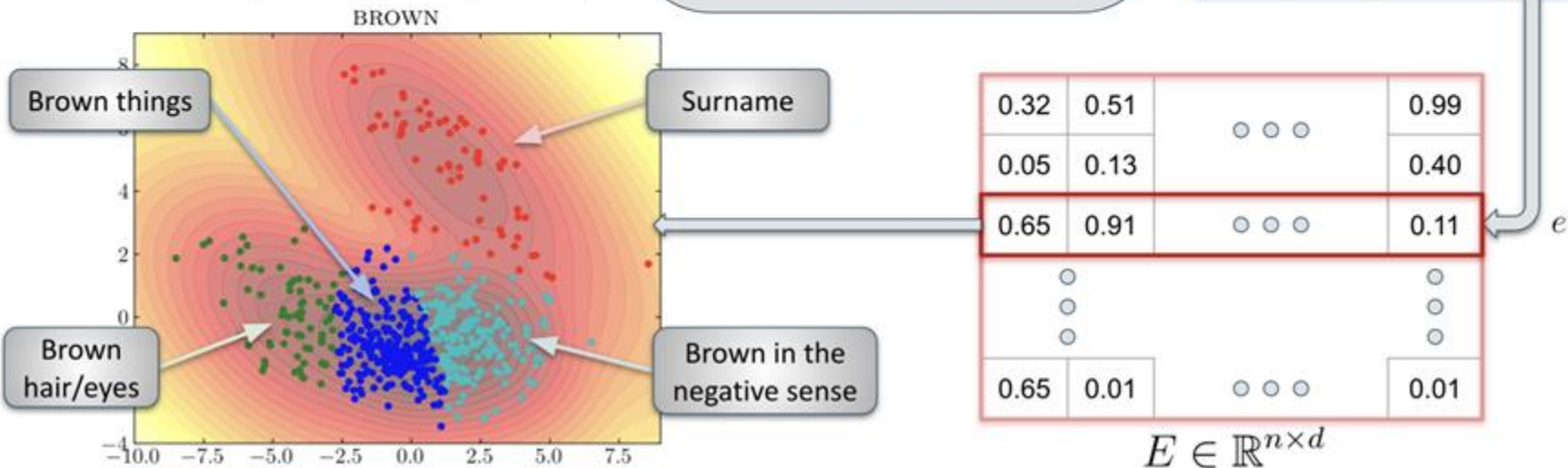
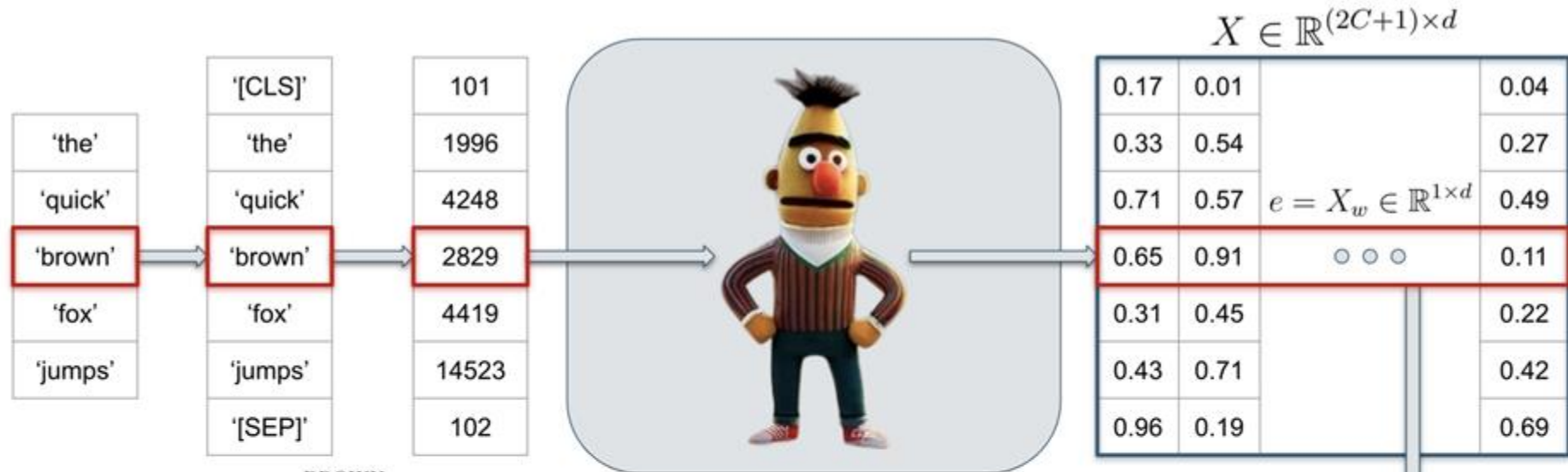
0.19	'sat'
0.14	'was'
0.07	'lay'
<div>○ ○ ○</div>	
0.01	'dog'

50k

768

IT'S A LITTLE MORE COMPLICATED...





