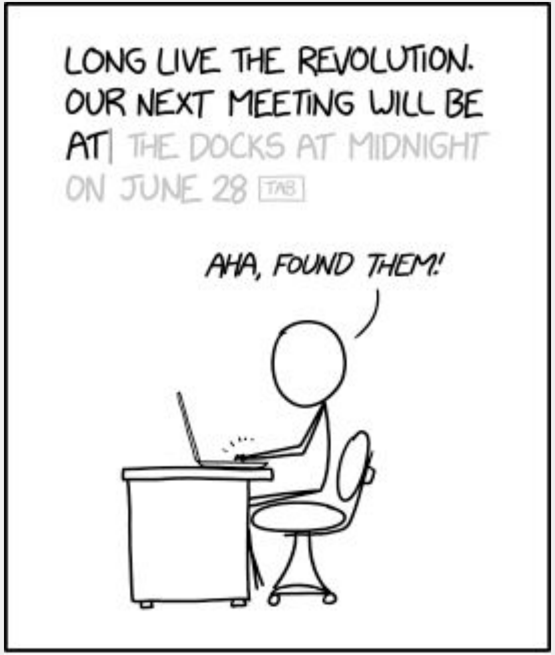
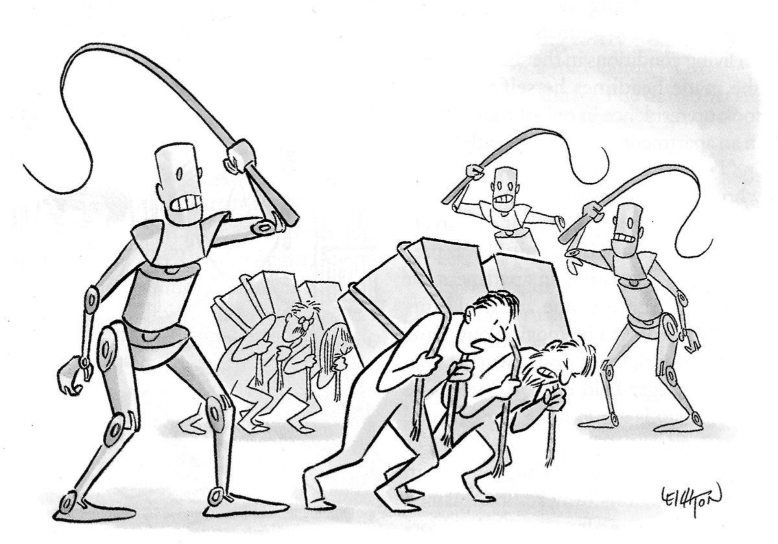
ML Models & Retrieval Augmented Generation

Tony Cui and Andrew Liu


LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28
TAB
AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.



"To think this all began with letting autocomplete finish our sentences."

Agenda

What is Machine Learning?

ML Terminology

How do language models work?

Agenda

Retrieval Augmented Generation (RAG)

- Preparation: Tokenize, Chunk
- **R**etrieve: Embedding Models, Vector Similarity
- **A**ugment: Prompt Engineering
- **G**enerate

Build a simple RAG app!

Gentle Intro to ML*

What is Machine Learn 👉🧠👈

Scenario: u are an engineer who needs to write a function that differentiates dogs and muffins

What is Machine Learn 👉🧠👈

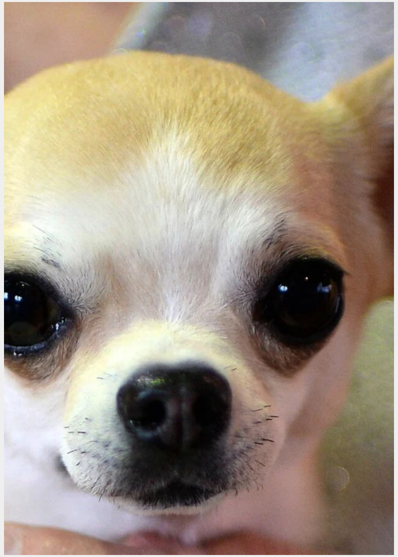
Scenario: u are an e ds to write a function that
differentiates dog

Muffin 😋

What is Machine Learn 👉🧠👈

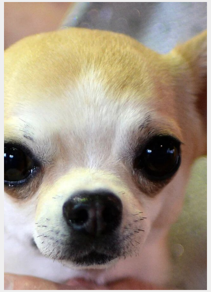
Scenario: u are an e ds to write a function that
differentiates dog

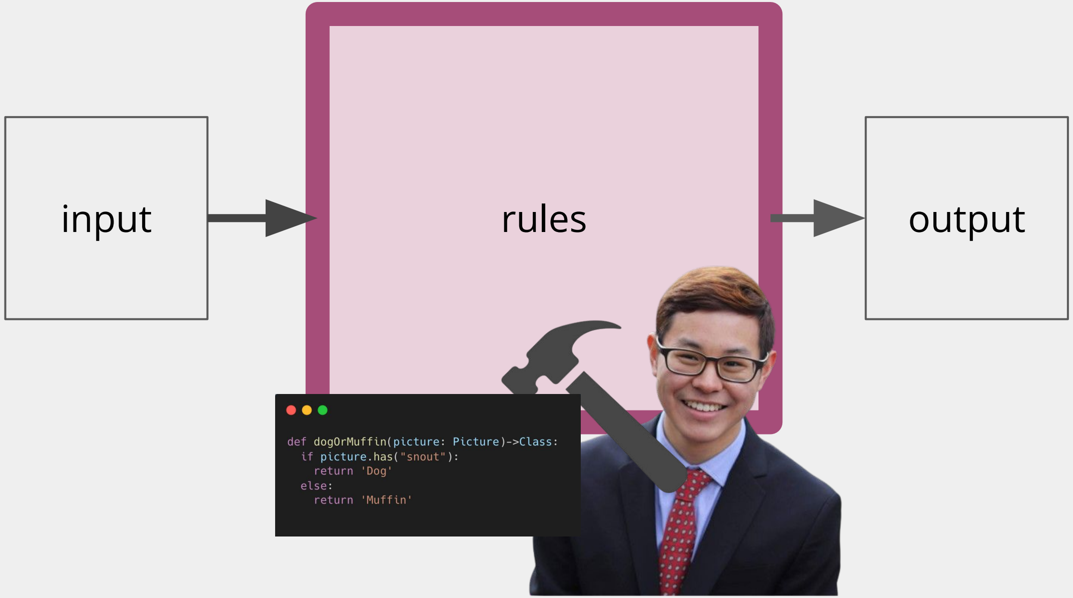
dog 😋?



What is Machine Learn 👉🧠👈

Traditional Programming:

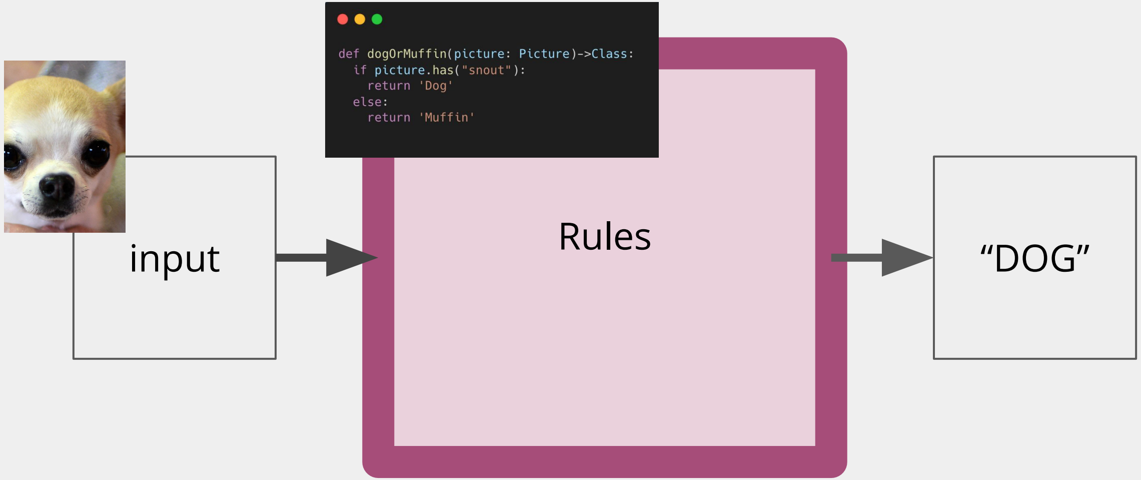




```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

What is Machine Learn 👉🧠👈

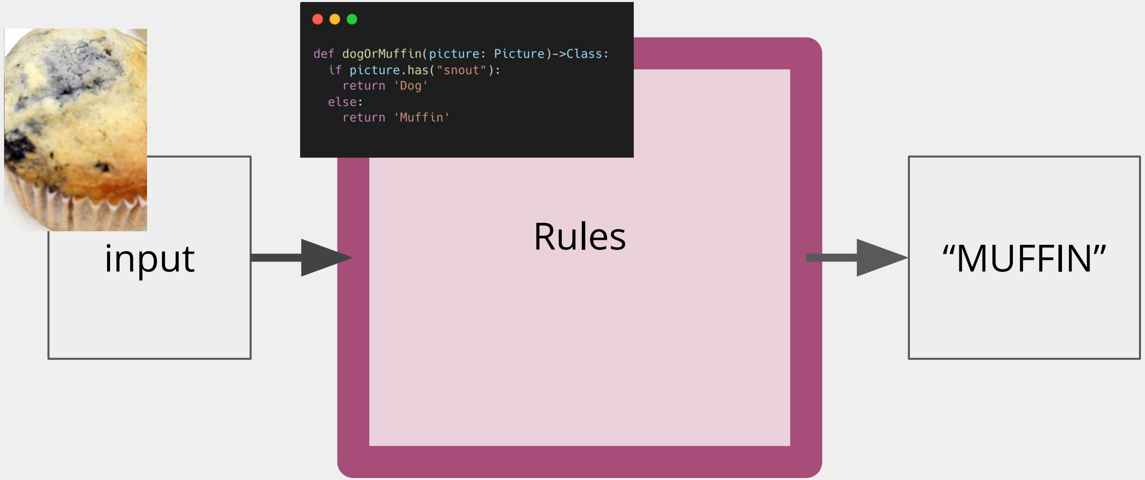
Scenario: u are an engineer who needs to write a function that differentiates dogs and muffins



```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

What is Machine Learn 👉🧠👈

Scenario: u are an engineer who needs to write a function that differentiates dogs and muffins

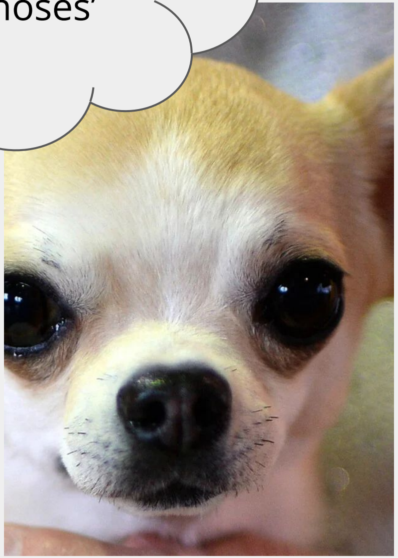


```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

What is Machine Learn 👉🧠👈

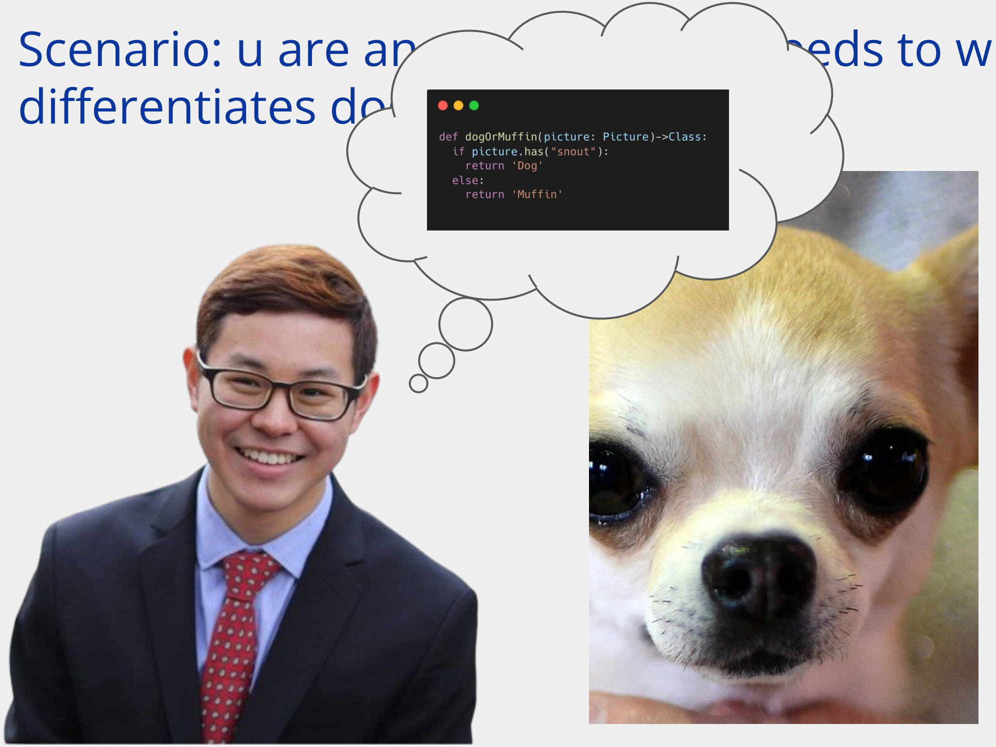
Scenario: u are an eeds to write a function that differentiates do

Hmm… dogs
have ‘noses’



What is Machine Learn 👉🧠👈

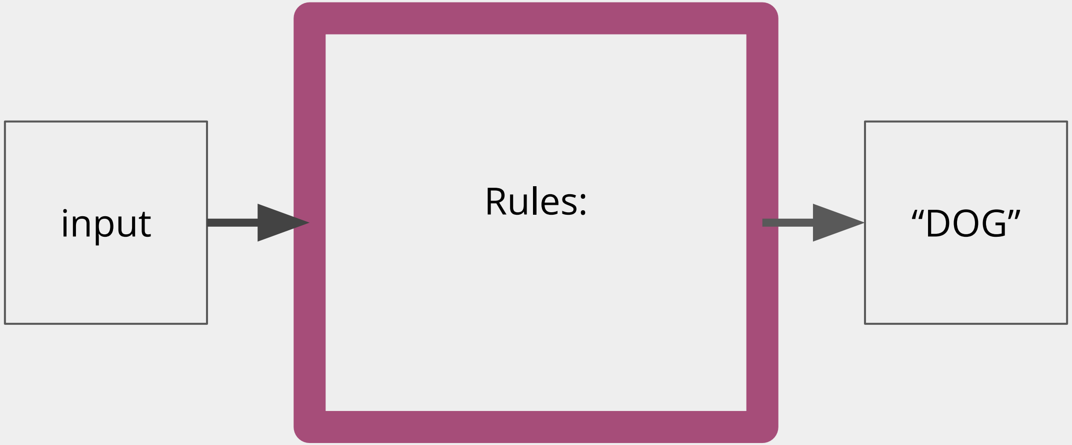
Scenario: u are an eds to write a function that
differentiates do



```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

What is Machine Learn 👉🧠👈

Scenario: u are an engineer who needs to write a function that differentiates dogs and muffins



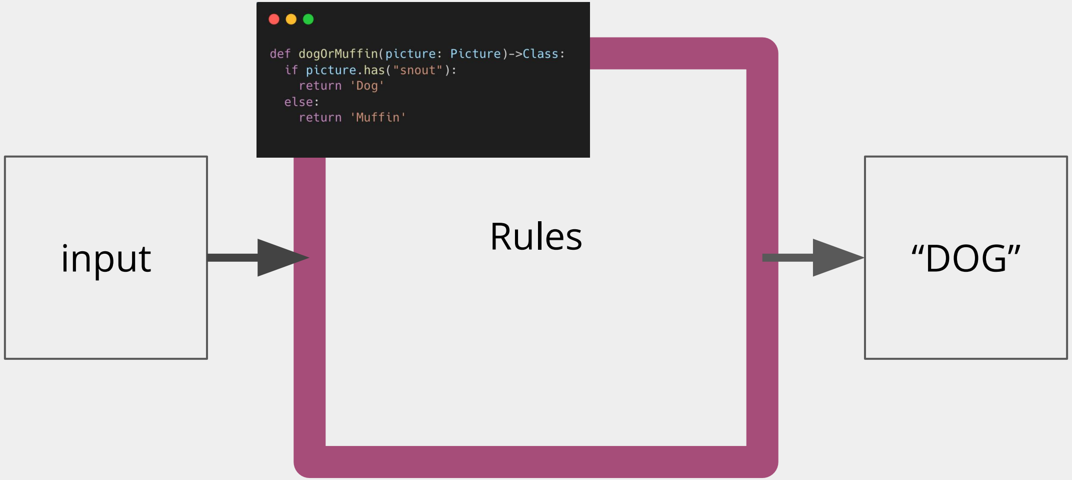
input

Rules:

“DOG”

