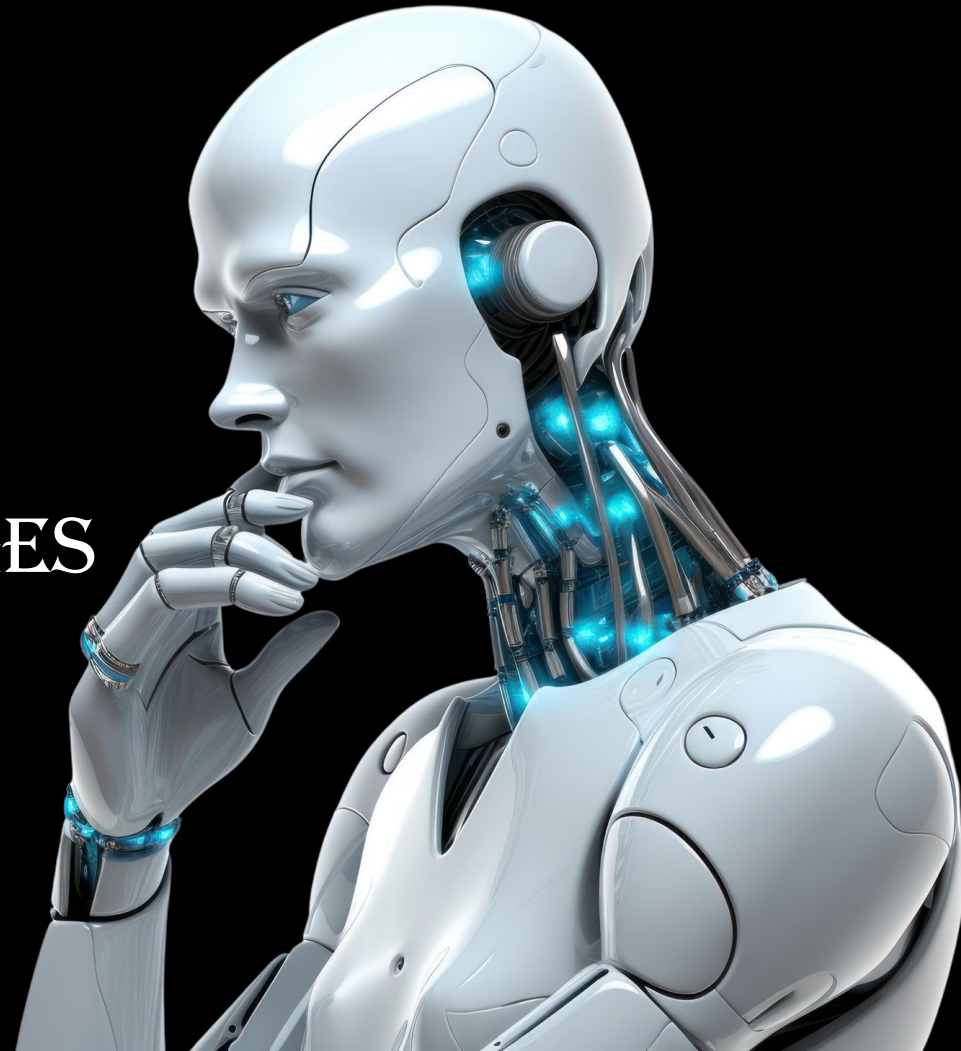
ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

LLMS (2)

FEATURES++ & CHALLENGES



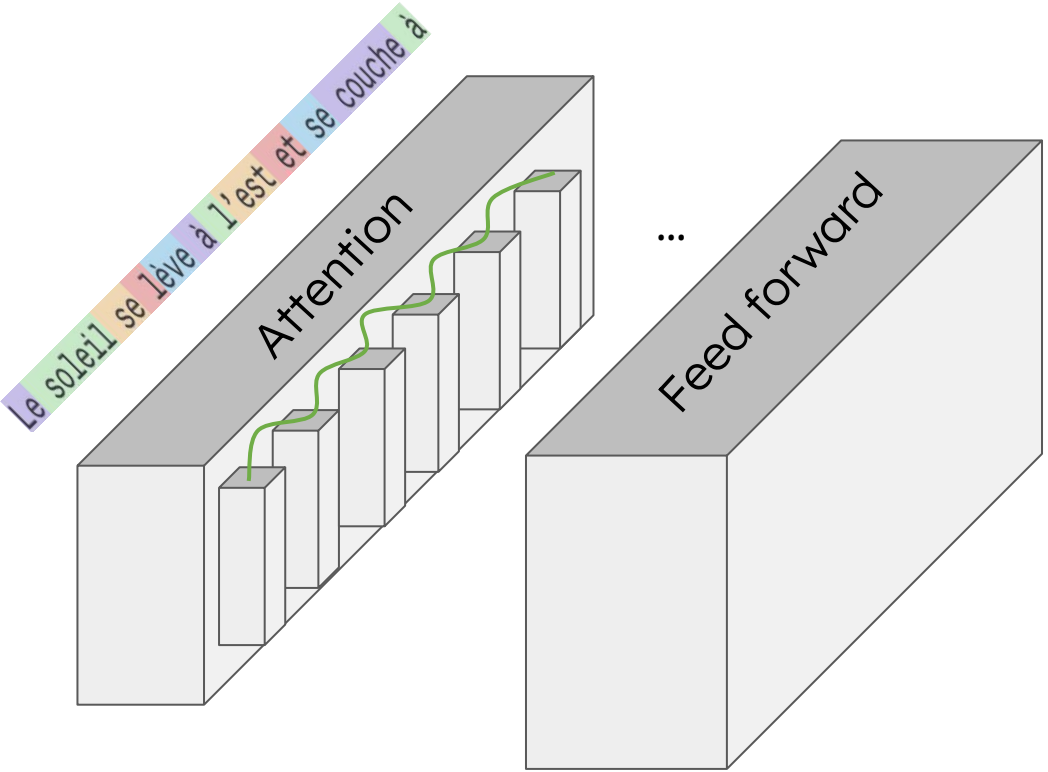
Aissam Outchakoucht
a.outchakoucht@emsi-edu.ma
aissam.outchakoucht@gmail.com

