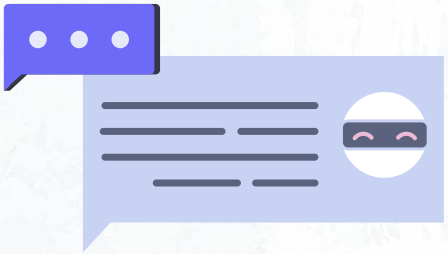
Large Language Models

CURRENT STATE

Rodrigo Gonzalez, PhD

(AI)



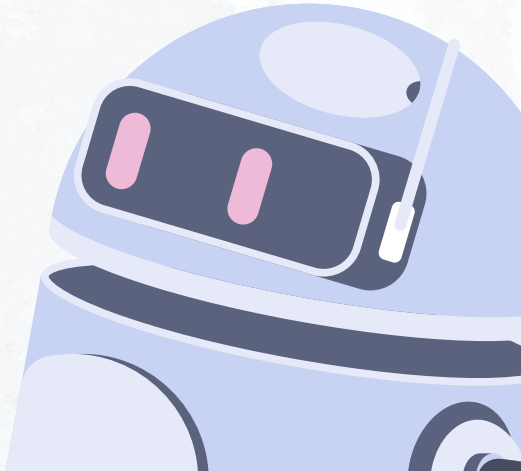


Table of contents

01 →

What is a Large Language Model (LLM)?

(AI)

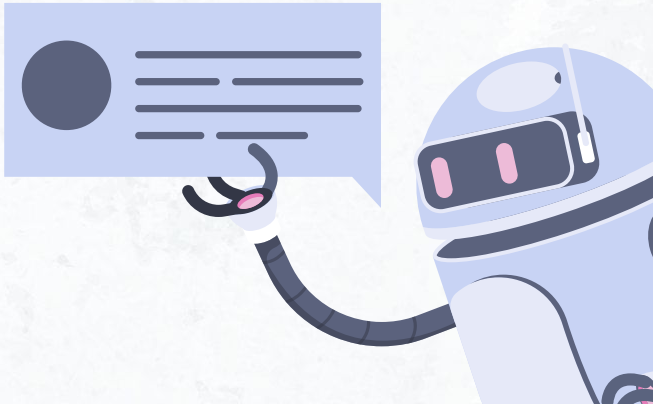
Large Language Model

A large language model (LLM) is a language model consisting of a neural network with many parameters (**billions**) traineded with immense amounts of texts using self-supervised learning.

LLMs are probabilistic models that attempt to map the probability of a sequence of words, given the surrounding context.

We don’t have to speak the language of computers anymore

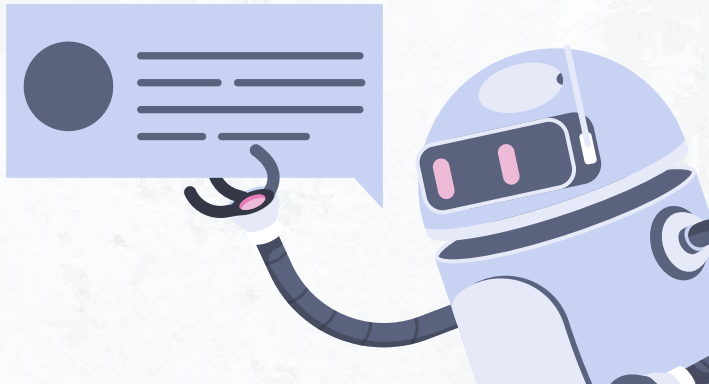
they can speak ours!



How does LLM work?

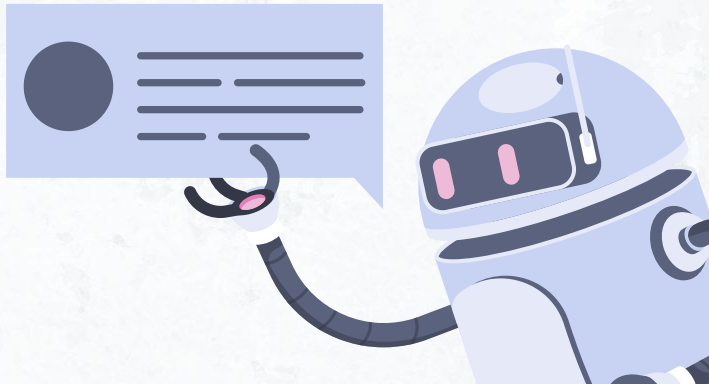
⏳

No nos une el amor ...



How does LLM work?

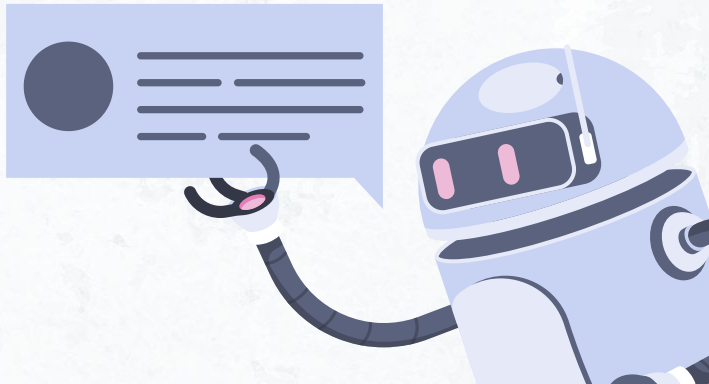
No nos une el amor sino



How does LLM work?

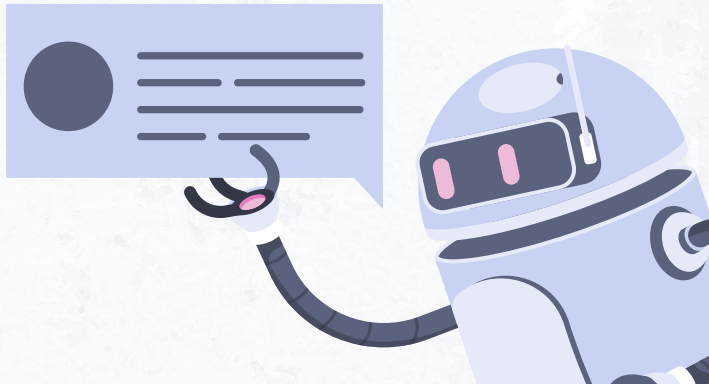
⏳

No nos une el amor sino ...



How does LLM work?

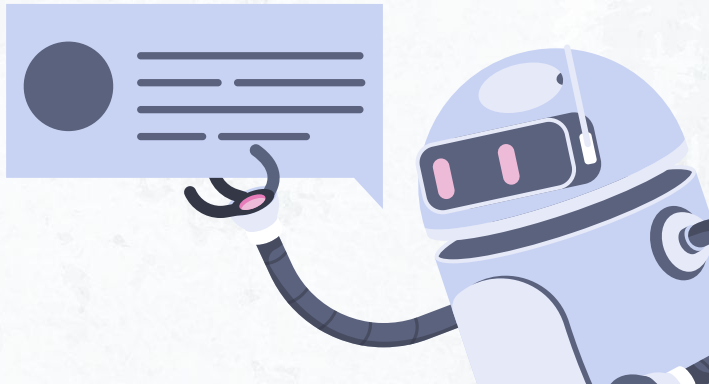
No nos une el amor sino el



How does LLM work?

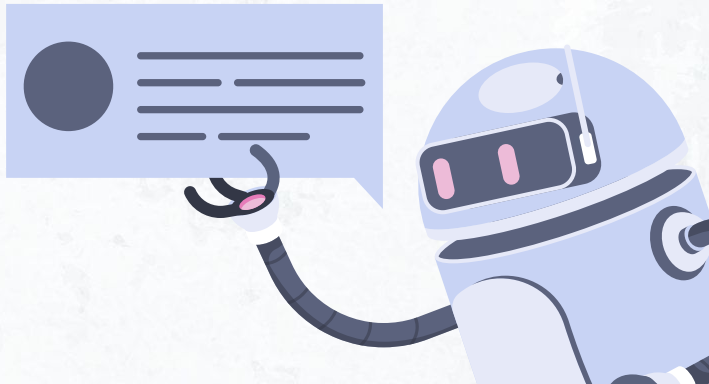
⏳

No nos une el amor sino el ...



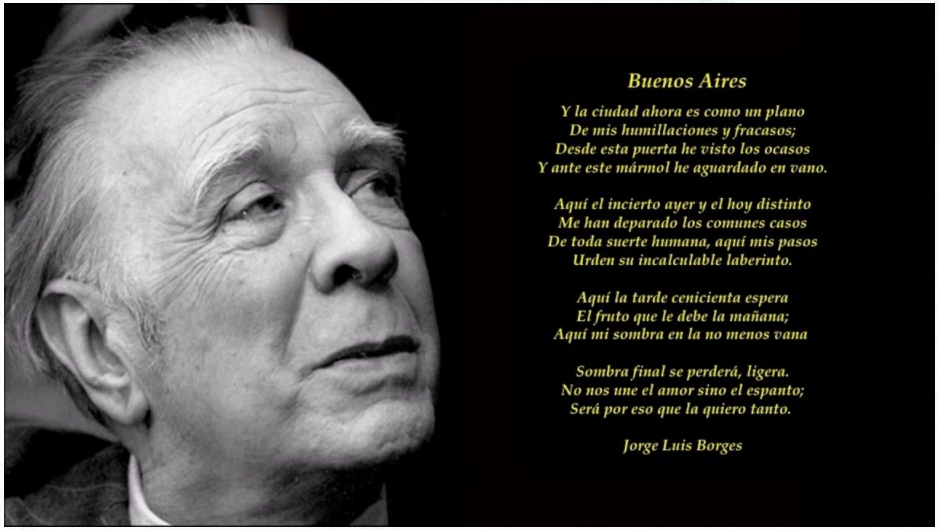
How does LLM work?

