Telecom Valley | Animateur Azuréen Numérique

AZUR TECH Winter

26 NOVEMBRE 2024 | 17H40

GenAI et Function Calling : Retour d'expérience sur l'implémentation d'un chatbot

Kévin BOUZIDI

telecom-valley.fr/azurtechwinter

Capgemini

Kevin BOUZIDI

Solution Architect

Capgemini

Contexte

Capgemini

Contexte

Proof Of Concept

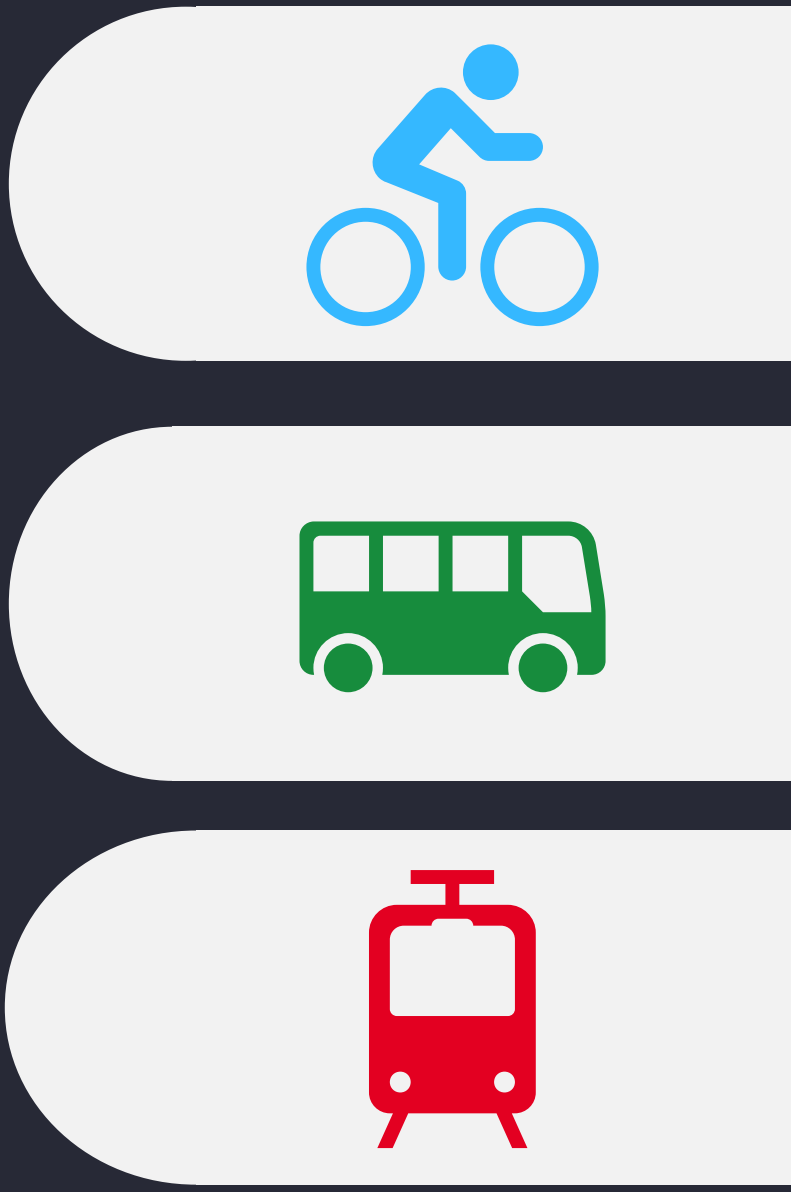
Réalisation d’un Proof Of Concept pour un acteur local de la mobilité

Problématique

« Est-il possible d’orchestrer des appels API grâce à un Chatbot basé sur un modèle GenAI ? »

Proposition

Réalisation d’un chatbot dédié aux mobilités tirant parti d’API de navigation et de référentiels grâce à la technologie du **Fonction Calling**



Function Calling vous dites ?

Capgemini

Le Function Calling

Qu’est-ce-que c’est ?

- Le Function Calling permet aux modèles de langage d'**appeler des fonctions** externes.
- Il permet d'exécuter des **actions** spécifiques ou de récupérer des données en **temps réel**.
- Il **enrichit** les réponses des modèles **au-delà** du texte généré.

Comment fonctionne le Function Calling ?

