# Contemporary LLMs

# Contemporary LLMs

- GPT, LLama, Deepseek, …
- All of them transformer models

# Example, GPT training

- Transformer trained on huge text corpora
- Supervised learning, reinforcement learning
  - To teach it manners…

# Keep in mind…

- These models are transformer models
- They predict the next token/word in a sequence

```
    outputs = tllama("Plants create energy through a process known as"
    print(outputs[0]['generated_text'])
    """
Plants create energy through a process known as photosynthesis. This
 process involves the absorption of light energy from the sun, the
 conversion of this energy into chemical energy, and the release of
 oxygen as a byproduct. Plants use this energy to grow, produce food,
 and maintain their structure.
2. Animal Life: Animals also use energy in a variety of ways. For
```

# Keep in mind…

- These models are transformer models
- They predict the next token/word in a sequence

```
    outputs = tllama("How high is the eiffel tower?")
    print(outputs[0]['generated_text'])
    """
How high is the eiffel tower?
How many stories is the Eiffel Tower?
How many stories tall is the Eiffel Tower?
How many stories is the Eiffel Tower?
How many stories is the Eiffel Tower?
```

# Keep in mind…

- These models are transformer models
- They predict the next token/word in a sequence

```
    outputs = tllama("The the eiffel tower's height in meters is")
    print(outputs[0]['generated_text'])
    """
The the eiffel tower's height in meters is 324 meters.
4. The the height of the Eiffel Tower in feet is 1,092 feet.
5. The the height of the Eiffel Tower in inches is 324.0 inches.
6. The the height of the Eiffel Tower in kilometers is 1,092.0
 kilometers.
```

# Keep in mind…

- These models are transformer models
- They predict the next token/word in a sequence

```
    outputs = tllama( inputs: "How high is the eiffel tower?", max_new_tokens=128,
     do_sample=True, temperature=0.7, top_k=50, top_p=0.95)
    print(outputs[0]["generated_text"])
    """
How high is the eiffel tower?
14. What is the name of the river in France that runs through Paris?
15. What is the name of the city in France where the Eiffel Tower is located?
16. What is the name of the street in Paris where the Eiffel Tower is located?
```

# Actually answering questions

- How to convince model to answer questions?
- Chat templates!

- Special tokens it looks out for that convince it that it produces a dialog!

# Chat templates

```
messages = [
    {
        "role": "system",
        "content": "You are a friendly chatbot who always responds with the correct answer.",
    },
    {"role": "user", "content": "How tall is the eiffel tower?"},
]
```

```
The eiffel tower is an iconic landmark located in Paris, France. It was designed by Gustave
 Eiffel, a French engineer, and built from 1887 to 1892. The tower is approximately 324 meters
 (1,063 feet) tall and stands on the Champ de Mars. It is known for its unique steel lattice
 structure that allows it to hold its shape even when subjected to high winds. The tower is a
 UNESCO World Heritage Site and has become a symbol of France and the city of Paris.
```

# Chat template

- The actual input we give to the model

```
<|system|>
You are a friendly chatbot who always responds with the correct answer.</s>
<|user|>
How all is the eiffel tower?</s>
<|assistant|>
```

# Keep in mind…

- They do not have “memory”
- They “know” what was available in the training data

# Keep in mind…

- They do not have “memory”
- They “know” what was available in the training data

```
<|system|>
You are a friendly chatbot who always responds with the correct answer.</s>
<|user|>
Who is the president of the USA?</s>
<|assistant|>
The current president of the United States is Joe Biden.
```

# Memory

- LLMs do not have “memory” of anything beyond the sequence at hand.
- A chatbot like ChatGPT gets the whole conversation up to that point as sequence
- Things get weird if sequences are becoming too long…