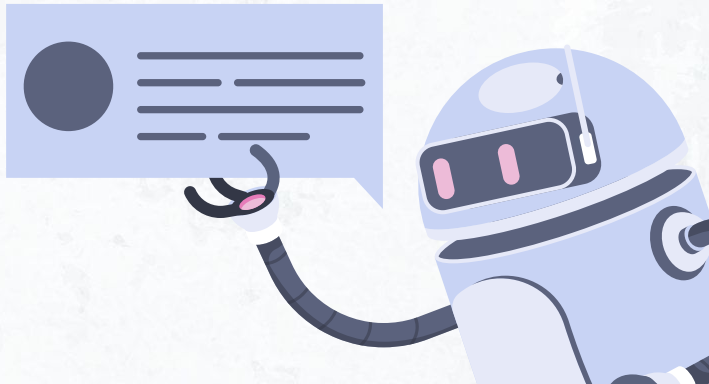
No nos une el amor sino el espanto



How does LLM work?

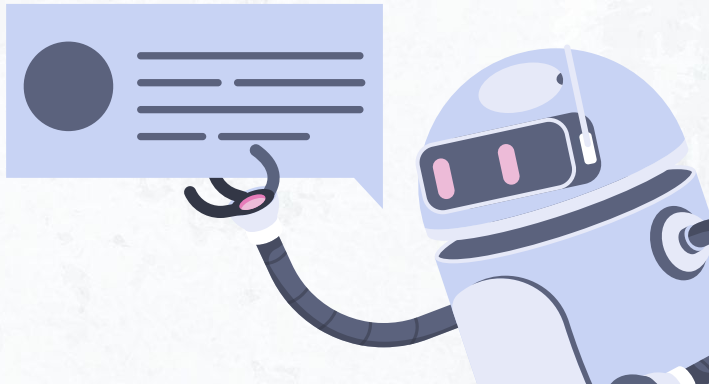
No nos une el amor sino el espanto





How does LLM work?

No nos une el amor sino el odio



02 →

LLMs, a very brief technical history

(AI)

LLMs, a very brief technical history

1. 2013, NLP, Embeddings (Word2Vec, Glove)
2. 2017, Transformers, Attention is all you need
3. 2018, GPT (117M params)
4. 2019, GPT-2 (1.5B)
5. 2020, GPT-3 (175B)
6. 2022, Chat GPT-3.5, more training (175B)

LLMs, a very brief technical history

7 Feb 2023, LLaMa (Meta), open source (70B)

8 March 2023, GPT-4, (1T?)

9 March 2023, Bard (Google) (137B)

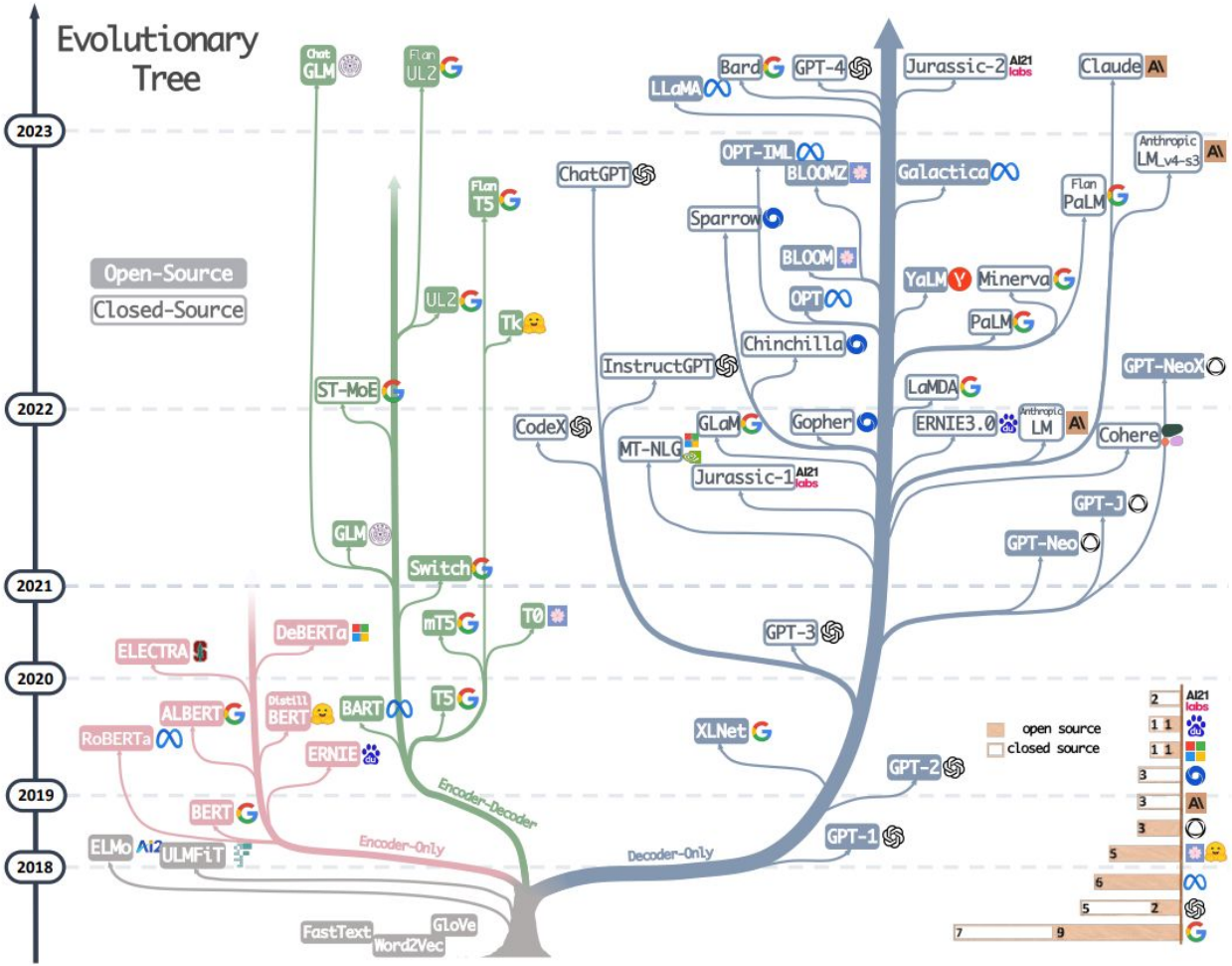
10 May 2023, QLoRa

11 May 2023, Falcon, open source (40B)

12 July 2023, Llama 2 (Meta), open source (70B)

Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond

26 Apr 2023

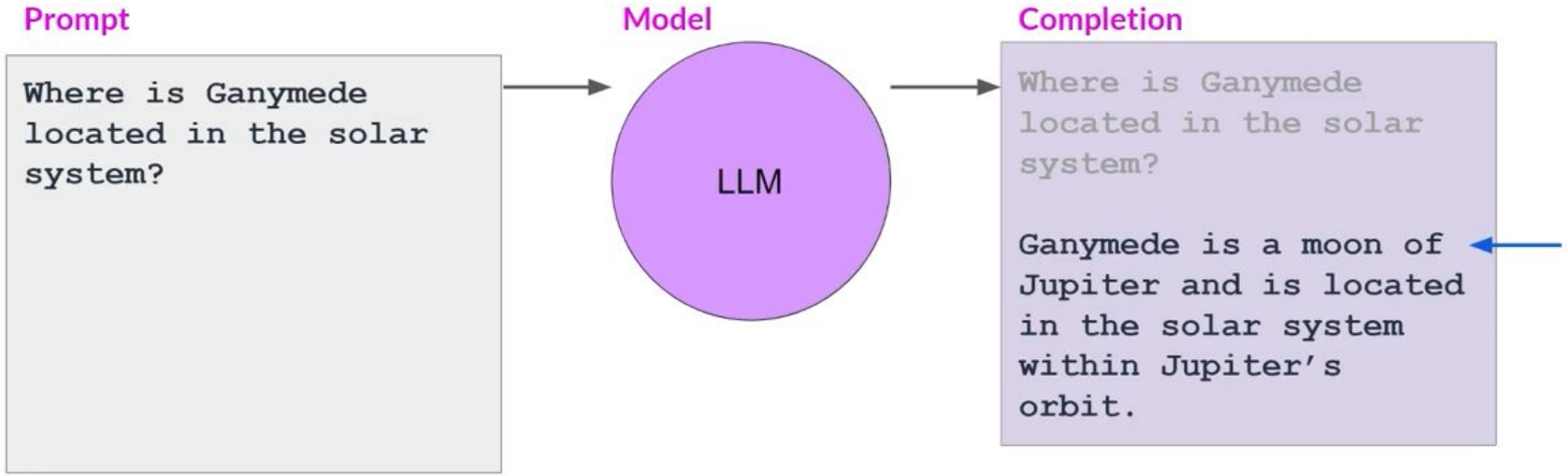


03 →

Prompt engineering

(AI)