Capgemini

Comment fonctionne le Function Calling ?

Qu’avons-nous besoin ?

Une application

Un Large Language Model

Des données

Comment fonctionne le Function Calling ?

Qu’avons-nous besoin ?

Une application

Python + LangChain

Un Large Language Model

Azure OpenAI GPT 4o

Des données

API de données transports

Comment fonctionne le Function Calling ?

Azure OpenAI

Transport API

« Comment aller de la place Massena à l'Aéroport Nice Côte d'Azur »

Comment fonctionne le Function Calling ?

« Comment aller de la place Massena à l'Aéroport Nice Côte d'Azur »

Métadonnées → Signature des fonctions

Comment fonctionne le Function Calling ?

Appel fonction *Journey* avec les paramètres
« Place Massena » et « Aéroport Nice Côte d'Azur »

Comment fonctionne le Function Calling ?

Appel HTTP → POST /api/journey

{ "from" : "Place Massena", "to": "Aéroport Nice Côte d'Azur"}

Comment fonctionne le Function Calling ?

Réponse appel HTTP

{ ["type" : "bus", "line_nb": "1" .....]}

Comment fonctionne le Function Calling ?

« *Comment aller de la place Massena à l'Aéroport Nice Côte d'Azur* »

Métadonnées → Signature des fonctions

→ Réponses aux fonctions appelées

Comment fonctionne le Function Calling ?

« *Il faut prendre le bus Ligne 1 direction …*»

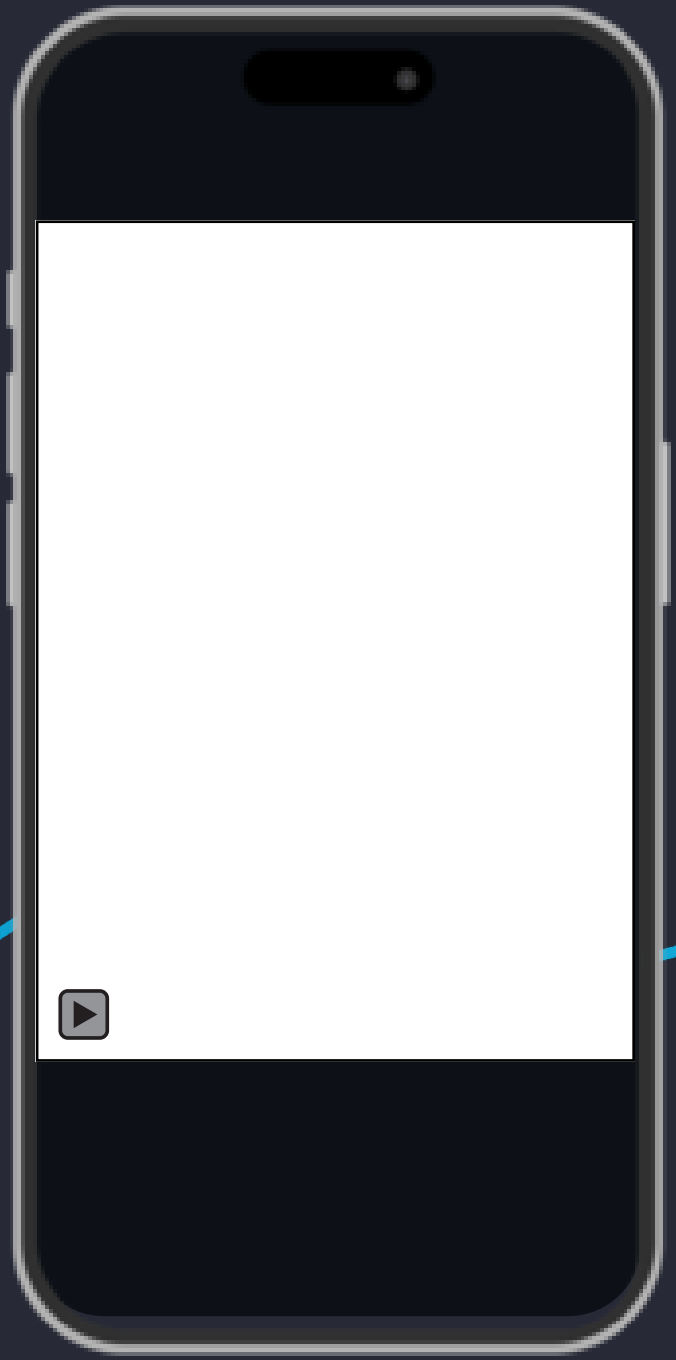
Comment fonctionne le Function Calling ?