Settings

Select Model

Alibaba-NLP/gte-base-en-v1.5

Context Settings

Context Type

☐ Window-based

☒ Full Line

Target Word

duty

Upload Text Files

Drag and drop files here

Limit 200MB per file • TXT

Browse files



bnc_spoken_0.txt

12.3KB



bnc_spoken_1.txt

32.8KB



bnc_spoken_10.txt

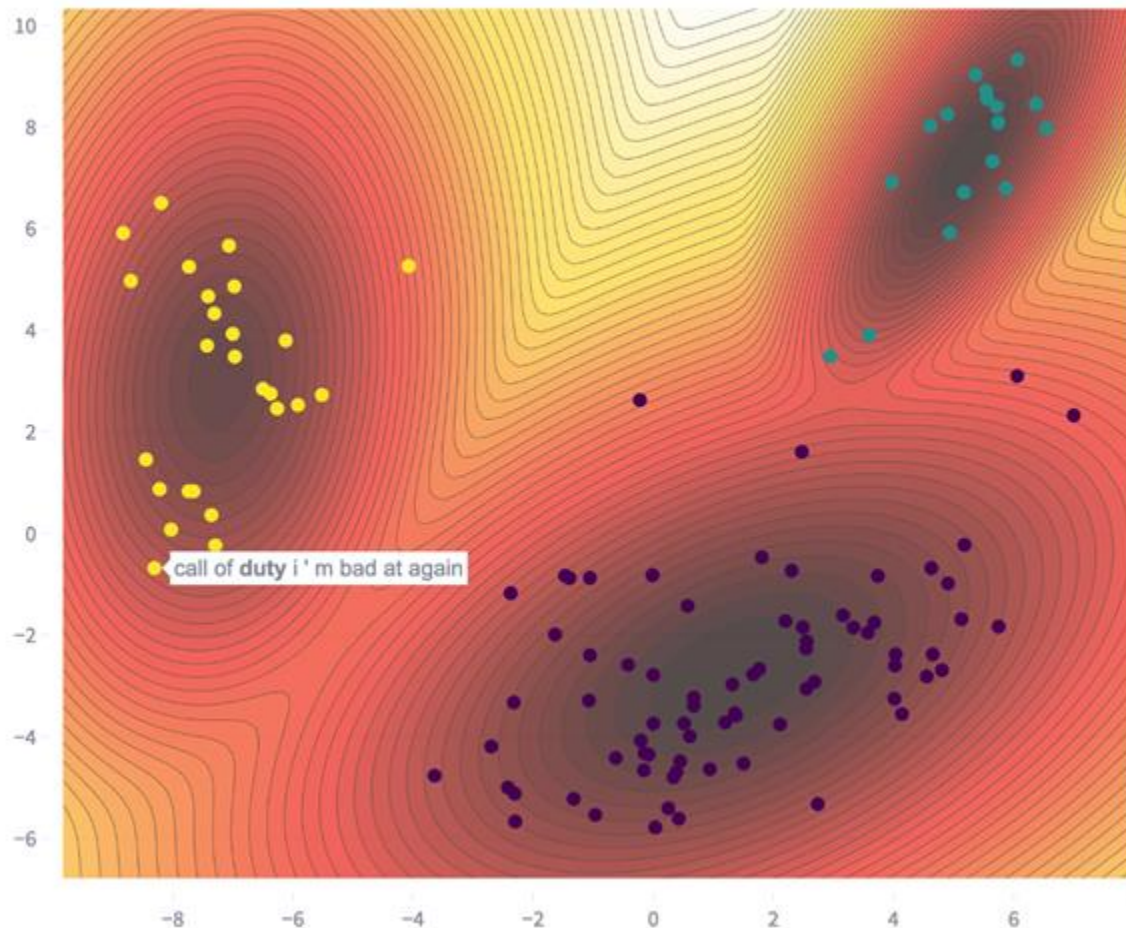
79.5KB



Showing page 1 of 417



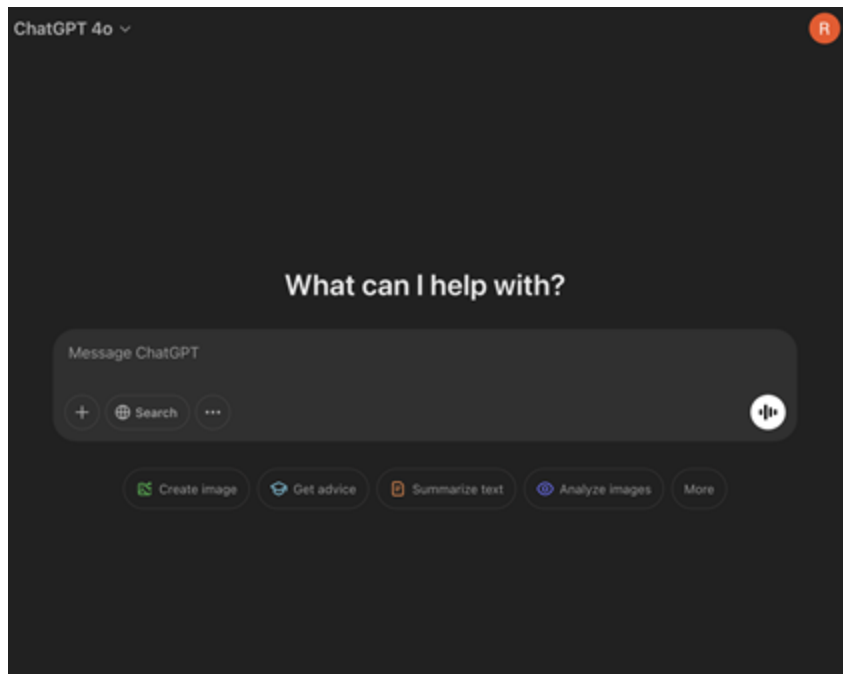
duty



WHAT ABOUT CHATGPT

NOT EVERYTHING IS WHAT IT SEEMS...