Prompts and completions



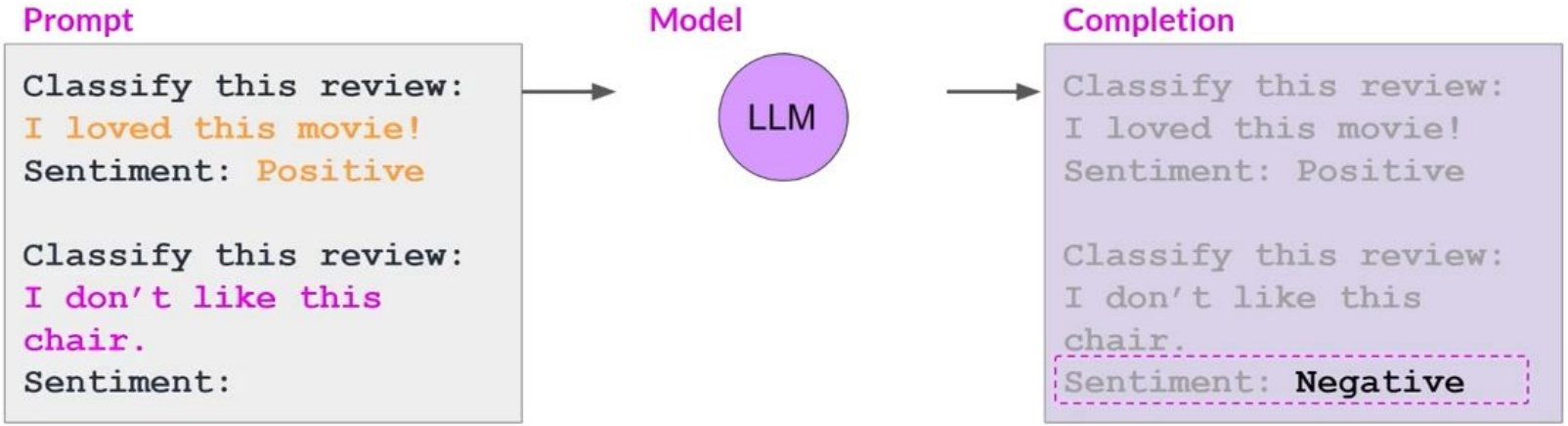
Context window

- typically a few 1000 words.

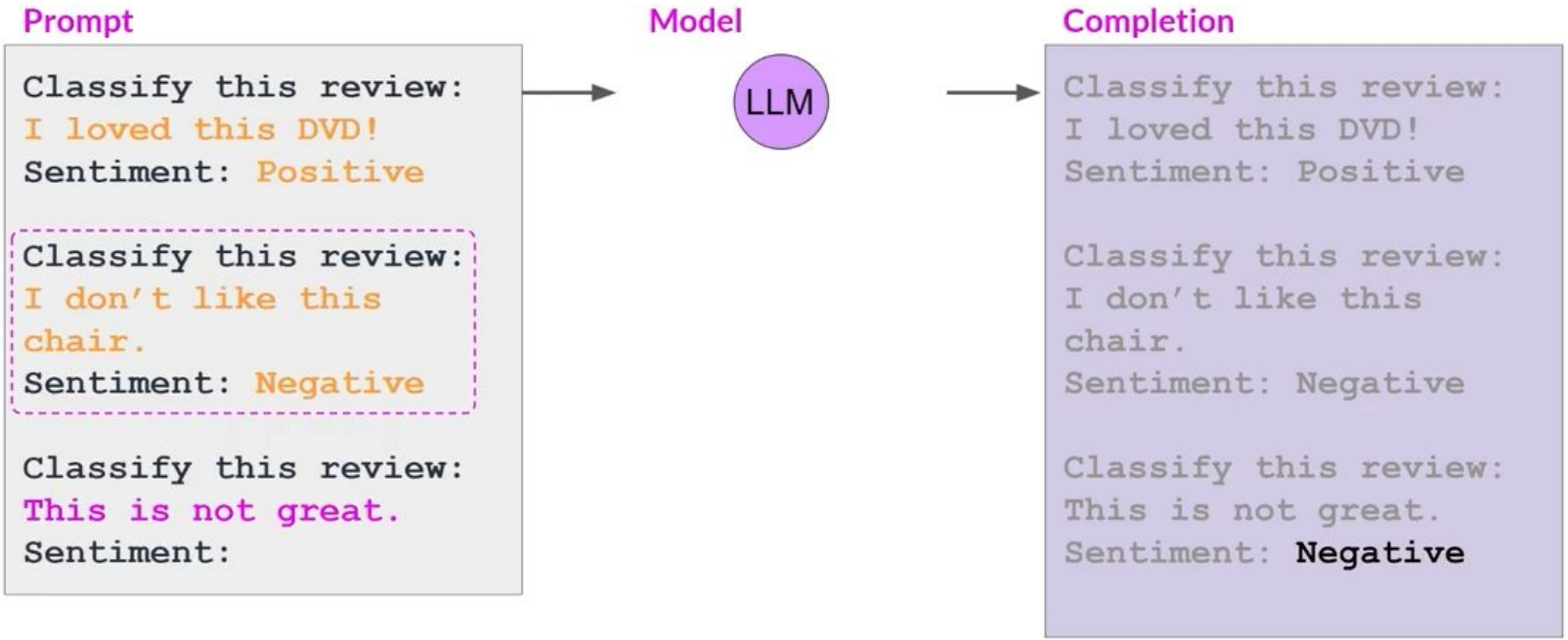
In-context learning (ICL) - zero shot inference

Zero-shot inference

In-context learning (ICL) - one shot inference



In-context learning (ICL) - few shot inference



Summary of in-context learning (ICL)

Prompt // Zero Shot

```
Classify this review:
I loved this movie!
Sentiment:
```

Prompt // One Shot

```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don't like this
chair.
Sentiment:
```

Prompt // Few Shot >5 or 6 examples

```
Classify this review:
I loved this movie!
Sentiment: Positive

Classify this review:
I don't like this
chair.
Sentiment: Negative

Classify this review:
Who would use this
product?
Sentiment:
```

Context Window
(few thousand words)

04 →

LLM fine-tuning

(AI)

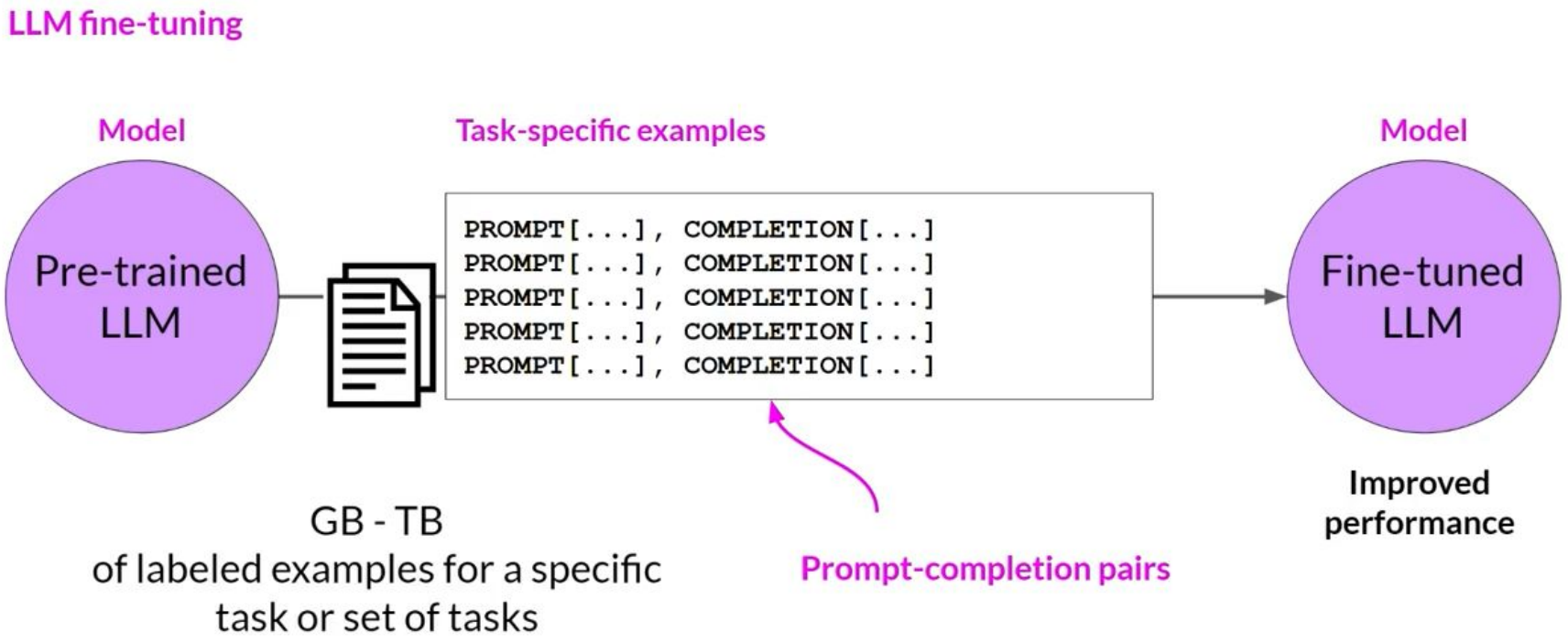
Limitations of in-context learning

Even with multiple examples

- In-context learning may not work for smaller models LLM
- Examples take up space in the context window