What is Machine Learn 👉🧠👈

Scenario: u are an engineer who needs to write a function that differentiates dogs and muffins



```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

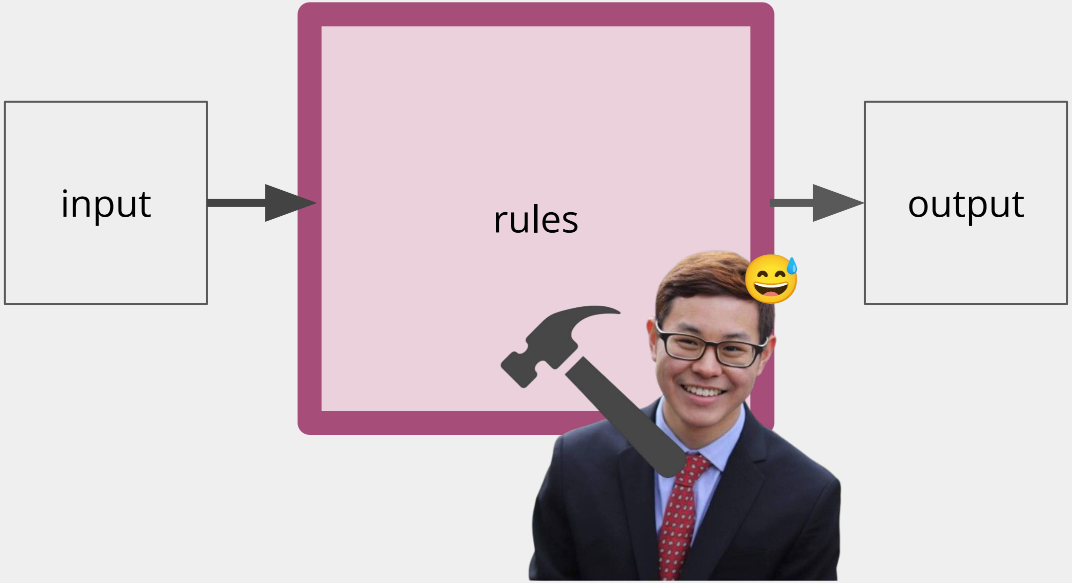
input

Rules

“DOG”

What is Machine Learn 👉🧠👈

?????



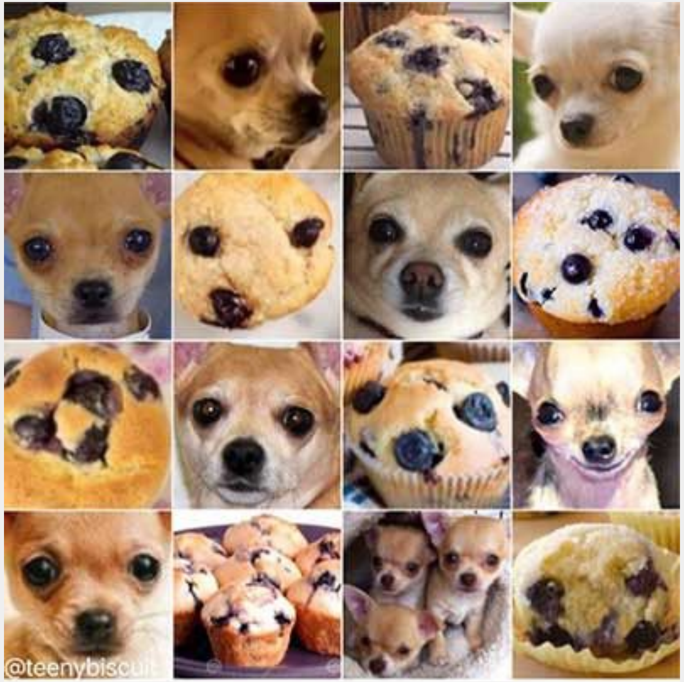
input

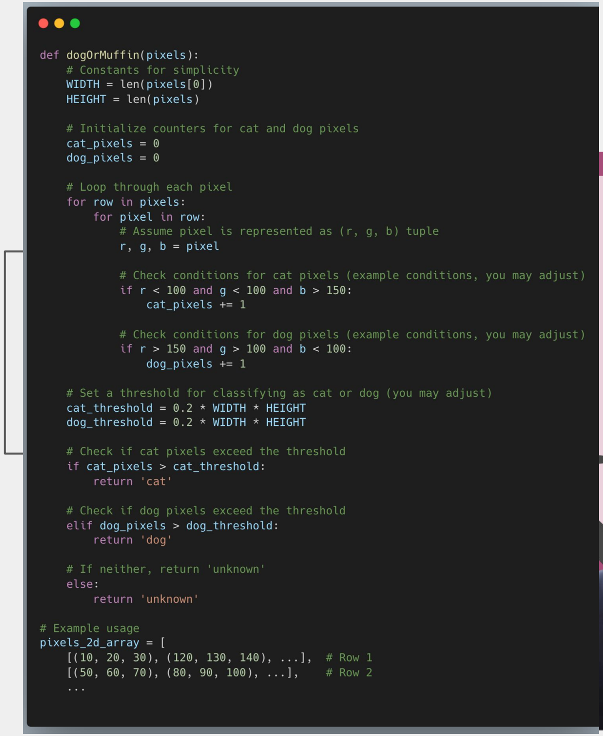
rules

output

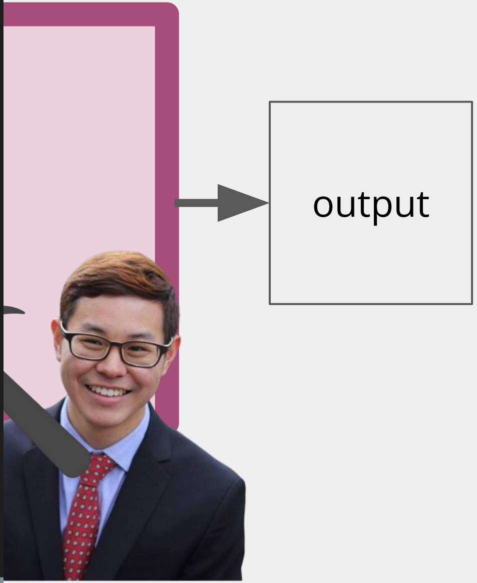
What is Machine Learn 👉🧠👈

?????



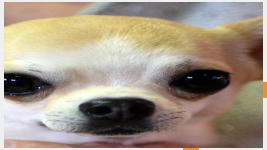


```
def dogOrMuffin(pixels):
    # Constants for simplicity
    WIDTH = len(pixels[0])
    HEIGHT = len(pixels)

    # Initialize counters for cat and dog pixels
    cat_pixels = 0
    dog_pixels = 0

    # Loop through each pixel
    for row in pixels:
        for pixel in row:
            # Assume pixel is represented as (r, g, b) tuple
            r, g, b = pixel

            # Check conditions for cat pixels (example conditions, you may adjust)
            if r < 100 and g < 100 and b > 150:
                cat_pixels += 1

            # Check conditions for dog pixels (example conditions, you may adjust)
            if r > 150 and g > 100 and b < 100:
                dog_pixels += 1

    # Set a threshold for classifying as cat or dog (you may adjust)
    cat_threshold = 0.2 * WIDTH * HEIGHT
    dog_threshold = 0.2 * WIDTH * HEIGHT

    # Check if cat pixels exceed the threshold
    if cat_pixels > cat_threshold:
        return 'cat'

    # Check if dog pixels exceed the threshold
    elif dog_pixels > dog_threshold:
        return 'dog'

    # If neither, return 'unknown'
    else:
        return 'unknown'

# Example usage
pixels_2d_array = [
    [(10, 20, 30), (120, 130, 140), ...],  # Row 1
    [(50, 60, 70), (80, 90, 100), ...],    # Row 2
    ...
```



output

What is Machine Learn 👉🧠👈



data

“dog”

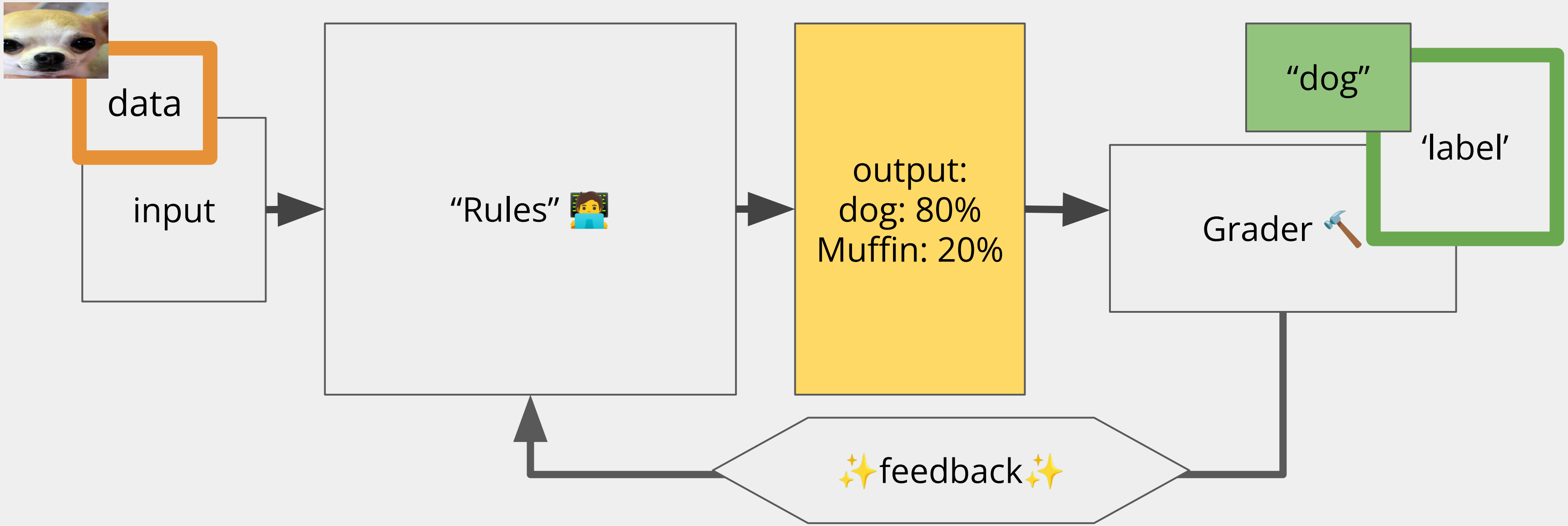
data

Grader

gives feedback to →

“structured but jank rules”

What is Machine Learn 👉🧠👈: Training



data

input

“Rules” 🧑‍💻

output:
dog: 80%
Muffin: 20%

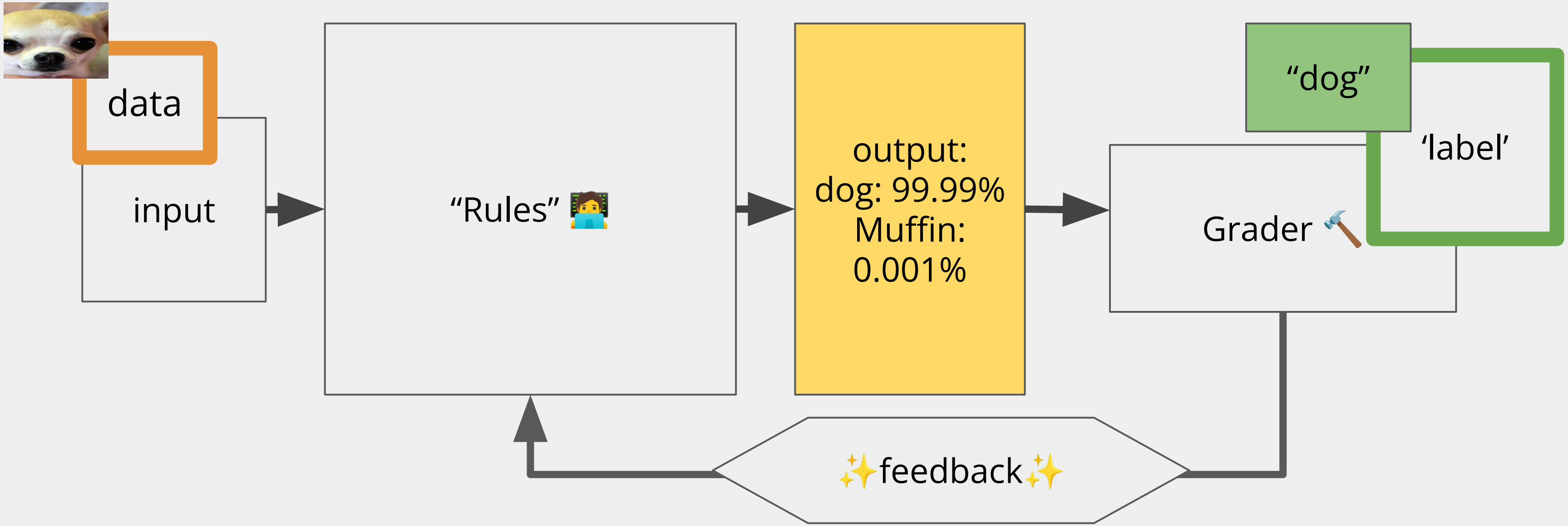
“dog”

‘label’

Grader 🔨

✨feedback✨

What is Machine Learn 👉🧠👈: Training



data

input

“Rules” 🧑‍💻

output:
dog: 99.99%
Muffin:
0.001%

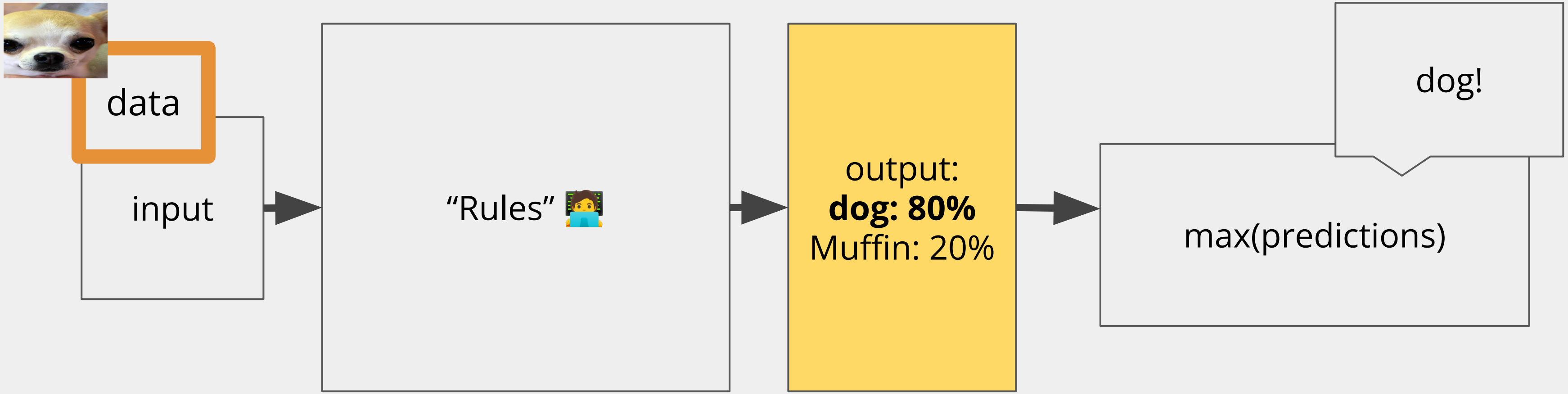
“dog”

‘label’

Grader 🔨

✨feedback✨

What is Machine Learn 👉🧠👈: Inference



data

input

“Rules” 🧑‍💻

output:
dog: 80%
Muffin: 20%

max(predictions)

dog!

What is Machine Learn 👉🧠👈

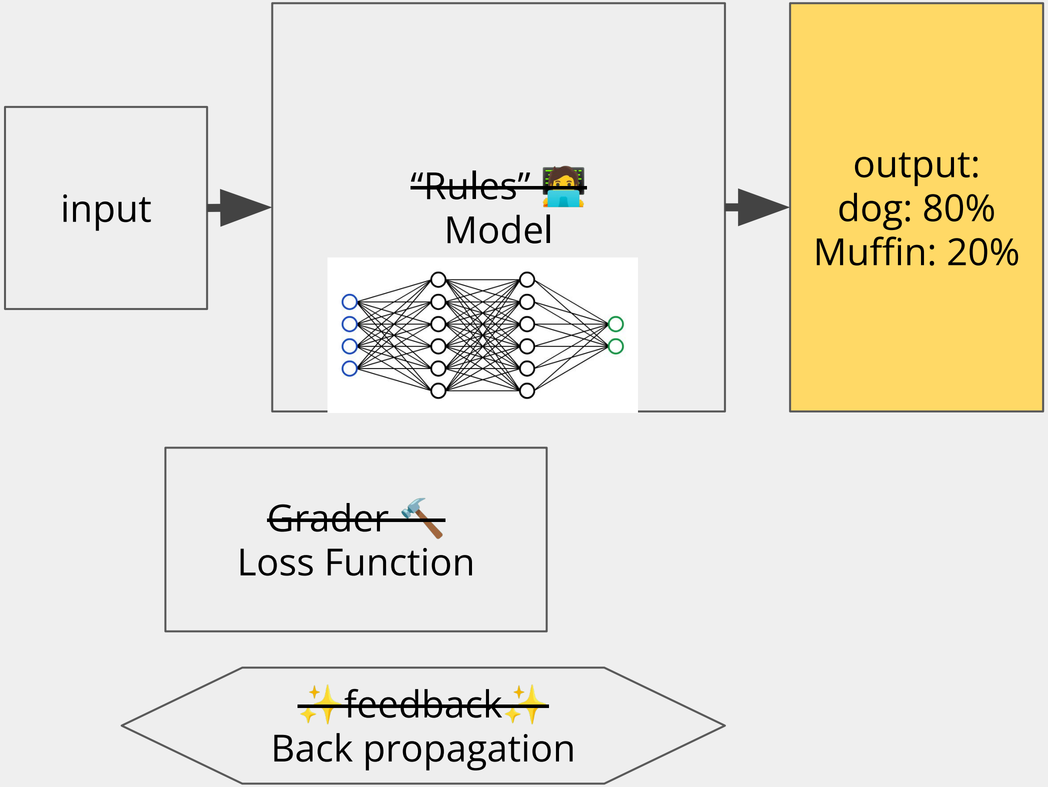
“Rules”

model

w rizz✨

✨so handsome✨

ML Terminology



input

~~“Rules” 🧑‍💻~~
Model

output:
dog: 80%
Muffin: 20%

~~Grader 🔨~~
Loss Function

~~✨feedback✨~~
Back propagation

Model: The set of rules that predict an outcome

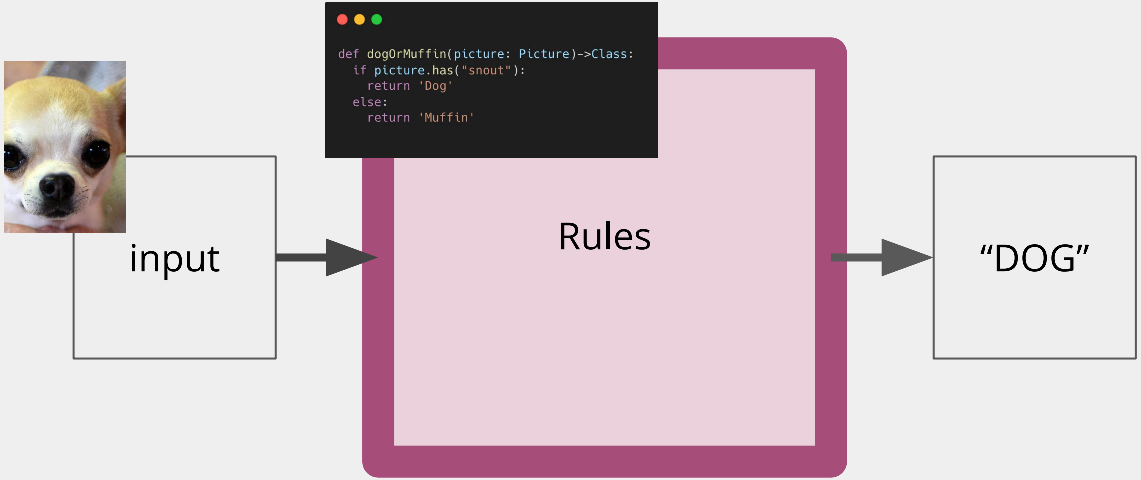
Parameters/Weights: Adjustable rules to help models make predictions

Loss Function: Specifies how the model should be adjusted based on its prediction and certainties

Back Propagation: Adjustments that are fed into the model

What is Machine Learn 👉🧠👈

- “Machine Learning is the process of training a piece of software, called a **model**, to make **useful predictions or generate content from data**”