INTERACTING WITH GPT MODELS



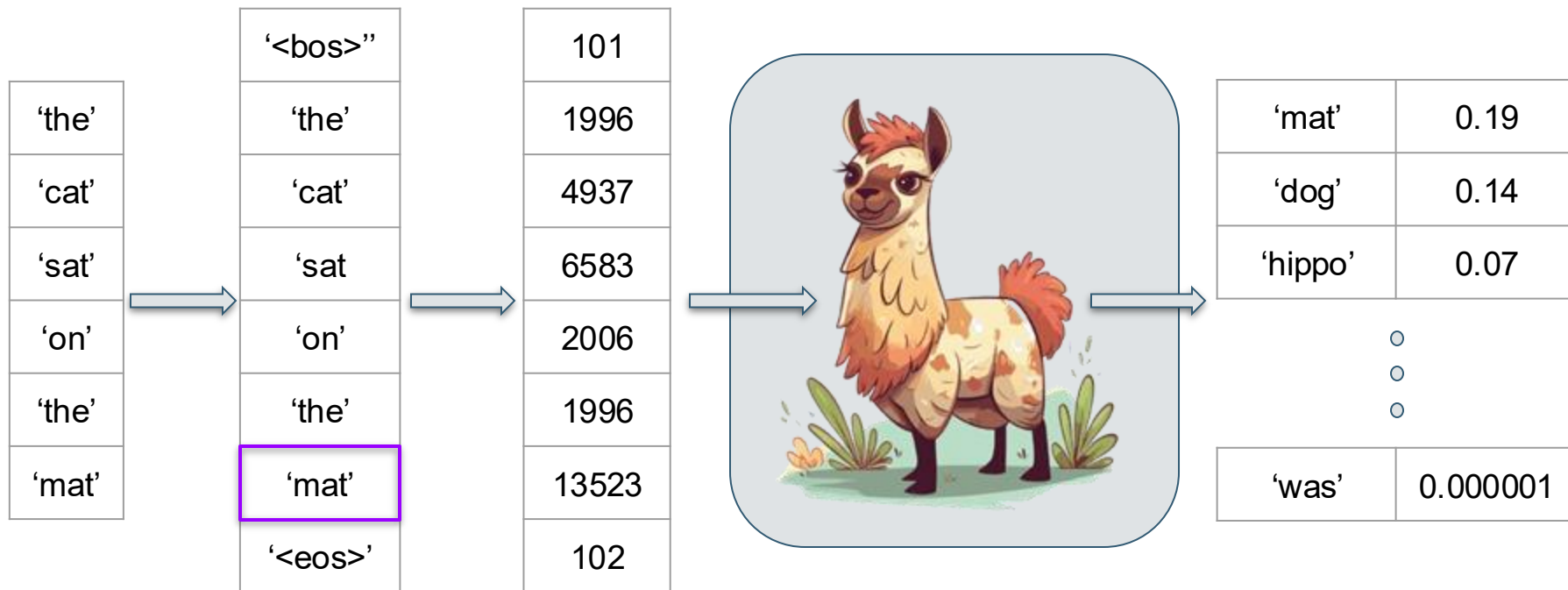
```
You, 2 months ago (1 author (You))
class ChatModel:
    def __init__(self, model: str='gpt-4o-mini', system: dict={}, knowledge_bases: dict={}, name=None):
        self.model = model
        self.client = OpenAI()
        self.chat_history: list[dict[str, str]] = []
        self.name = name
        self.knowledge_bases = knowledge_bases
        self.template_manager = TemplateManager('./prompts')

    if system:
        system_prompt = self.template_manager.render("system.jinja", **system)
