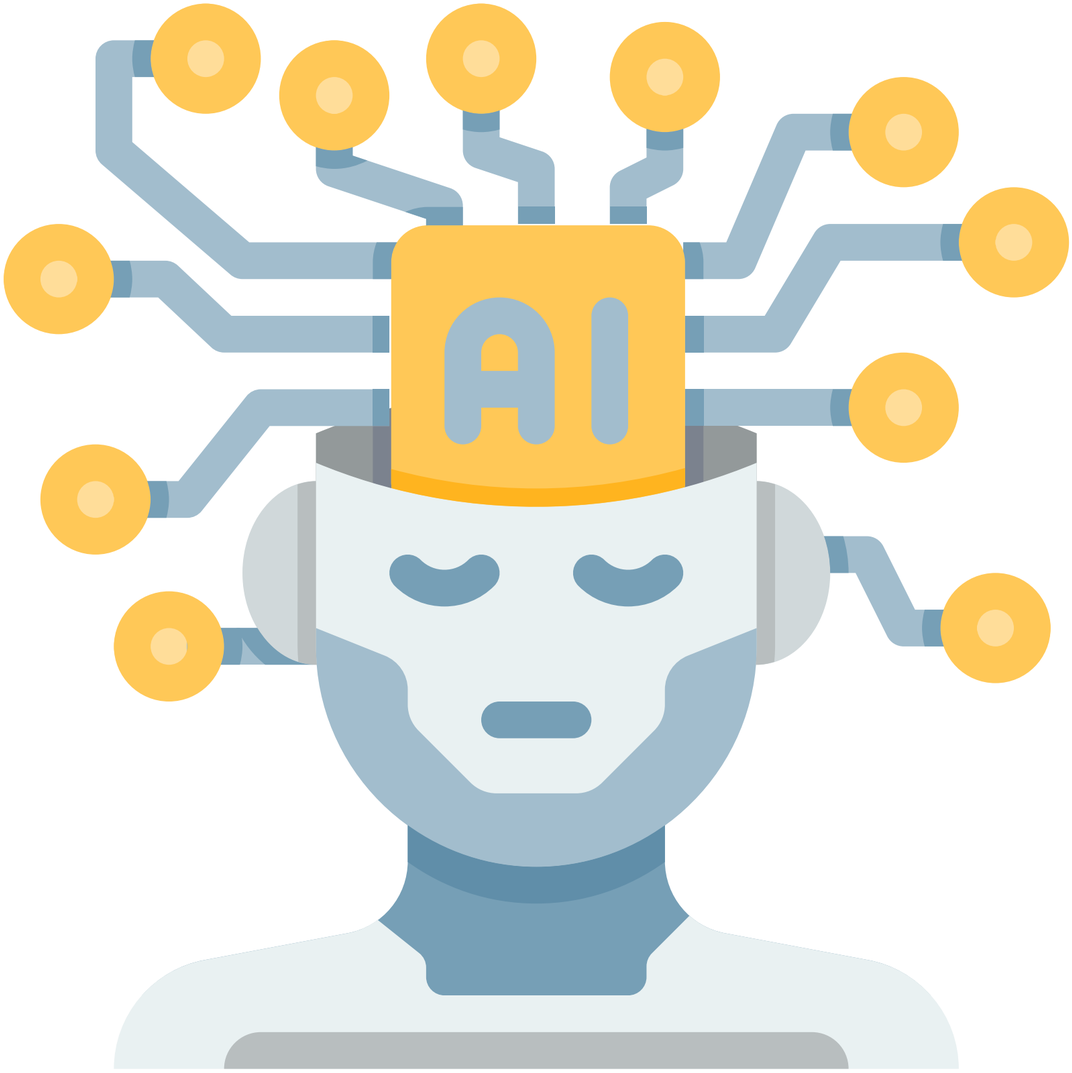
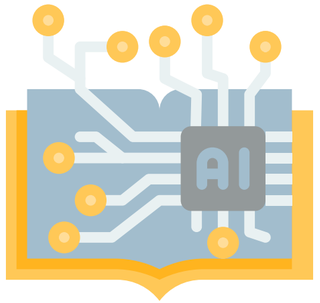
Retrieval Augmented Generation

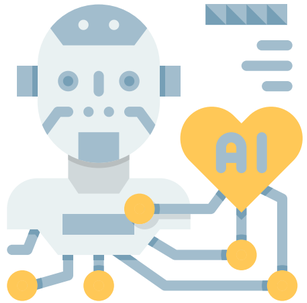


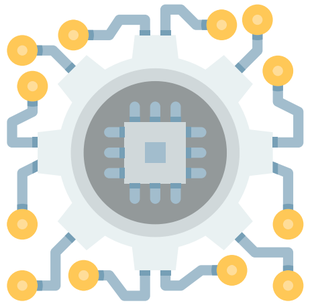
Who am I?

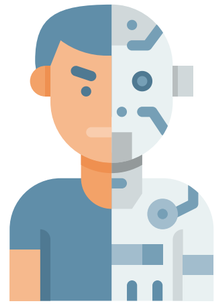
Gardner Bickford is a seasoned software developer and engineering leader with a career spanning continents and decades. From an early start at Borland to impactful roles at Adobe and Canva, Gardner has combined deep technical expertise with a passion for social innovation. His global journey; from coding collectives in New Zealand to sailing humanitarian missions in the Pacific; reflects a lifelong commitment to learning, adventure, and building technology that makes a difference.

