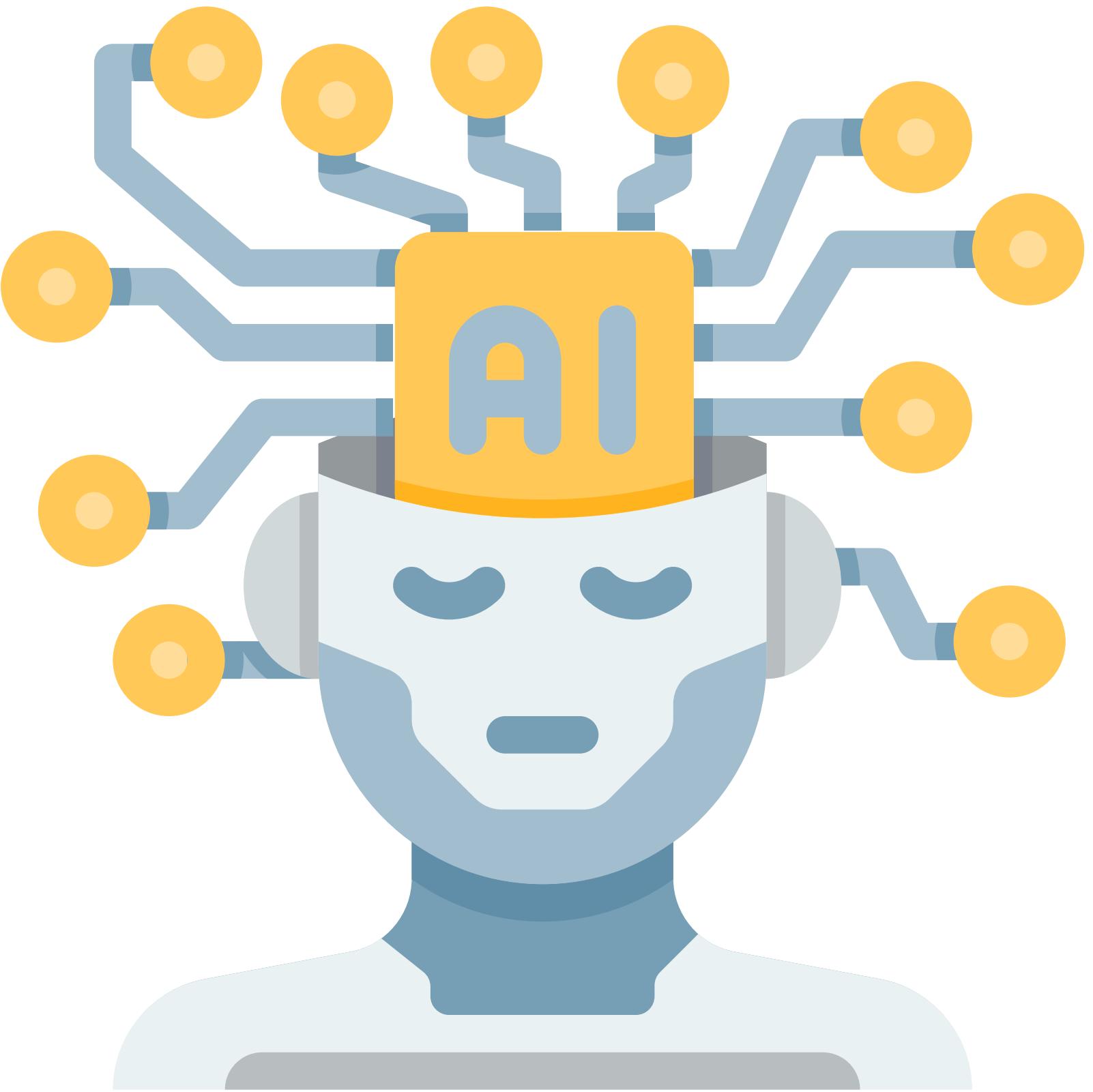
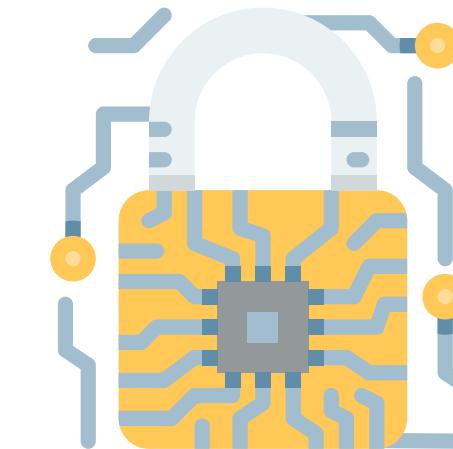
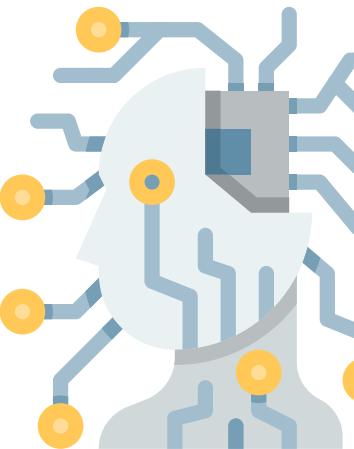
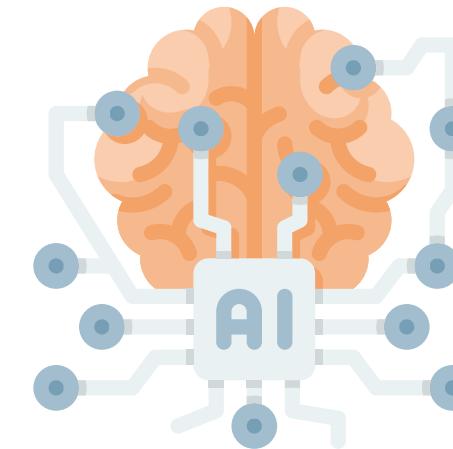
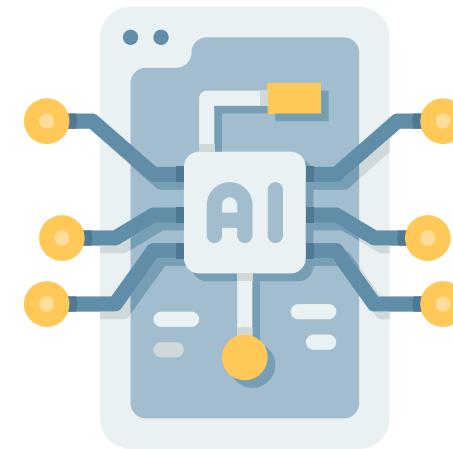
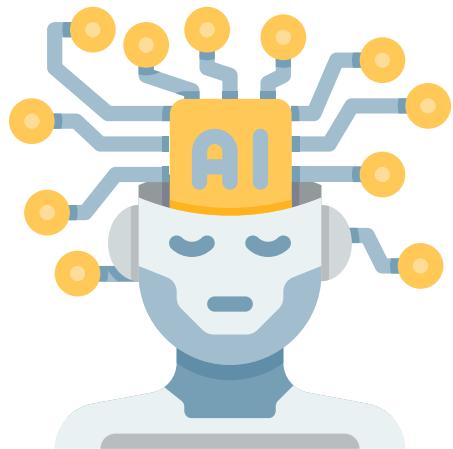