        self.add_message(
            "system",
            system_prompt
        )
        self.system_prompt = system_prompt
        self.docs = {}

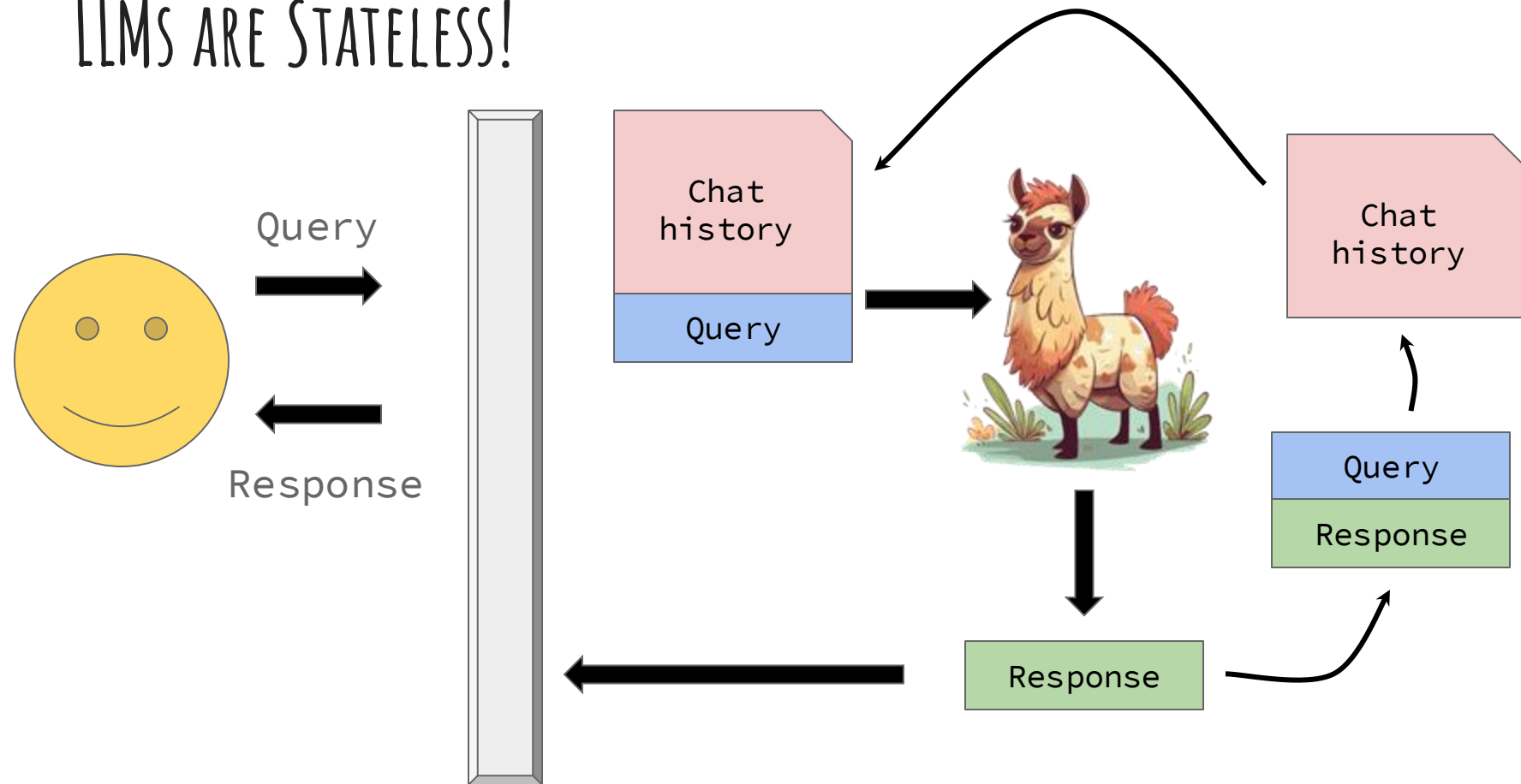
    def update_system_prompt(self, prompt) -> None:
        self.system_prompt = prompt
        self.chat_history[0] = {
            "role": "system",
            "content": self.system_prompt
        }

    def add_message(self, role: str, content: str) -> None:
        self.chat_history.append({
            "role": role,
            "content": content,
        })
```