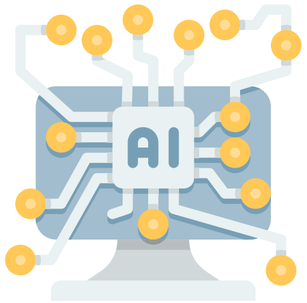
What to understand

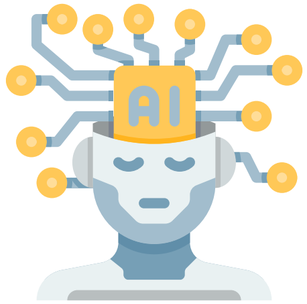


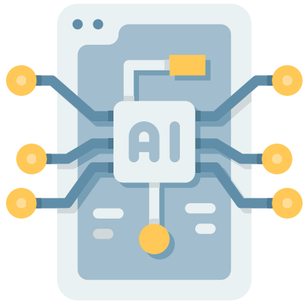


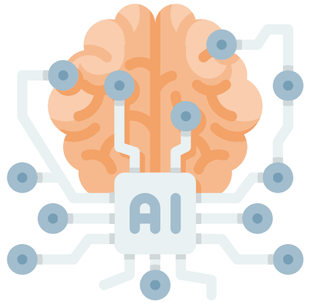


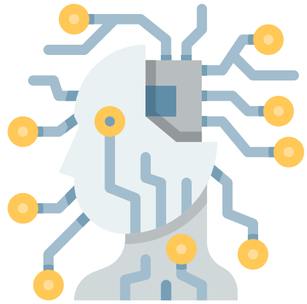


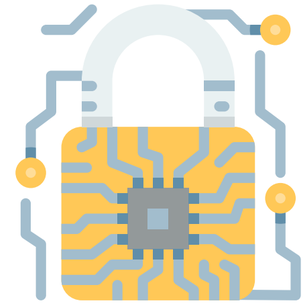


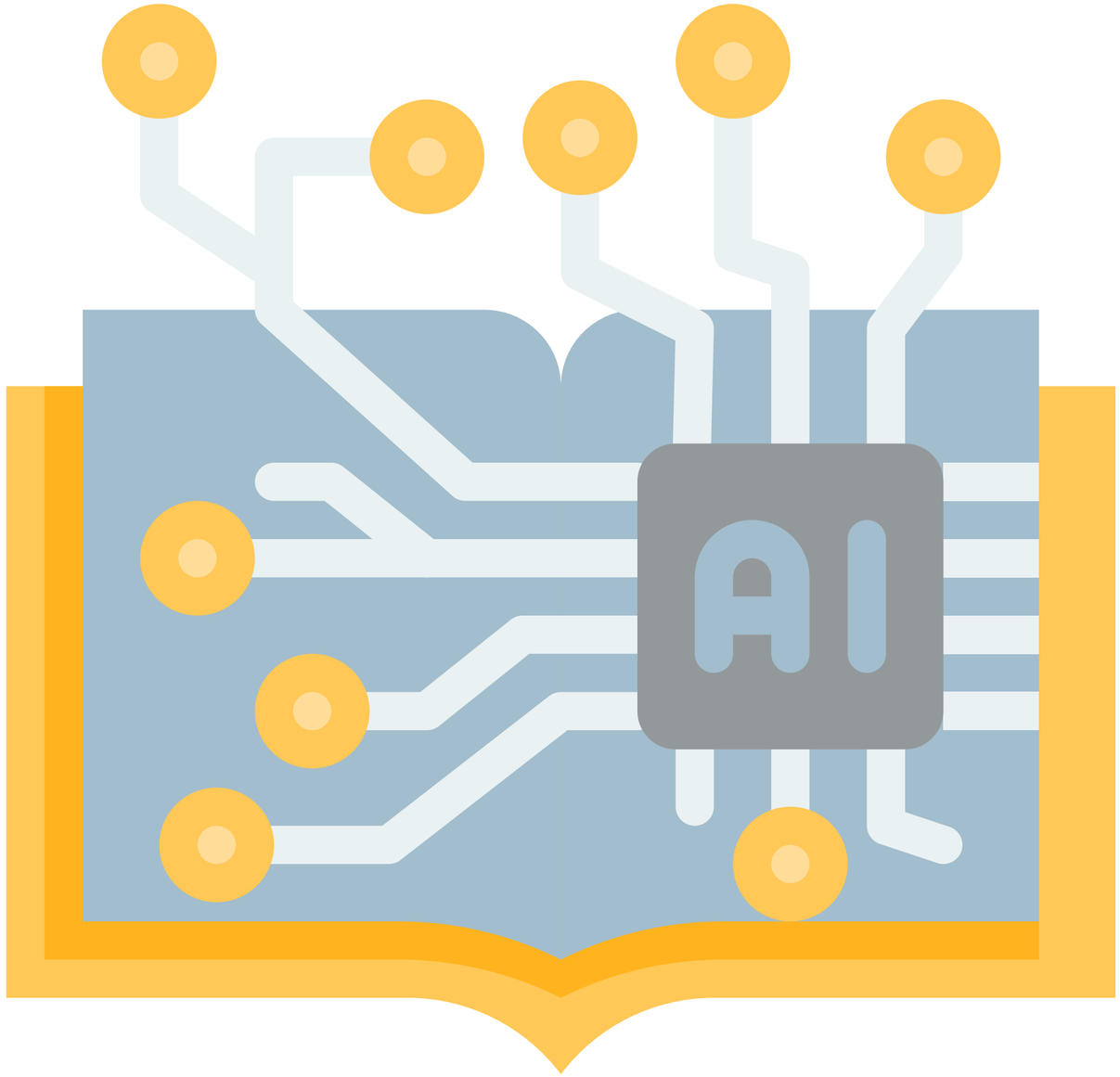












How does ChatGPT work?

Anyone wanna take a stab at explaining it?

Tokens


T omato e s are one of the most popular plants for vegetable gardens .
Ti p for success : If you select varieties that are resistant to disease and
pest s , growing tomatoes can be quite easy . For experienced garden ers
looking for a challenge , there are endless heir loom and specialty
varieties to cultiv ate . T omato plants come in a range of sizes .

What is ChatGPT

and why does it work?

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

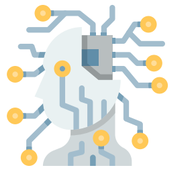
https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

What are AI hallucinations?