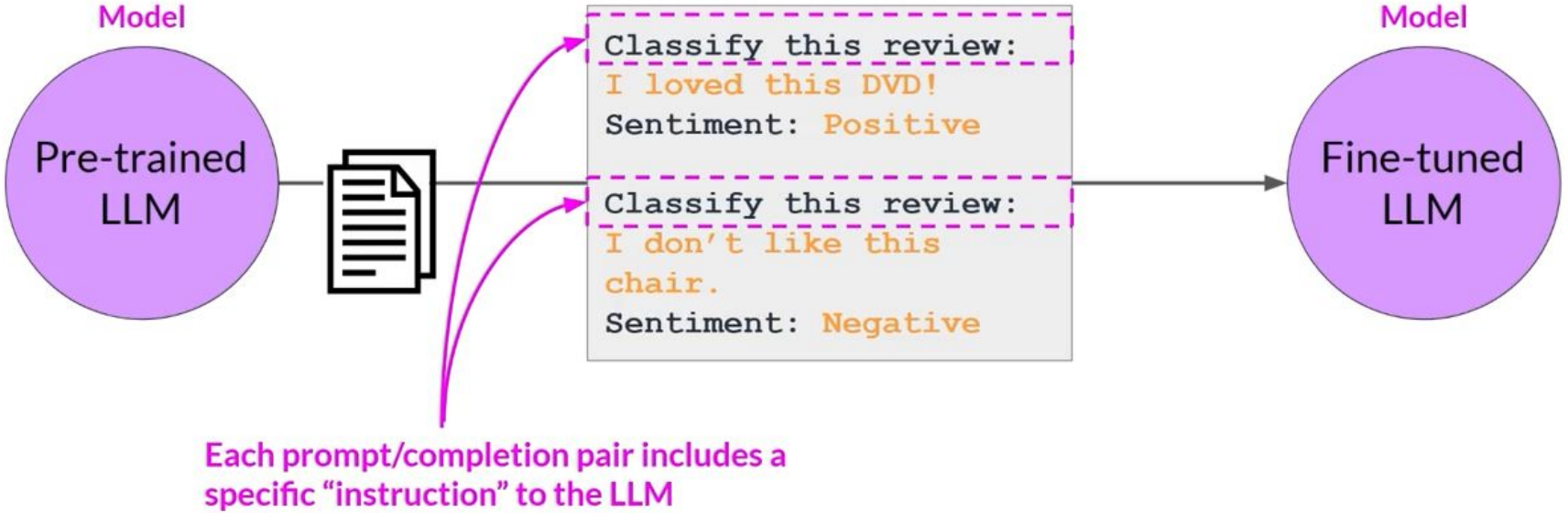
Instead, try **fine-tuning** the model

LLM fine-tuning at a high level



Using prompts to fine-tune LLMs with instruction

LLM fine-tuning



Sample prompt instruction templates

Classification / sentiment analysis

```
jinja: "Given the following review:\n{{review_body}}\npredict the associated rating\
  \ from the following choices (1 being lowest and 5 being highest)\n- {{ answer_choices\
  \ | join('\\n- ') }} \n|||\n{{answer_choices[star_rating-1]}}"
```

Text generation

```
jinja: Generate a {{star_rating}}-star review (1 being lowest and 5 being highest)
  about this product {{product_title}}.        |||        {{review_body}}
```

Text summarization

```
jinja: "Give a short sentence describing the following product review:\n{{review_body}}\
  \ \n|||\n{{review_headline}}"
```

Source: https://github.com/bigscience-workshop/promptsource/blob/main/promptsource/templates/amazon_polarity/templates.yaml

LLM fine-tuning process

LLM fine-tuning

Prepared instruction dataset

Training splits

```
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
```

Training

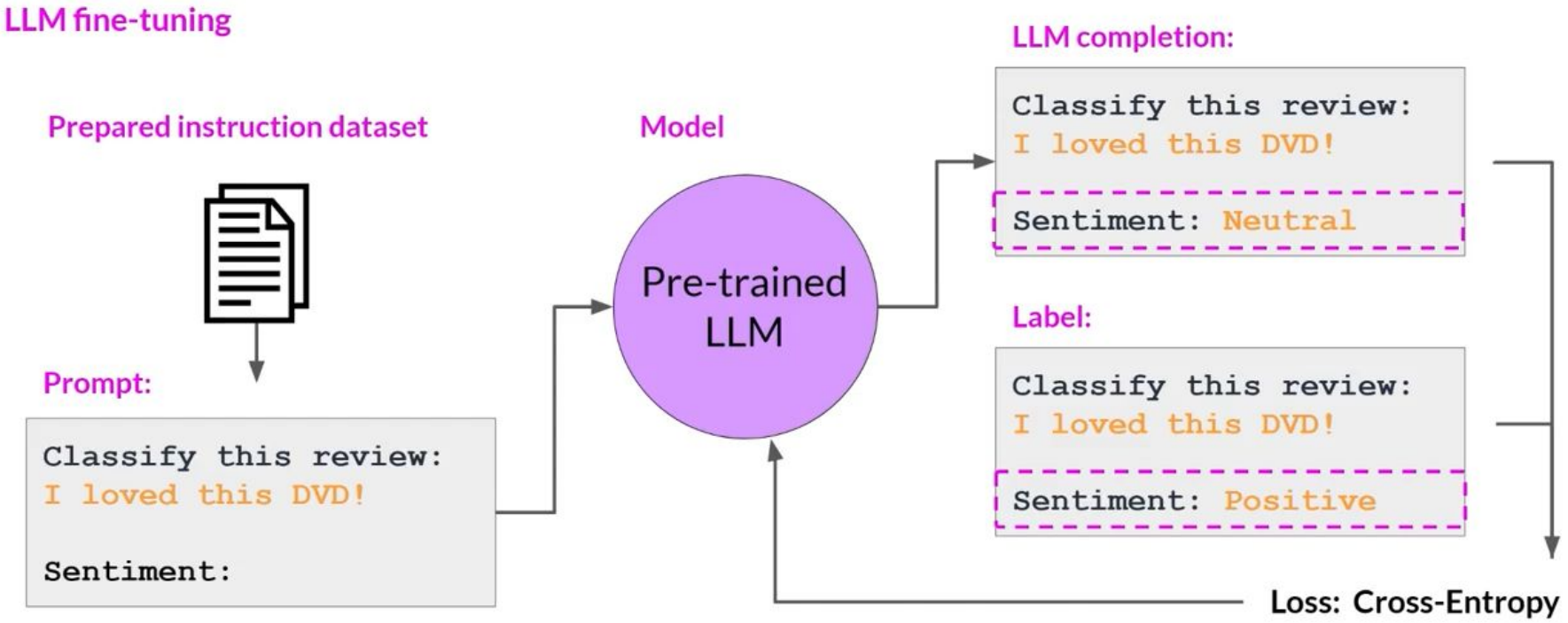
```
PROMPT[...], COMPLETION[...]
...
```

Validation

```
PROMPT[...], COMPLETION[...]
...
```

Test

LLM fine-tuning process



LLM fine-tuning process

LLM fine-tuning

Prepared instruction dataset

Training splits

```
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
```

Training

```
PROMPT[...], COMPLETION[...]
...
```

Validation

validation_accuracy

```
PROMPT[...], COMPLETION[...]
...
```

Test

LLM fine-tuning process

LLM fine-tuning

Prepared instruction dataset

Training splits

```
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
PROMPT[...], COMPLETION[...]
```

Training

```
PROMPT[...], COMPLETION[...]
...
```

Validation

```
PROMPT[...], COMPLETION[...]
...
```

Test

test_accuracy

LLM fine-tuning process

05 →

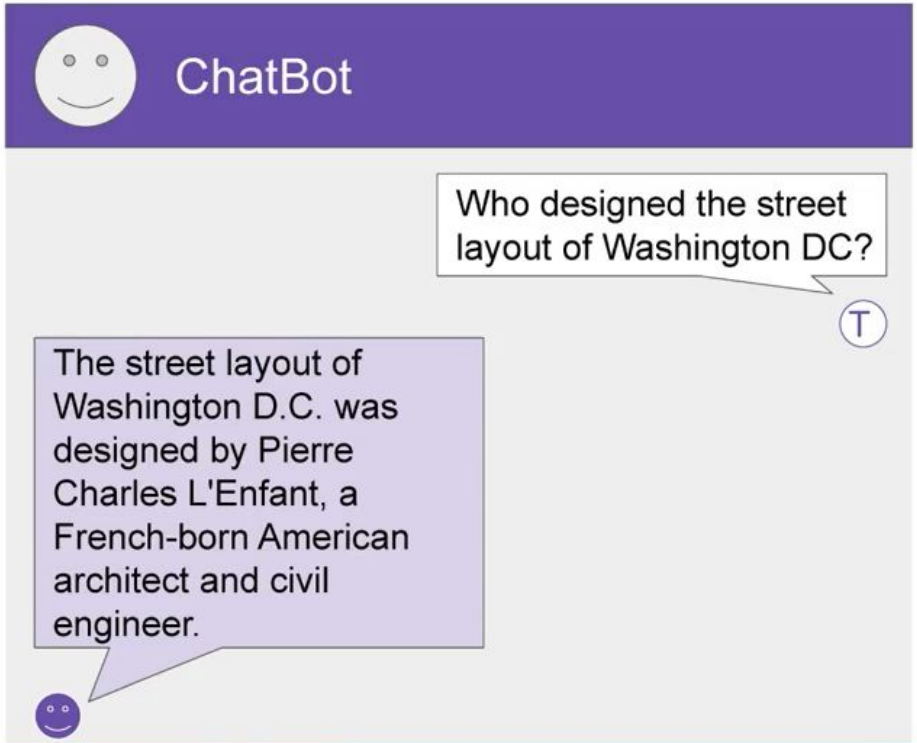
LLM main use cases

(AI)