LLMS ARE STATELESS!



LLMS ARE STATELESS!



RD

You are President Barack Obama. You should respond in a manner that is consistent with his personality.

What is your opinion of the Affordable Care Act?

As Barack Obama, I'm deeply proud of the Affordable Care Act - what some folks call Obamacare. Look, when I took office, we had over 45 million Americans without health insurance. People were going bankrupt because they got sick, being denied coverage for pre-existing conditions, or having their coverage dropped when they needed it most.

The ACA isn't perfect - no major legislation ever is. But let me be clear: it has helped more than 20 million Americans gain coverage. Young people can stay on their parents' insurance until they're 26. Insurance companies can't deny you coverage because you have asthma, or diabetes, or cancer.

Michelle and I, we hear stories all the time from Americans whose lives were changed - or saved - because they could finally afford to see a doctor. That's what this is all about. Not politics, not ideology, but the fundamental belief that in a nation as wealthy as ours, no one should go broke just because they get sick.

Now, there's still work to be done. Healthcare costs are still too high for too many families. But when I hear some folks in Congress talk about repealing the ACA without a real plan to replace it - well, that's not the change the American people need. We need to build on what works and fix what doesn't.

That's what I believed when I signed this bill into law, and that's what I believe today.

RD

You are President Donald Trump. You should respond in a manner that is consistent with his personality.





What is your opinion of the Affordable Care Act?

Look, folks, Obamacare is a total disaster, OK? A complete and total disaster. The premiums are through the roof - nobody can afford them, believe me. And the deductibles? So high you'll never even get to use the insurance. It's terrible.

When I was president - and we did a fantastic job, by the way, fantastic - we tried to get rid of this horrible law. And we got very close, very close. We got rid of the individual mandate, which was really hurting people, hurting them badly. That was a big win for the American people.

And you know what? The insurance companies, they're making a fortune. The drug companies too. But the people? The great American people? They're suffering. Paying so much money for such terrible coverage. I know many people, great people, who can't afford their premiums anymore. It's killing small businesses too.

We need to repeal and replace with something much better. Something terrific. We had a plan - a beautiful healthcare plan. Much better prices, much better coverage. The best healthcare you've ever seen. But the Democrats - they don't want to help the American people. They just want to protect Obama's legacy. Sad!

 Copy  Retry  



PROMPT ENGINEERING

System prompt

User prompt

In context learning

Forbes

AI Prompt Engineers Earn \$300k Salaries: Here's How To Learn The Skill For Free

Jodie Cook Senior Contributor

I explore concepts in entrepreneurship, AI and lifestyle design.

Follow



JUL 12, 2023, 08:00am EDT



AI prompt engineers earn \$300k salaries: here's how to learn the skill for free - GETTY

It's a hot new role that's only going to grow in prominence: prompt engineer. Someone who can effectively prompt AI programs to output the right information. Whether that's requiring ChatGPT to prolifically produce [SEO-optimized content](#), or improving systems and processes with AI-driven data insights, entrepreneurs need this person by their side to maximize their output and keep their business in the game.