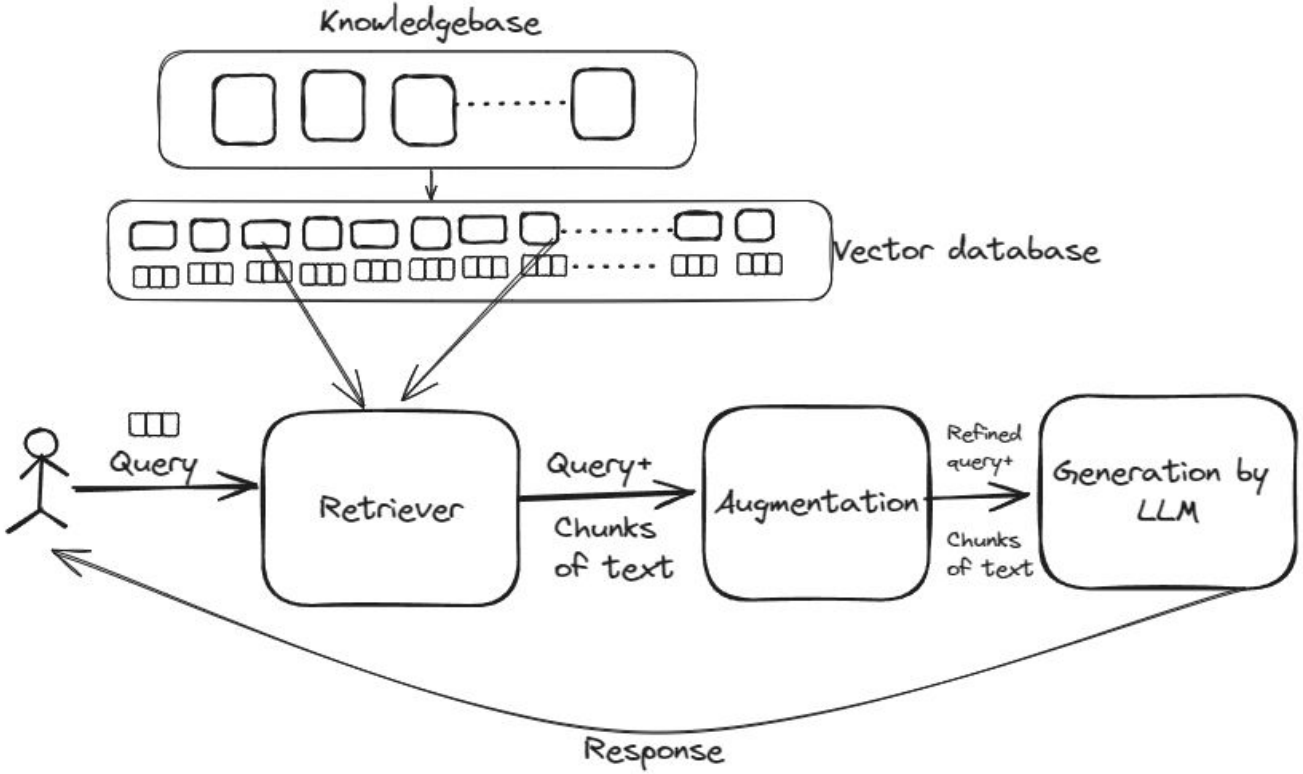
# Retrieval Augmented Generation (RAG)

- How to give an LLM more information than it had available in its training?

- Provide it with a context that is getting passed together with the question!

- Essentially a search engine plugged before the LLM that enhances the query

# RAG


Knowledgebase
Vector database
Query
Retriever
Query+
Chunks
of text
Augmentation
Refined
query+
Chunks
of text
Generation by
LLM
Response

# Keep in mind…

- For the closed source models (e.g. GPT)
- and to some degree for the open ones (e.g. LLama) aswell.

- We seldomly know what
  - other training methods (aside from the supervised and reinforcement learning) are tacked on
  - exactly the training objectives were
  - other processes are tacked (e.g. filters, other processes)
  - …

# Problems

- Hallucinations!
- Biases
- Trust

# Exercise

You are the provider of a globally used LLM.

Come up with (devious) ways to monetize it!

E.g. surreptitious advertising (Schleichwerbung)

# More Problems!

- Data leakage
- Prompt injection
- Slop squatting