Hallucinations

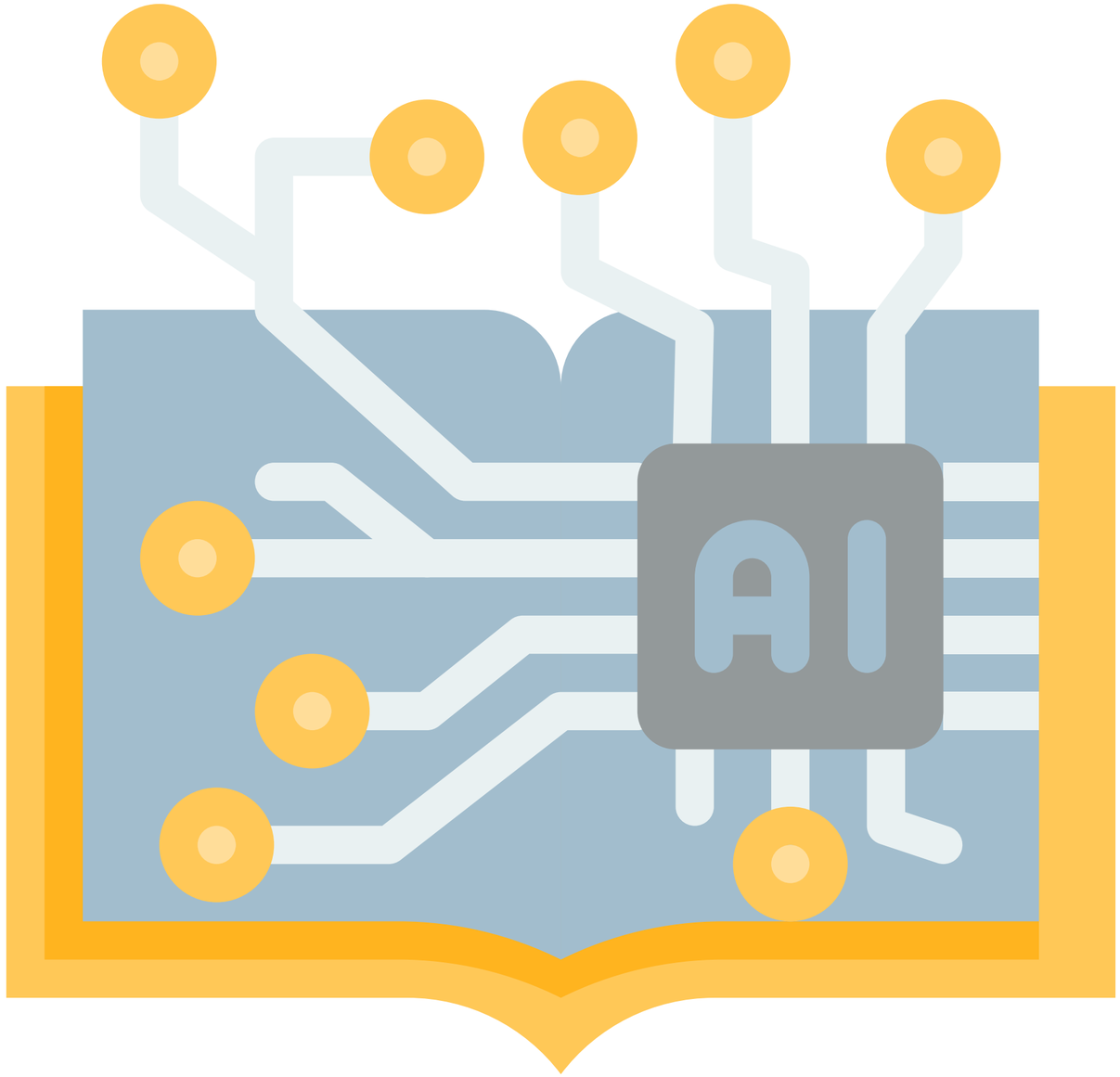
Caused largely by “Next Token”

AI hallucination is a phenomenon where, in a large language model (LLM) often a generative AI chatbot or computer vision tool, perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.



