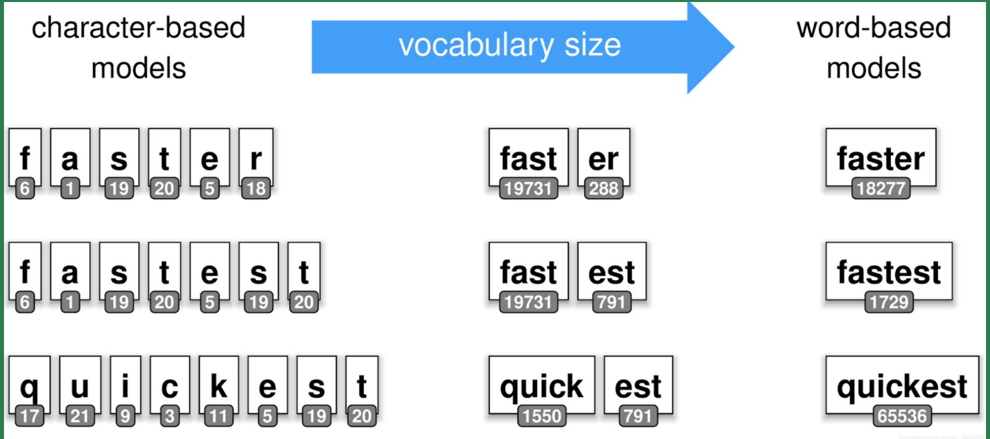
LLM
2025

ML Basics

1. What is the input?
2. What is the output?
3. What is the model architecture?
4. What is the learning algorithm?

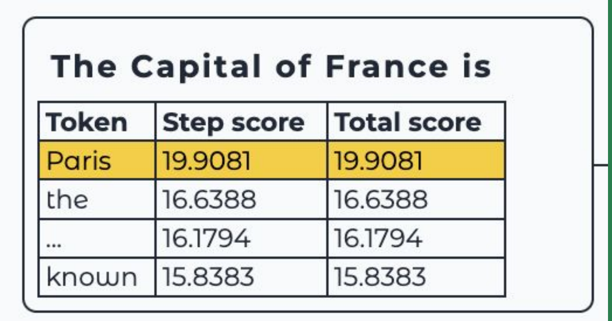
LLM
NEXT WORD PREDICTORS

What is the input?



https://tiktokenizer.vercel.app/

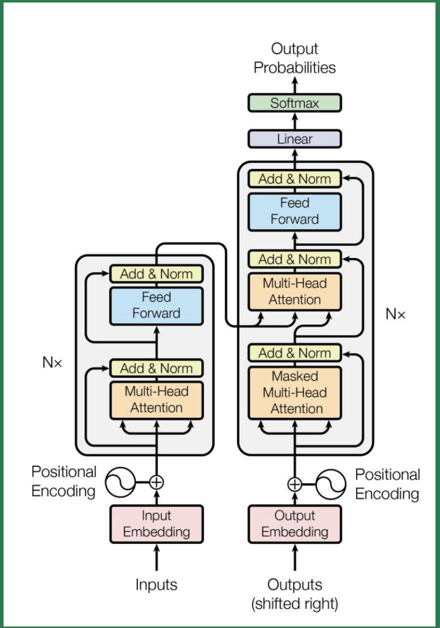
What is the output?



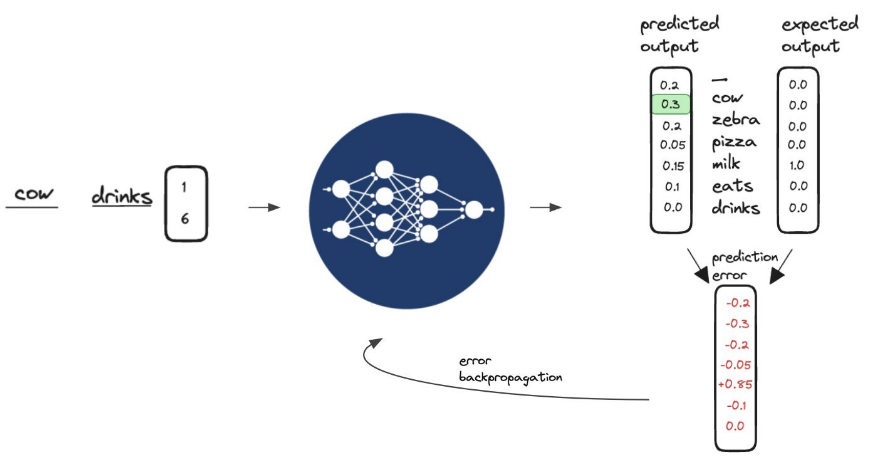
The Capital of France is

Token	Step score	Total score
Paris	19.9081	19.9081
the	16.6388	16.6388
...	16.1794	16.1794
known	15.8383	15.8383

What is the model arch algorithm?