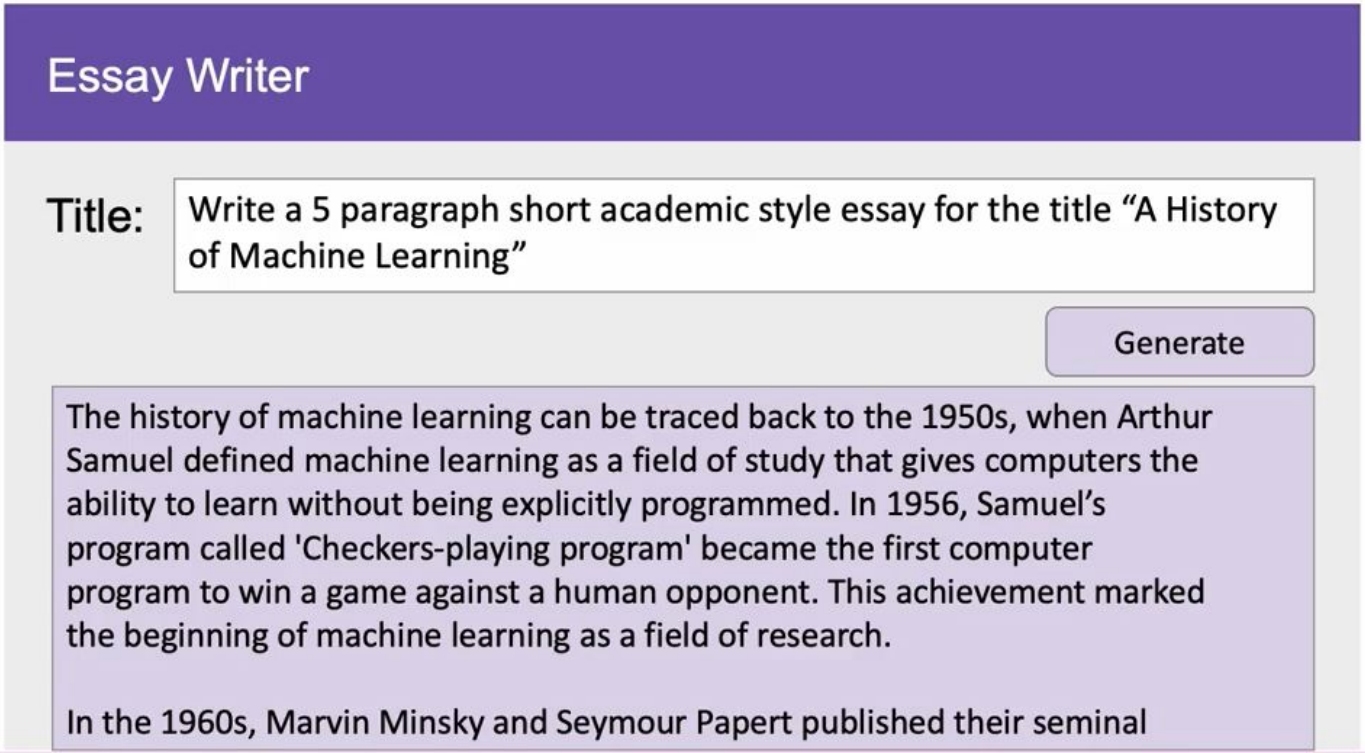
Main use cases

- Chatbots and virtual assistants 🤖
- Content generation
- Transforming (translation, parsing)
- Editing and proofreading
- Sentiment analysis and text classification
- Question answering systems (RAG)
- Code generation and debugging
- Summarization
- Agents
- Automation! 🤖

LLM chatbot


ChatBot
Who designed the street layout of Washington DC?
T
The street layout of Washington D.C. was designed by Pierre Charles L'Enfant, a French-born American architect and civil engineer.

LLM use cases & tasks



LLM use cases & tasks

Summarize

Text file:  support.txt

Generate

In the chat session, Support efficiently and effectively assists Alex, who was initially unable to access their account due to issues with a password reset email, leading to a positive customer service experience.

LLM use cases & tasks

LLM use cases & tasks

Code AI

Prompt:

Write some python code that will return the mean of every column in a dataframe.

Generate

Code:

```
import pandas as pd

df = pd.DataFrame({
    'A': [1, 2, 3, 4, 5],
    'B': [2, 3, 4, 5, 6],
    'C': [3, 4, 5, 6, 7]
})

mean_values = df.mean()
```

LLM use cases & tasks

Entity Extraction

Input:

Scientist Dr. Evangeline Starlight of Technopolis announced a breakthrough in quantum computing at Nova University. Mayor Orion Pulsar commended her. The discovery will be shared at the Galactic Quantum Computing Symposium in Cosmos.

Extract

LLM use cases & tasks

Entity Extraction

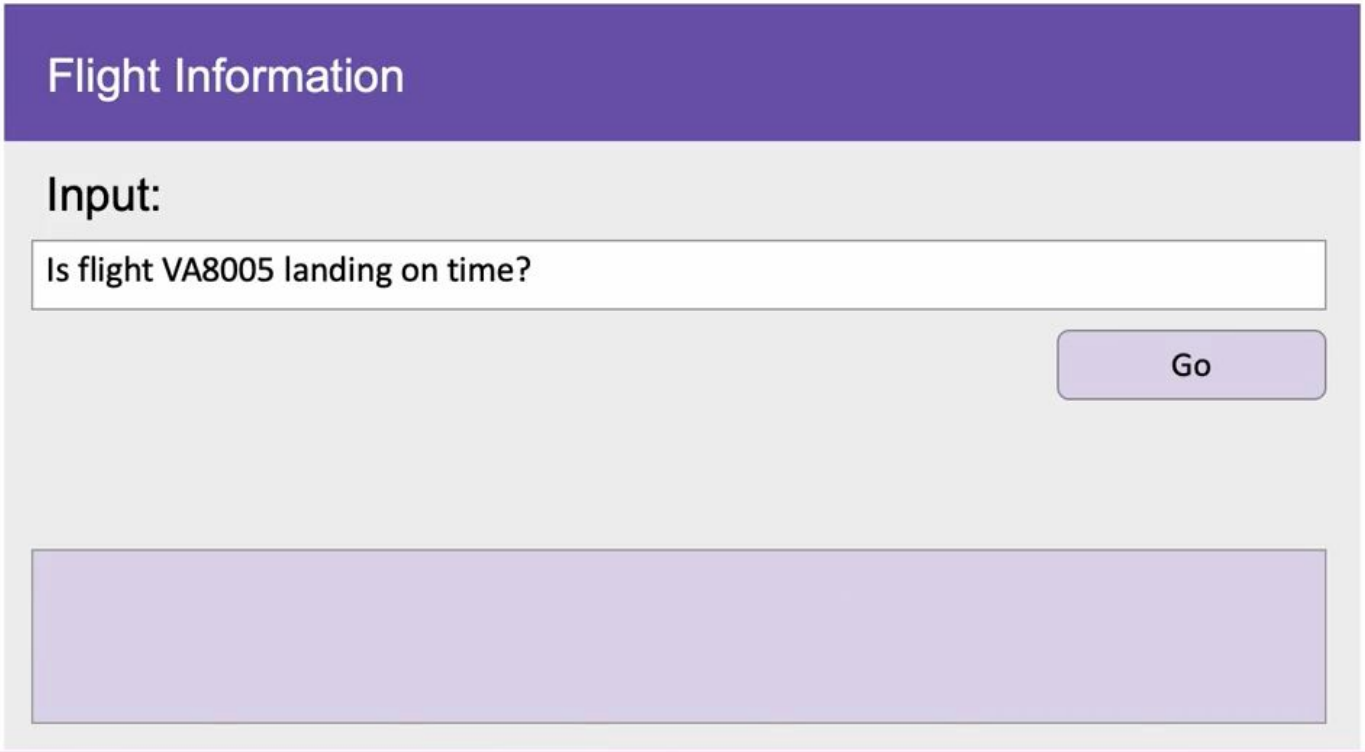
Input:

Scientist Dr. Evangeline Starlight of Technopolis announced a breakthrough in quantum computing at Nova University. Mayor Orion Pulsar commended her. The discovery will be shared at the Galactic Quantum Computing Symposium in Cosmos.