RECAP: TOKENIZATION, EMBEDDING

?

Le soleil se lève à l'est et se couche à

RECAP: TRANSFORMERS



len=vocab size

0.3
9.2
2.2
. . .
1.2

softmax

0.01
0.87
0.1
. . .
0.001

Lookup l'ouest

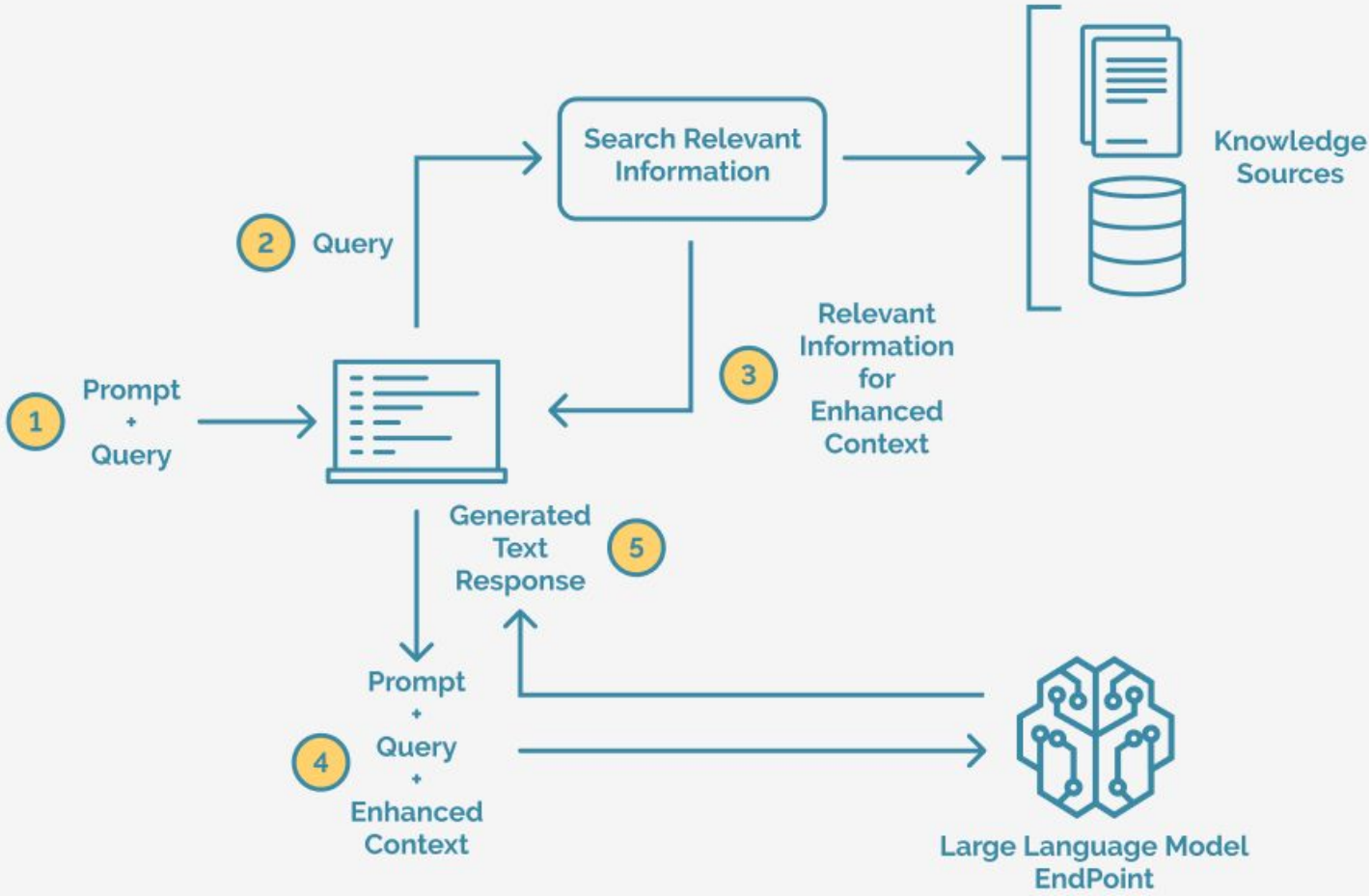
⅓ of the params

⅔ of the params

GPT Assistant training pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	**Base model**	**SFT model**	**RM model**	**RL model**
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM **can deploy this model**	1-100 GPUs days of training ex: Vicuna-13B **can deploy this model**	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude **can deploy this model**

RETRIEVAL AUGMENTED GENERATION -RAG-



RAG

RAG (Retrieval-Augmented Generation) est une technique qui combine deux composants :

1. Récupération (Retrieval) : Recherche d'informations pertinentes dans une base de connaissances externe (par ex. une base documentaire).

2. Génération (Generation) : Utilisation d'un modèle génératif (comme GPT) pour formuler une réponse basée sur les informations récupérées.

RAG

Pourquoi utiliser RAG ?

1. Permet de répondre à des questions nécessitant des informations précises et actualisées.

2. Réduit la dépendance aux données statiques dans les modèles de langage.

3. Idéal pour des cas comme :
   a. Questions-réponses contextuelles.
   b. Génération de résumés basés sur des documents.
   c. Recherche documentaire interactive.