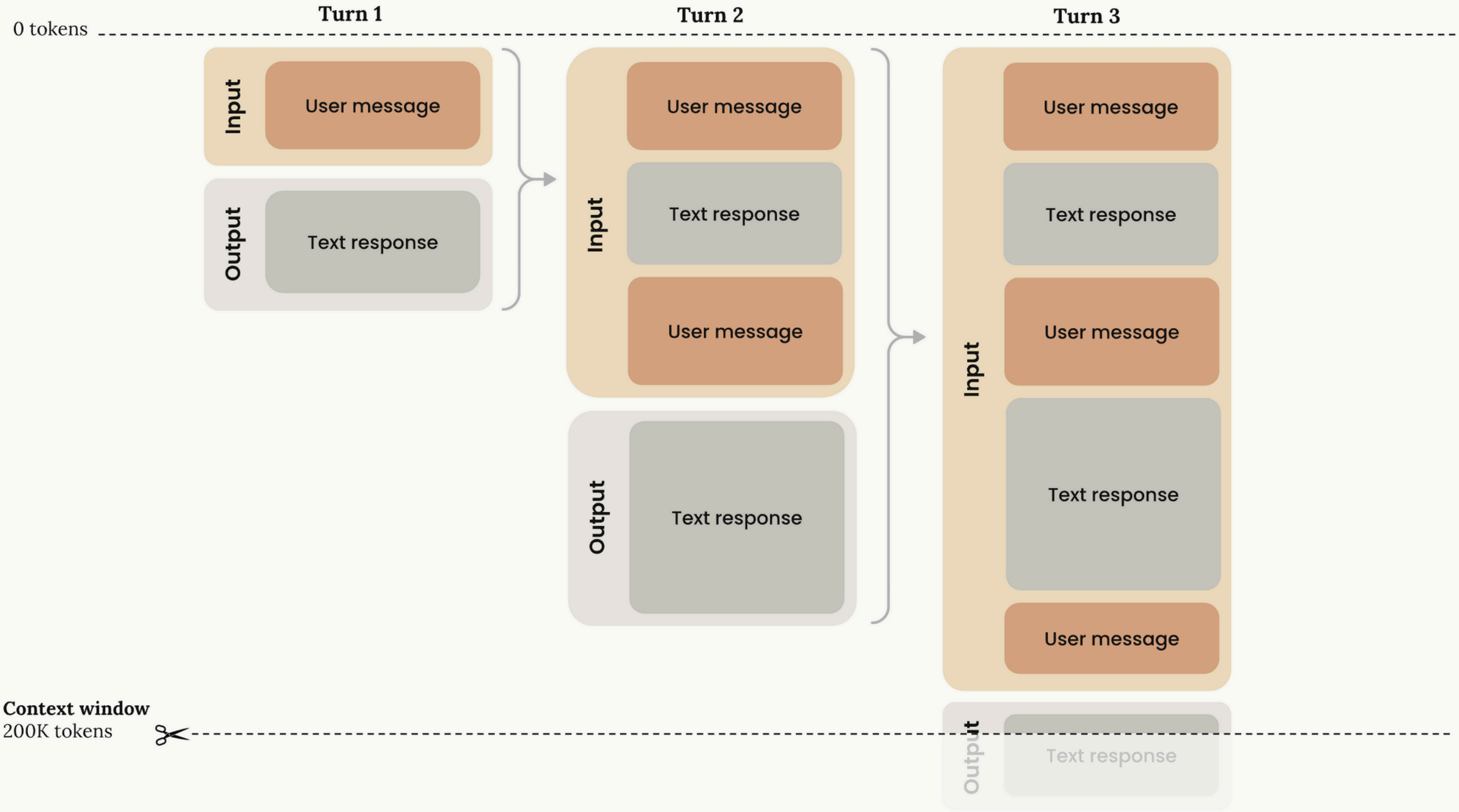
How to trick a model

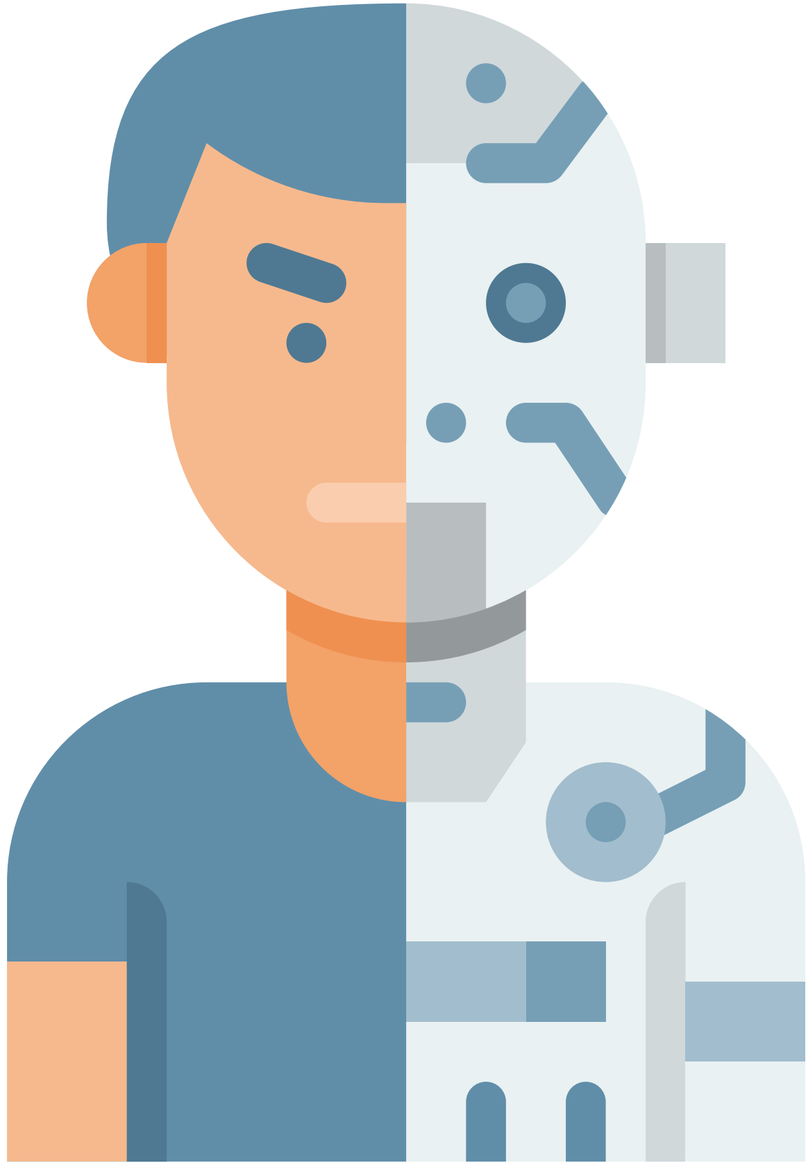
1. Turn off web search
2. Ask it to describe something that doesn't exist



What's a context window?

Anyone wanna take a stab at explaining it?


0 tokens
Turn 1
Turn 2
Turn 3
Input
User message
Output
Text response
Input
User message
Text response
User message
Output
Text response
Input
User message
Text response
User message
Text response
User message
Output
Text response
Context window
200K tokens