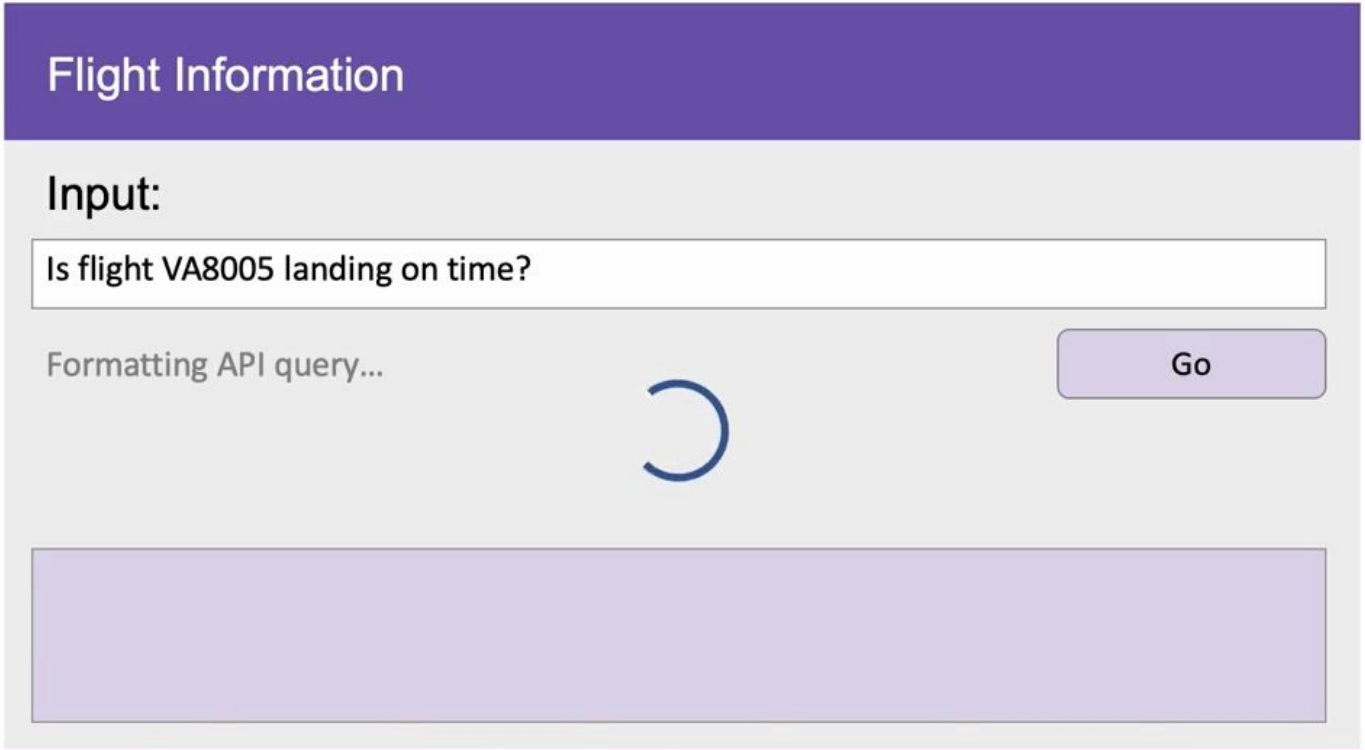
The named entities in this shorter text are "Dr. Evangeline Starlight", "Technopolis", "quantum computing", "Nova University", "Mayor Orion Pulsar", "Galactic Quantum Computing Symposium", and "Cosmos".