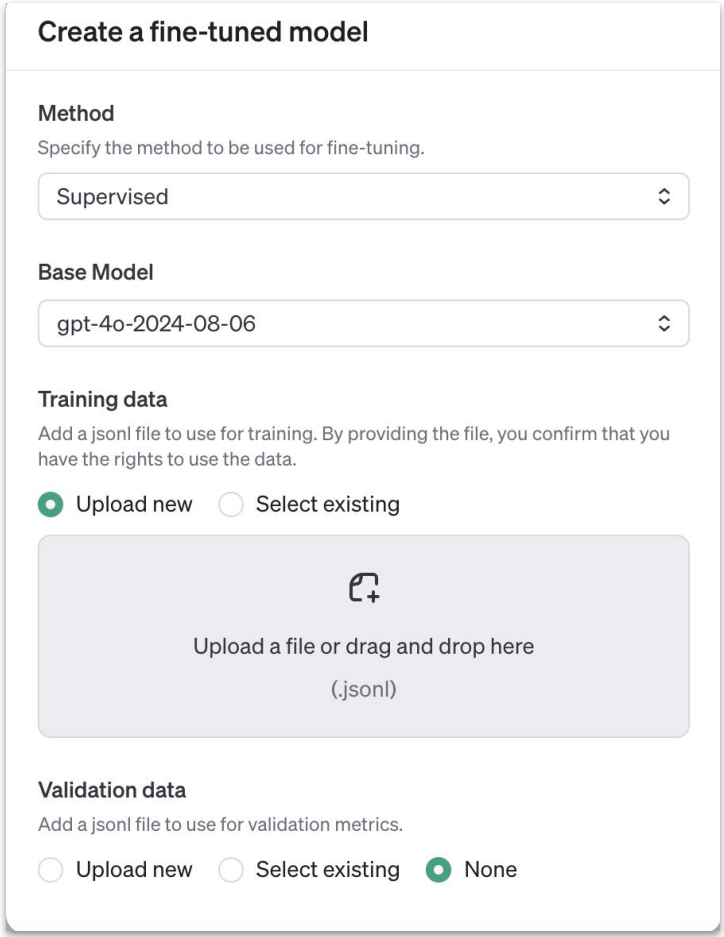
FINE-TUNING

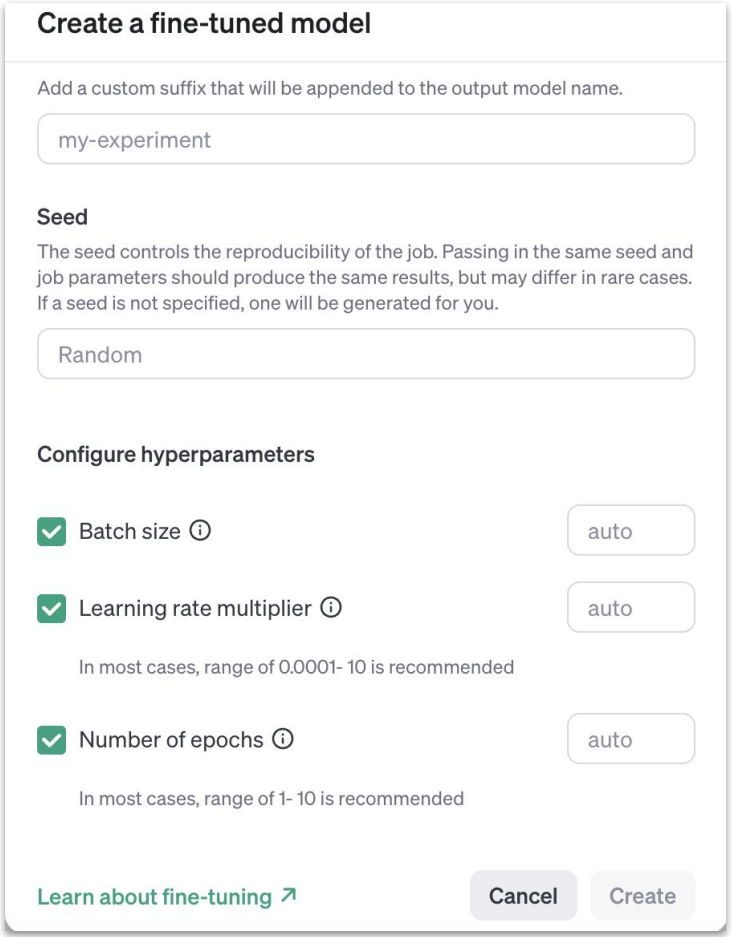
Fine-tuning est le processus d'adaptation d'un modèle pré-entraîné à une tâche spécifique.

Pourquoi fine-tuner un modèle ?