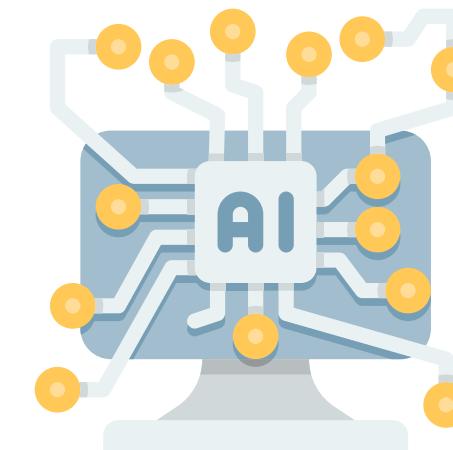
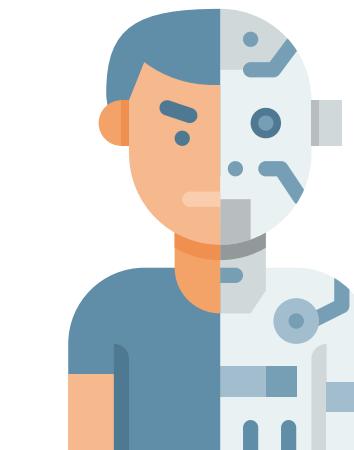
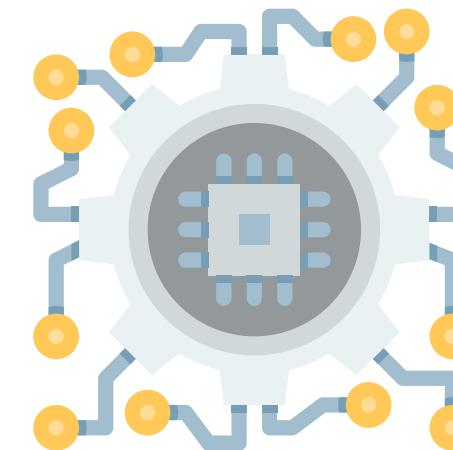
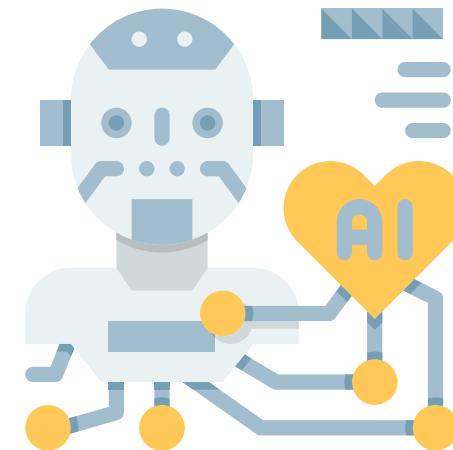
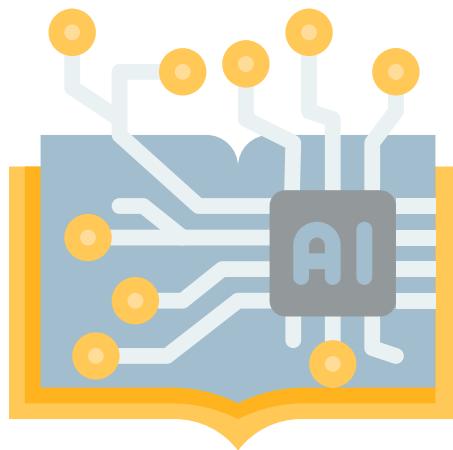
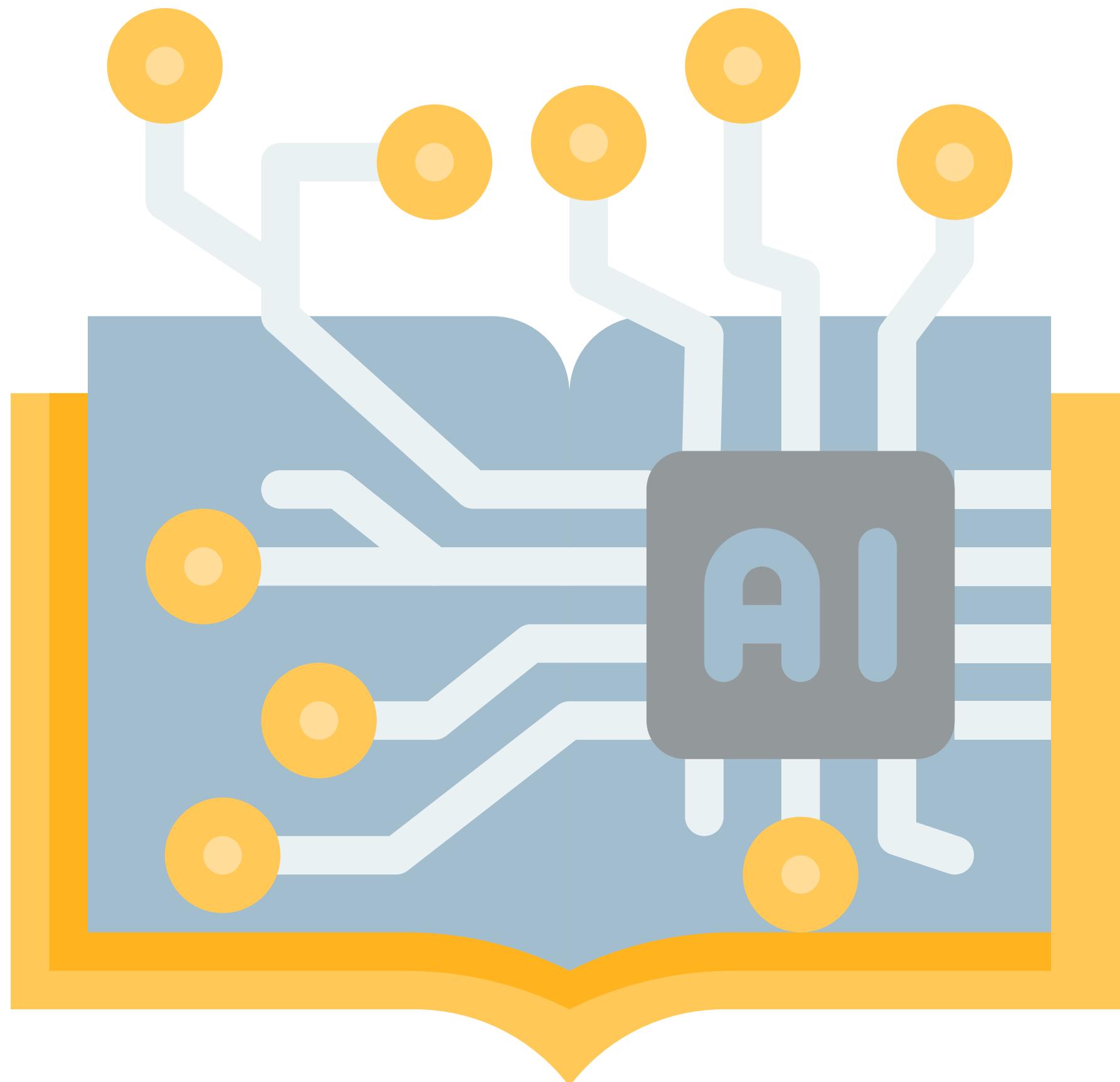