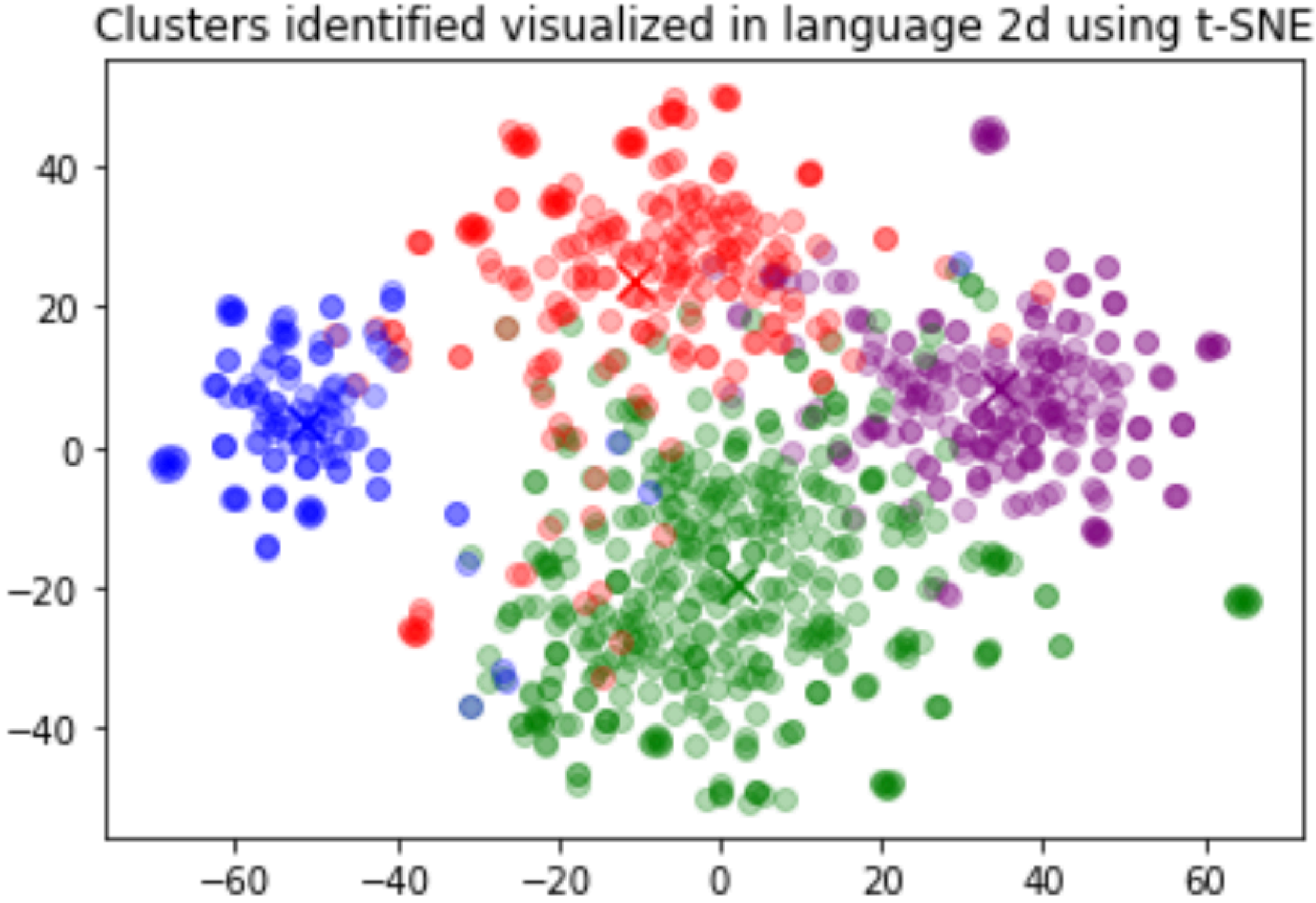


Embeddings

Are numerical representations of how an LLMs interpret **semantic meaning**.

Every time you enter text into an LLM, the first thing it does is break up the text into tokens, using the tokenizer.

Then it converts the tokens to numbers using the special layers which act as a lookup table created during pre-training.


Clusters identified visualized in language 2d using t-SNE
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

Example: Getting embeddings

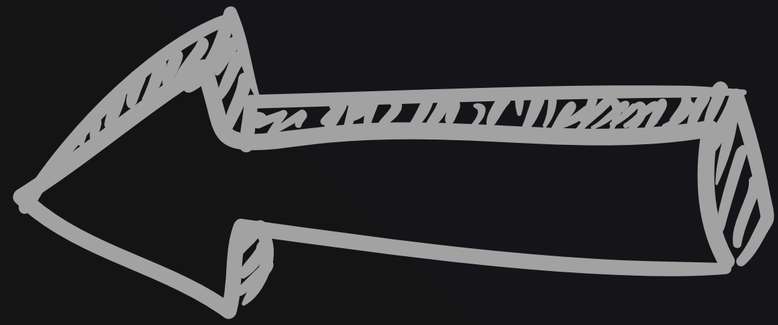
```
import OpenAI from "openai";
const openai = new OpenAI();

const embedding = await openai.embeddings.create({
  model: "text-embedding-3-small",
  input: "Your text string goes here",
  encoding_format: "float",
});

console.log(embedding);
```

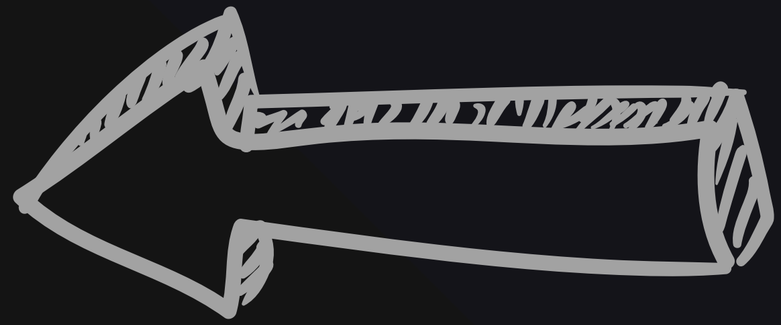
```
{
  "object": "list",
  "data": [
    {
      "object": "embedding",
      "index": 0,
      "embedding": [
        -0.006929283495992422,
        -0.0053364220075309 28,
        -4.547132266452536e-05,
        -0.02404750511050224 3
      ],
    }
  ],
  "model": "text-embedding-3-small",
  "usage": {
    "prompt_tokens": 5,
    "total_tokens": 5
  }
}
```