Extract

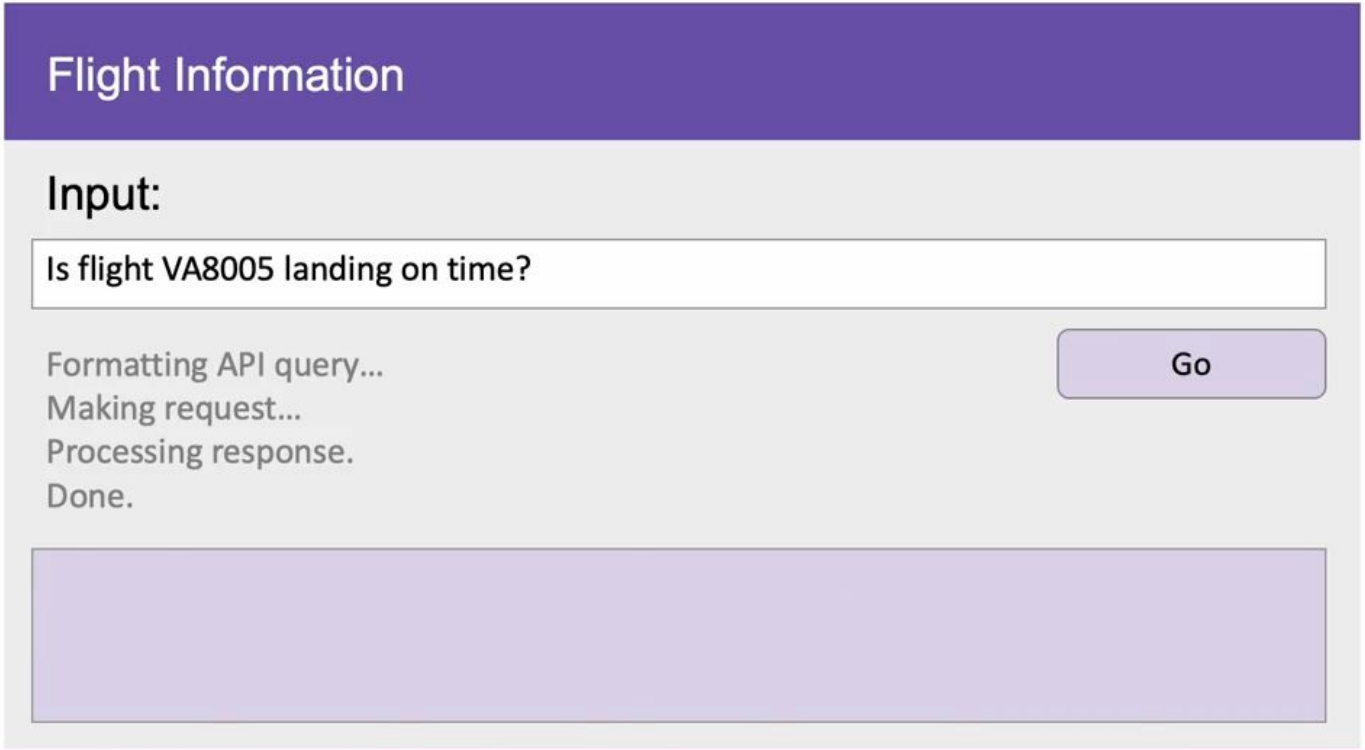
LLM use cases & tasks



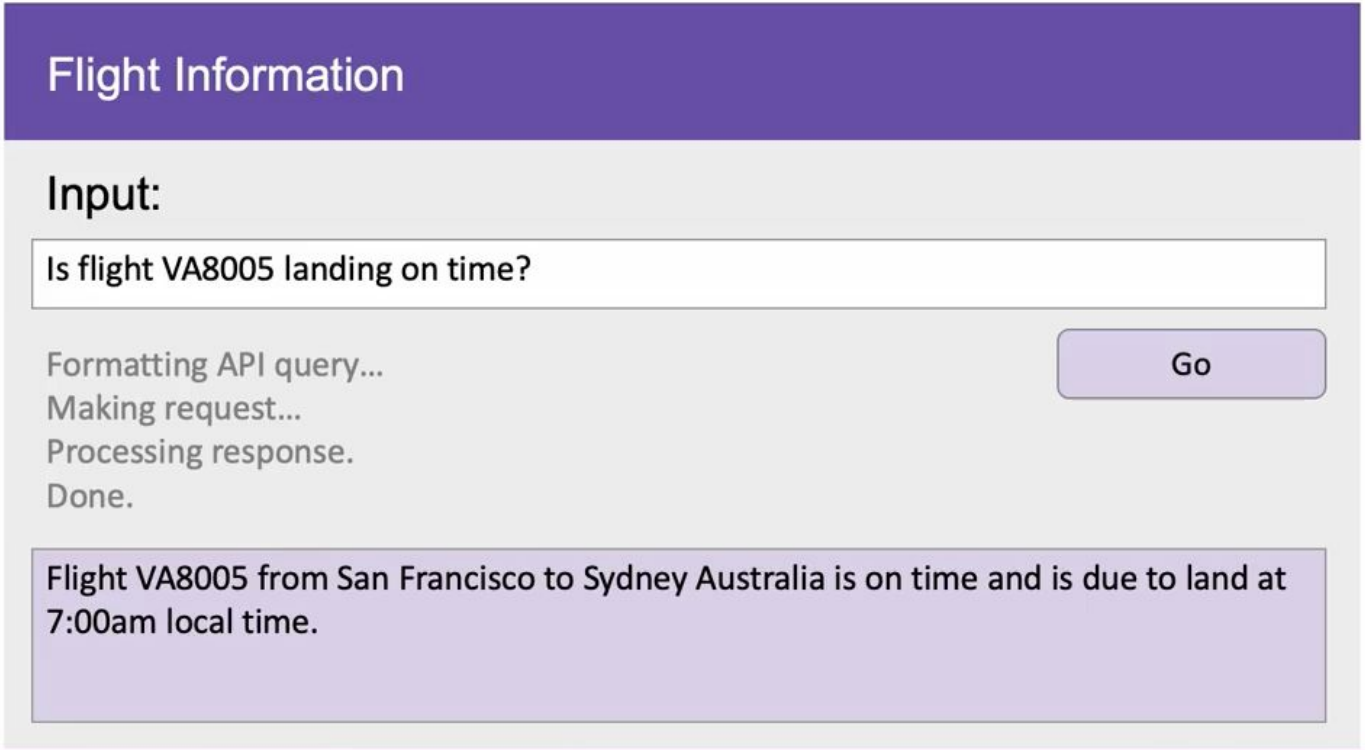
LLM use cases & tasks



LLM use cases & tasks



LLM use cases & tasks

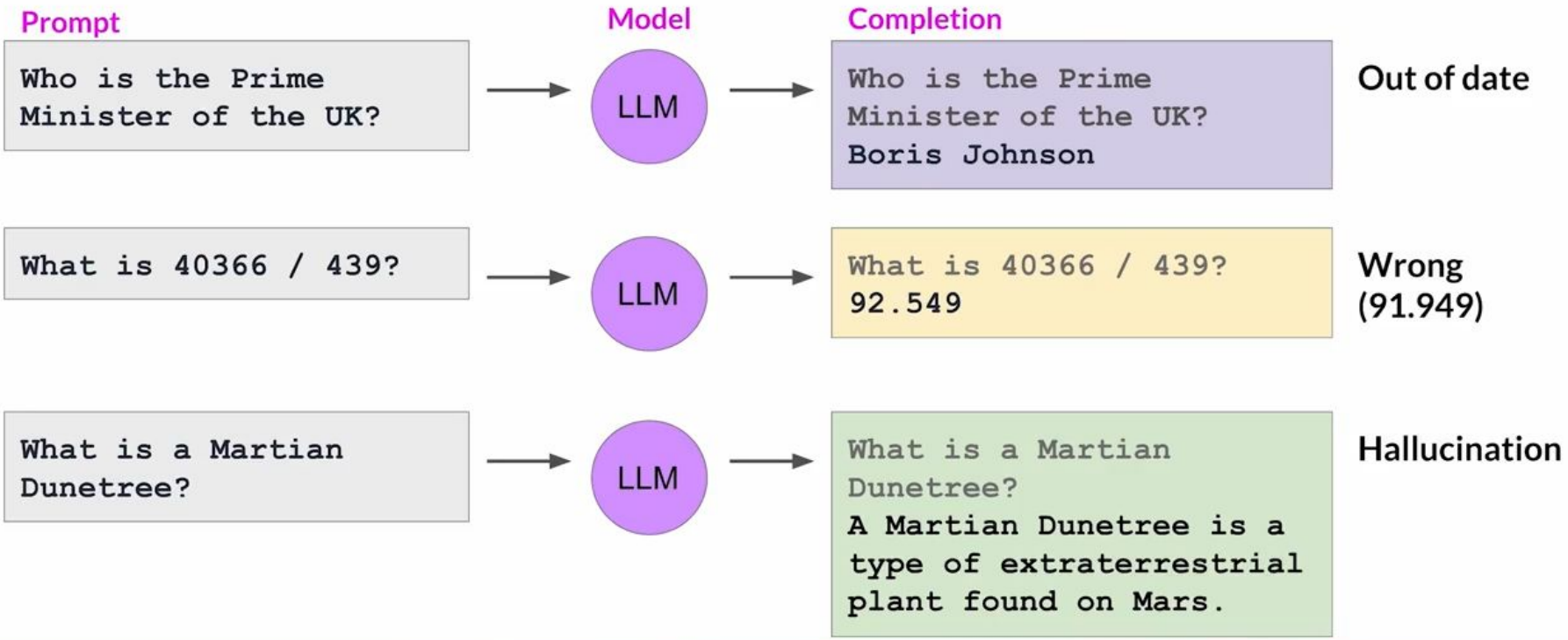


06 →

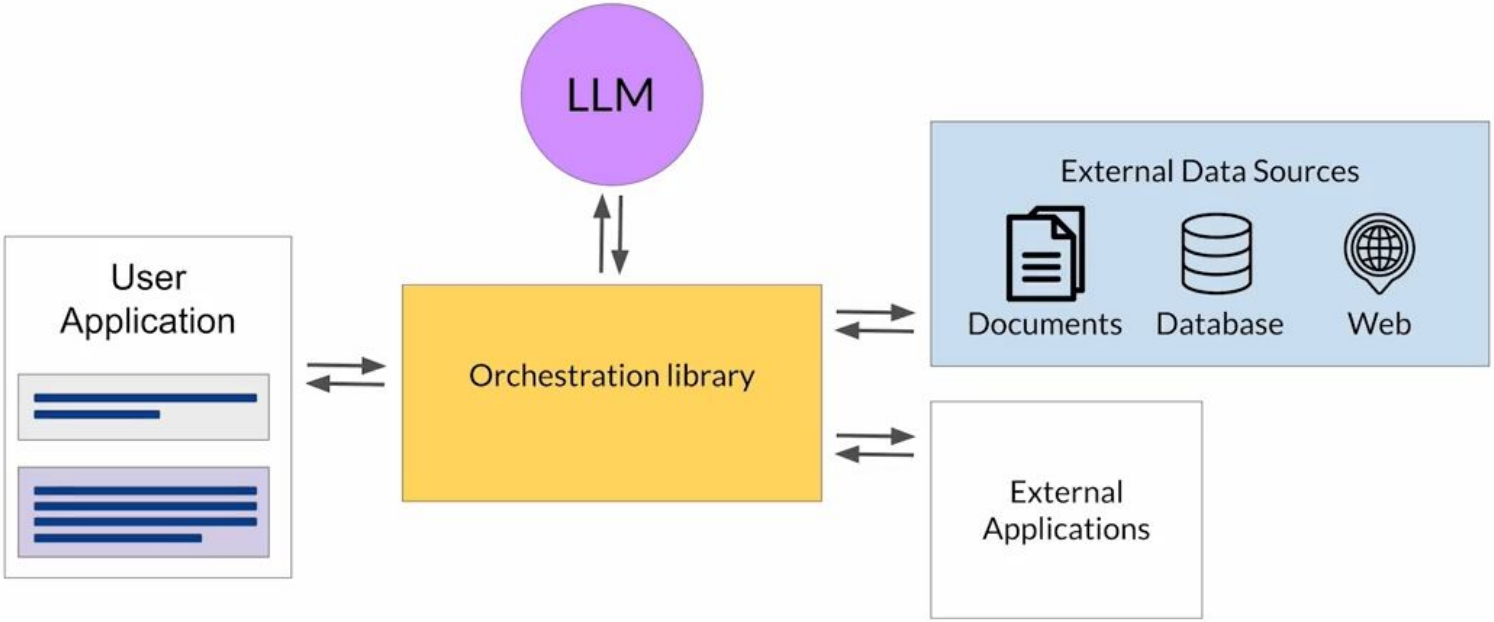
Using LLM in applications

(AI)