Retrieval Augmented Generation



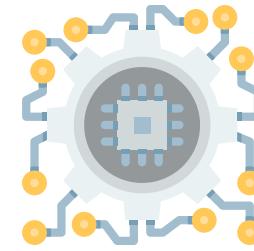
What to understand



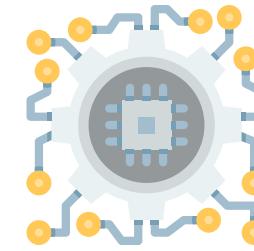


How does ChatGPT work?

Anyone wanna take a stab at
explaining it?

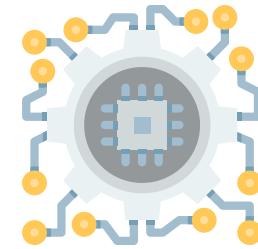


Tokens

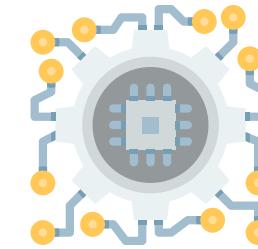


Tomatoes are one of the most popular plants for vegetable gardens.

Tip for success : If you select varieties that are resistant to disease and pests, growing tomatoes can be quite easy. For experienced gardeners looking for a challenge, there are endless heirloom and specialty varieties to cultivate. Tomato 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%



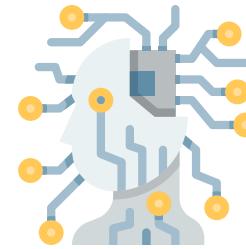
What are AI hallucinations?



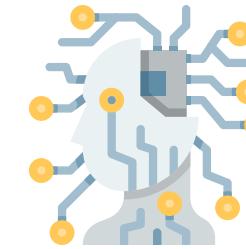
Hallucinations

Caused largely by “Next Token”

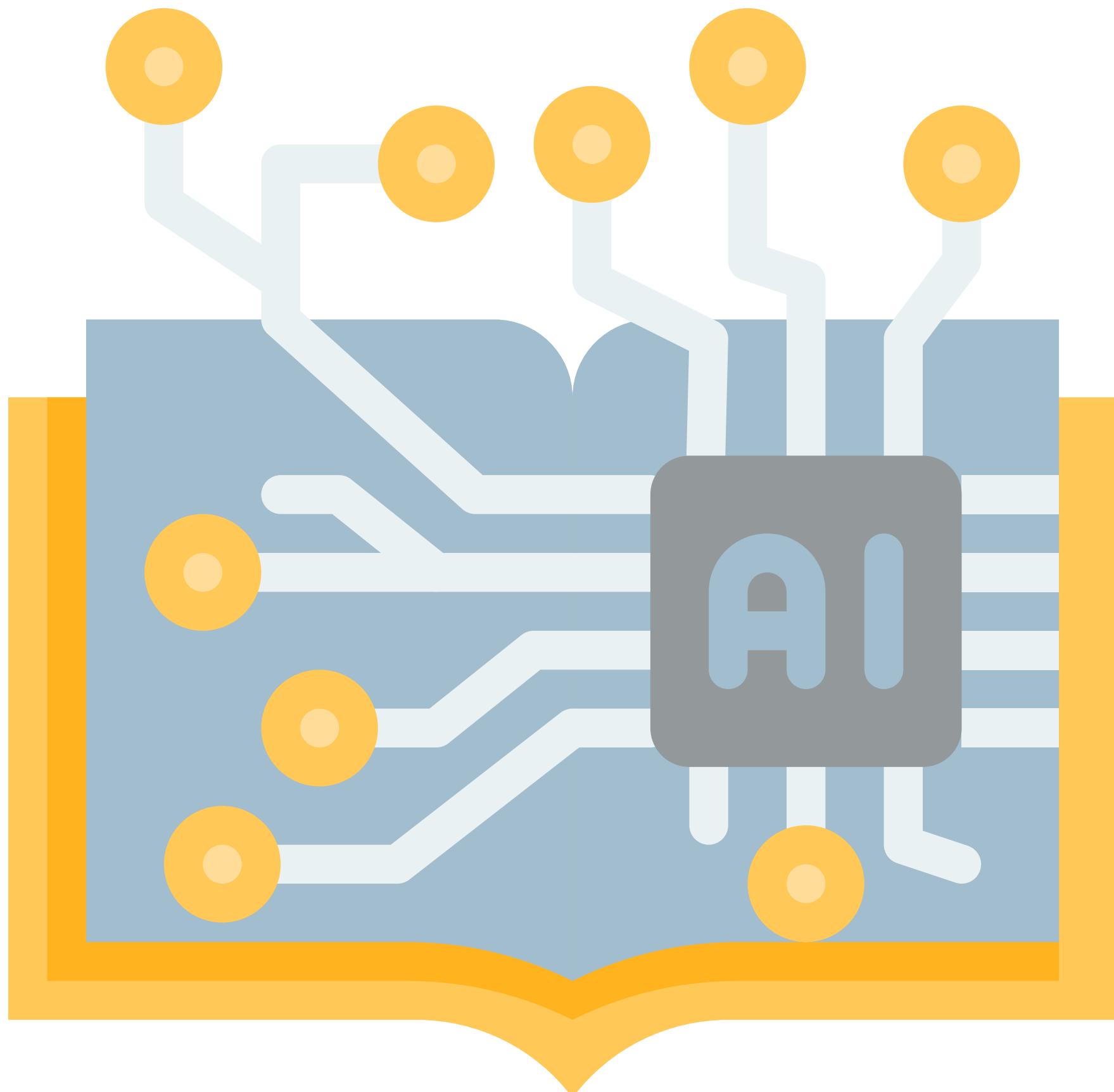
AI hallucination is a phenomenon where, in a large language model (LLM) often a generative AI chatbot or computer vision tool, perceives patterns or objects that are nonexistent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate.