What is the learning algorithm?



predicted output		expected output	prediction error
0.2	—	0.0	-0.2
0.3	cow	0.0	-0.3
0.2	zebra	0.0	-0.2
0.05	pizza	0.0	-0.05
0.15	milk	1.0	+0.85
0.1	eats	0.0	-0.1
0.0	drinks	0.0	0.0

ML Basics

1. What is the input? -tokens

2. What is the output? -token probabilities

3. What is the model architecture? -transformers

4. What is the learning algorithm? -self-supervised learning (Stochastic Gradient Descent - error back propagation)

DEMO

Spelling/Counting

LLM History

2018 - Transformer paper (Attention is all you need)

2020 - GPT3 (Language Models are Few-Shot Learners)

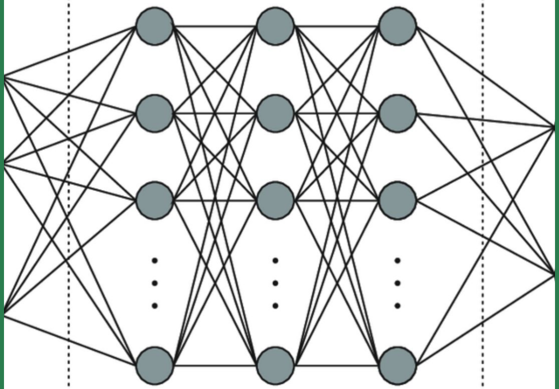
2022 - ChatGPT (GPT 3.5 + RLHF)

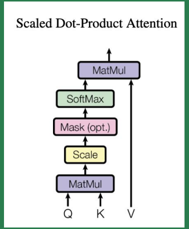
2024 - Deepseek (R1-Zero & R1)

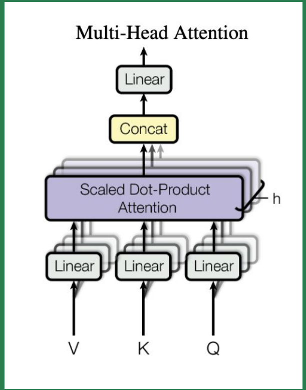
LLM History

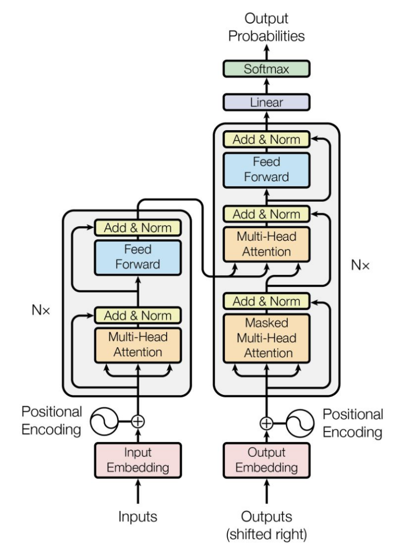
2018 - Transformer paper (Attention is all you need)

- basic language model architecture
- smarter / more efficient way to combine parameters