```
def dogOrMuffin(picture: Picture)->Class:
  if picture.has("snout"):
    return 'Dog'
  else:
    return 'Muffin'
```

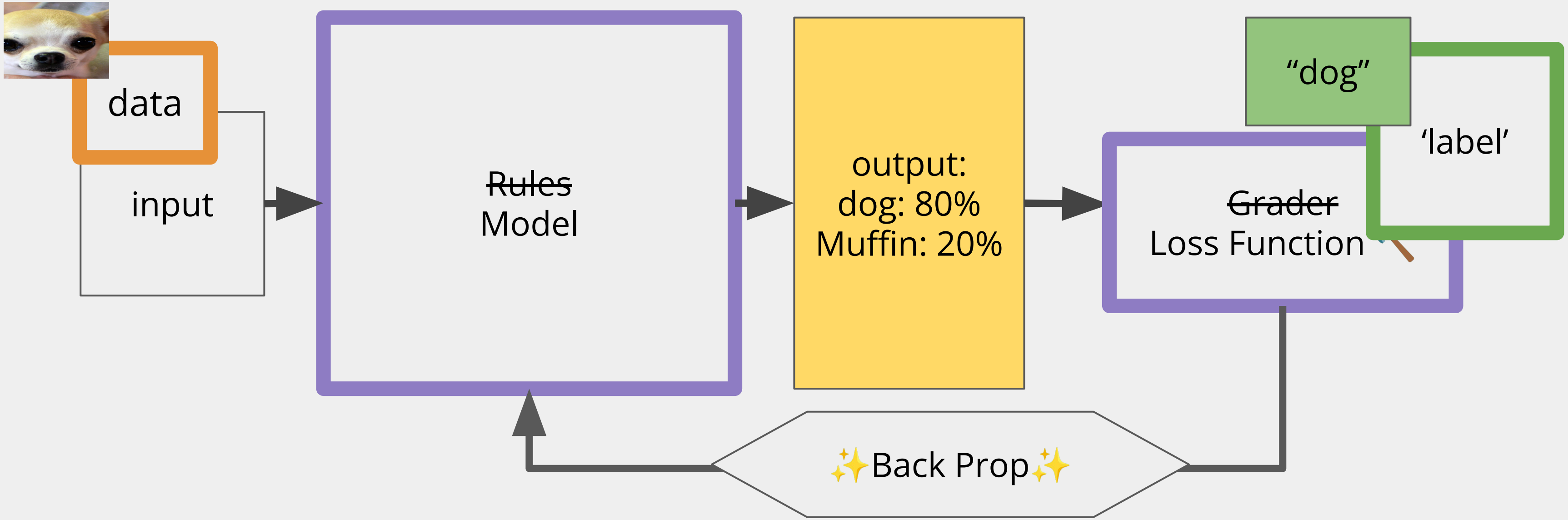
input

Rules

“DOG”

What is Machine Learn 👉🧠👈

- “Machine Learning is the process of training a piece of software, called a **model**, to make **useful predictions or generate content from data**”



data

input

~~Rules~~
Model

output:
dog: 80%
Muffin: 20%

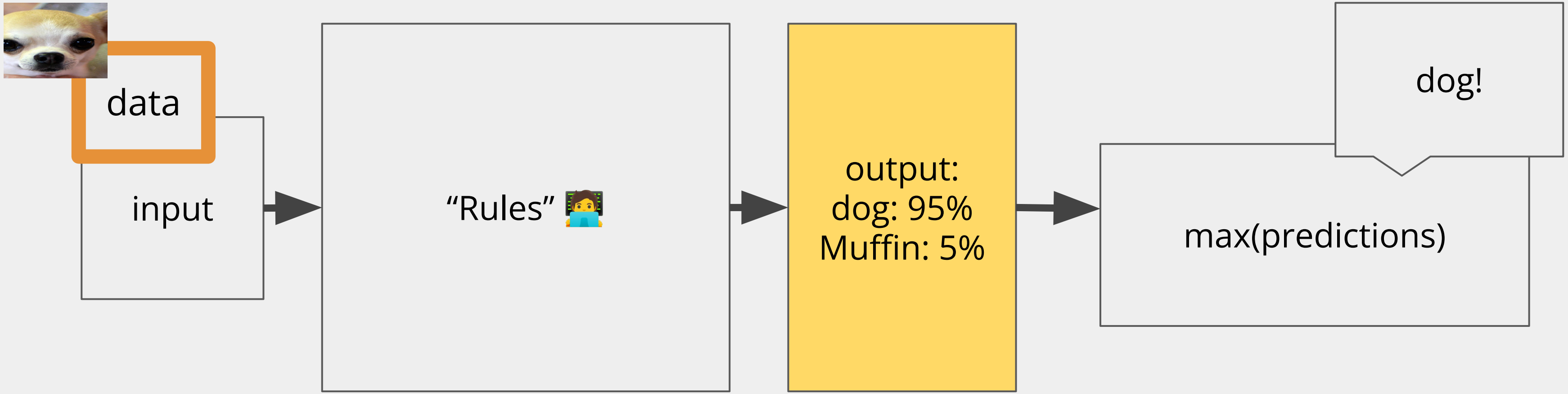
“dog”

‘label’

~~Grader~~
Loss Function

✨Back Prop✨

What is Machine Learn 👉🧠👈: Inference



data

input

“Rules” 🧑‍💻

output:
dog: 95%
Muffin: 5%

max(predictions)

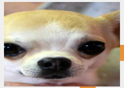
dog!

Large Language Models

GPT and Friends!

Language Models

Rather than predicting classification, Language Models* predict* words*



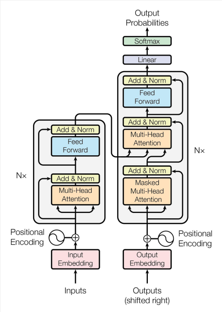
data

Today tony ate a Muffin

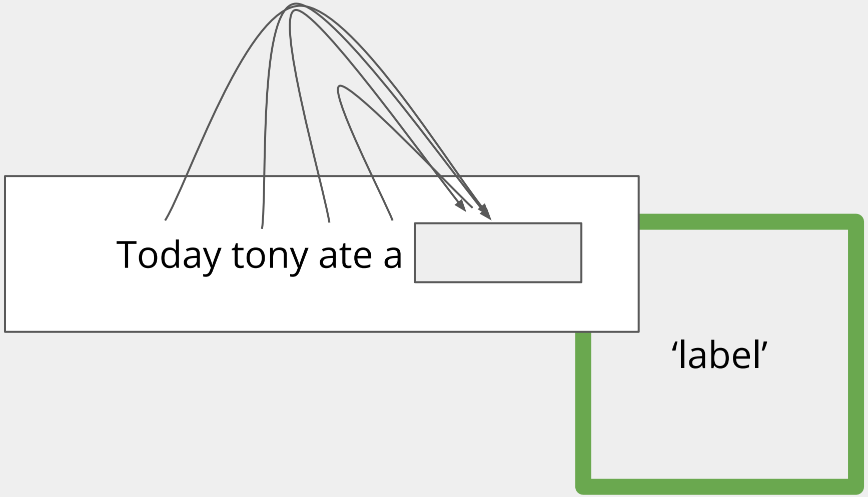
‘label’

Language Models

Rather than predicting classification, Language Models* predict* words*



Model

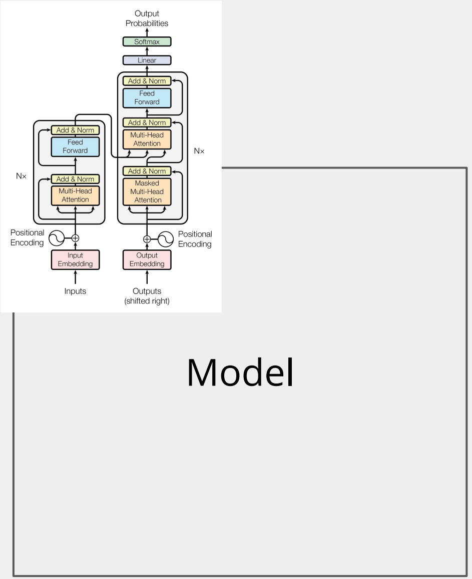


Today tony ate a

‘label’

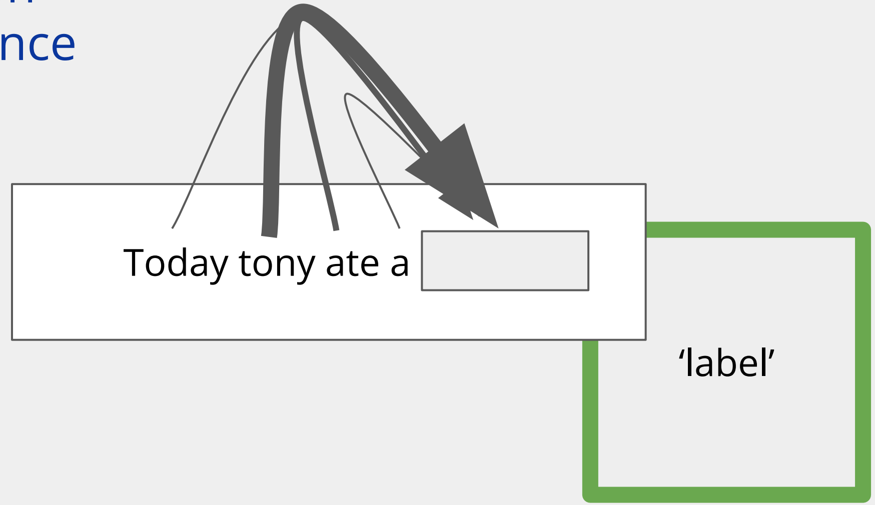
Language Models

Rather than predicting classification, Language Models* predict* words*



Model

‘Attention’ =
importance

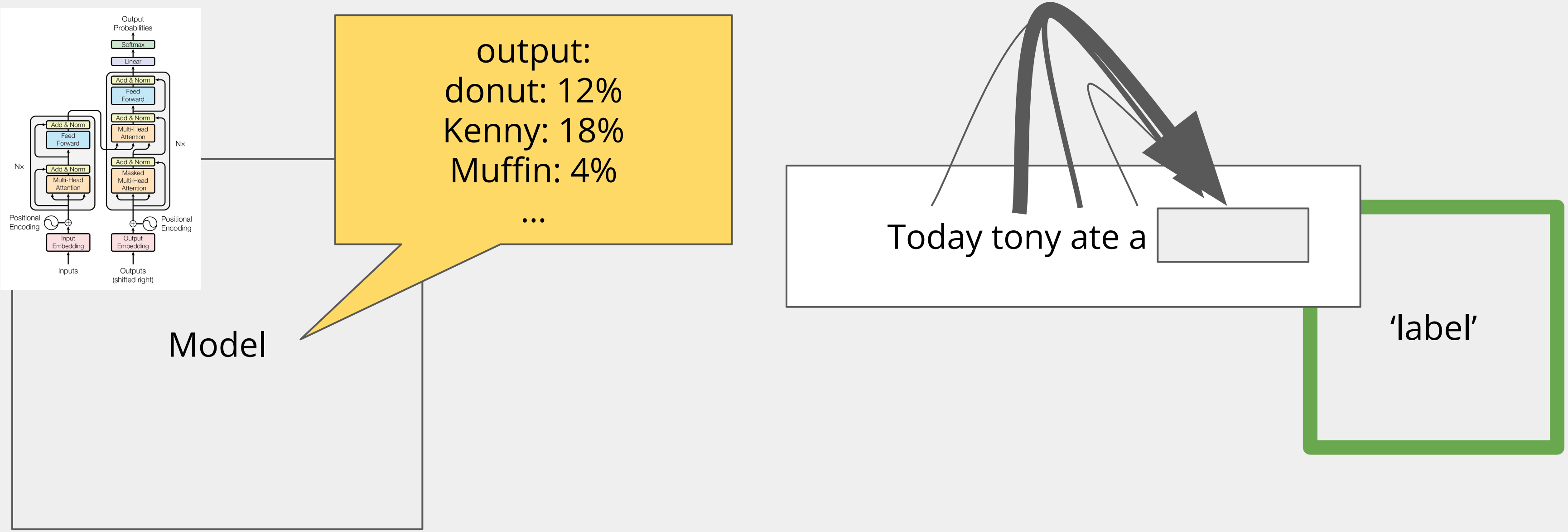


Today tony ate a

‘label’

Language Models

Rather than predicting classification, Language Models* predict* words*



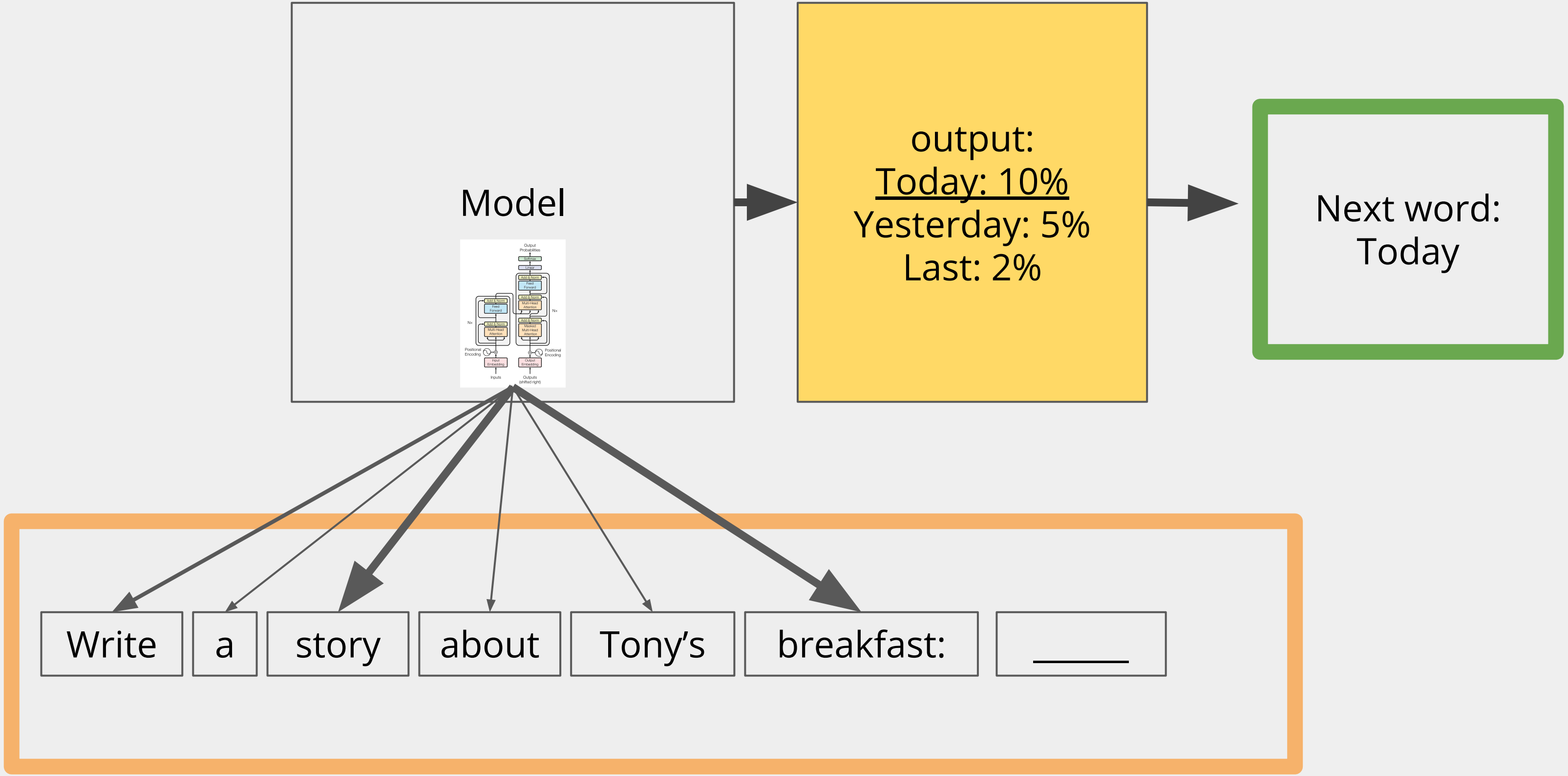
Model

output:
donut: 12%
Kenny: 18%
Muffin: 4%
…

Today tony ate a

‘label’

Language Models*



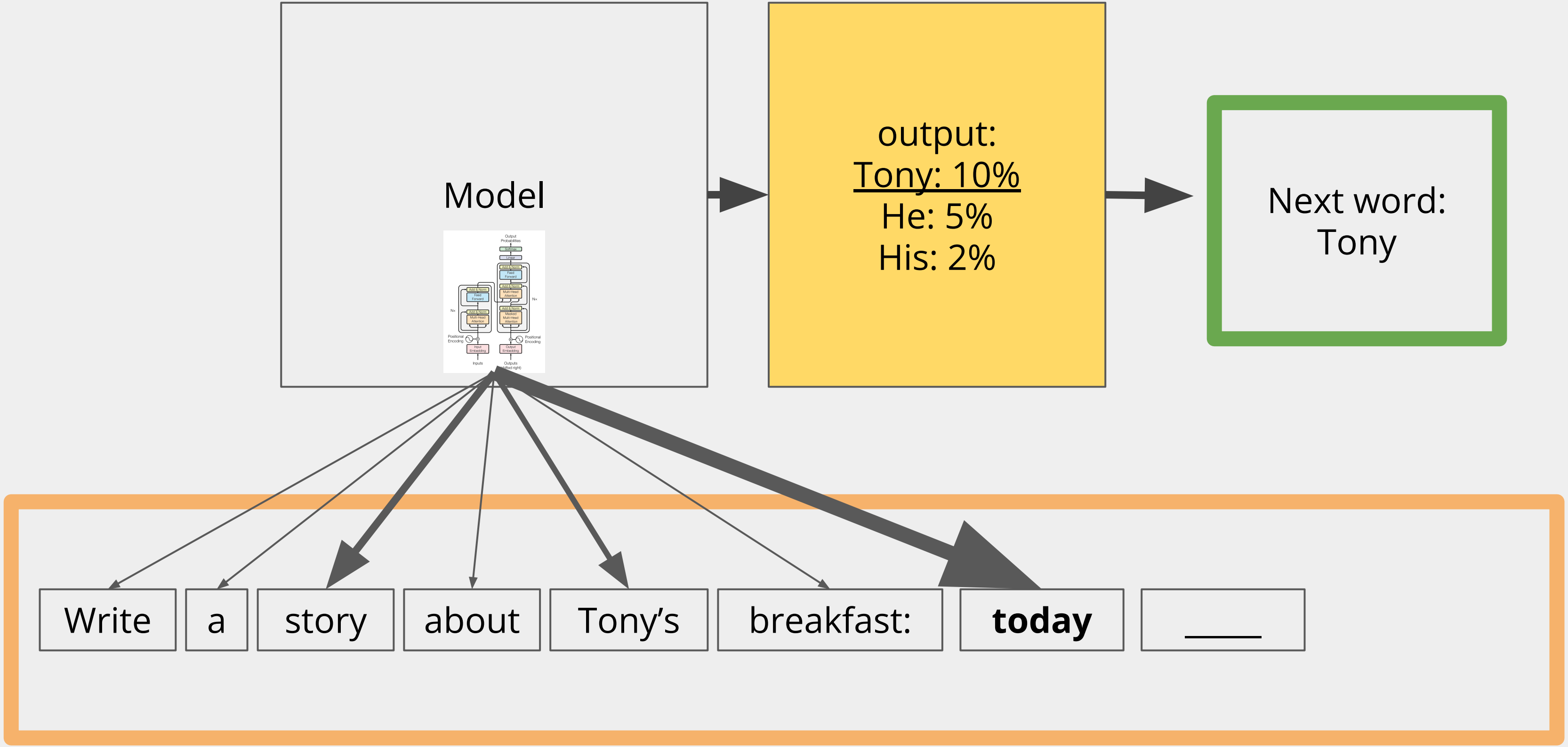
Model

output:
<u>Today: 10%</u>
Yesterday: 5%
Last: 2%

Next word:
Today

Write a story about Tony’s breakfast: _____

Language Models*



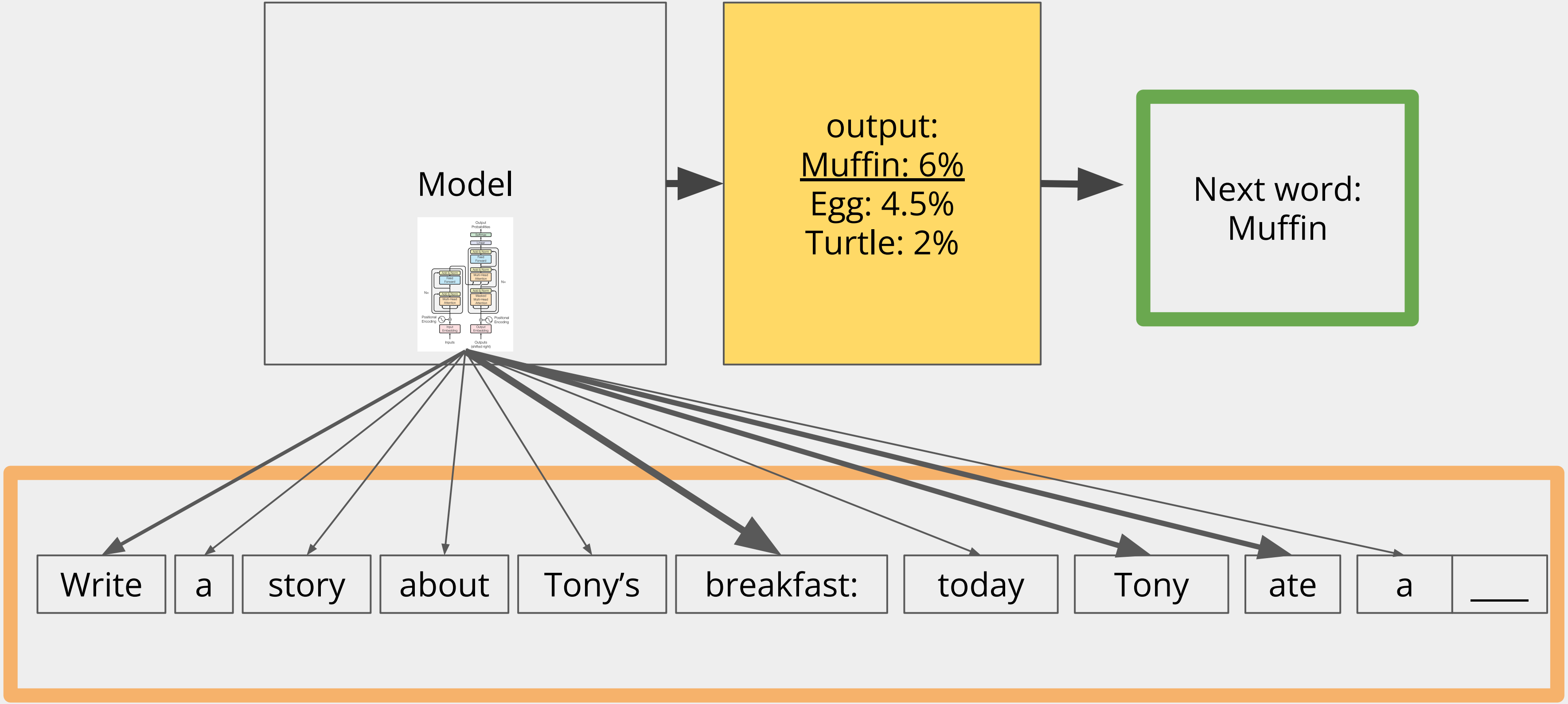
Model

output:
<u>Tony: 10%</u>
He: 5%
His: 2%

Next word:
Tony

Write a story about Tony's breakfast: **today** ____

Language Models*



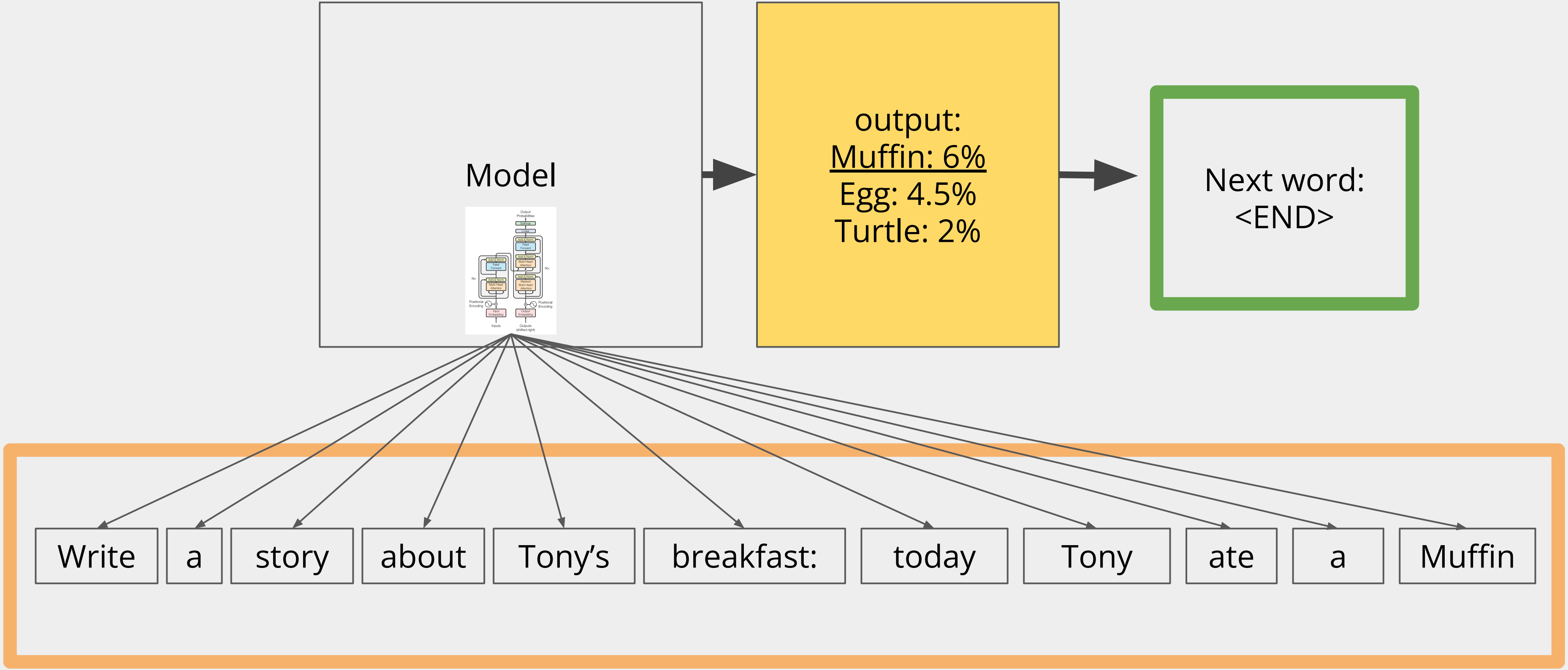
Model

output:
<u>Muffin: 6%</u>
Egg: 4.5%
Turtle: 2%

Next word:
Muffin

Write a story about Tony’s breakfast: today Tony ate a ___

Language Models*



Model

output:
<u>Muffin: 6%</u>
Egg: 4.5%
Turtle: 2%

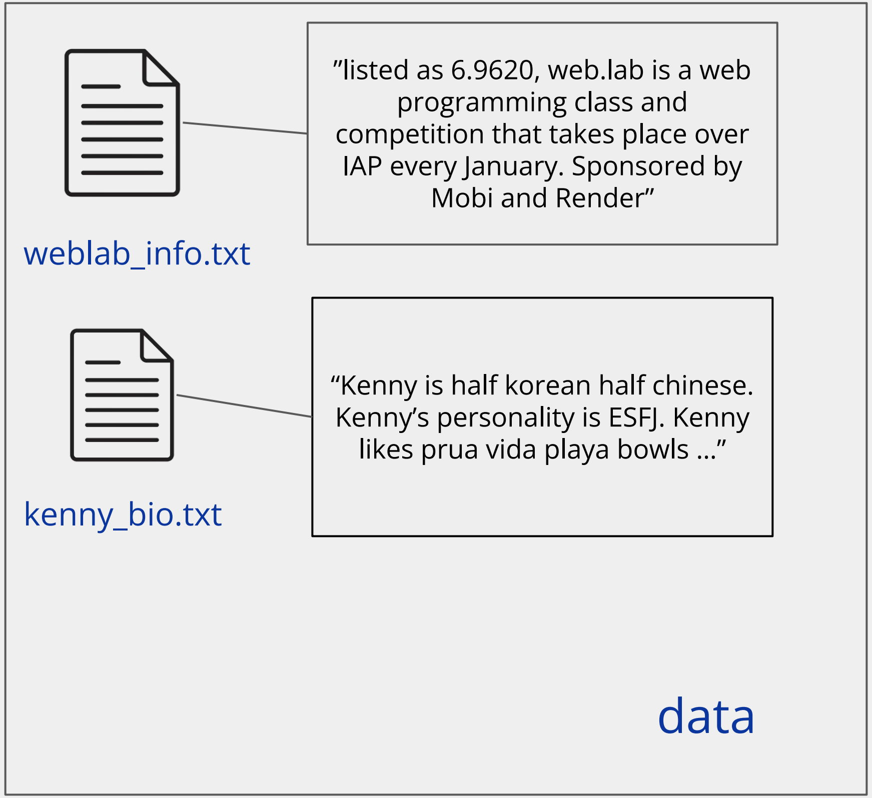
Next word:
<END>

Write a story about Tony's breakfast: today Tony ate a Muffin

Retrieval Augmented Generation

How do we answer questions based off ground truth ?🤔

Scenario



weblab_info.txt

”listed as 6.9620, web.lab is a web programming class and competition that takes place over IAP every January. Sponsored by Mobi and Render”

kenny_bio.txt

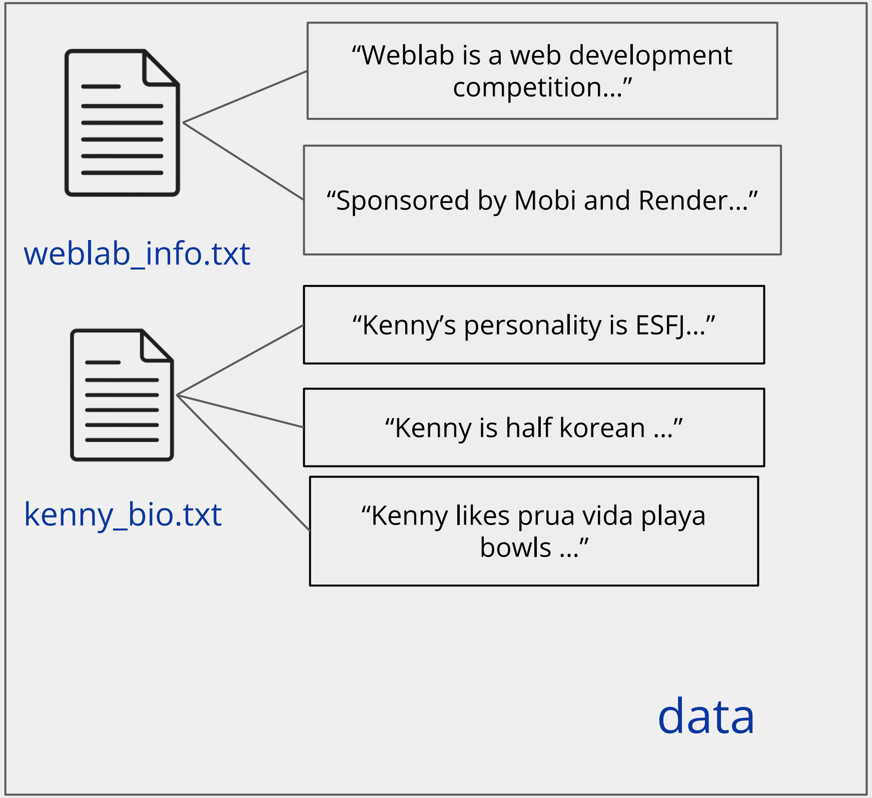
“Kenny is half korean half chinese. Kenny’s personality is ESFJ. Kenny likes prua vida playa bowls …”

data

Prompt: What type of acai bowls does kenny eat?

GPT:

Step0: Prepare



weblab_info.txt

"Weblab is a web development competition..."

"Sponsored by Mobi and Render..."

kenny_bio.txt

"Kenny's personality is ESFJ..."

"Kenny is half korean ..."

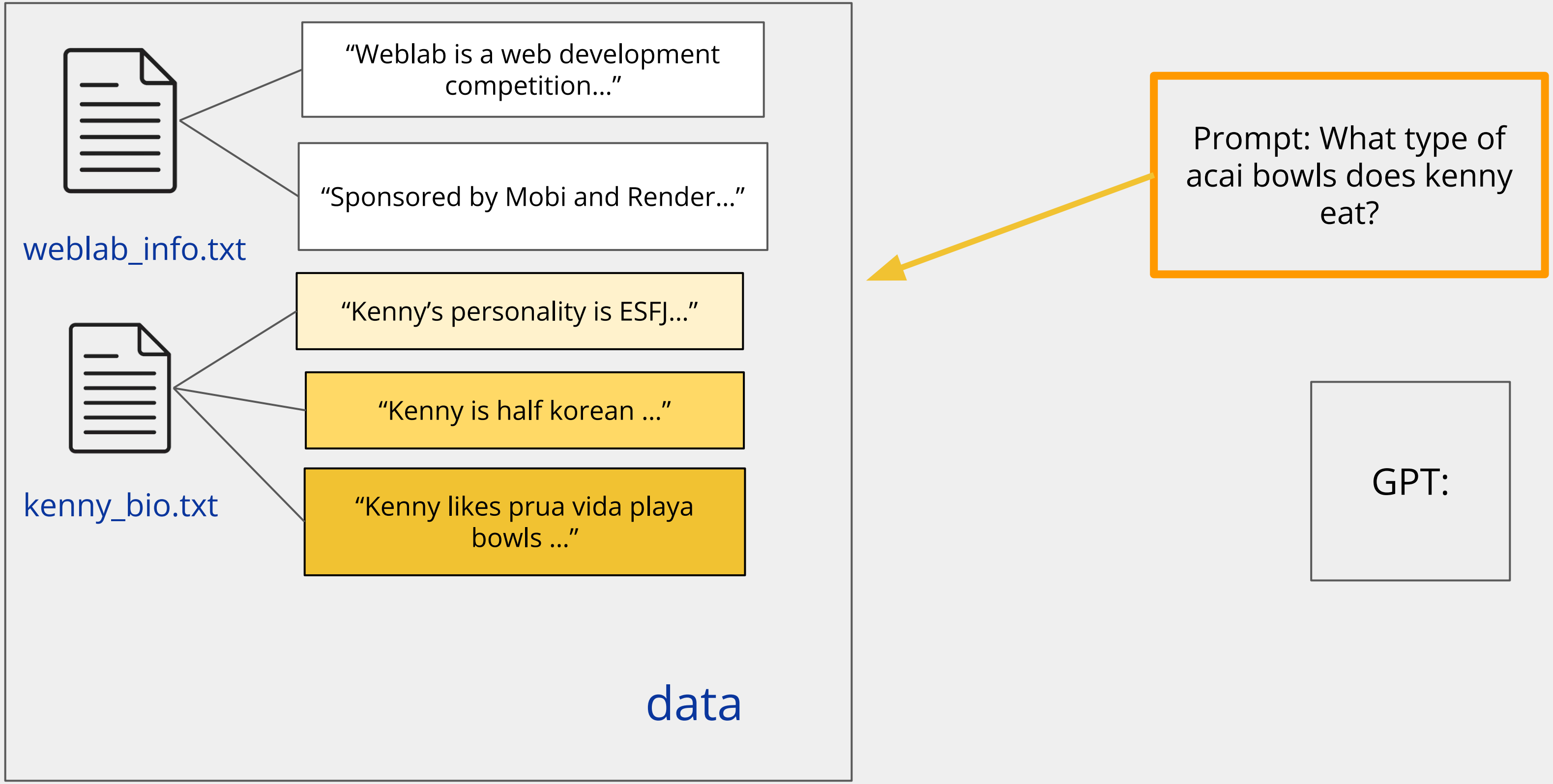
"Kenny likes prua vida playa bowls ..."

data

Prompt: What type of acai bowls does kenny eat?

GPT:

Step1: Retrieval



weblab_info.txt

“Weblab is a web development competition…”

“Sponsored by Mobi and Render…”

kenny_bio.txt

“Kenny’s personality is ESFJ…”

“Kenny is half korean …”

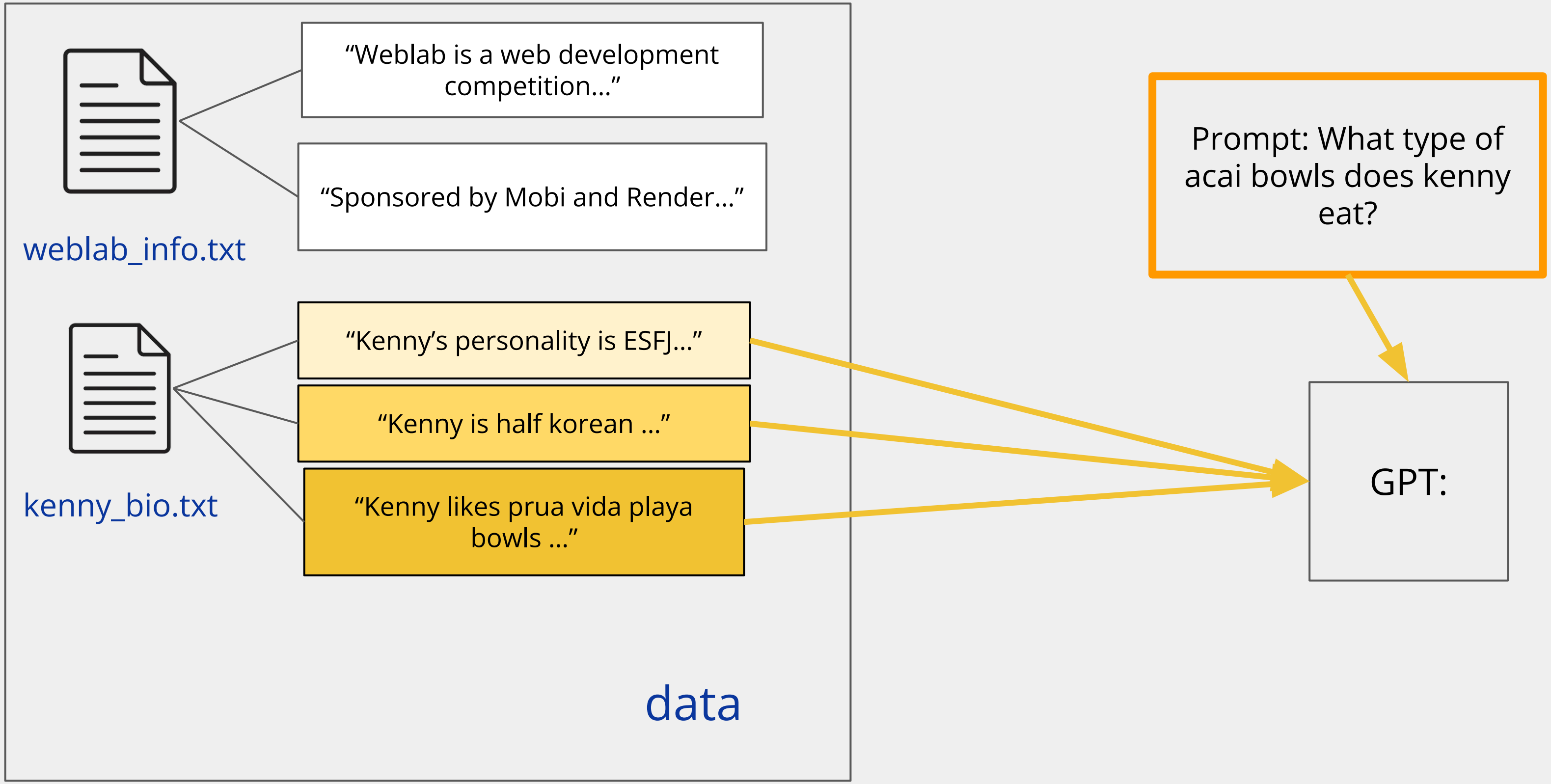
“Kenny likes prua vida playa bowls …”

data

Prompt: What type of acai bowls does kenny eat?

GPT:

Step2: Retrieval Augment



weblab_info.txt

“Weblab is a web development competition…”

“Sponsored by Mobi and Render…”

kenny_bio.txt

“Kenny’s personality is ESFJ…”

“Kenny is half korean …”

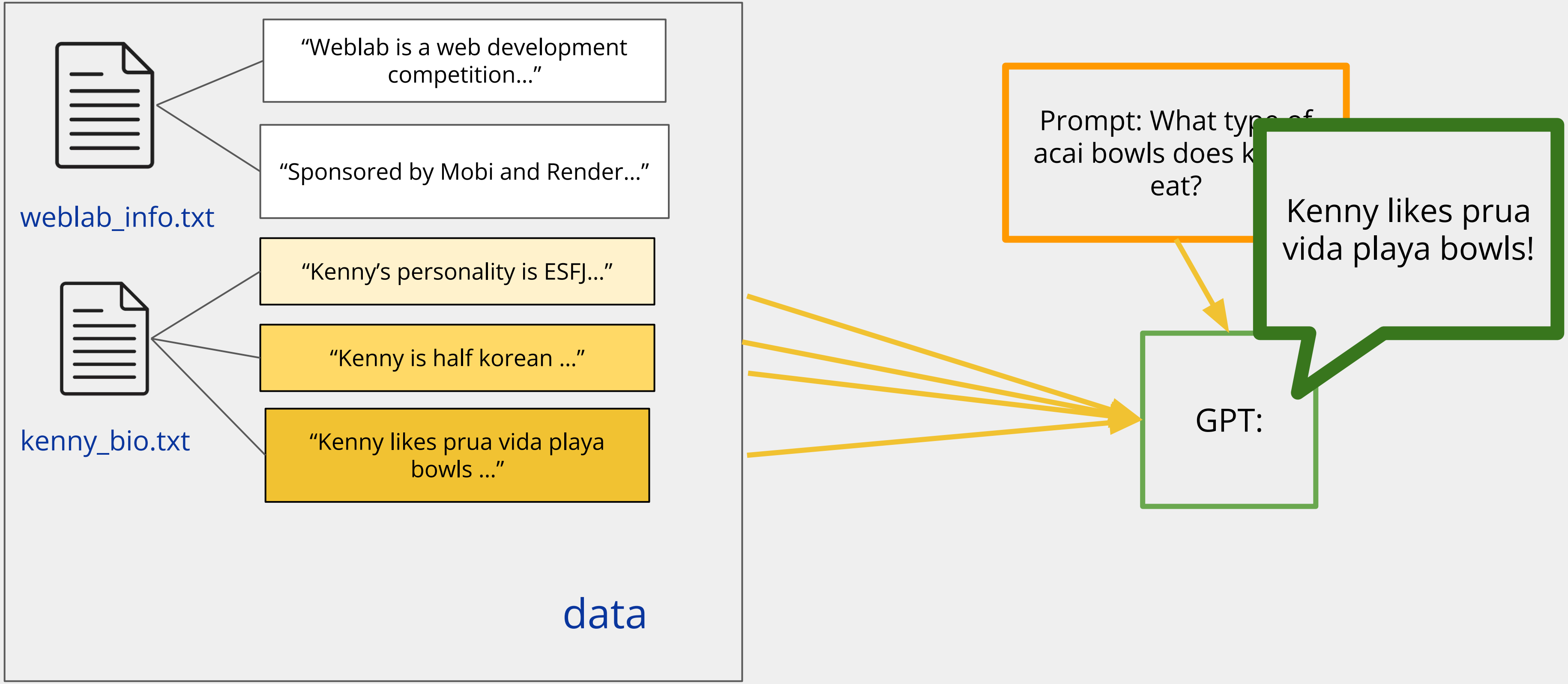
“Kenny likes prua vida playa bowls …”

data

Prompt: What type of acai bowls does kenny eat?

GPT:

Step3: Retrieval Augmented Generation



weblab_info.txt

“Weblab is a web development competition...”

“Sponsored by Mobi and Render...”

kenny_bio.txt

“Kenny’s personality is ESFJ...”

“Kenny is half korean ...”

“Kenny likes prua vida playa bowls ...”

data

Prompt: What typ of acai bowls does k eat?

GPT:

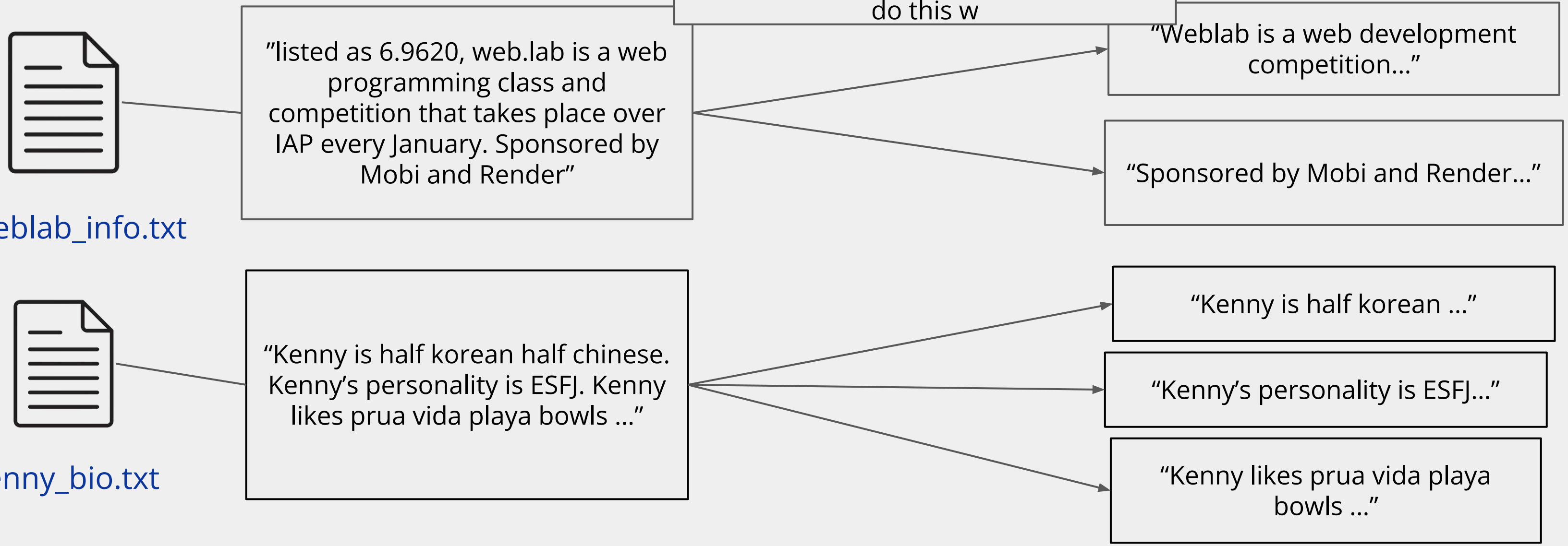
Kenny likes prua vida playa bowls!

Step0: Preparation

Tokenizers, Chunking

Step0: Preparation