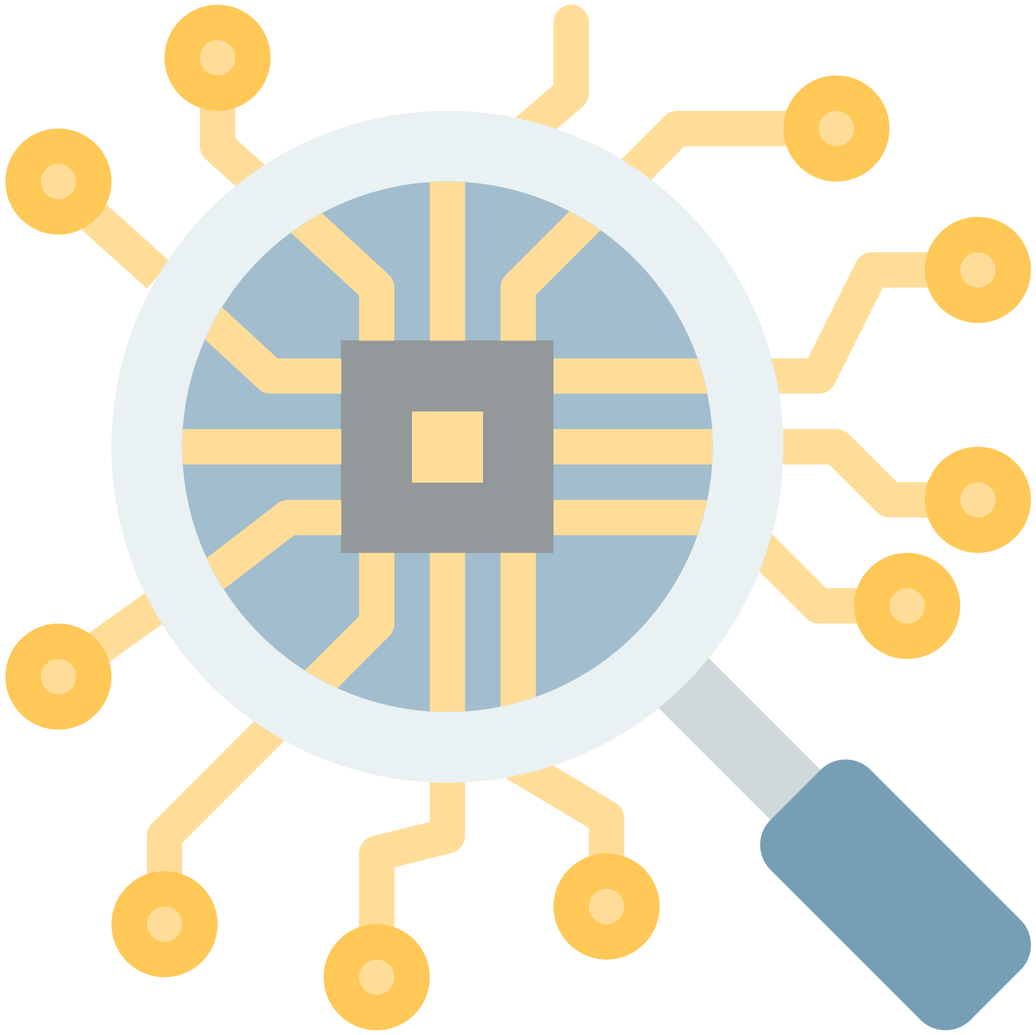

Vector Embeddings



```
{
  "object": "list",
  "data": [
    {
      "object": "embedding",
      "index": 0,
      "embedding": [
        -0.006929283495992422,
        -0.0053364220075309 28,
        -4.547132266452536e-05,
        -0.02404750511050224 3
      ],
    }
  ],
  "model": "text-embedding-3-small",
  "usage": {
    "prompt_tokens": 5,
    "total_tokens": 5
  }
}
```


Where do they come from?

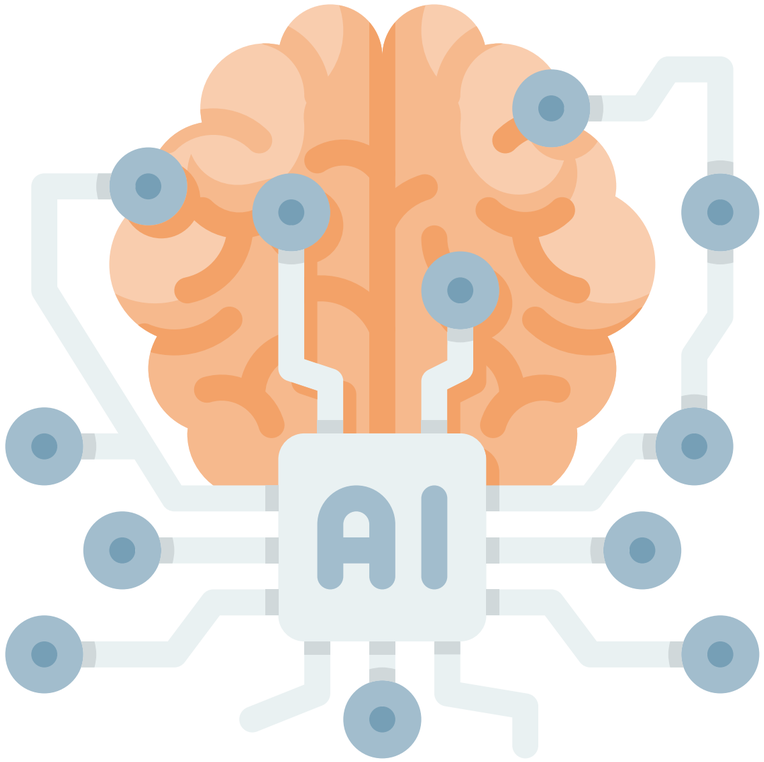




Where do embeddings come from?

Embeddings are generated from machine learning models that learn to represent data (like text, images, or audio) as numerical vectors. These vectors, known as embeddings, capture meaningful relationships and characteristics of the data, allowing for tasks like similarity searches and pattern recognition.

What problem does RAG solve?



What problem does RAG solve?

Context window limitations

What problem does RAG solve?

What problem does RAG solve?