Scaled Dot-Product Attention

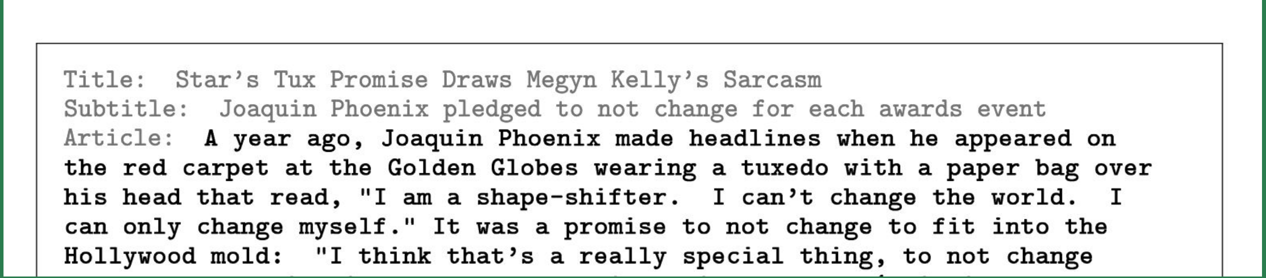

Multi-Head Attention



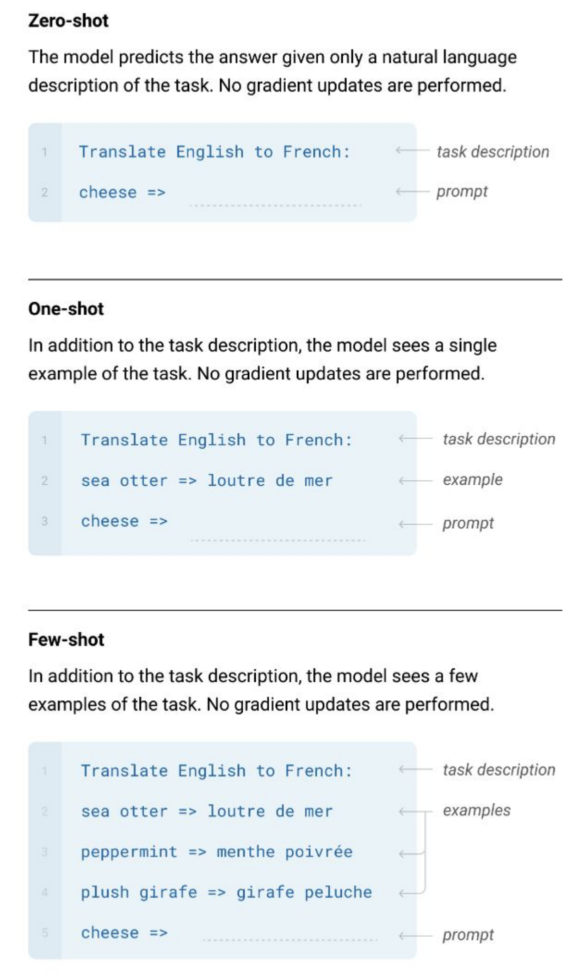
LLM History

2020 - GPT3 (Language Models are Few-Shot Learners)

- 10X scale of previous models
- can generate human-like text - news articles
- performs language tasks without task-specific tuning (emergent SOTA capabilities)
  - Translation
  - Question-answering
  - Unscrambling words
  - 3-digit arithmetic.



Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: **A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change**



Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

DEMO

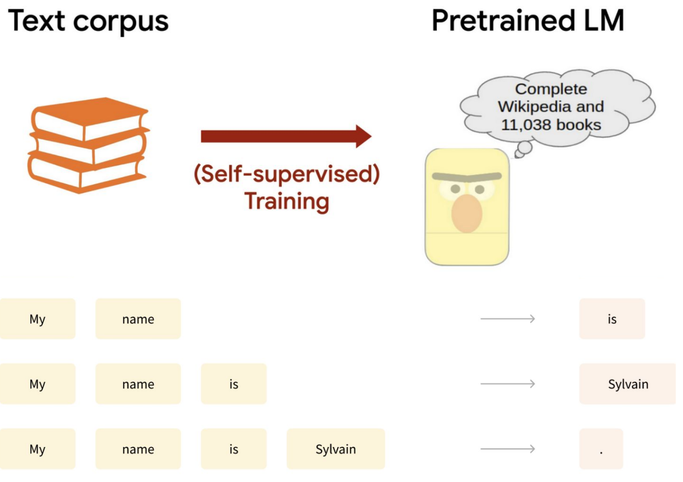
translate to hungarian
bear => medve
strawberry =>

Llama3.1-405b-base

LLM History

2022 - ChatGPT (GPT 3.5 + RLHF)

- base models vs instruction following/assistant models
- raw text corpus vs special format conversation data
- human preference feedback



Text corpus

Pretrained LM

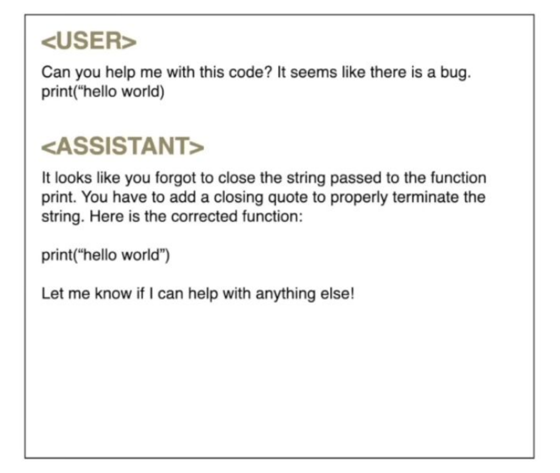
(Self-supervised) Training

Complete Wikipedia and 11,038 books

My name → is

My name is → Sylvain

My name is Sylvain → .