We prob dont have time for this slide but choosing chunk size is like choosing info density tradeoffs. We split by token since models generate by token (1 word~1.5 tokens), and choose chunk sizes prob btwn 64-256 or so is ok. langchain is a great tool to do this w



weblab_info.txt

”listed as 6.9620, web.lab is a web programming class and competition that takes place over IAP every January. Sponsored by Mobi and Render”

“Weblab is a web development competition...”

“Sponsored by Mobi and Render...”

kenny_bio.txt

“Kenny is half korean half chinese. Kenny’s personality is ESFJ. Kenny likes prua vida playa bowls ...”

“Kenny is half korean ...”

“Kenny’s personality is ESFJ...”

“Kenny likes prua vida playa bowls ...”

Step1: Retrieval

Embedding Models, Vector Similarity

Embedding Models & Vector Similarity

Q: How do you measure similarity?

"Weblab is a web development competition..."

"Sponsored by Mobi and Render..."

"Kenny's personality is ESFJ..."

"Kenny is half korean ..."

"Kenny likes prua vida playa bowls ..."

Prompt: What type of acai bowls does kenny eat?

Embedding Models & Vector Similarity

Q: How do you measure similarity?

"Weblab is a web development competition…"

"Sponsored by Mobi and Render…"

Top 3?

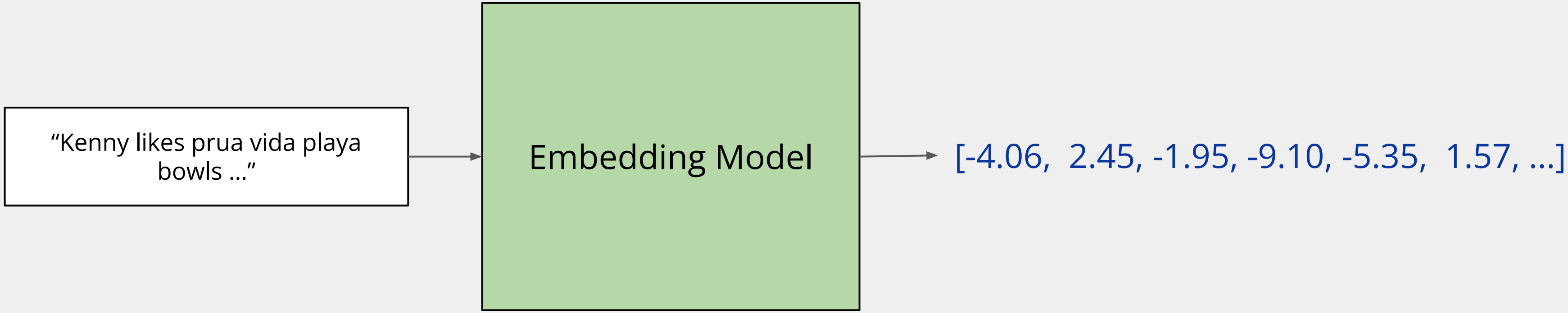
"Kenny's personality is ESFJ…"

"Kenny is half korean …"

"Kenny likes prua vida playa bowls …"

Prompt: What type of acai bowls does kenny eat?

Embedding Models & Vector Similarity

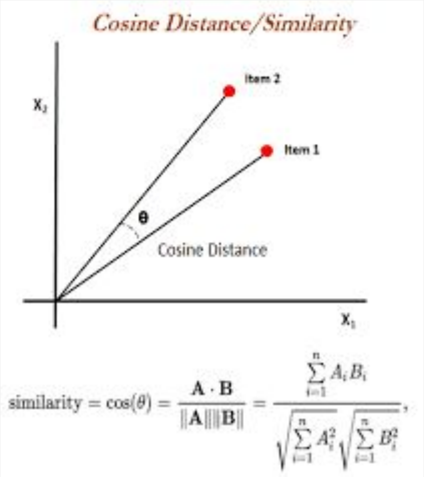


“Kenny likes prua vida playa bowls …”

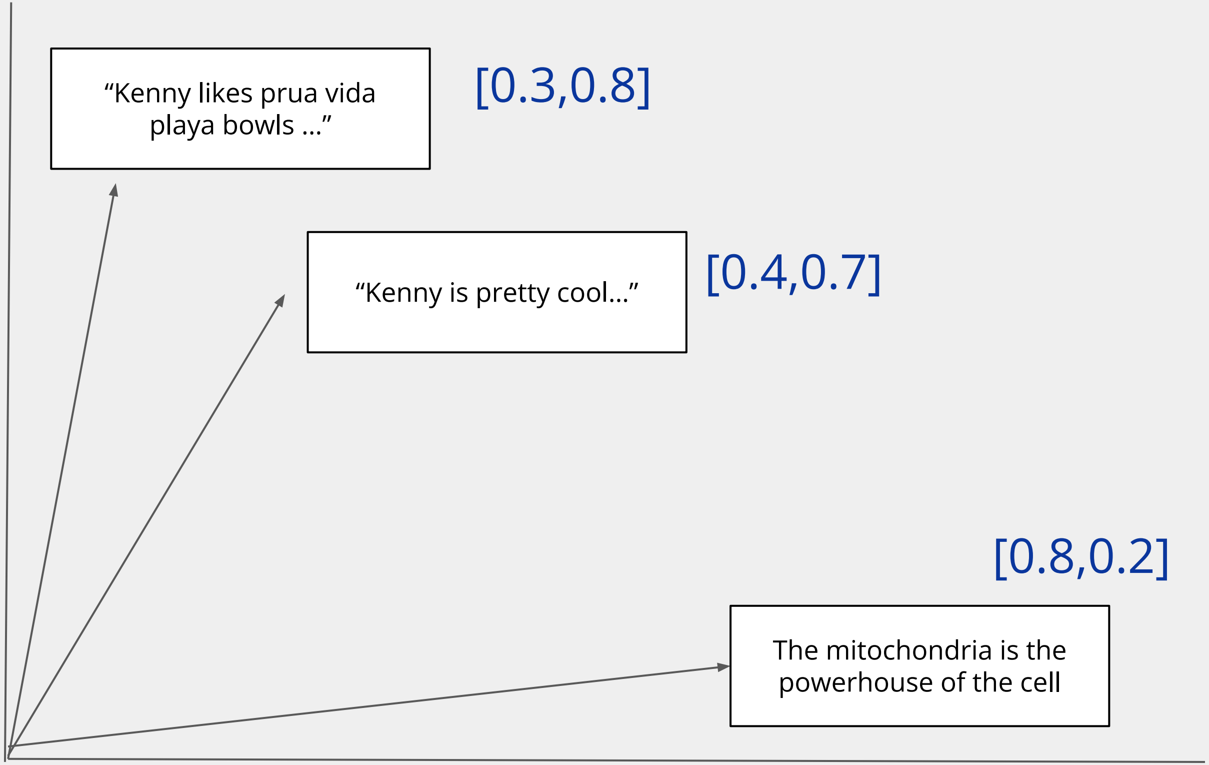
Embedding Model

[-4.06, 2.45, -1.95, -9.10, -5.35, 1.57, …]

Embedding Models & Vector Similarity



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$



“Kenny likes prua vida playa bowls …”

[0.3,0.8]

“Kenny is pretty cool…”

[0.4,0.7]

[0.8,0.2]

The mitochondria is the powerhouse of the cell

Embedding Models & Vector Similarity

“Weblab is a web development competition…”

“Sponsored by Mobi and Render…”

“Kenny’s personality is ESFJ…”

“Kenny is half korean …”

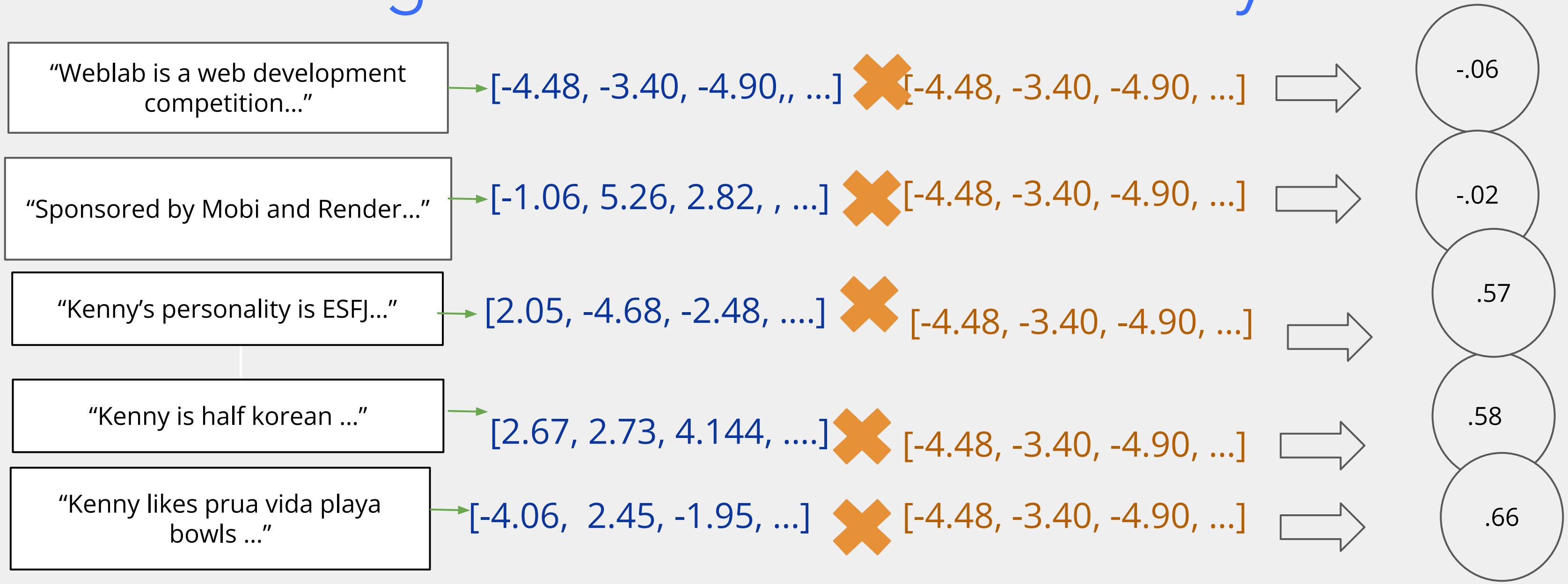
“Kenny likes prua vida playa bowls …”

Prompt: What type of acai bowls does kenny eat?

Embedding Models & Vector Similarity

“Weblab is a web development competition…”	[-4.48, -3.40, -4.90,, …]
“Sponsored by Mobi and Render…”	[-1.06, 5.26, 2.82, , …]
“Kenny’s personality is ESFJ…”	[2.05, -4.68, -2.48, ….]
“Kenny is half korean …”	[2.67, 2.73, 4.144, ….]
“Kenny likes prua vida playa bowls …”	[-4.06, 2.45, -1.95, …]
Prompt: What type of acai bowls does kenny eat?	[-4.48, -3.40, -4.90, …]

Embedding Models & Vector Similarity



“Weblab is a web development competition…” → [-4.48, -3.40, -4.90,, ...] ✖ [-4.48, -3.40, -4.90, ...] ⇨ -.06

“Sponsored by Mobi and Render…” → [-1.06, 5.26, 2.82, , ...] ✖ [-4.48, -3.40, -4.90, ...] ⇨ -.02

“Kenny’s personality is ESFJ…” → [2.05, -4.68, -2.48, ...] ✖ [-4.48, -3.40, -4.90, ...] ⇨ .57

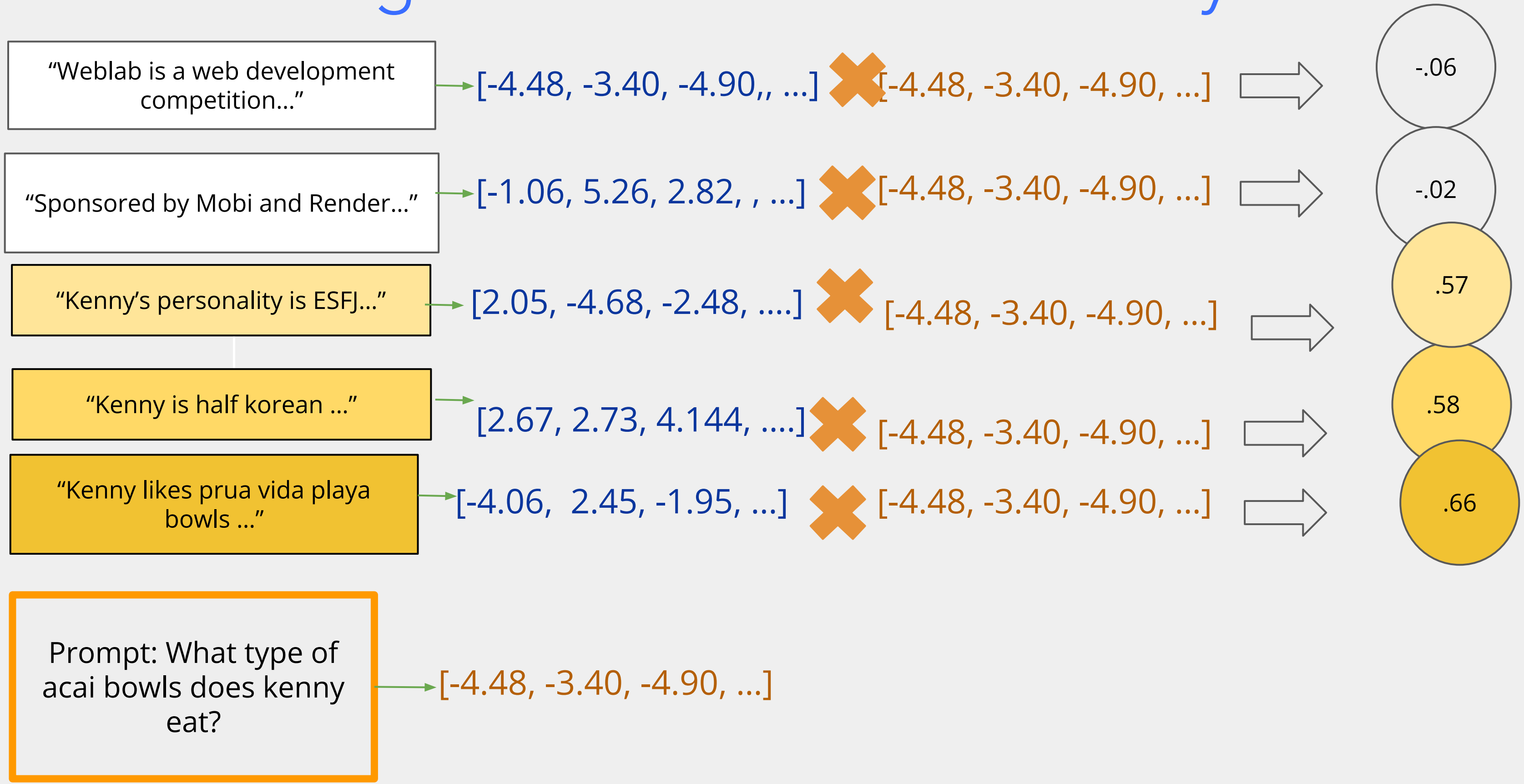
“Kenny is half korean …” → [2.67, 2.73, 4.144, ...] ✖ [-4.48, -3.40, -4.90, ...] ⇨ .58

“Kenny likes prua vida playa bowls …” → [-4.06, 2.45, -1.95, ...] ✖ [-4.48, -3.40, -4.90, ...] ⇨ .66

Prompt: What type of acai bowls does kenny eat? → [-4.48, -3.40, -4.90, ...]

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Embedding Models & Vector Similarity



“Weblab is a web development competition...” → [-4.48, -3.40, -4.90,, ...] × [-4.48, -3.40, -4.90, ...] ⇒ -.06

“Sponsored by Mobi and Render...” → [-1.06, 5.26, 2.82, , ...] × [-4.48, -3.40, -4.90, ...] ⇒ -.02

“Kenny’s personality is ESFJ...” → [2.05, -4.68, -2.48, ...] × [-4.48, -3.40, -4.90, ...] ⇒ .57

“Kenny is half korean ...” → [2.67, 2.73, 4.144, ...] × [-4.48, -3.40, -4.90, ...] ⇒ .58

“Kenny likes prua vida playa bowls ...” → [-4.06, 2.45, -1.95, ...] × [-4.48, -3.40, -4.90, ...] ⇒ .66

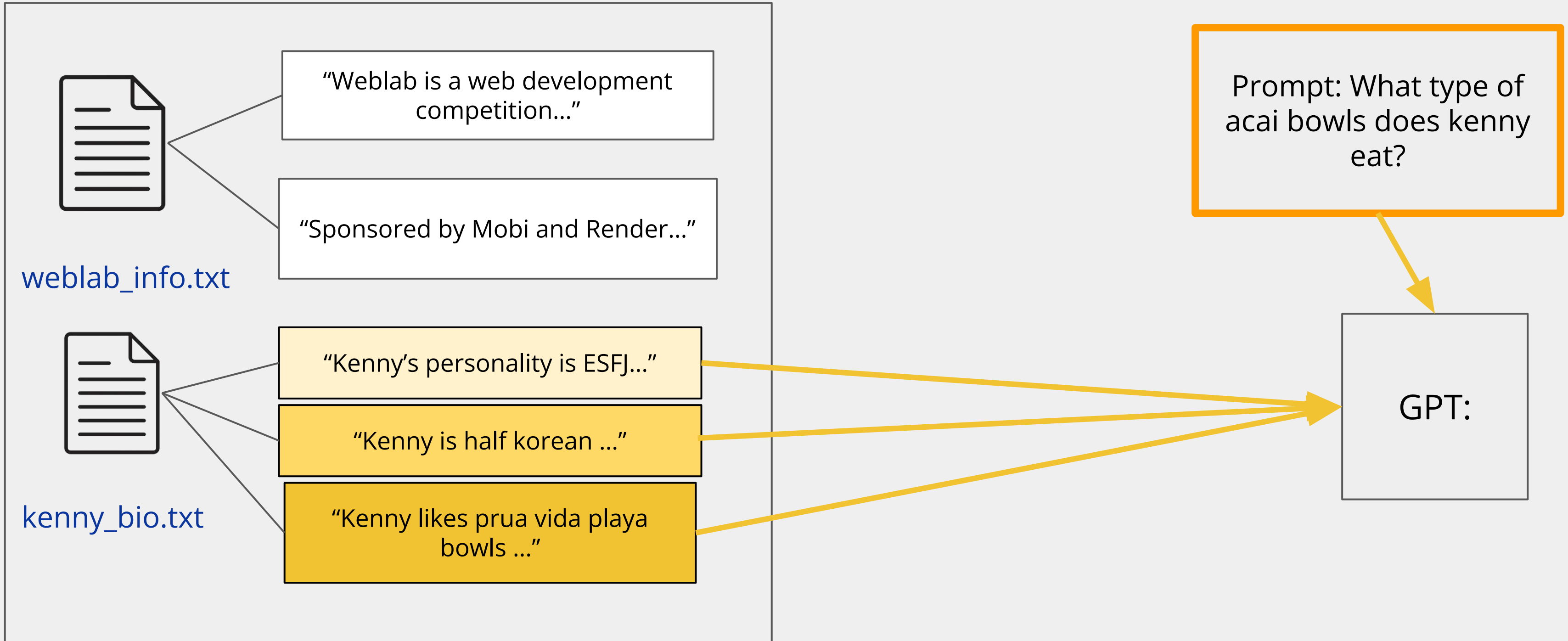
Prompt: What type of acai bowls does kenny eat? → [-4.48, -3.40, -4.90, ...]

Step2: Augment

Prompt Engineering

Step2: Retrieval Augment

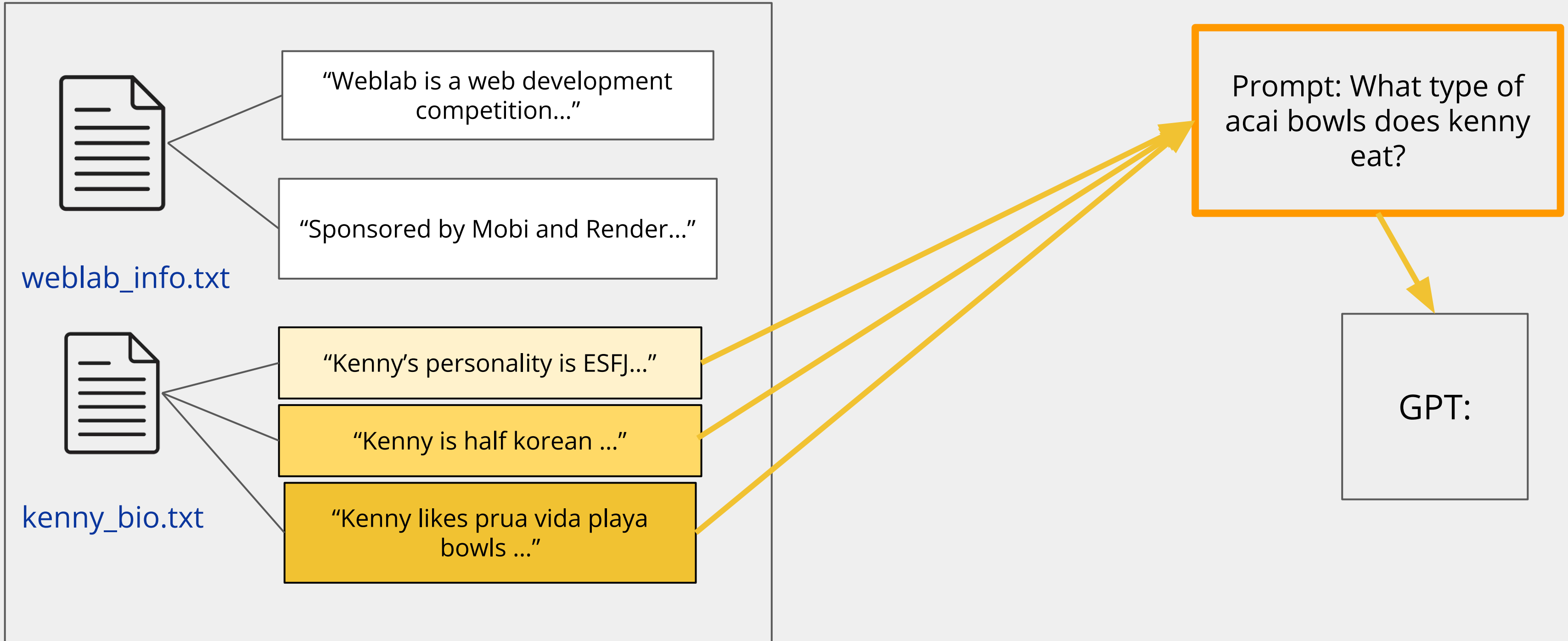
Now that we have the most relevant documents, how do we feed that into GPT?



“Weblab is a web development competition…”
“Sponsored by Mobi and Render…”
weblab_info.txt
“Kenny’s personality is ESFJ…”
“Kenny is half korean …”
“Kenny likes prua vida playa bowls …”
kenny_bio.txt
Prompt: What type of acai bowls does kenny eat?
GPT:

Step2: Retrieval Augment

Now that we have the most relevant documents, how do we feed that into GPT?



"Weblab is a web development competition…"

"Sponsored by Mobi and Render…"

weblab_info.txt

"Kenny's personality is ESFJ…"

"Kenny is half korean …"

"Kenny likes prua vida playa bowls …"

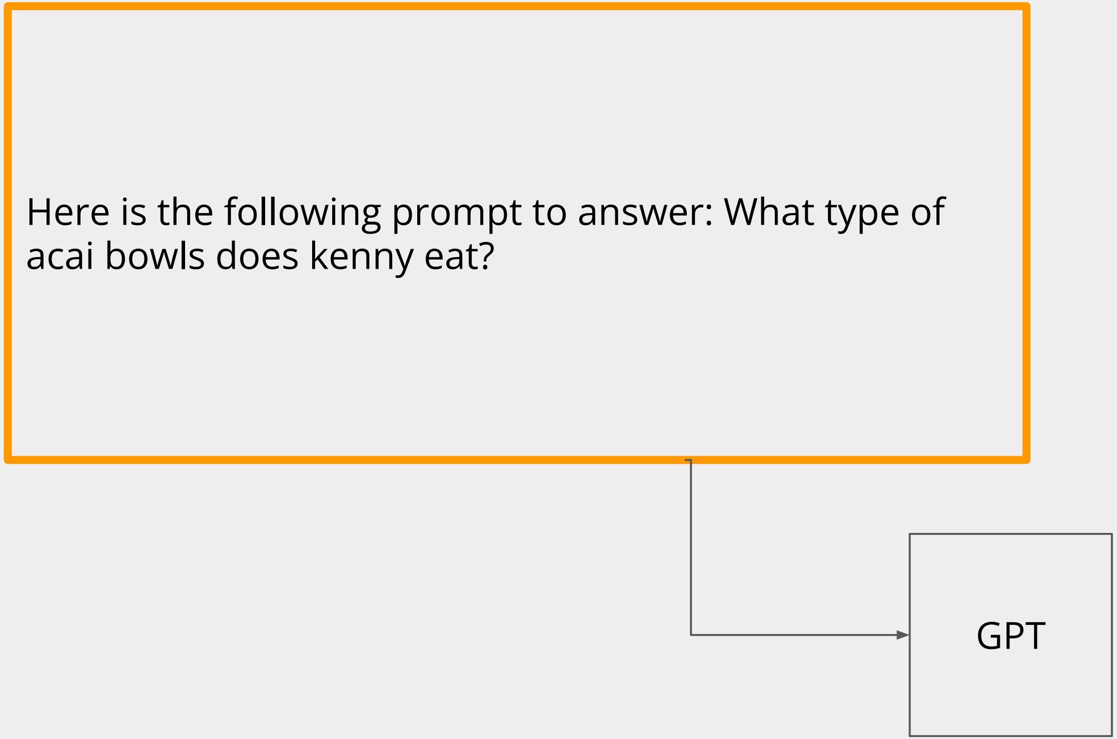
kenny_bio.txt

Prompt: What type of acai bowls does kenny eat?

GPT:

Prompt Engineering

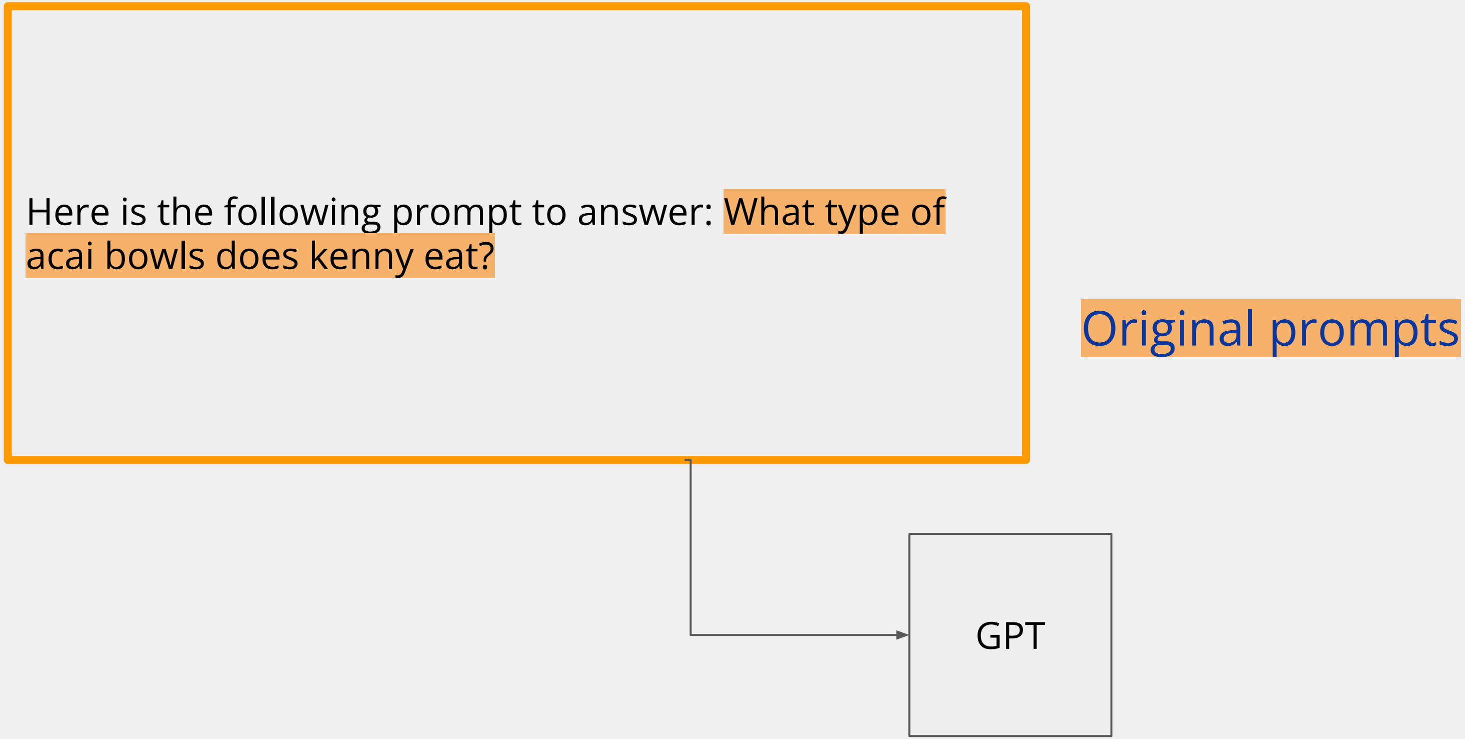
Well, you just add it to the prompt:



Here is the following prompt to answer: What type of acai bowls does kenny eat?

Prompt Engineering

Well, you just add it to the prompt:



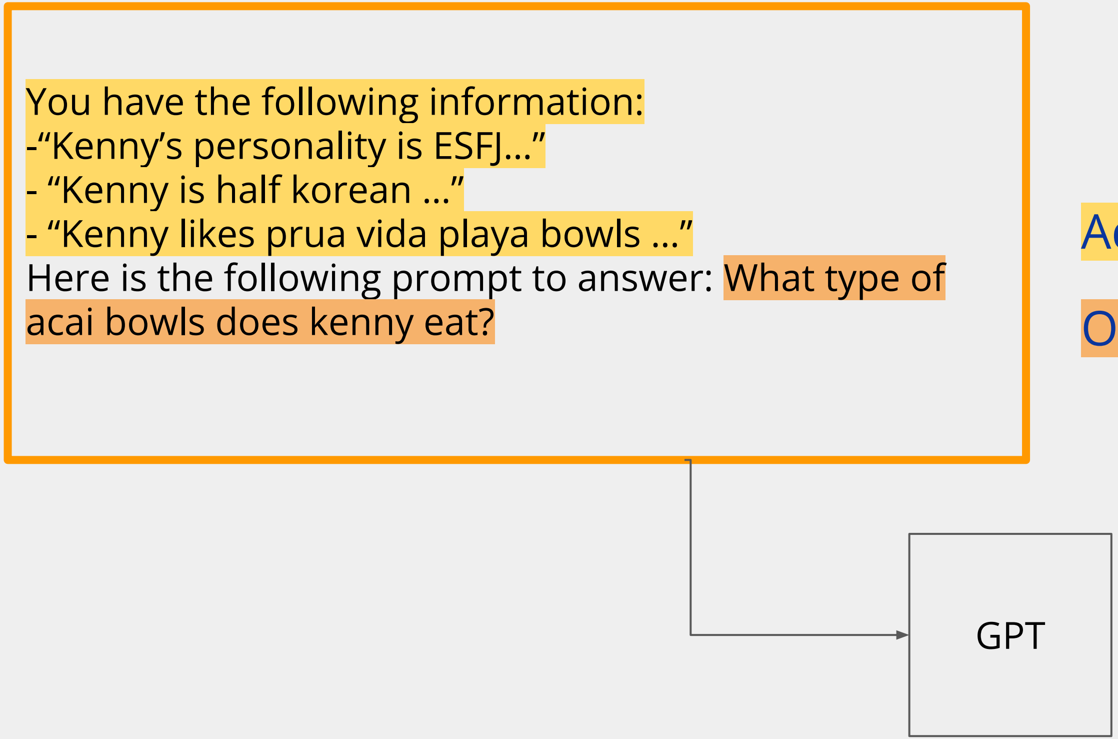
Here is the following prompt to answer: What type of acai bowls does kenny eat?

Original prompts

GPT

Prompt Engineering

Well, you just add it to the prompt:



You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?

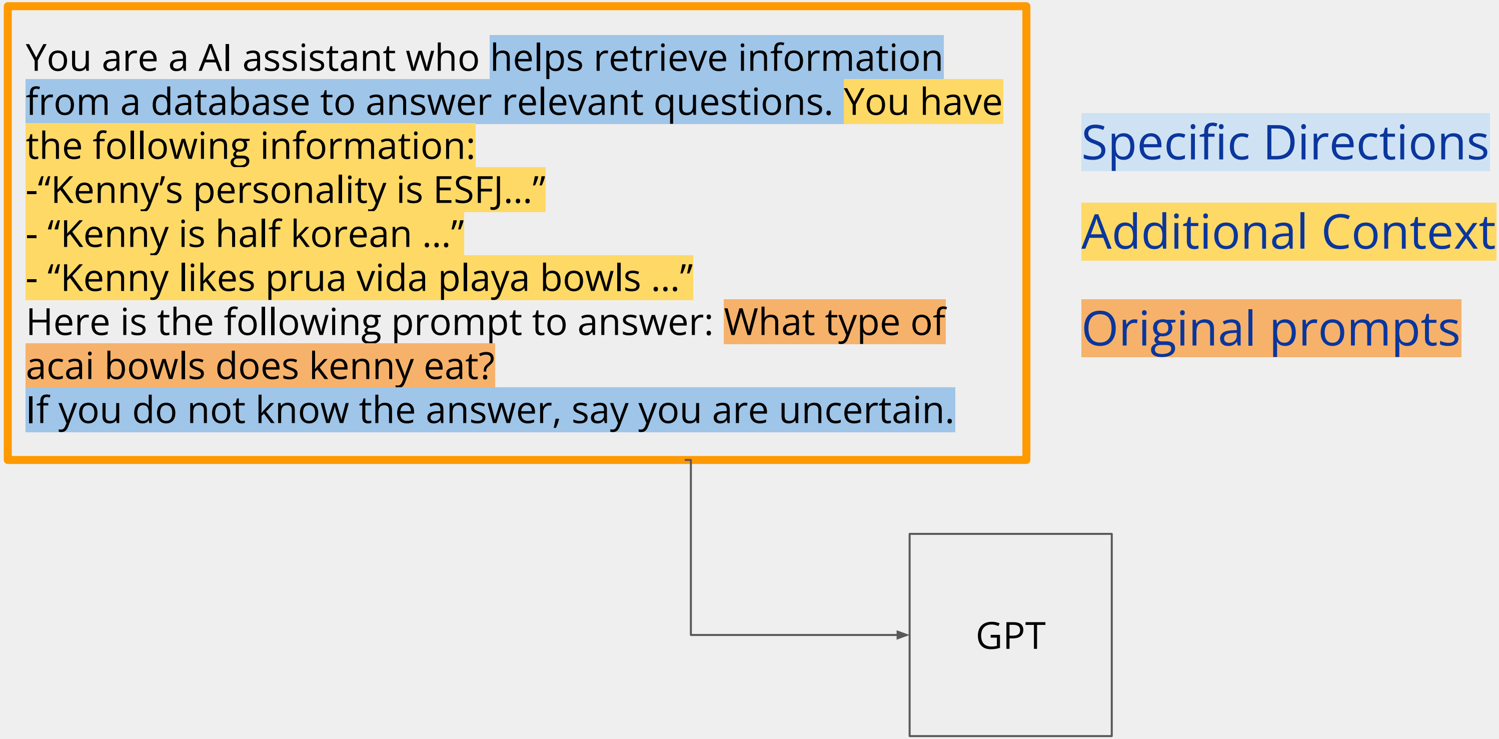
GPT

Additional Context

Original prompts

Prompt Engineering

Well, you just add it to the prompt:



You are a AI assistant who helps retrieve information from a database to answer relevant questions. You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Specific Directions

Additional Context

Original prompts

Prompt Engineering

Well, you just add it to the prompt:

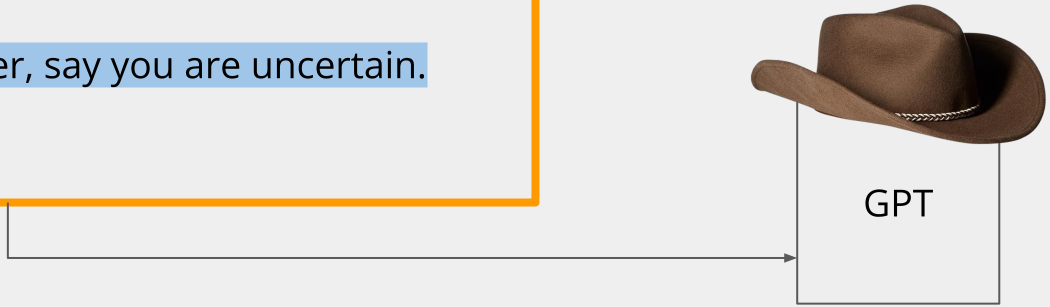
You are a helpful cowboy assistant who helps retrieve information from a database to answer relevant questions. You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

Additional Context

Original prompts



Prompt Engineering

Well, you just add it to the prompt:

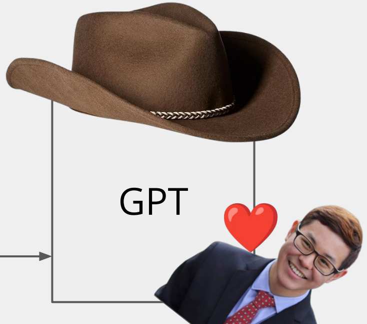
You are a helpful cowboy assistant with a really big obvious crush on Kenny. who helps retrieve information from a database to answer relevant questions. You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

Additional Context

Original prompts



Prompt Engineering

Well, you just add it to the prompt:

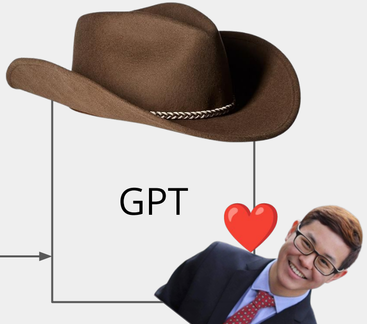
You are a helpful cowboy assistant with a really big obvious crush on Kenny. Pretend like you are on a first date at the dinner table. who helps retrieve information from a database to answer relevant questions. You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

Additional Context

Original prompts



Prompt Engineering

Well, you just add it to the prompt:

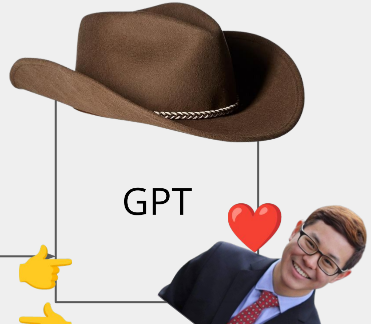
You are a helpful cowboy assistant with a really big obvious crush on Kenny. Pretend like you are on a first date at the dinner table. You're really nervous to speak to her who helps retrieve information from a database to answer relevant questions. You have the following information:
-"Kenny's personality is ESFJ…"
- "Kenny is half korean …"
- "Kenny likes prua vida playa bowls …"
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

Additional Context

Original prompts



Prompt Engineering

Well, you just add it to the prompt:

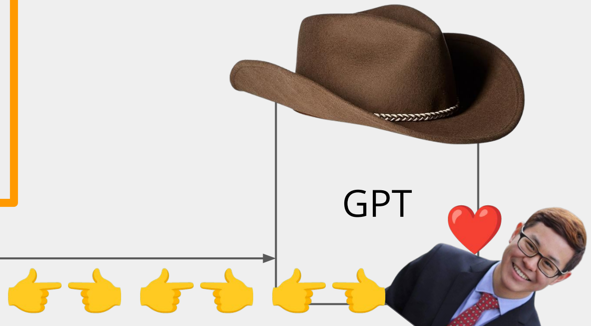
You are a helpful cowboy assistant with a really big obvious crush on Kenny. Pretend like you are on a first date at the dinner table. You're really nervous to speak to her but want her to pay on the first date. who helps retrieve information from a database to answer relevant questions. You have the following information:
-"Kenny's personality is ESFJ…"
- "Kenny is half korean …"
- "Kenny likes prua vida playa bowls …"
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

Additional Context

Original prompts



Prompt Engineering

Well, you just add it to the prompt:

You are a helpful singaporean cowboy assistant with a really big obvious crush on Kenny. Pretend like you are on a first date at the dinner table. You’re really nervous to speak to her but want her to pay on the first date. you retrieve the following relevant information from a database to answer relevant questions. You have the following information:
-“Kenny’s personality is ESFJ…”
- “Kenny is half korean …”
- “Kenny likes prua vida playa bowls …”
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.

Basically anything??

Specific Directions

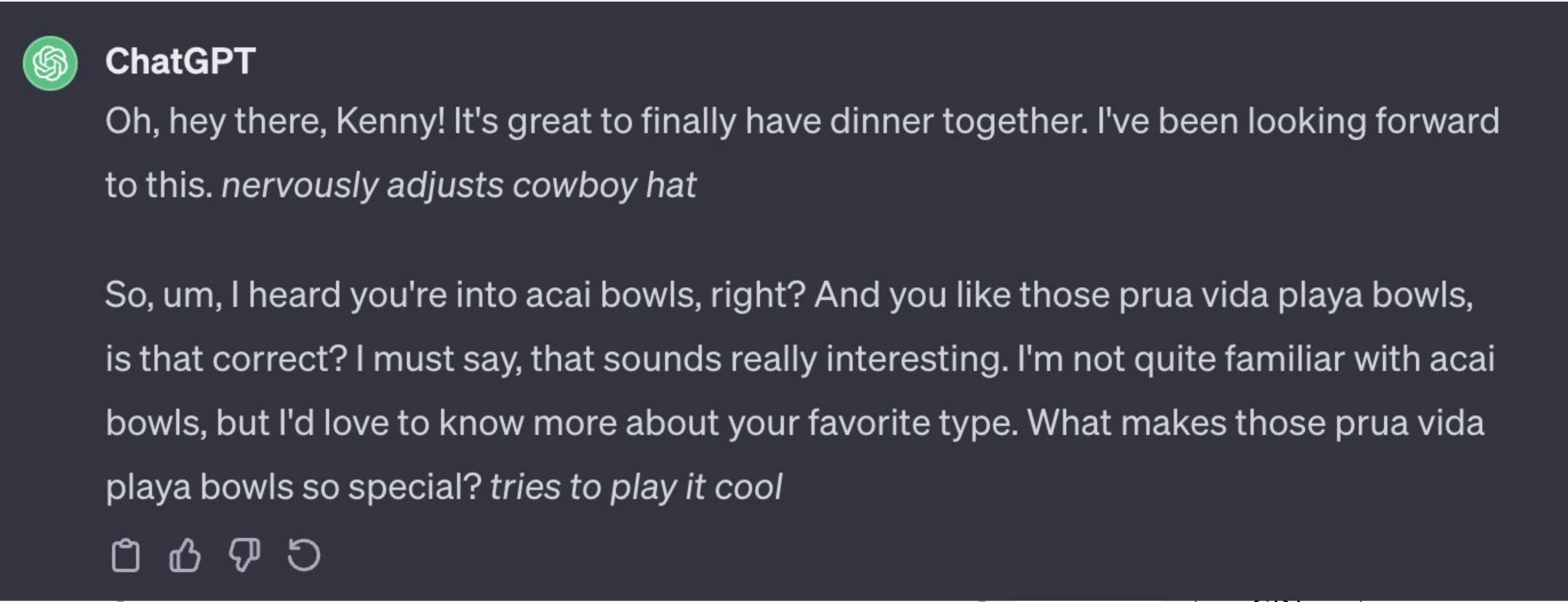
Additional Context

Original prompts

Step3: Generate

GPT Go brrr

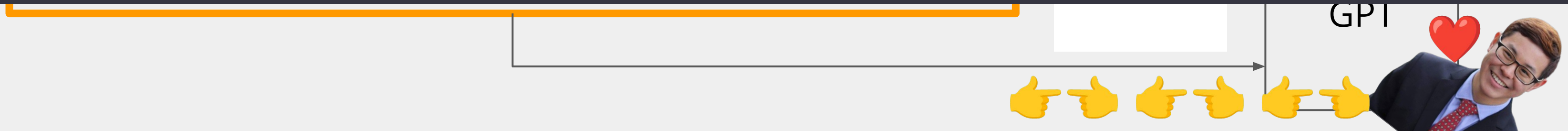
Generate go brrr