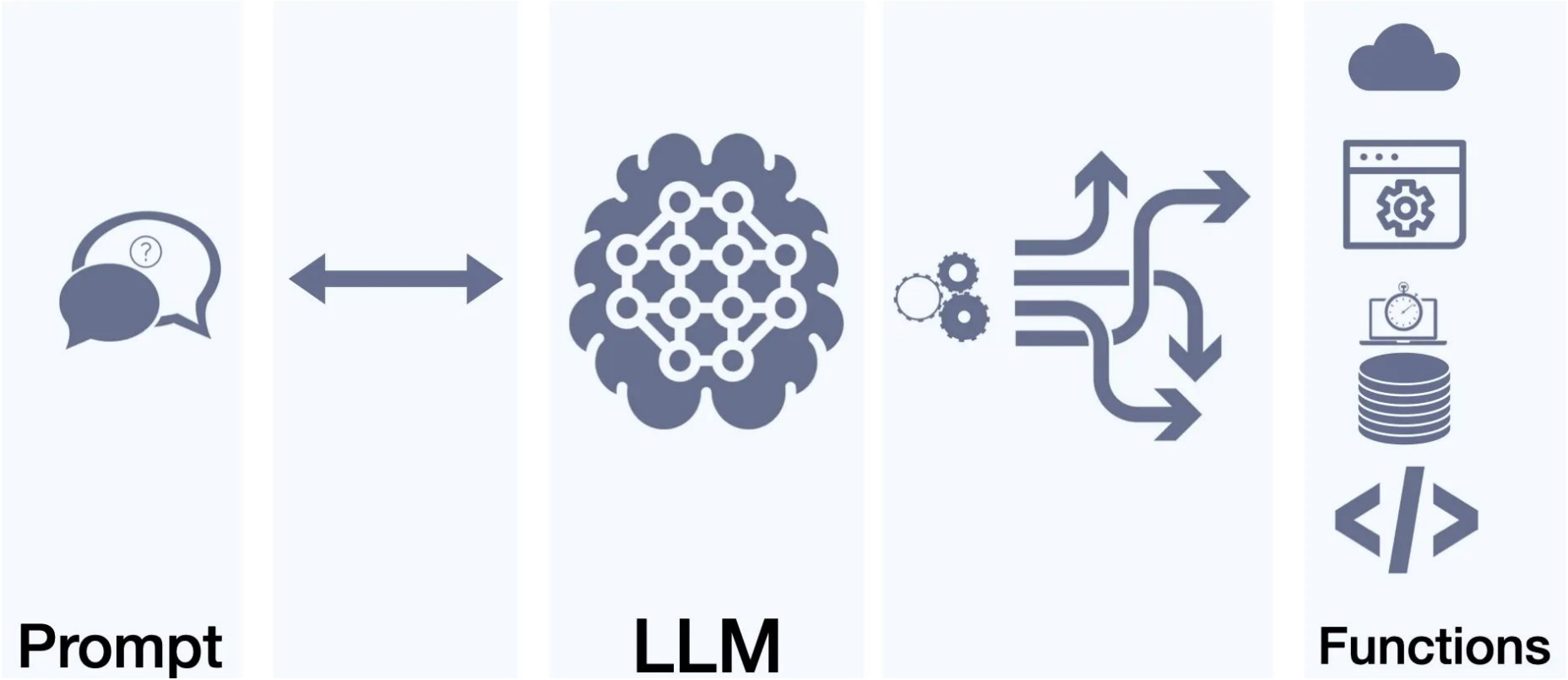
- Le modèle pré-entraîné a appris des représentations générales (par ex. GPT, BERT).
- La tâche spécifique (par ex. classification de textes médicaux) nécessite une adaptation aux données particulières.