How to trick a model

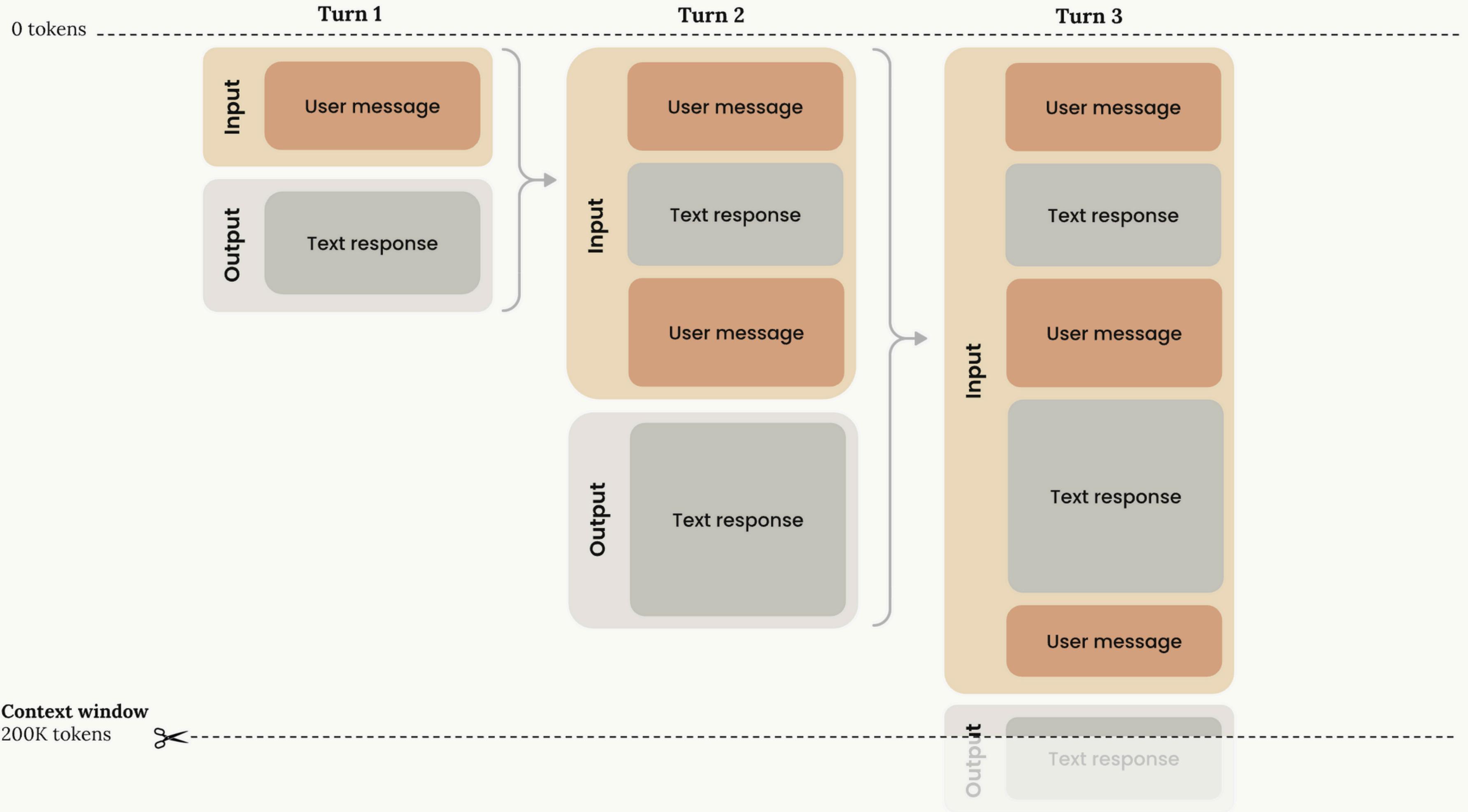


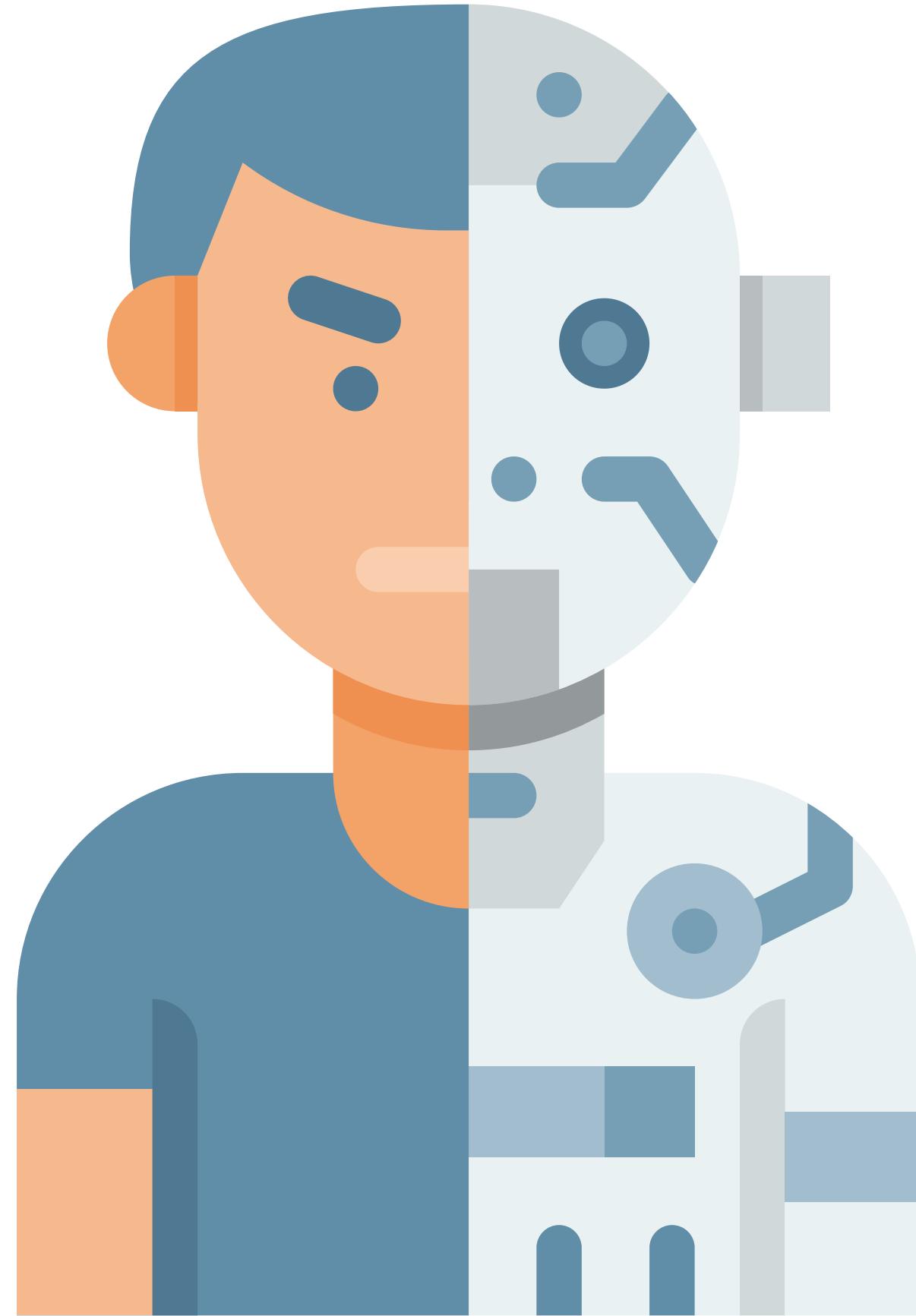
1. Turn off web search
2. Ask it to describe something that doesn't exist



What's a context window?

Anyone wanna take a stab at
explaining it?