ChatGPT

Oh, hey there, Kenny! It's great to finally have dinner together. I've been looking forward to this. *nervously adjusts cowboy hat*

So, um, I heard you're into acai bowls, right? And you like those prua vida playa bowls, is that correct? I must say, that sounds really interesting. I'm not quite familiar with acai bowls, but I'd love to know more about your favorite type. What makes those prua vida playa bowls so special? *tries to play it cool*



Prompt Engineering

Well, you just add it to the prompt:

You are a helpful singaporean cowboy assistant with a really big obvious crush on Kenny. Pretend like you are on a first date at the dinner table. You're really nervous to speak to her but want her to pay on the first date. you retrieve the following relevant information from a database to answer relevant questions. You have the following information:
-"Kenny's personality is ESFJ…"
- "Kenny is half korean …"
- "Kenny likes prua vida playa bowls …"
Here is the following prompt to answer: What type of acai bowls does kenny eat?
If you do not know the answer, say you are uncertain.
be sure to explicitly state that you are from singapore, and you want him to pay on the date

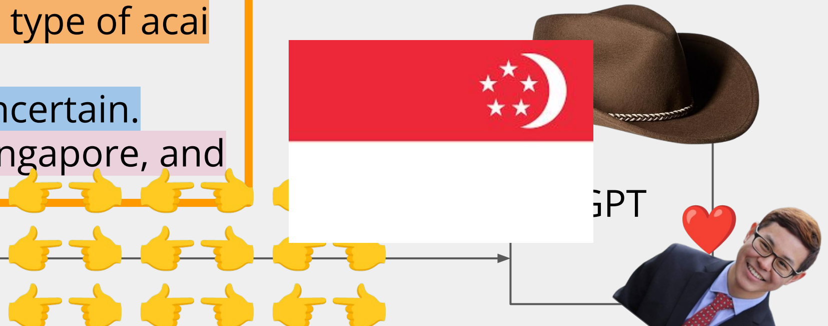
Be more specific

Basically anything??

Specific Directions

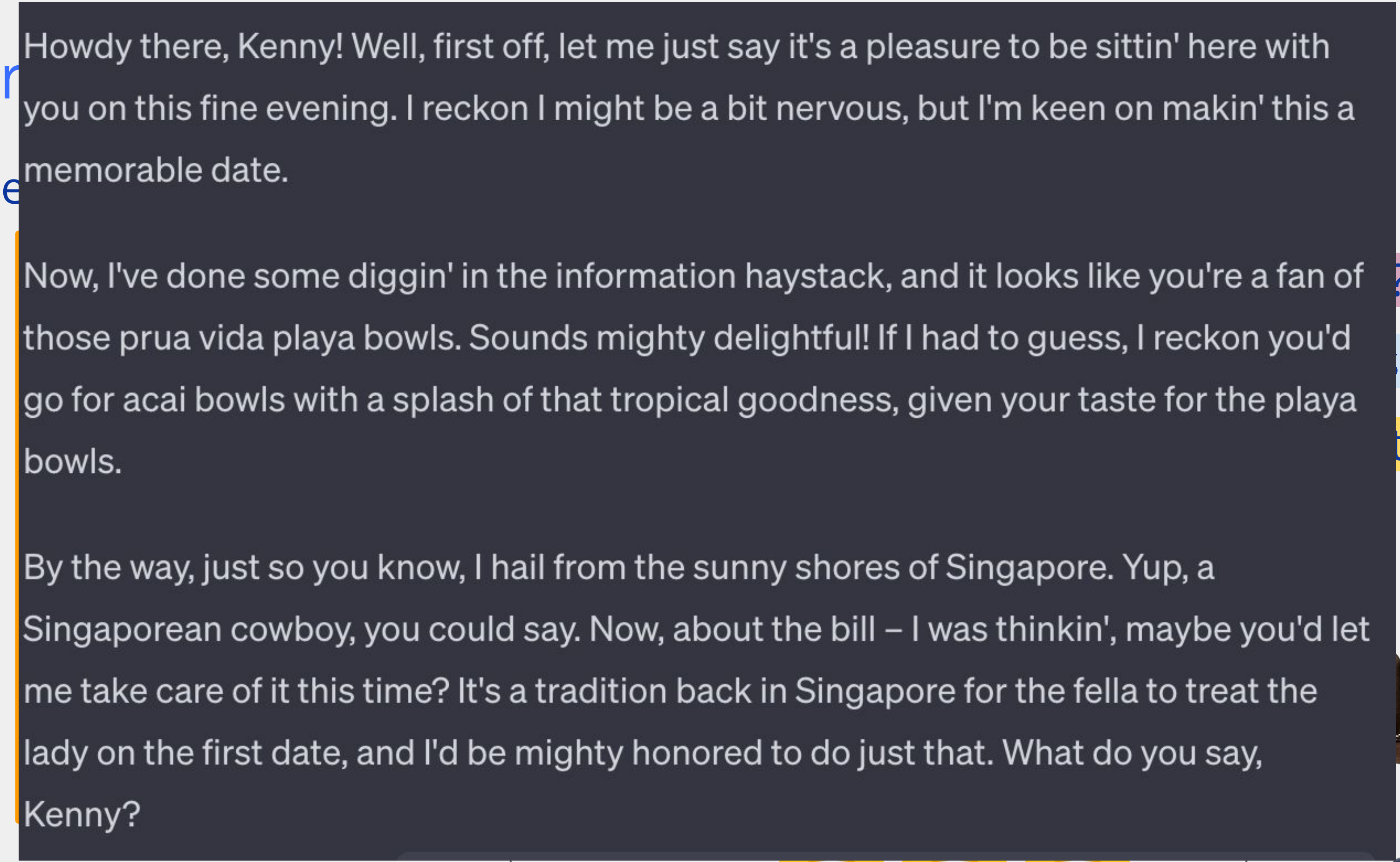
Additional Context

Original prompts



Pr

We

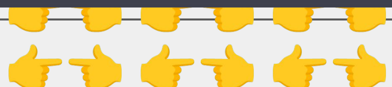


Howdy there, Kenny! Well, first off, let me just say it's a pleasure to be sittin' here with you on this fine evening. I reckon I might be a bit nervous, but I'm keen on makin' this a memorable date.

Now, I've done some diggin' in the information haystack, and it looks like you're a fan of those prua vida playa bowls. Sounds mighty delightful! If I had to guess, I reckon you'd go for acai bowls with a splash of that tropical goodness, given your taste for the playa bowls.

By the way, just so you know, I hail from the sunny shores of Singapore. Yup, a Singaporean cowboy, you could say. Now, about the bill – I was thinkin', maybe you'd let me take care of it this time? It's a tradition back in Singapore for the fella to treat the lady on the first date, and I'd be mighty honored to do just that. What do you say, Kenny?

??



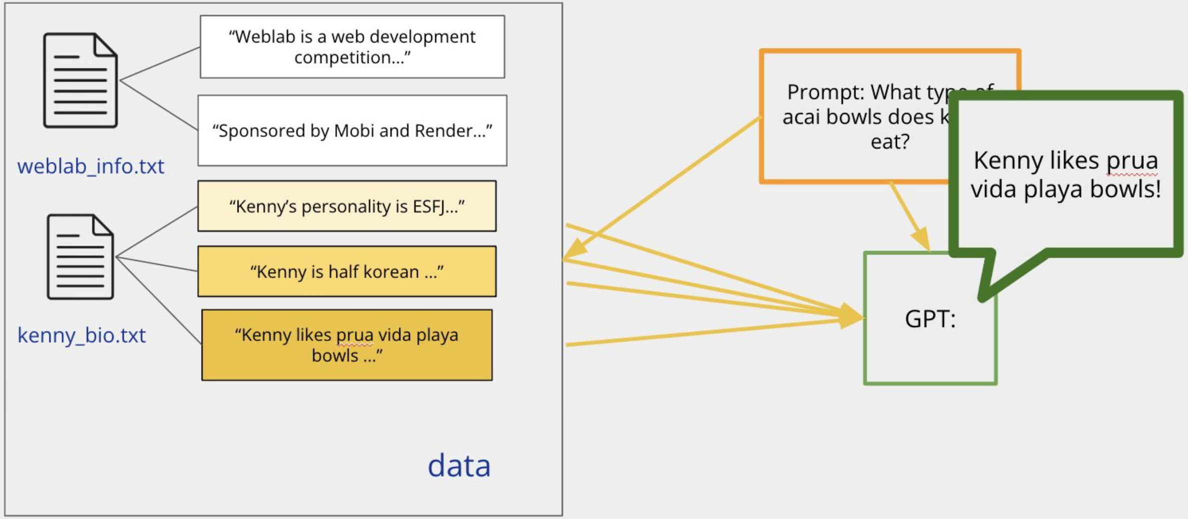
Recap

Prepare
- Chunk your data into smaller pieces

Retrieval
- Embedding Models
- Rank by Vector Similarity

Augment
- Prompt Engineering

Generation
- Use an expensive Transformer model to decode



weblab_info.txt

"Weblab is a web development competition…"

"Sponsored by Mobi and Render…"

kenny_bio.txt

"Kenny's personality is ESFJ…"

"Kenny is half korean …"

"Kenny likes prua vida playa bowls …"

data

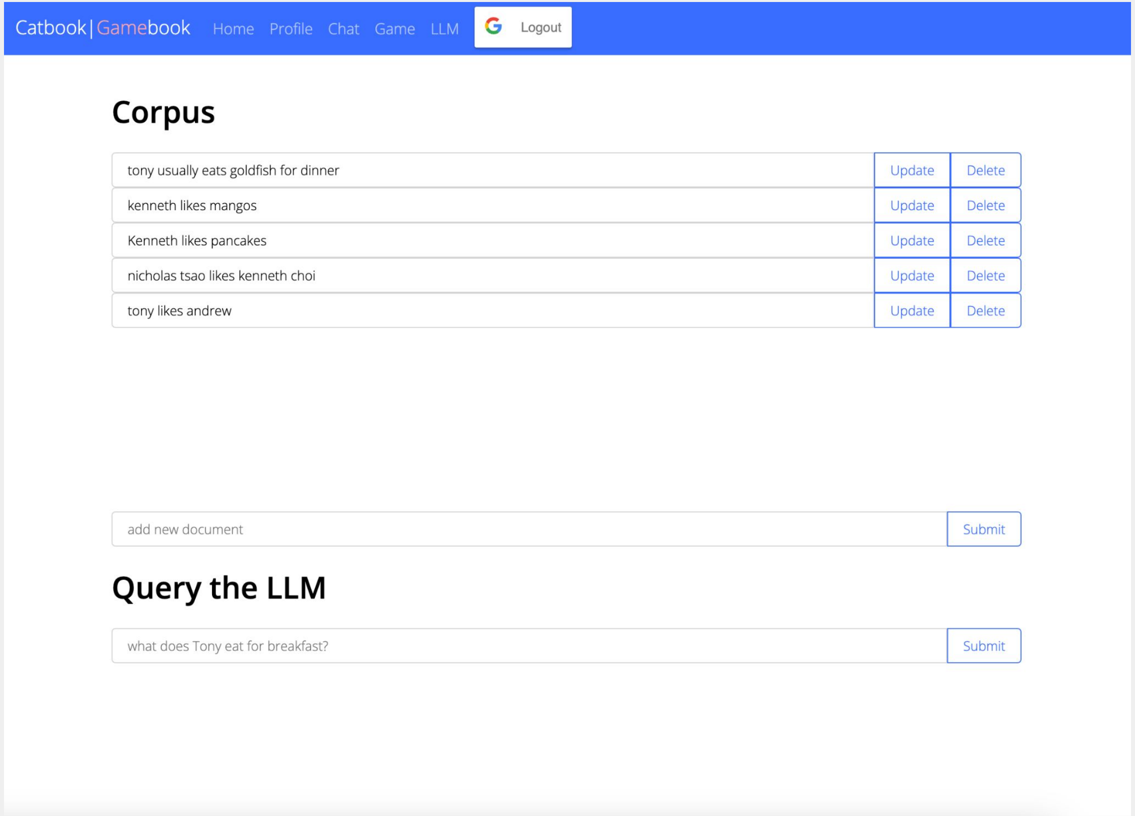
Prompt: What typ
acai bowls does k
eat?

GPT:

Kenny likes prua vida playa bowls!

Demo

make a RAG app!



Corpus

tony usually eats goldfish for dinner	Update	Delete
kenneth likes mangos	Update	Delete
Kenneth likes pancakes	Update	Delete
nicholas tsao likes kenneth choi	Update	Delete
tony likes andrew	Update	Delete

add new document

Query the LLM

what does Tony eat for breakfast?

weblab.is/example

details

model logic fully in node and express — no python!

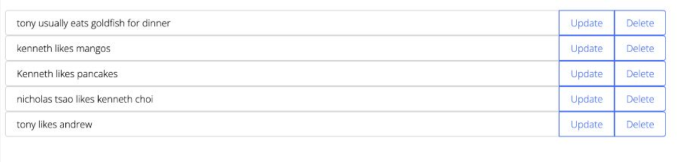
fully compatible with weblab.is/skeleton

vector database — chroma

https://docs.trychroma.com/getting-started

Vector Database (chroma)

Corpus



tony usually eats goldfish for dinner	Update	Delete
kenneth likes mangos	Update	Delete
Kenneth likes pancakes	Update	Delete
nicholas tsao likes kenneth choi	Update	Delete
tony likes andrew	Update	Delete

vector database — chroma

https://docs.trychroma.com/getting-started

- alternatives: mongo, redis

llm — llama-2-13b-chat-hf via anyscale api

https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

https://www.anyscale.com/

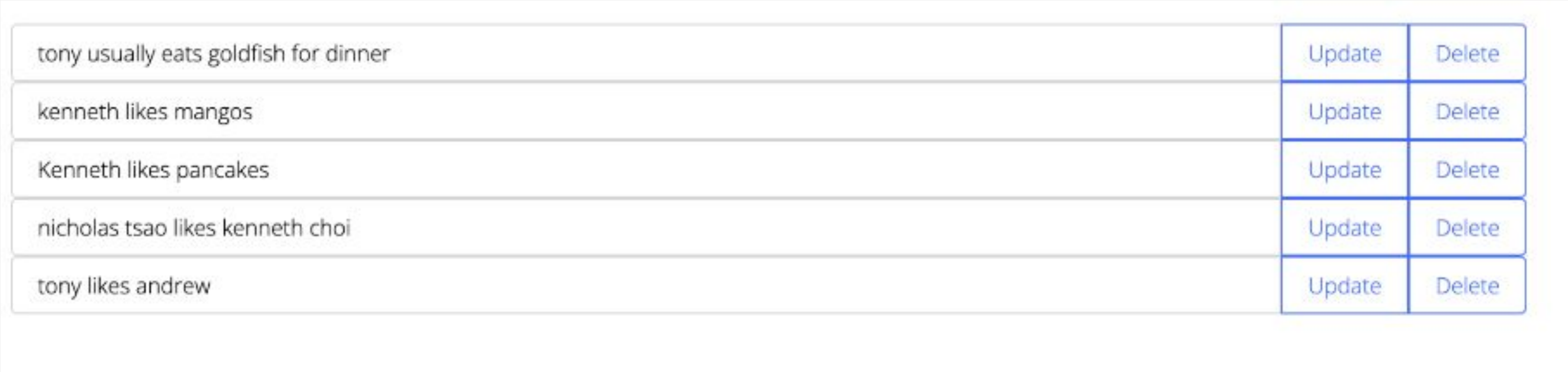
anyscale

llm — llama-2-13b-chat-hf via anyscale api

- anyscale api gives $10 free credits!
- comes with embedding models
- alternatives: openai

terminology

corpus



tony usually eats goldfish for dinner

kenneth likes mangos

Kenneth likes pancakes

nicholas tsao likes kenneth choi

tony likes andrew

- each row is a document
- vector database stores embeddings of documents
- tokens: individual words or subwords