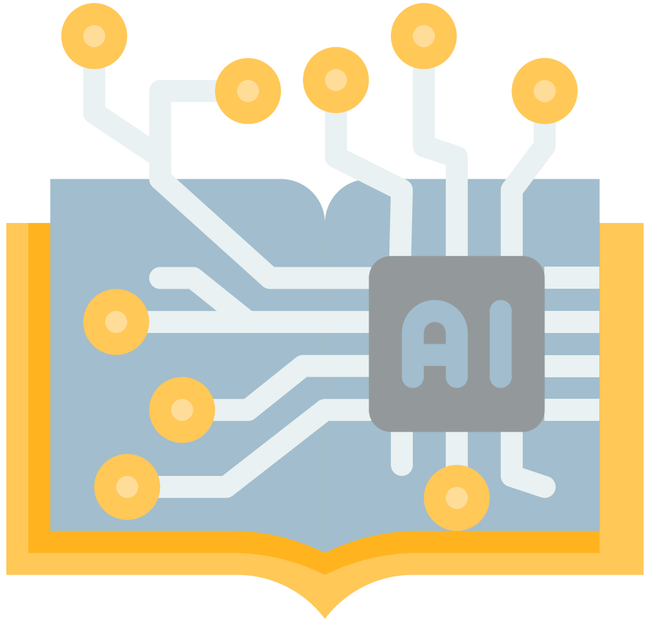
hallucinations

by “grounding” the model in truth

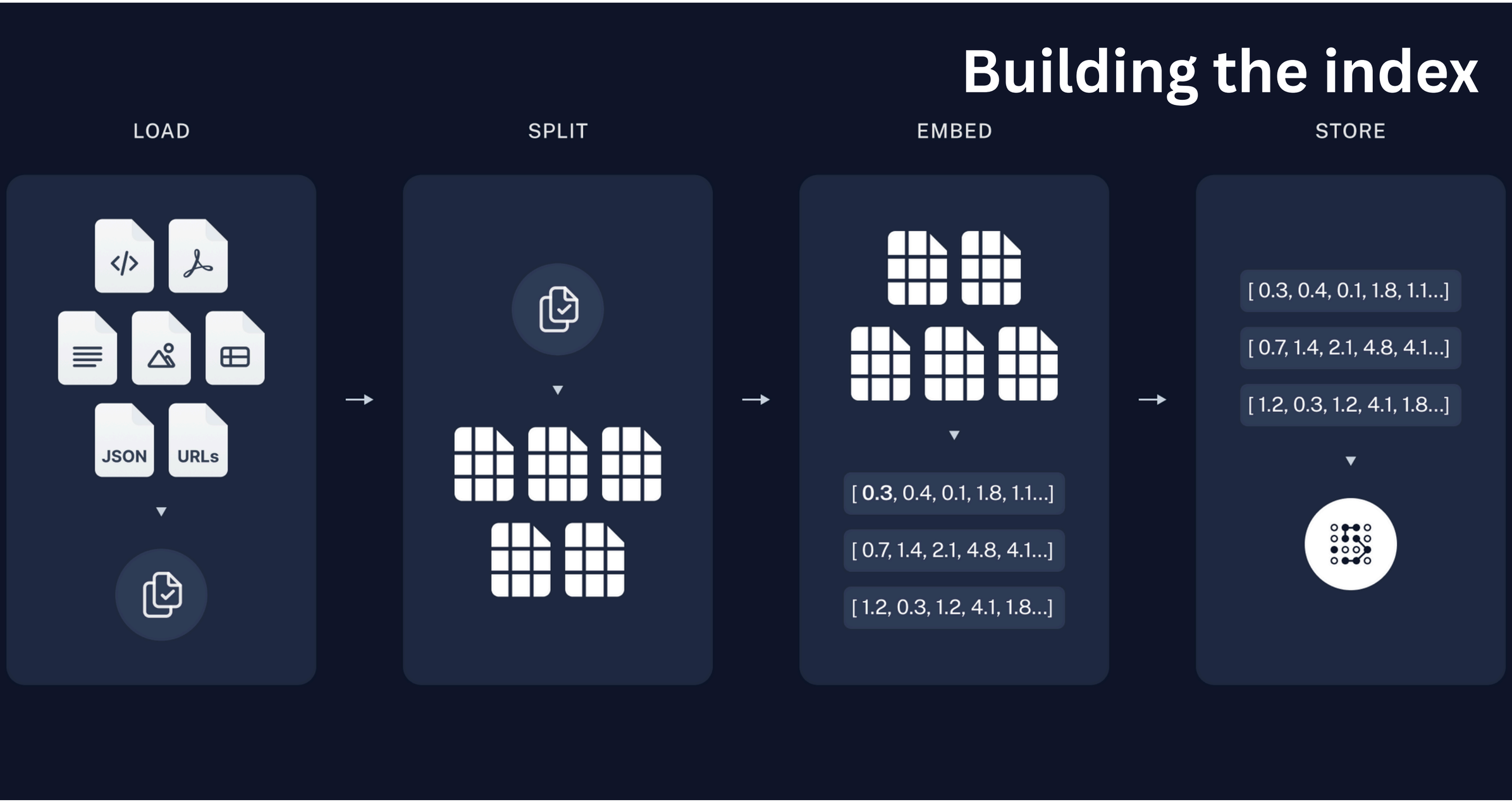
What problem does RAG solve?

Up-to-date “facts”