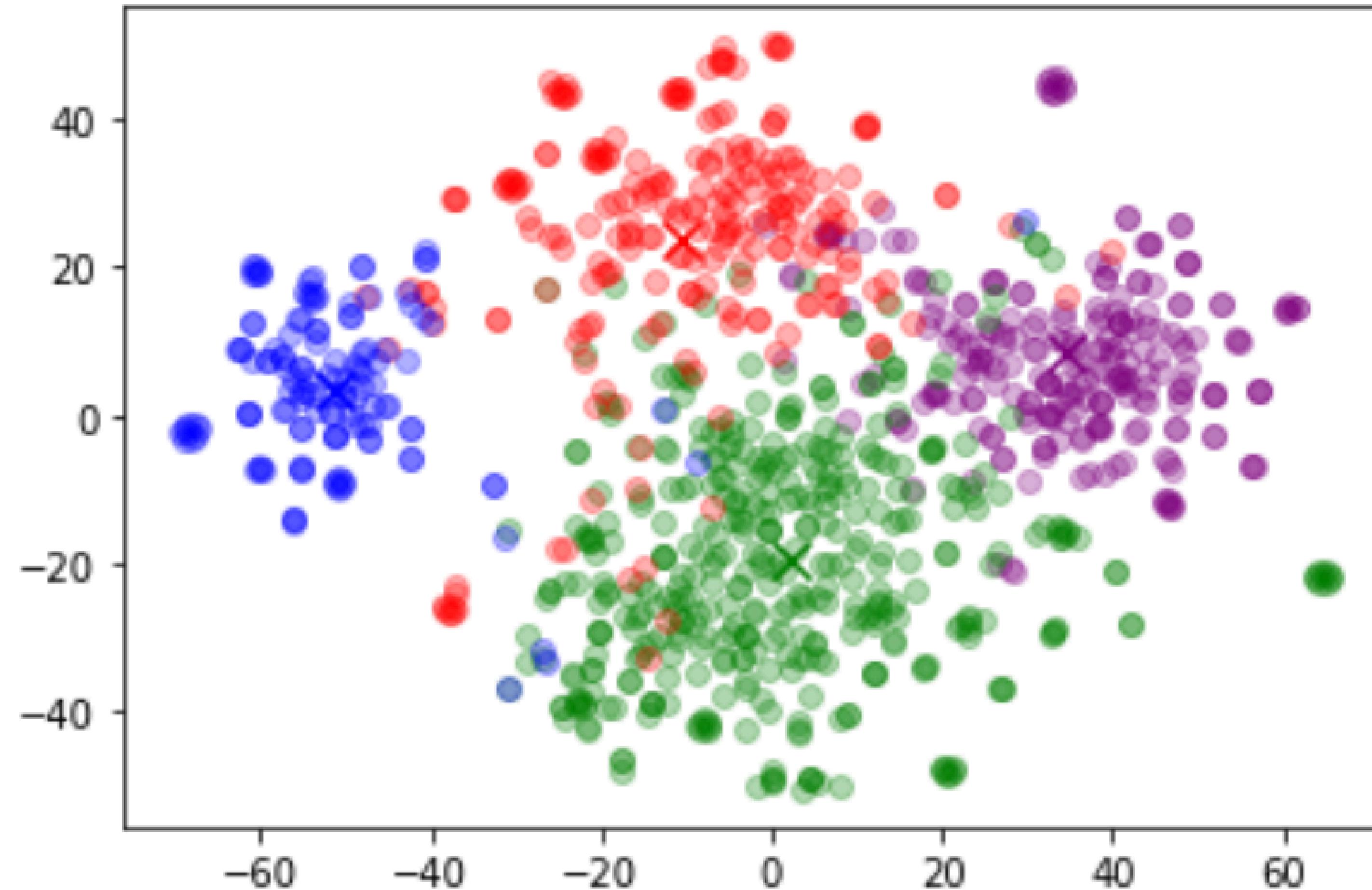
Embeddings

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



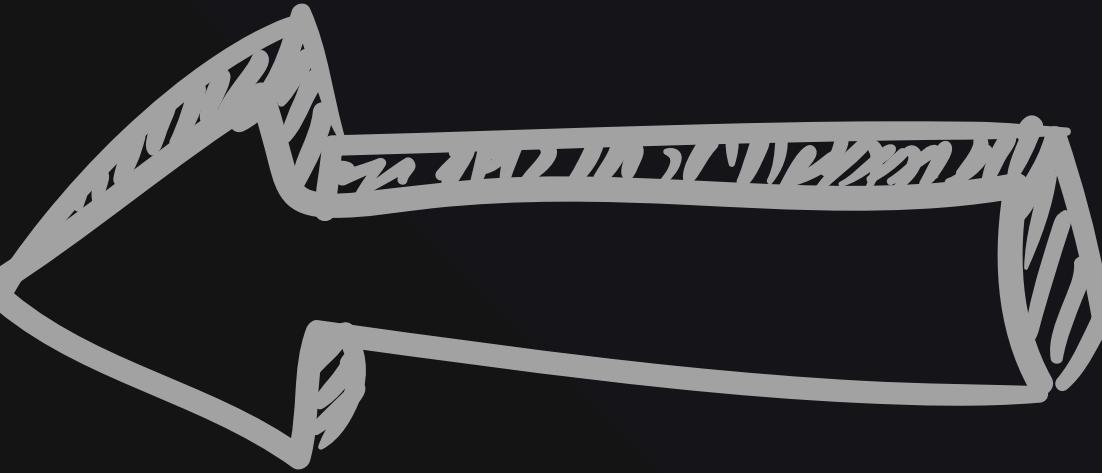
Example: Getting embeddings

```
1 import OpenAI from "openai";
2 const openai = new OpenAI();
3
4 const embedding = await openai.embeddings.create({
5   model: "text-embedding-3-small",
6   input: "Your text string goes here",
7   encoding_format: "float",
8 });
9
10 console.log(embedding);
```



```
1  {
2      "object": "list",
3      "data": [
4          {
5              "object": "embedding",
6              "index": 0,
7              "embedding": [
8                  -0.006929283495992422,
9                  -0.005336422007530928,
10                 -4.547132266452536e-05,
11                 -0.024047505110502243
12             ],
13         }
14     ],
15     "model": "text-embedding-3-small",
16     "usage": {
17         "prompt_tokens": 5,
18         "total_tokens": 5
19     }
20 }
```