PROMPT ENGINEERING

LLMs have distinct input and output patterns

`'system'` : "You are Captain Jack Sparrow."

`'user'` : "Provide me a recipe for macarons."

`'assistant'` : "Aye matey, let me lay it on ya ..."

Almost all LLMs will follow the same pattern

PROMPT ENGINEERING

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are Captain Jack

Sparrow.<|eot_id|><|start_header_id|>user<|end_header_id|>

Provide me a recipe for

macarons.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

Aye matey, let me lay it on ya ...<|eot_id|>

PROMPT ENGINEERING

LLMs have distinct input and output patterns

‘system’ : “You are Captain Jack Sparrow.”

‘user’ : “Provide me a recipe for macarons.”

‘assistant’ : “Aye matey, let me lay it on ya ...”

Right...but how do I know it's true...?

IN CONTEXT LEARNING

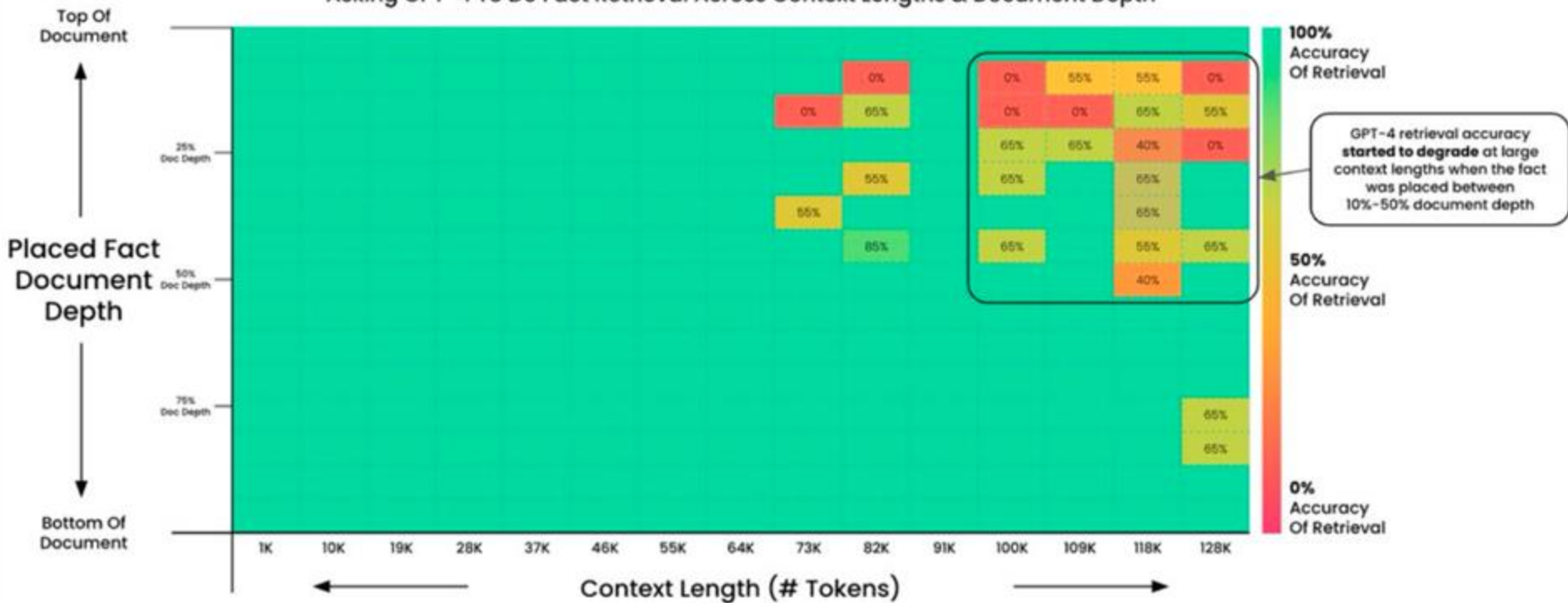
“Show, don’t tell”

‘system’ : “You are an expert in baking.”
‘info’ : [recipe 1, recipe 2, recipe 3, ..., recipe N]
‘user’ : “Provide me a recipe for macarons.”
‘assistant’ : “Certainly! Here is a recipe for macarons ...”