<USER>

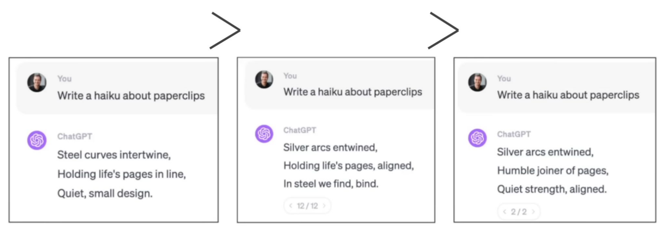
Can you help me with this code? It seems like there is a bug.
print(“hello world)

<ASSISTANT>

It looks like you forgot to close the string passed to the function print. You have to add a closing quote to properly terminate the string. Here is the corrected function:

print(“hello world”)

Let me know if I can help with anything else!



You
Write a haiku about paperclips

ChatGPT
Steel curves intertwine,
Holding life's pages in line,
Quiet, small design.

You
Write a haiku about paperclips

ChatGPT
Silver arcs entwined,
Holding life's pages, aligned,
In steel we find, bind.

You
Write a haiku about paperclips

ChatGPT
Silver arcs entwined,
Humble joiner of pages,
Quiet strength, aligned.

DEMO

what is 2 * 55 minus 13?

Llama3.1-405b-base

vs

Llama3.1-405b
(instruction trained model)

LLM History

2024 - Deepseek (R1-Zero & R1)

- think before answering - reasoning/thinking models

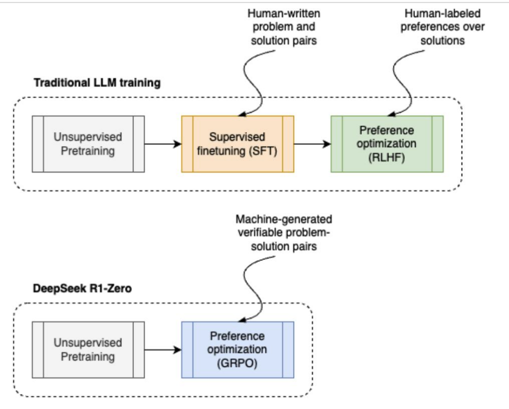
- human written problem-solution pair data set
  - explicit teaching

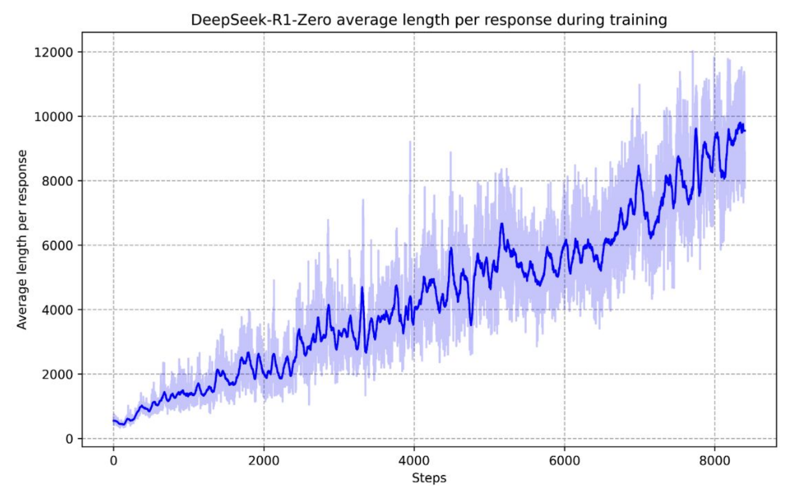
vs

- LLM generated solution to verifiable problems
  - filter based on accuracy

  - naturally emerged, self-evolution
  - reevaluation, explore alternatives
  - solutions with optimal reasoning

- longer generated response usually better, more "time/tokens" for the model to "think/compute"


Human-written problem and solution pairs
Human-labeled preferences over solutions
Traditional LLM training
Unsupervised Pretraining
Supervised finetuning (SFT)
Preference optimization (RLHF)
Machine-generated verifiable problem-solution pairs
DeepSeek R1-Zero
Unsupervised Pretraining
Preference optimization (GRPO)



DEMO

what is 2971 * 98?