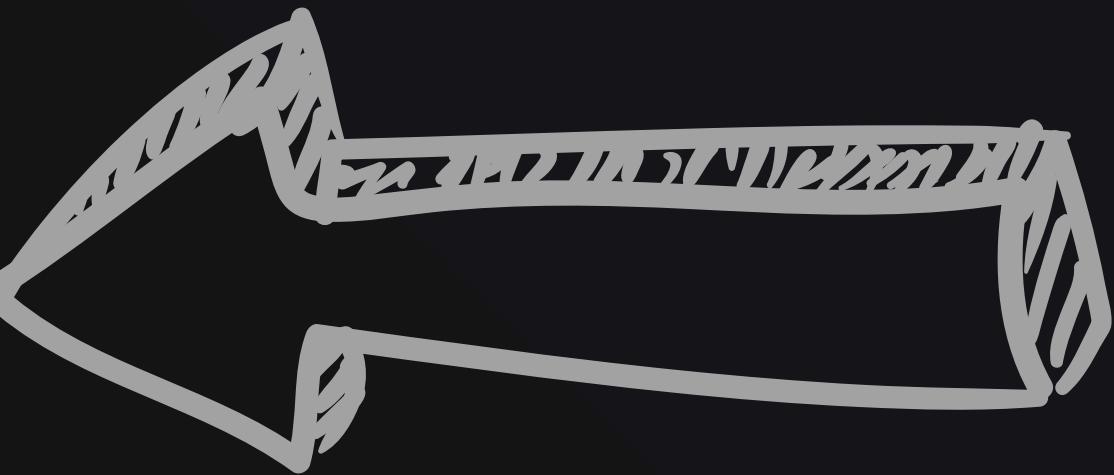
Vector Embeddings

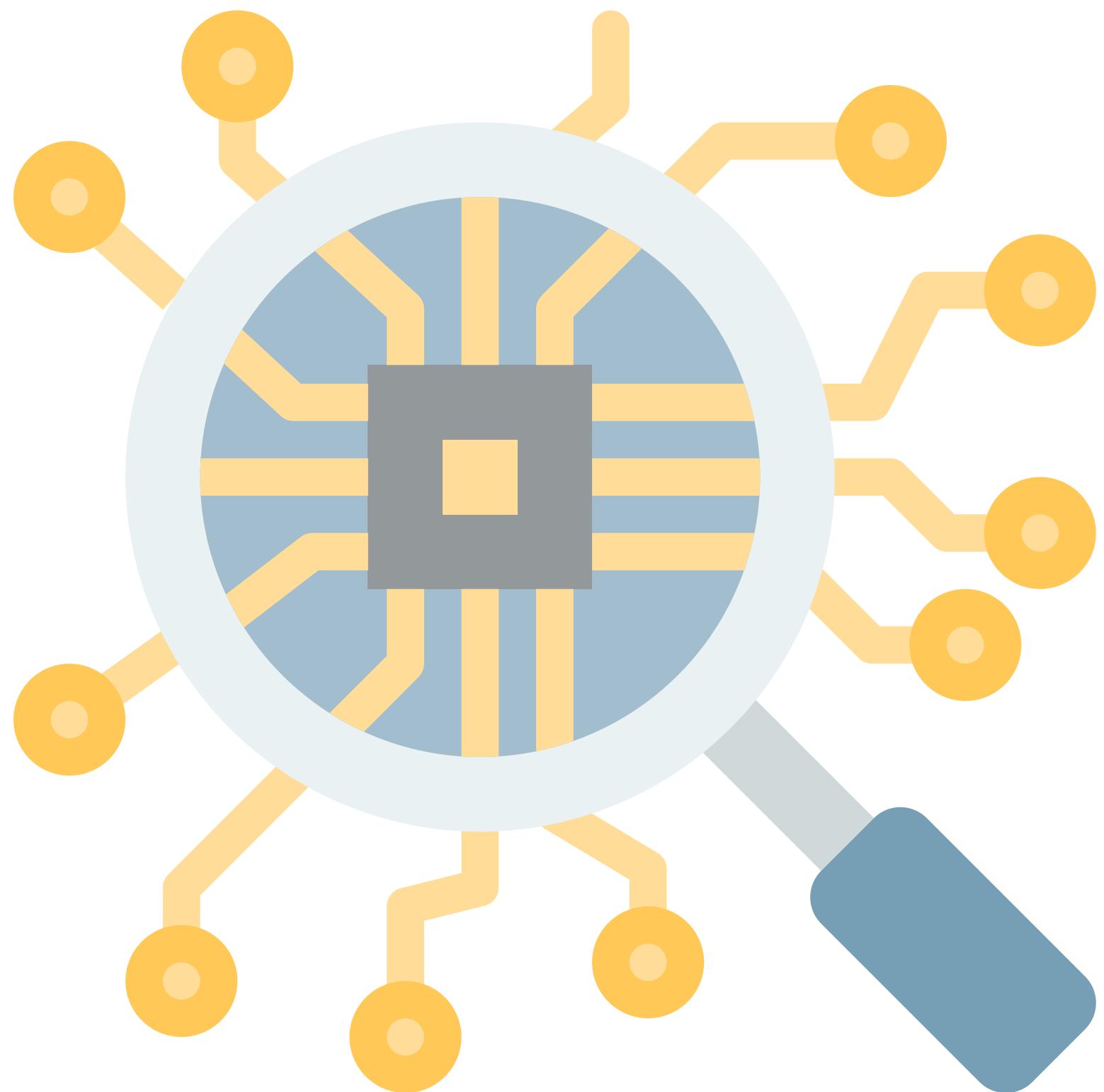




```
1  {
2      "object": "list",
3      "data": [
4          {
5              "object": "embedding",
6              "index": 0,
7              "embedding": [
8                  -0.006929283495992422,
9                  -0.005336422007530928,
10                 -4.547132266452536e-05,
11                 -0.024047505110502243
12             ],
13         }
14     ],
15     "model": "text-embedding-3-small",
16     "usage": {
17         "prompt_tokens": 5,
18         "total_tokens": 5
19     }
20 }
```

Where do they come from?

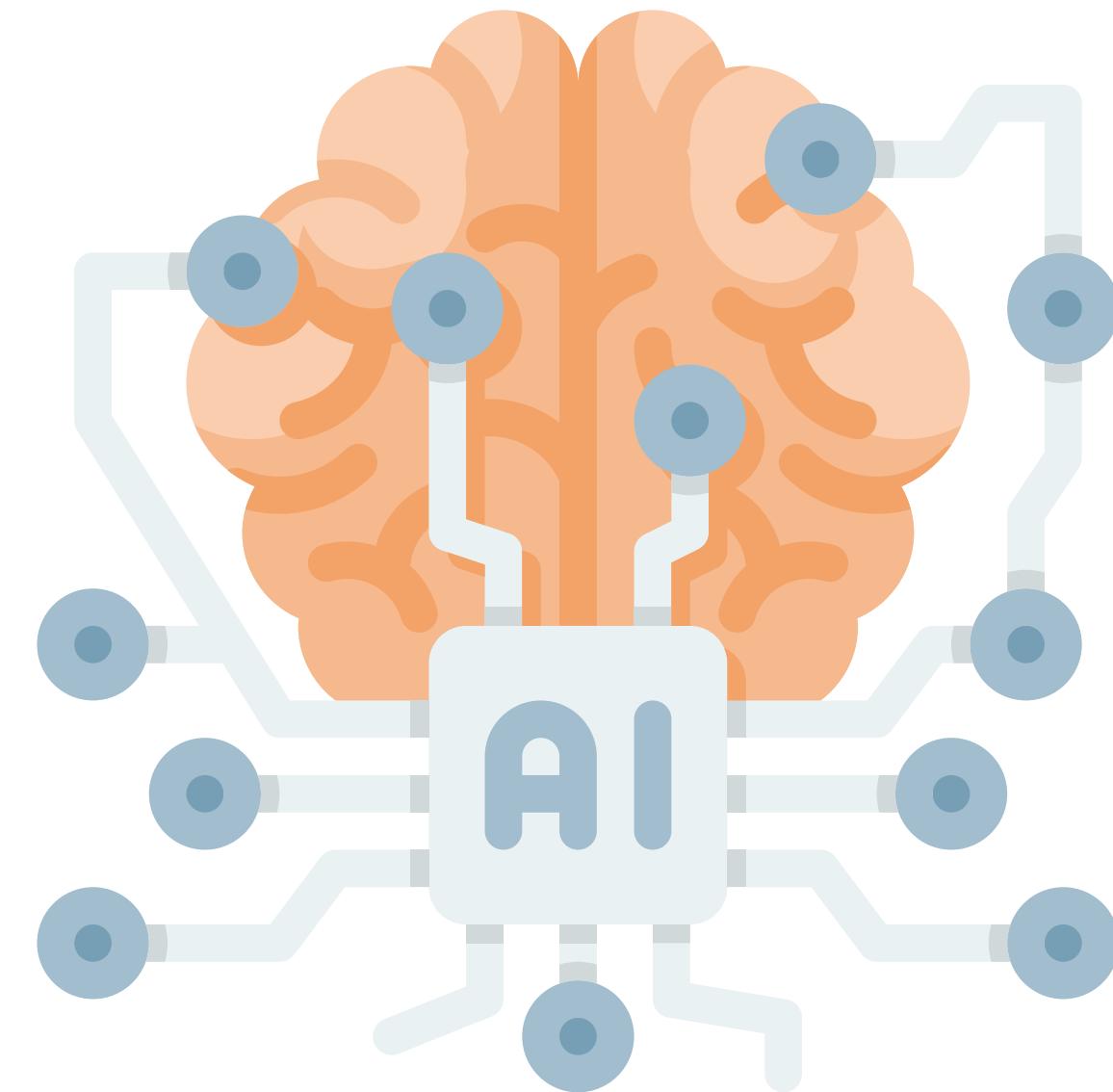




Where do embeddings come from?

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?



hallucinations

What problem does RAG solve?