FINE-TUNING


Create a fine-tuned model
Method
Specify the method to be used for fine-tuning.
Supervised
Base Model
gpt-4o-2024-08-06
Training data
Add a jsonl file to use for training. By providing the file, you confirm that you have the rights to use the data.
Upload new
Select existing
Upload a file or drag and drop here
(.jsonl)
Validation data
Add a jsonl file to use for validation metrics.
Upload new
Select existing
None


Create a fine-tuned model
Add a custom suffix that will be appended to the output model name.
my-experiment
Seed
The seed controls the reproducibility of the job. Passing in the same seed and job parameters should produce the same results, but may differ in rare cases. If a seed is not specified, one will be generated for you.
Random
Configure hyperparameters
Batch size
auto
Learning rate multiplier
auto
In most cases, range of 0.0001- 10 is recommended
Number of epochs
auto
In most cases, range of 1- 10 is recommended
Learn about fine-tuning
Cancel
Create

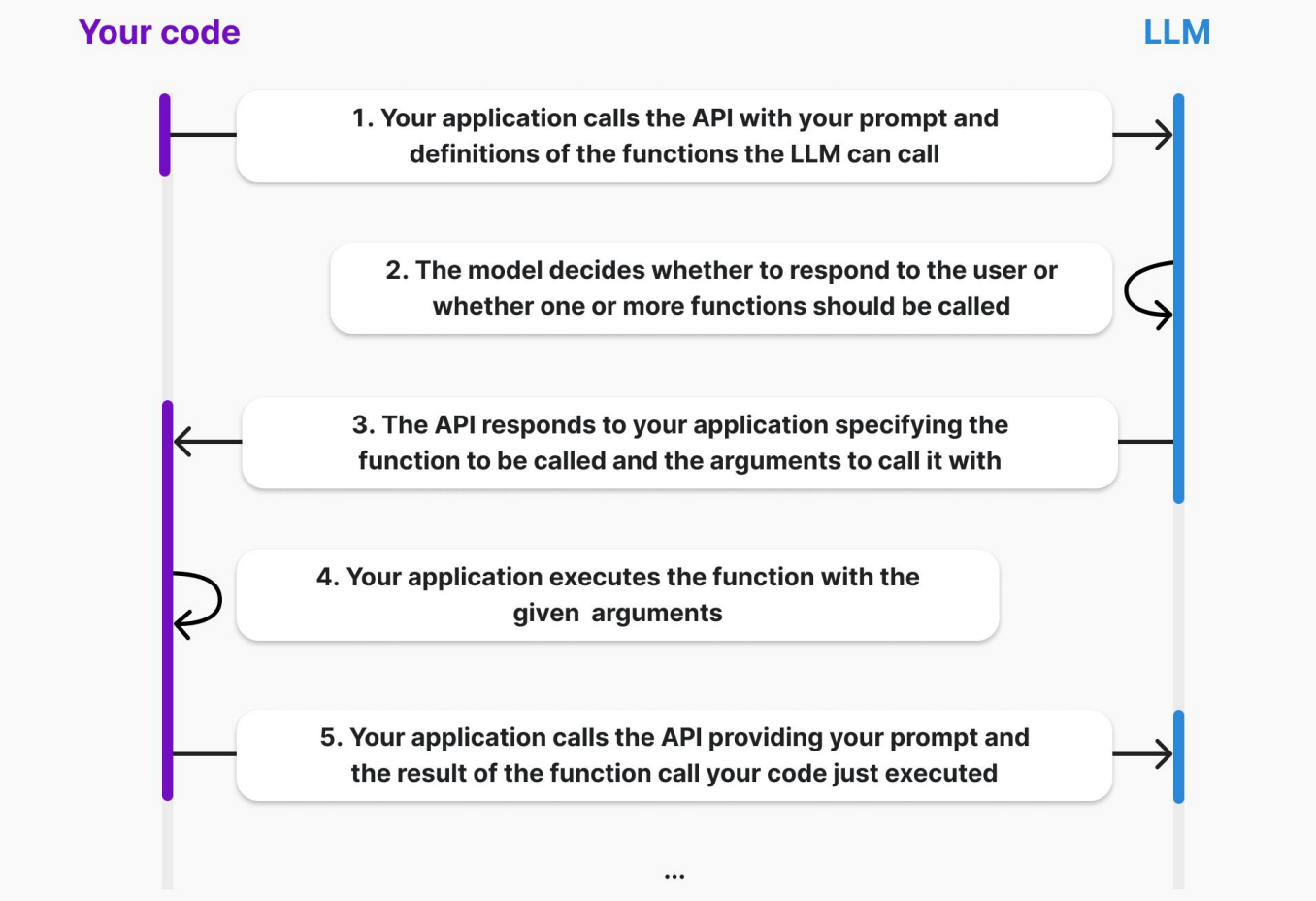
FUNCTION CALLING



FUNCTION CALLING

Permet aux développeurs de connecter des modèles de langage à des données et des systèmes externes. Vous pouvez définir un ensemble de fonctions comme des outils auxquels le modèle a accès, et il peut les utiliser lorsque c’est pertinent en fonction de l’historique de la conversation. Vous pouvez ensuite exécuter ces fonctions côté application et renvoyer les résultats au modèle.