Models having difficulty

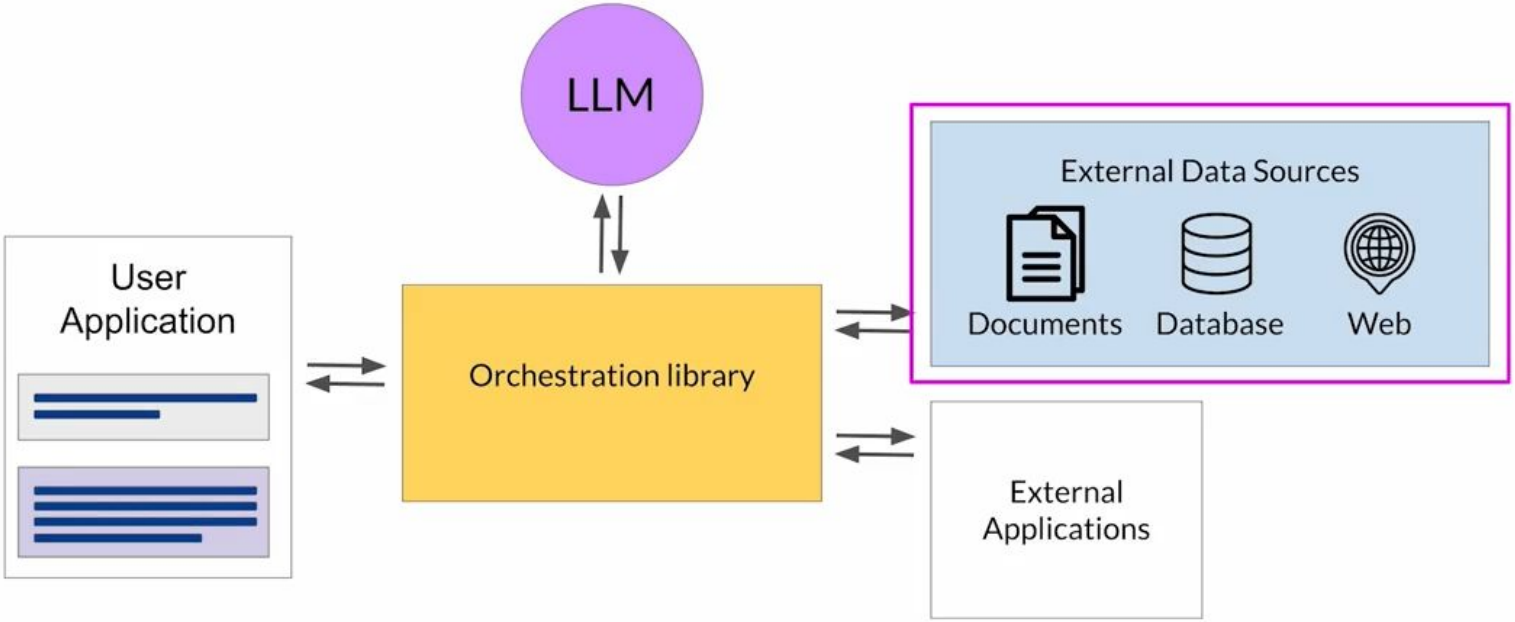


LLM-powered applications

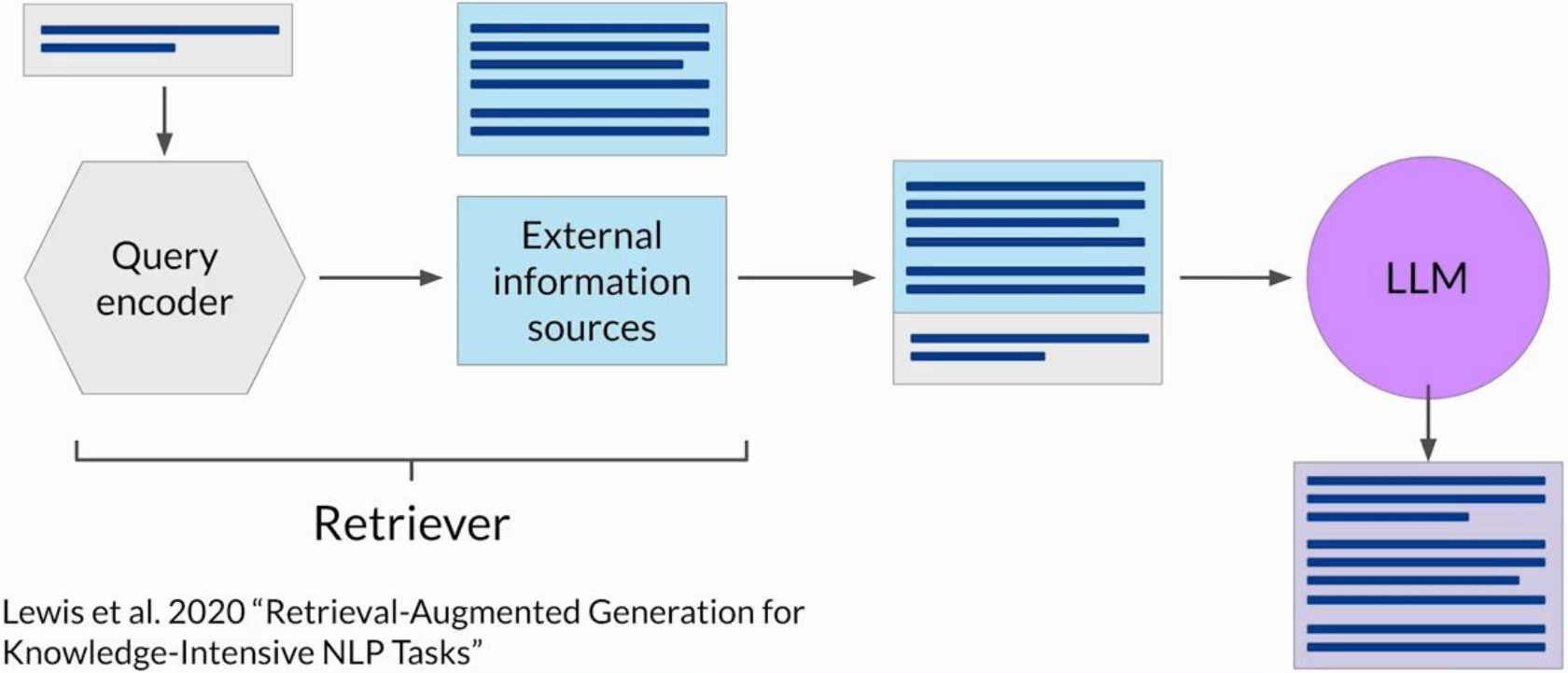


Retrieval augmented generation (RAG)

LLM-powered applications

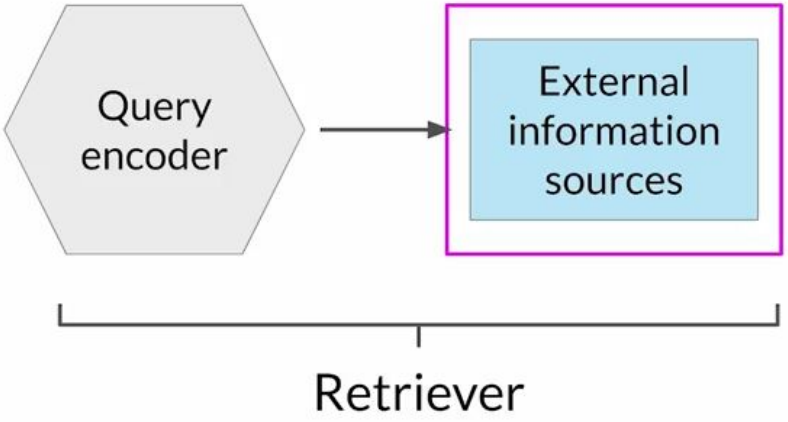


Retrieval Augmented Generation (RAG)



Lewis et al. 2020 “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”

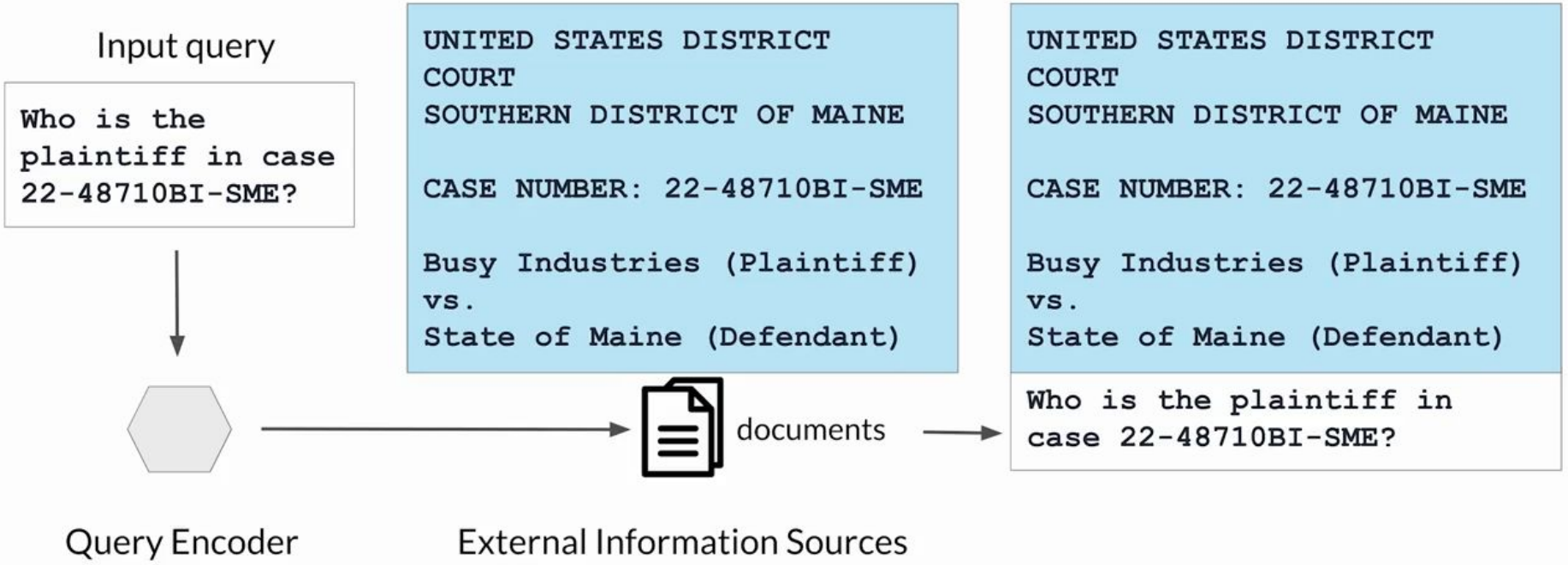
RAG integrates with many types of data sources



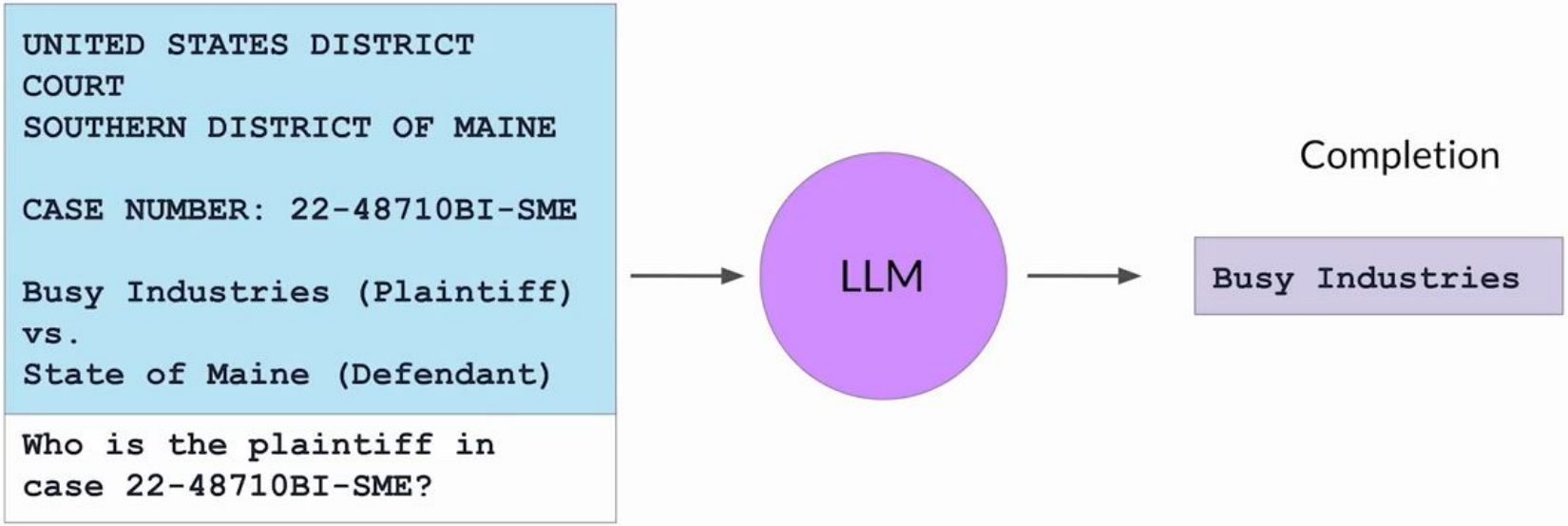
External Information Sources

- Documents
- Wikis
- Expert Systems
- Web pages
- Databases
- Vector Store

Example: Searching legal documents



Example: Searching legal documents



Thanks! →

Any questions?

Rodrigo Gonzalez, PhD

rodrigo.gonzalez@ingenieria.uncuyo.edu.ar