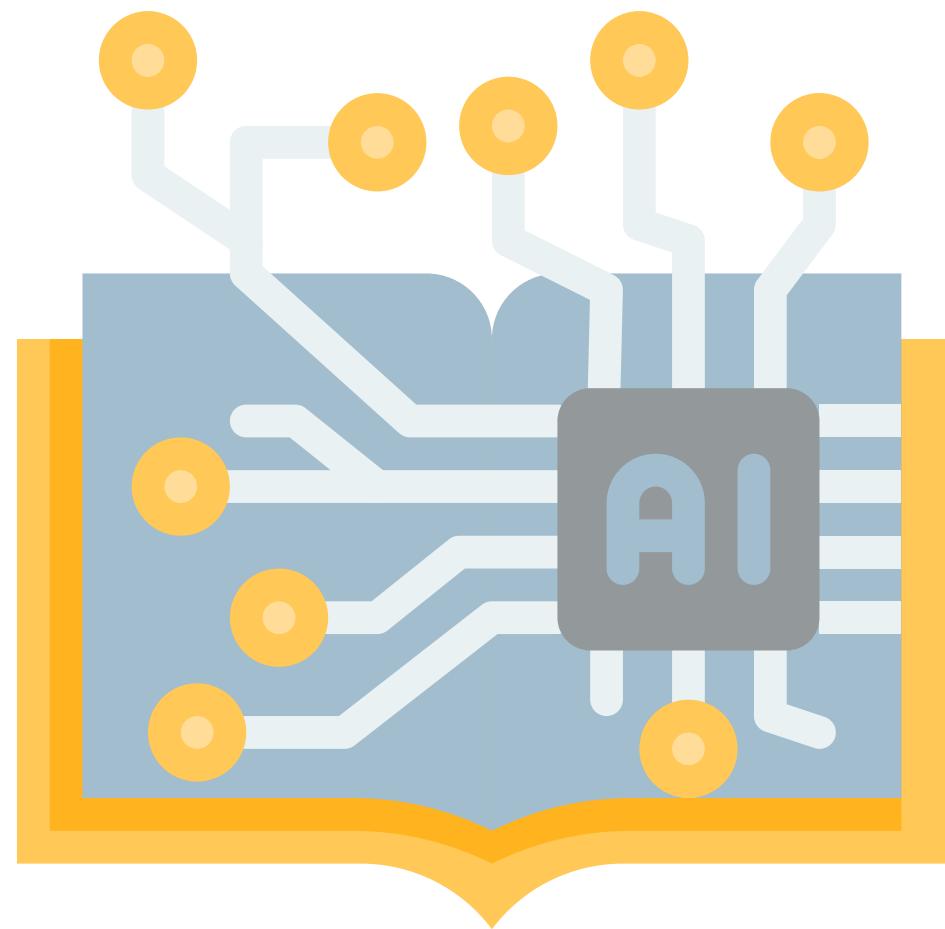
hallucinations

by “grounding” the model in truth

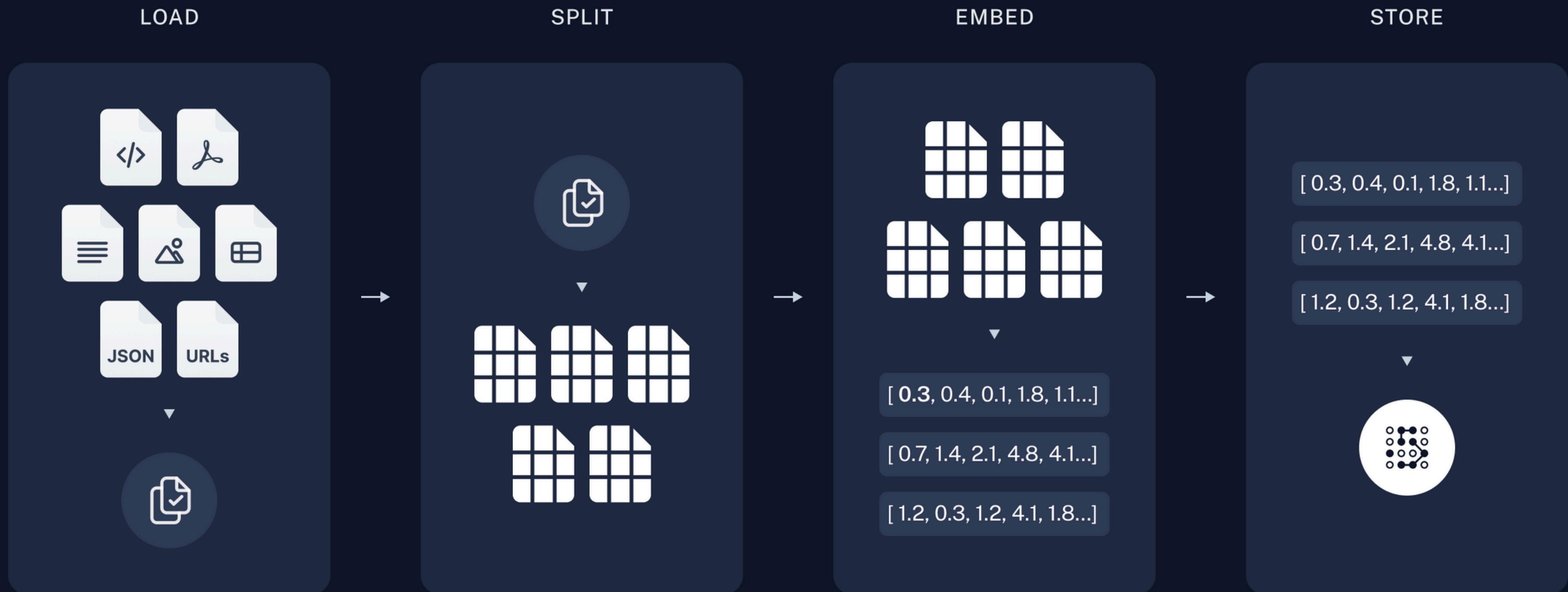
What problem does RAG solve?

Up-to-date “facts”