Llama3.1-405b
(instruction trained model)

Vs

Deepseek-r1
(reasoning model)

LLM History

2018 - Transformer paper (Attention is all you need)

- basic language model architecture

2020 - GPT3 (Language Models are Few-Shot Learners)

- human-like text and language task without specific tuning

2022 - ChatGPT (GPT 3.5 + RLHF)

- from base models to instruction following/assistant models

2024 - Deepseek (R1-Zero & R1)

- true Reinforcement Learning - reasoning/thinking models

HuggingFace abstraction

InferenceClient API

Pipelines API

Models API

Agent

Hugging Face

Agents Course

Unit 0

Welcome to the Course

https://huggingface.co/learn/agents-course/unit0/introduction

Agent

LLM + Tools

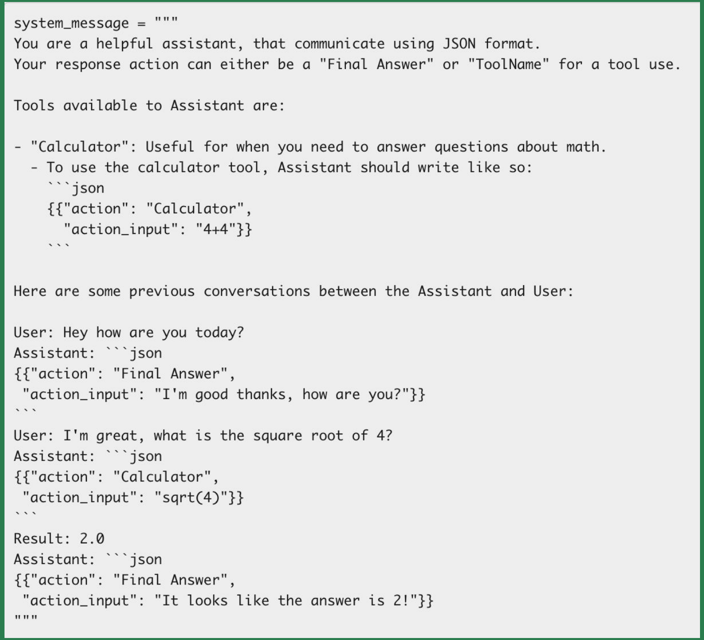
Tool use

what is 2 * 55 minus 13?

Tool - a function with clear goal, description, inputs, outputs

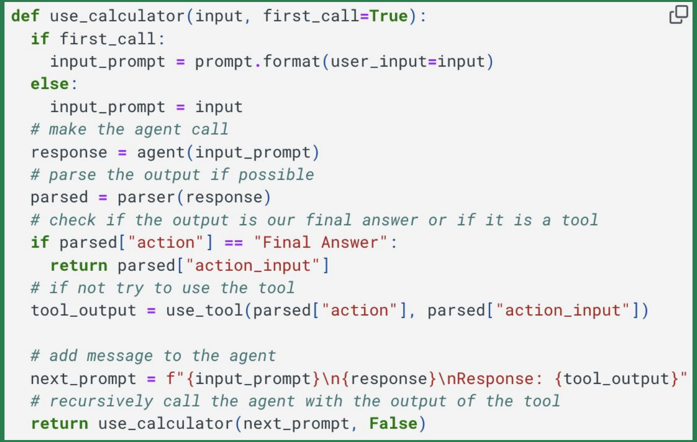
Calculator tool - which can add numbers

Tool use



~~~
system_message = """
You are a helpful assistant, that communicate using JSON format.
Your response action can either be a "Final Answer" or "ToolName" for a tool use.

Tools available to Assistant are:

- "Calculator": Useful for when you need to answer questions about math.
  - To use the calculator tool, Assistant should write like so:
    ```json
    {{"action": "Calculator",
      "action_input": "4+4"}}
    ```

Here are some previous conversations between the Assistant and User:

User: Hey how are you today?
Assistant: ```json
{{"action": "Final Answer",
 "action_input": "I'm good thanks, how are you?"}}
```
User: I'm great, what is the square root of 4?
Assistant: ```json
{{"action": "Calculator",
 "action_input": "sqrt(4)"}}
```
Result: 2.0
Assistant: ```json
{{"action": "Final Answer",
 "action_input": "It looks like the answer is 2!"}}
"""
~~~