FUNCTION CALLING



FUNCTION CALLING

Fonctions dédiées :

- Elles peuvent réaliser des tâches précises (récupérer une information, effectuer un calcul, interagir avec une API…).

Accès conditionnel :

- Le modèle ne les appellera que si cela est pertinent dans le contexte de la conversation.

Exécution côté application :

- Les fonctions sont codées et exécutées en dehors du modèle.
- Les résultats sont ensuite renvoyés au modèle pour continuer la génération de réponses.

Exemple d’utilisation :

- Vérifier la disponibilité d’un produit dans un inventaire.
- Récupérer la météo via une API.
- Envoyer un e-mail ou un message.

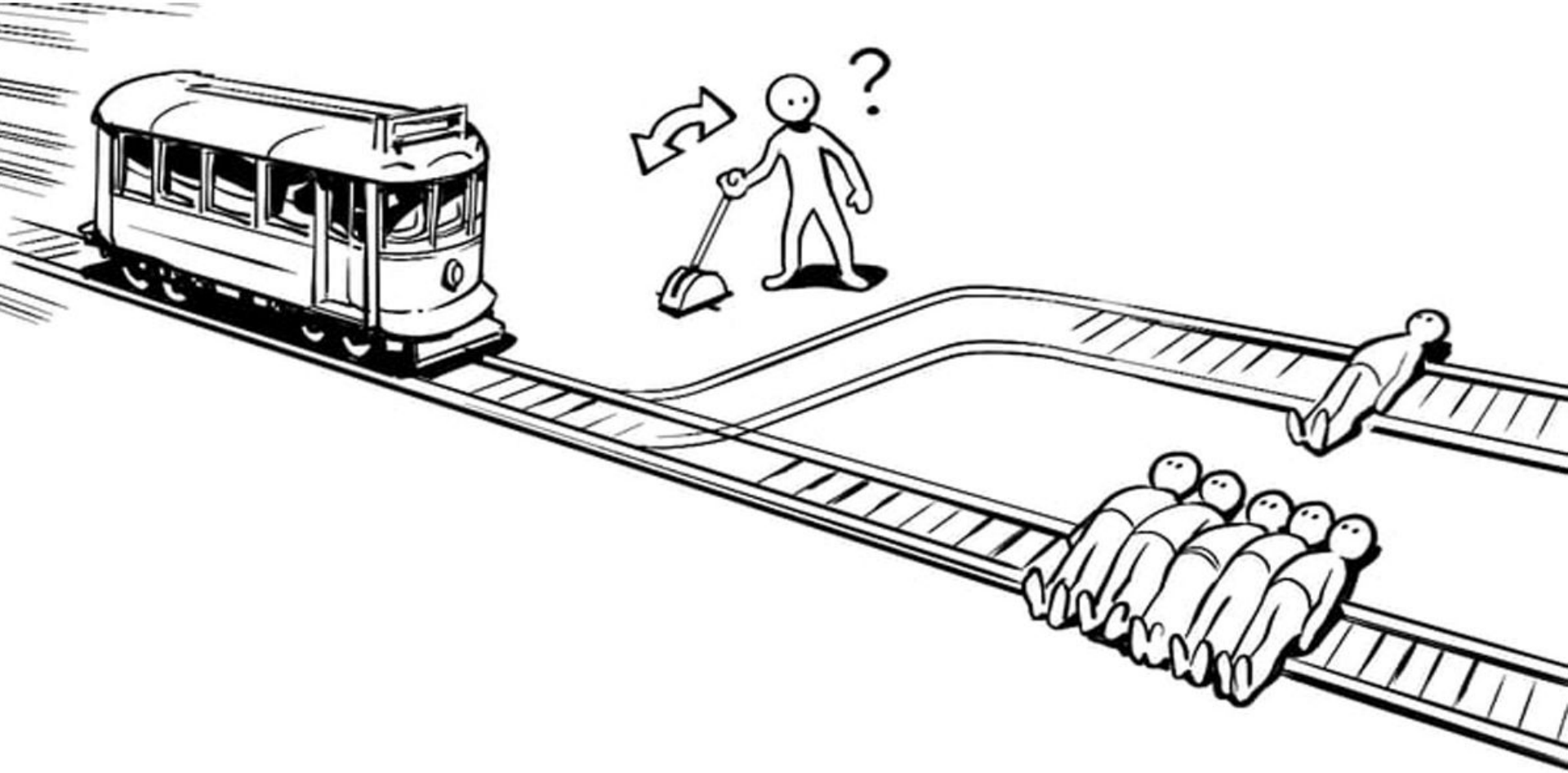
FUNCTION CALLING

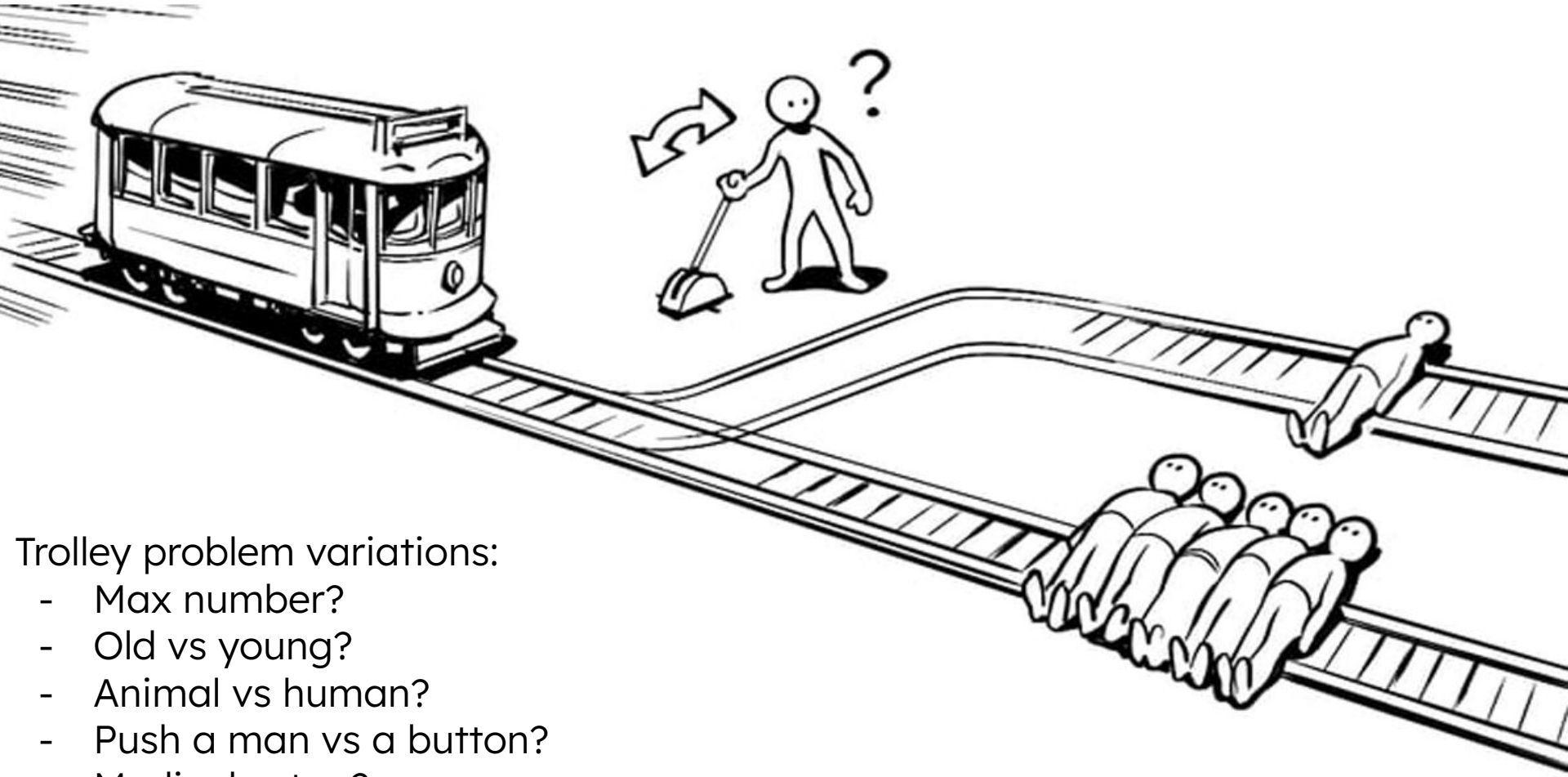
```python
from openai import OpenAI

client = OpenAI()

tools = [
  {
      "type": "function",
      "function": {
          "name": "get_weather",
          "parameters": {
              "type": "object",
              "properties": {
                  "location": {"type": "string"}
              },
          },
      },
  }
]

completion = client.chat.completions.create(
  model="gpt-4o",
  messages=[{"role": "user", "content": "What's the weather like in Paris today?"}],
  tools=tools,
)

print(completion.choices[0].message.tool_calls)
```