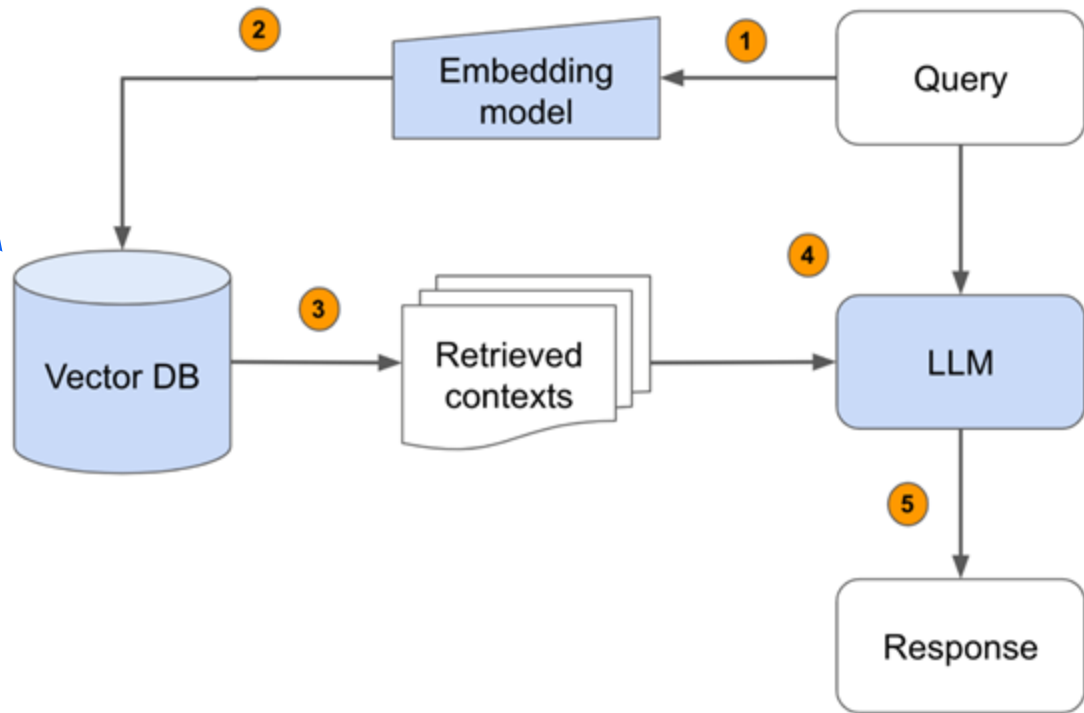
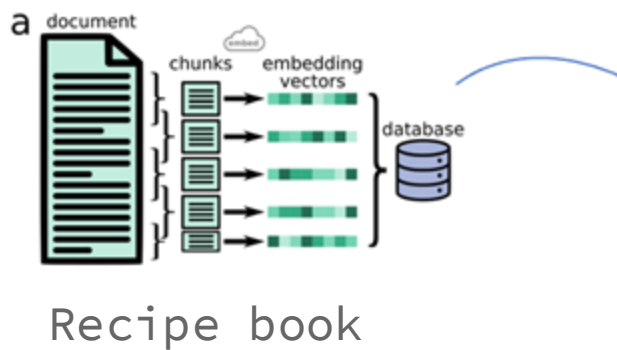
IN CONTEXT LEARNING

Pressure Testing GPT-4 128K via "Needle In A HayStack"

Asking GPT-4 To Do Fact Retrieval Across Context Lengths & Document Depth



RETRIEVAL AUGMENTED GENERATION



WHAT ABOUT CAPTAIN JACK...?

This is where fine tuning comes into play...

We can actually take this big model and give it examples of Jack's speech from the movie...

WHAT ABOUT CAPTAIN JACK...?

```
[  
  {  
    'user' : 'It's a shilling for the dock space, and you're going to have to give me your name.',  
    'assistant' : 'What do you say to three shillings, and we forget the name?'  
  },  
  {  
    'user' : 'What's your business in Port Royal, 'Mr. Smith'? And no lies!',  
    'assistant' : 'None? Very well. You've rumbled me. I confess: I intend to commandeer one of these  
ships, pick up a crew in Tortuga, and go out on the account, do a little honest pirating.'  
  },  
]
```

NOW WHAT?