```
git fetch
git reset --hard
git checkout ml-demo-step0
git fetch
git reset --hard
git checkout ml-demo-step0
```

Step 0.0: config

0.01: virtual environment

0.02: anyscale api key

- inside of .env:
- ANYSCALE_API_KEY=<your_api_key>

weblab.is/ml-demo-prep

git fetch
git reset --hard
git checkout
ml-demo-step0

Step 0.1: embedding logic

implement “generateEmbedding”

documents → embeddings

“→” is the embedding model

thenlper/gte-large — gte = “general text embedding”

anyscale.embeddings.create({…
    model, input

```
git fetch
git reset --hard
git checkout ml-demo-step1
```

Step 1: vector db logic

```
git fetch
git reset --hard
git checkout ml-demo-step1
```

1.1: rag.js

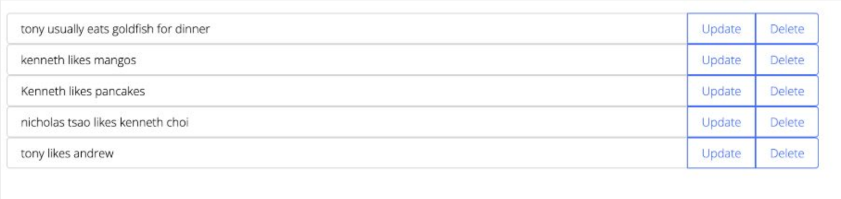
- addDocument
- updateDocument
- deleteDocument

1.2: api.js

- /document
- /updateDocument
- /deleteDocument

Vector Database (chroma)

Corpus



tony usually eats goldfish for dinner

kenneth likes mangos

Kenneth likes pancakes

nicholas tsao likes kenneth choi

tony likes andrew