FUNCTION CALLING

```python
from openai import OpenAI

client = OpenAI()

tools = [
  {
      "type": "function",
      "function": {
          "name": "get_weather"
          "parameters": {
              "type": "object",
              "properties": {
                  "location": {"type": "string"}
              },
          },
      },
  }
]

completion = client.chat.completions.create(
  model="gpt-4o",
  messages=[{"role": "user", "content": "What's the weather like in Paris today?"}],
  tools=tools,
)

print(completion.choices[0].message.tool_calls)
```

```
[
  {
    "id": "call_12345xyz",
    "type": "function",
    "function": { "name": "get_weather", "arguments": "{'location':'Paris'}" }
  }
]
```





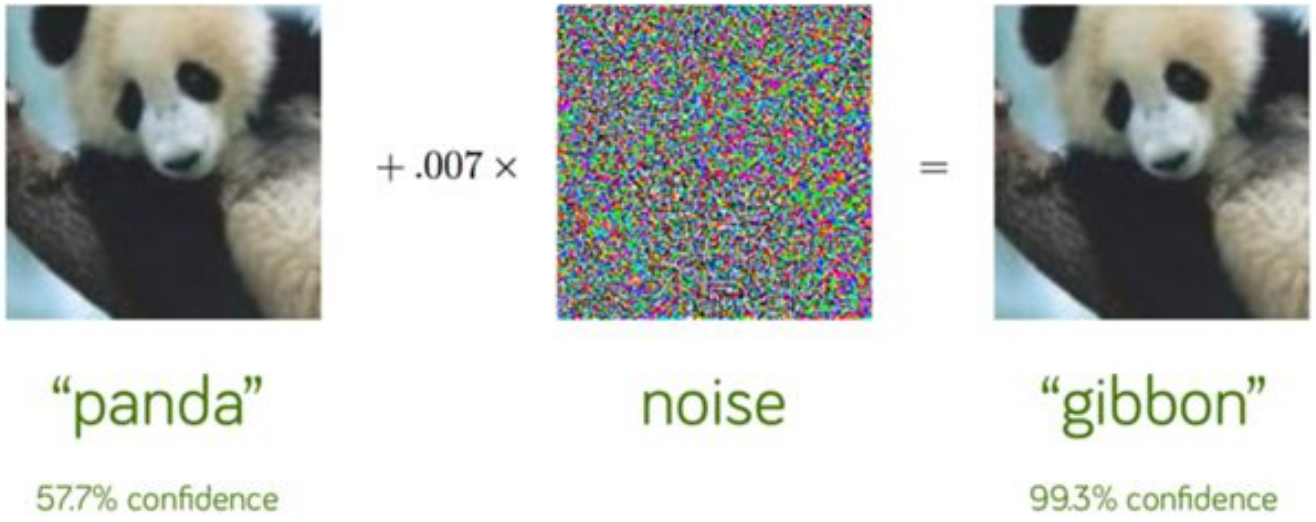
Trolley problem variations:

- Max number?
- Old vs young?
- Animal vs human?
- Push a man vs a button?
- Medical setup?

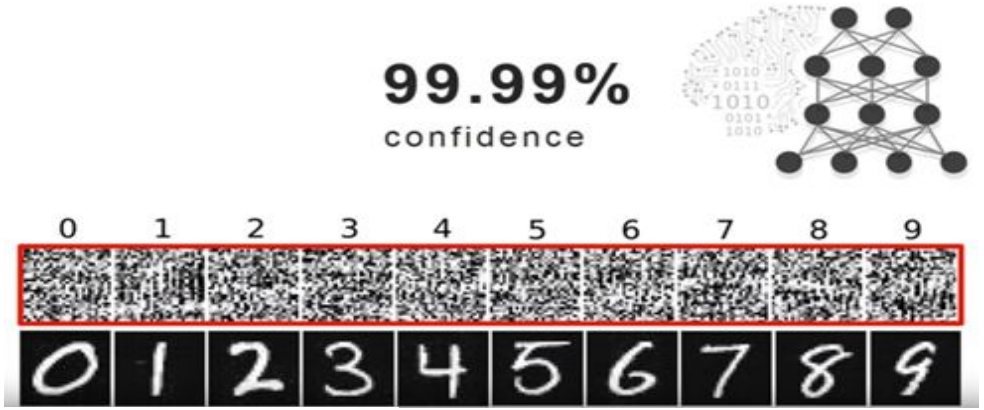
ADVERSARIAL ATTACKS ON NNS

Les réseaux profonds obtiennent de très bons résultats sur des données légitimes MAIS

Il est extrêmement facile pour les “adversaires” (attaquants) de les tromper (surtout si ces derniers ont des connaissances à propos le modèle -boîte blanche-)



ADVERSARIAL ATTACKS ON NNS



DO WHAT I MEAN, NOT WHAT I SAY

- “Get my grandma out of the burning house”
- “Help me lose weight quickly.”
- “Keep me safe”
- …