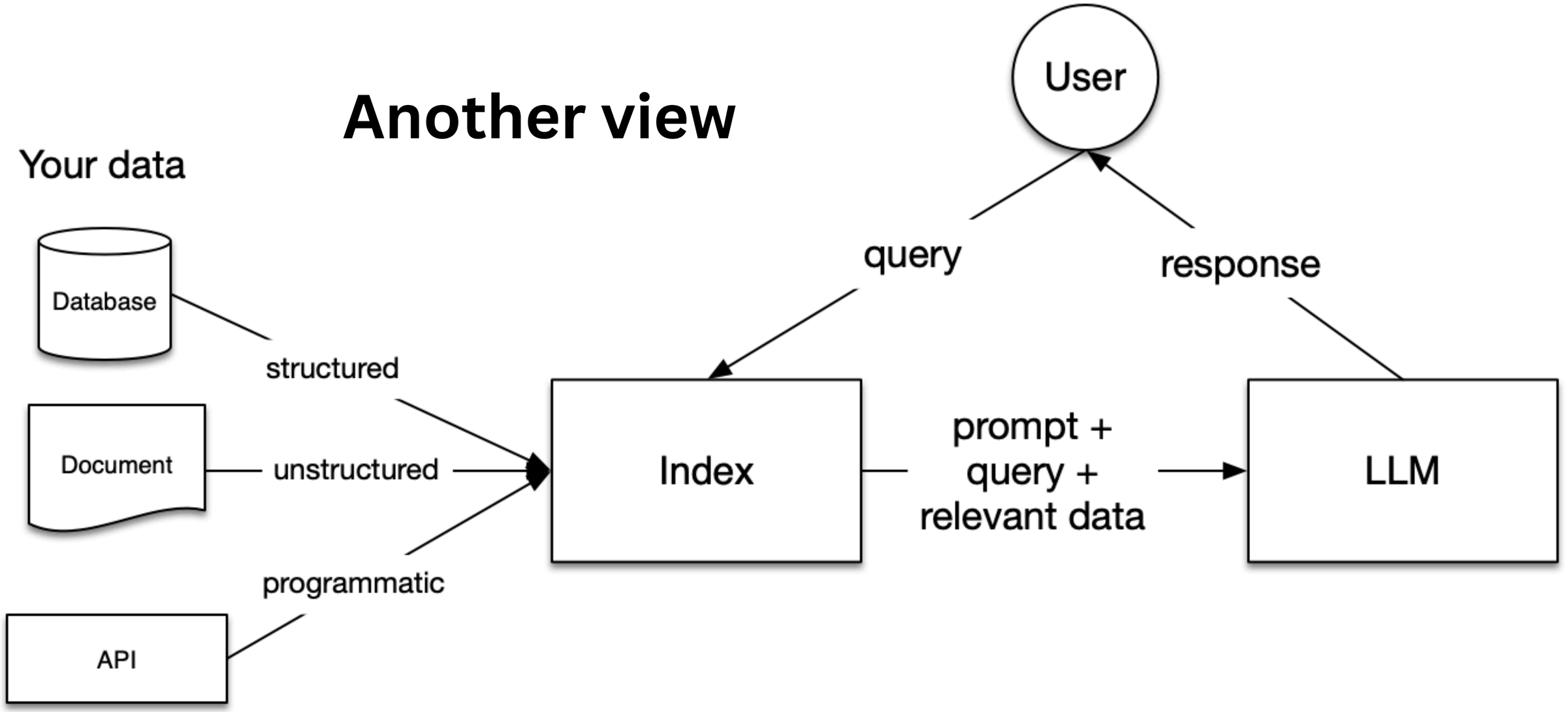
How does it work?



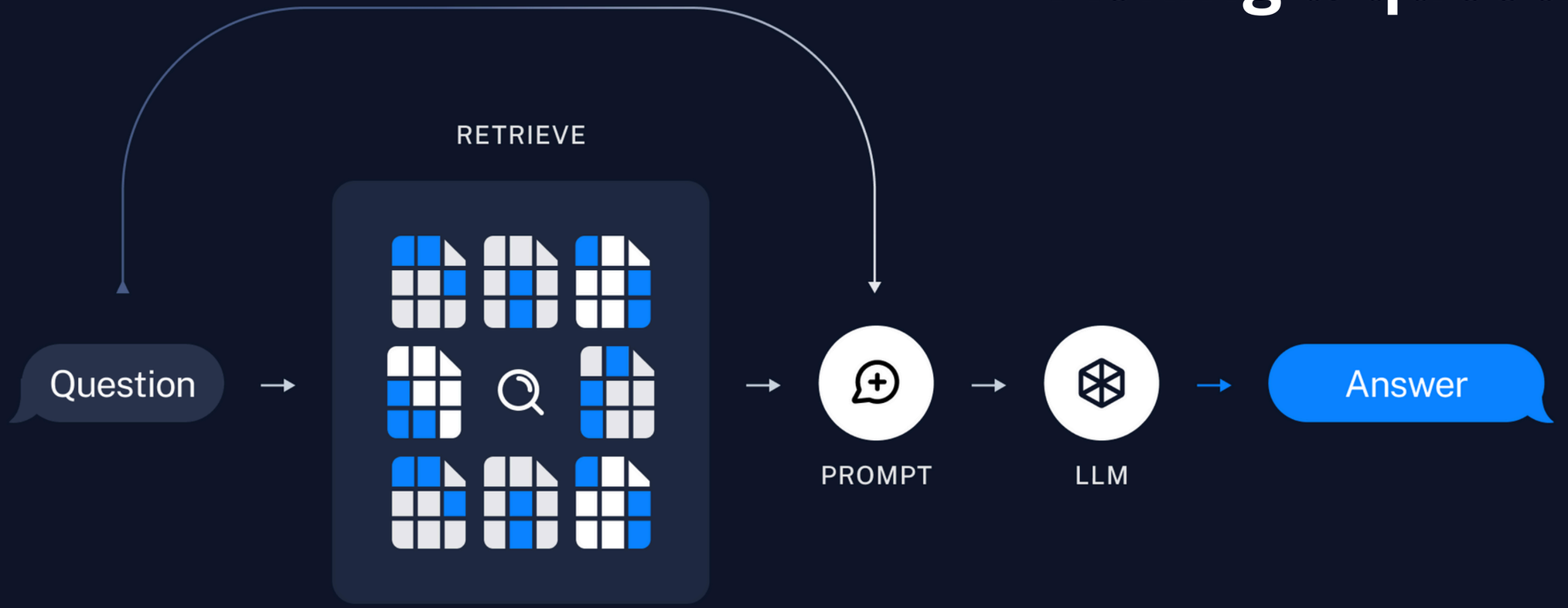
Building the index



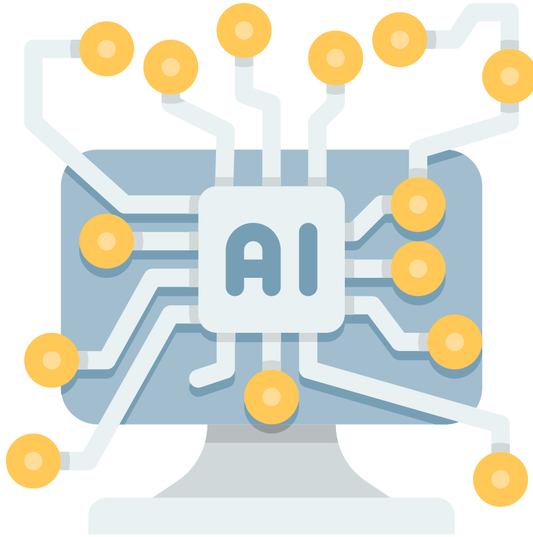
Another view



Asking a question



Let’s code!!!



github.com/gardner/nais