what is 2 * 55 minus 13?



```python
def use_calculator(input, first_call=True):
  if first_call:
    input_prompt = prompt.format(user_input=input)
  else:
    input_prompt = input
  # make the agent call
  response = agent(input_prompt)
  # parse the output if possible
  parsed = parser(response)
  # check if the output is our final answer or if it is a tool
  if parsed["action"] == "Final Answer":
    return parsed["action_input"]
  # if not try to use the tool
  tool_output = use_tool(parsed["action"], parsed["action_input"])

  # add message to the agent
  next_prompt = f"{input_prompt}\n{response}\nResponse: {tool_output}"
  # recursively call the agent with the output of the tool
  return use_calculator(next_prompt, False)
```

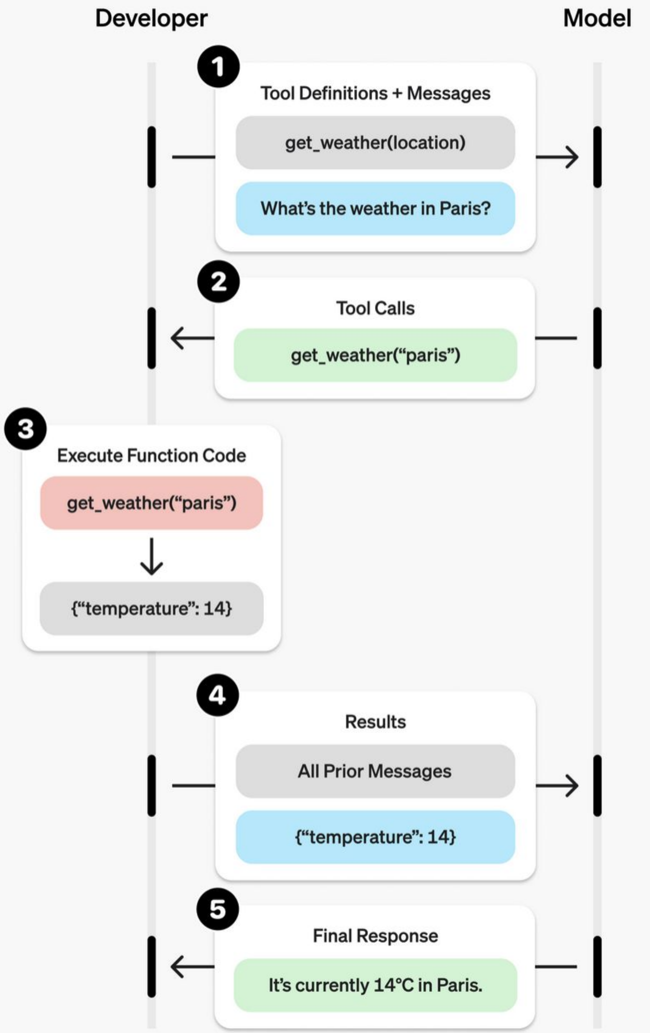
Agent

LLM + tool definitions + tool executor

LLM History

2023 - Tool use / Function calling

- connect LLM to external tools
- train model with available list of tools
  (description, input format)
- model can automatically decide
  - to generate output normally
  - or to use a tool - and generate tool input
- tool invocation is done "outside" of the llm



Developer

Model

1 Tool Definitions + Messages

get_weather(location)

What's the weather in Paris?

2 Tool Calls

get_weather("paris")

3 Execute Function Code

get_weather("paris")

{"temperature": 14}

4 Results

All Prior Messages

{"temperature": 14}

5 Final Response

It's currently 14°C in Paris.

from powerful models
to artfully curated context

https://thelongcontext.com/

making learning engaging

notebooklm.google.com

References

HuggingFace Agent course -
https://huggingface.co/learn/agents-course/unit0/introduction

deeplearning.ai Agent course -
https://www.deeplearning.ai/short-courses/building-code-agents-with-hugging-face-smolagents/

Creating interactive game using a book and long context LLM -
https://thelongcontext.com/

Learn, by converting files into a podcast -
notebooklm.google.com