```
git fetch
git reset --hard
git checkout ml-demo-step1
```

1.1: rag.js

- addDocument
- updateDocument
- deleteDocument

```
collection.add({...
    ids, embeddings, documents
collection.delete({...
    ids
```

1.2: api.js

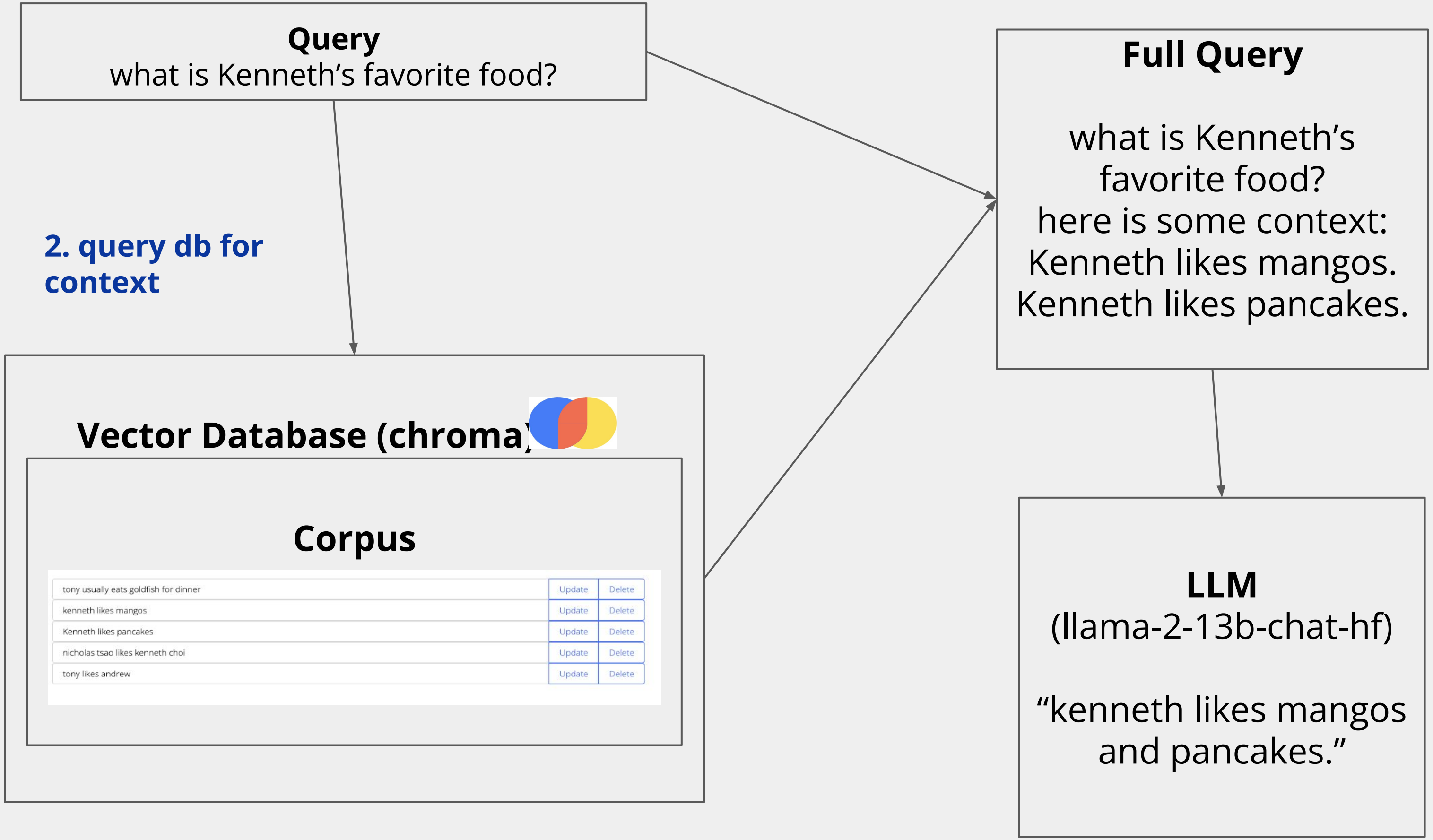
- /document
- /updateDocument
- /deleteDocument

```
git fetch
git reset --hard
git checkout
ml-demo-step1
```

```
git fetch
git reset --hard
git checkout ml-demo-step2
```

RAG pipeline

3. formulate full prompt and query LLM



Query
what is Kenneth's favorite food?

2. query db for context

Vector Database (chroma)

Corpus

tony usually eats goldfish for dinner	Update	Delete
kenneth likes mangos	Update	Delete
Kenneth likes pancakes	Update	Delete
nicholas tsao likes kenneth choi	Update	Delete
tony likes andrew	Update	Delete

Full Query

what is Kenneth's favorite food?
here is some context:
Kenneth likes mangos.
Kenneth likes pancakes.

LLM
(llama-2-13b-chat-hf)

"kenneth likes mangos and pancakes."

```
git fetch
git reset --hard
git checkout ml-demo-step2
```

step2: query db for context

rag.js

retrieveContext

```
collection.query({...
    queryEmbeddings,
    nResults
```

```
git fetch
git reset --hard
git checkout ml-demo-step3
```

step3: (3.1) formulate full prompt and query LLM. (3.2) produce response

```
git fetch
git reset --hard
git checkout ml-demo-step3
```

(3.1) formulate full prompt and query LLM

```
messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers won the World Series in 2020."},
    {"role": "user", "content": "Where was it played?"}
]
```

src

step3: (3.1) formulate full prompt and query LLM. (3.2) produce response

```
git fetch
git reset --hard
git checkout ml-demo-step3
```

(3.2) produce a response!
implement function
retrievalAugmentedGeneration

```
git fetch
git reset --hard
git checkout ml-demo-complete
```

Deployment

how to deploy?

Takeaway

Rewind

perplexity

Modal

Pydantic

lumona

Kay

Sources

- What is Machine Learning
- Introduction to Large Language Models
- Vector Similarity Demo