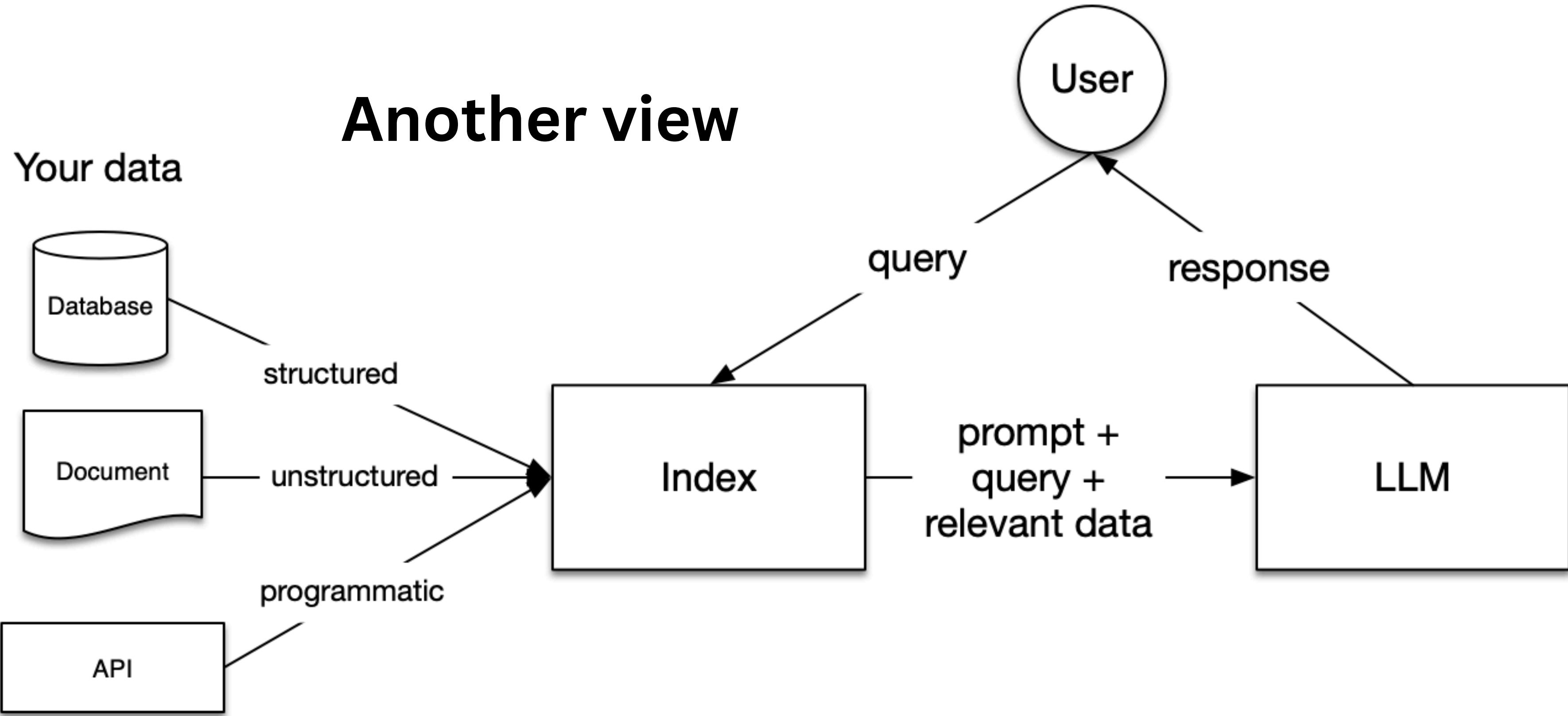
How does it work?



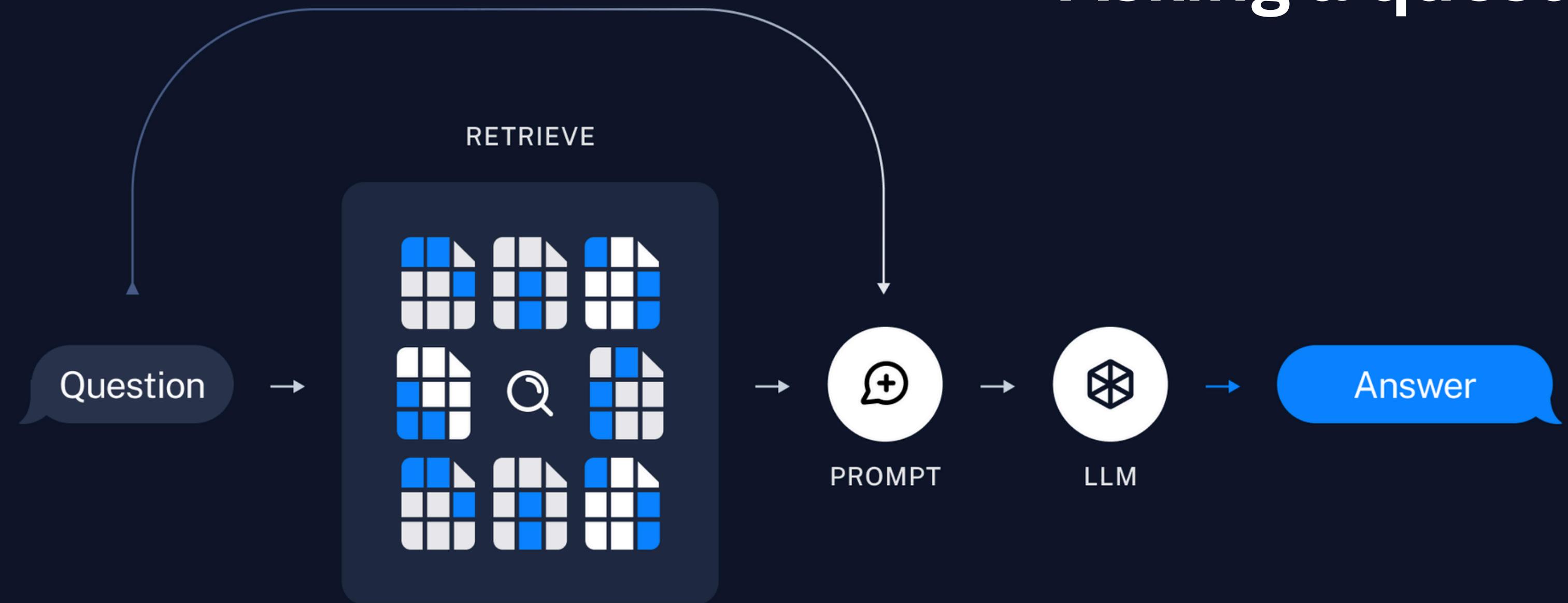
Building the index



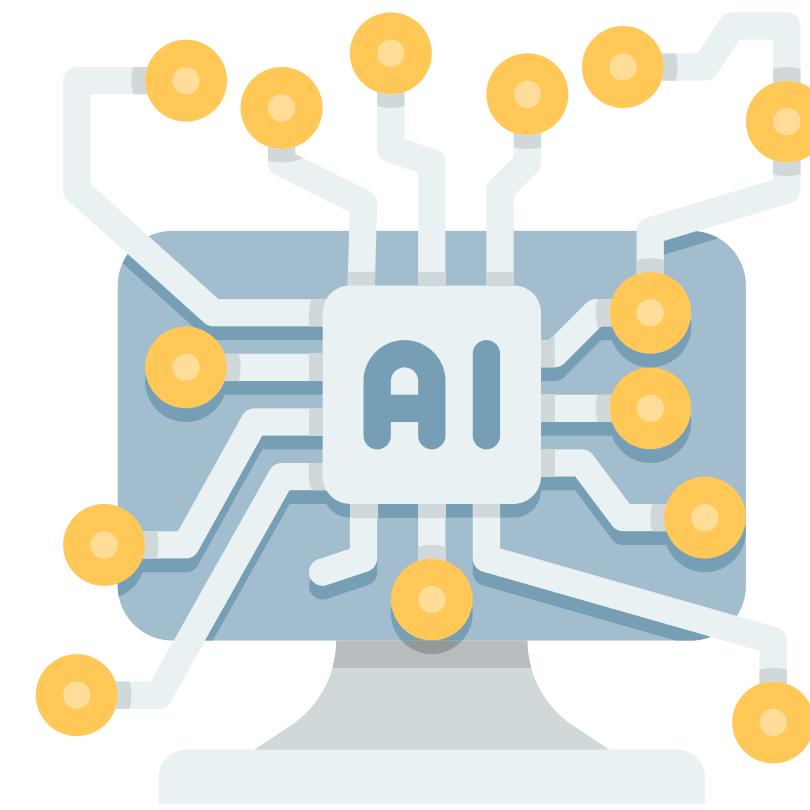
Another view



Asking a question



Let's code!!!



github.com/gardner/nais