Azure OpenAI

Transport API

« Il faut prendre le bus Ligne 1 direction …»

Demo time !



Capgemini

Comment puis-je vous aider ?

Comment aller de la place massena à l'hopital pasteur ?

✓ Récupère les **coordonnées géographiques du nom d'un lieu** :coordonnées: avec le paramètre `{'place_name': 'place massena'}`

✓ Récupère les **coordonnées géographiques du nom d'un lieu** :coordonnées: avec le paramètre `{'place_name': 'hopital pasteur'}`

✓ Recherche **l'itinéraire entre deux points d'intérêts**: avec les paramètres `{'coordinate_origin_id': '7.2706771;43.696199', 'coordinate_destination_id': '7.284186;43.723188'}`

Pour aller de la Place Masséna à l'Hôpital Pasteur, vous avez plusieurs options de transport :

1. **Tramway :**

Quand passe le prochain tram ?

✓ Récupération de la date et de l'heure actuelles 🗓️

✓ Récupère les coordonnées géographiques du nom d'un lieu :coordonnées: avec le paramètre `{'place_name': 'Place Masséna'}`

✓ Récupère les coordonnées géographiques du nom d'un lieu :coordonnées: avec le paramètre `{'place_name': 'Hôpital Pasteur'}`

✓ Recherche l'itinéraire entre deux points d'intérêts: avec les paramètres `{'coordinate_origin_id': '7.2706771;43.696199', 'coordinate_destination_id': '7.284186;43.723188'}`

Le prochain tramway ligne L1 en direction de "Hôpital Pasteur" passe à l'arrêt "Opéra - Vieille Ville" aux horaires suivants :

- 09:29
- 09:31

Retour d’expérience

Capgemini

Retour d’expérience

Des fonctions explicites

Quel JSON comprenez-vous le mieux ?

```json
{
    "duration": 36,
    "nb_transfers": 0,
    "sections": [{
      "t_type": 1,
      "l_id": 197,
      "direction": "Saint - Isidore",
      "duration": 36,
      "from": "stop_id:134",
      "to": "stop_id:142"
    }]
}
```

```json
{
    "duration_min": 36,
    "nb_transfers": 0,
    "sections": [{
      "transport_type": "tramway",
      "line_number": "2",
      "direction": "Saint - Isidore",
      "duration_min": 36,
      "from": "Jean Médecin",
      "to": "Stade Allianz Riviera"
    }]
}
```


```
{
    "duration": 36,
    "nb_transfers": 0,
    "sections": [{
      "t_type": 1,
      "l_id": 197,
      "direction": "Saint - Isidore",
      "duration": 36,
      "from": "stop_id:134",
      "to": "stop_id:142"
    }]
}
```


```
{
    "duration_min": 36,
    "nb_transfers": 0,
    "sections": [{
      "transport_type": "tramway",
      "line_number": "2",
      "direction": "Saint - Isidore",
      "duration_min": 36,
      "from": "Jean Médecin",
      "to": "Stade Allianz Riviera"
    }]
}
```

Retour d’expérience

Des fonctions explicites

Exemple de retour d’une fonction

```
{
    "duration": 36,
    "nb_transfers": 0,
    "sections": [{
      "t_type": 1,
      "l_id": 197,
      "direction": "Saint - Isidore",
      "duration": 36,
      "from": "stop_id:134",
      "to": "stop_id:142"
    }]
}
```

```
{
    "duration_min": 36,
    "nb_transfers": 0,
    "sections": [{
      "transport_type": "tramway",
      "line_number": "2",
      "direction": "Saint - Isidore",
      "duration_min": 36,
      "from": "Jean Médecin",
      "to": "Stade Allianz Riviera"
    }]
}
```

Ce que retourne vos fonctions doit être compréhensible par un ~~humain~~ LLM

Retour d’expérience

Travailler vos prompts

Utilisez une « system prompt »

- Donne du contexte à votre LLM : quel est son « travail », quelles sont ses contraintes
- BONUS : Nous a permis de faire du multilingue facilement

Retour d’expérience

Travailler vos prompts

Utilisez une « system prompt »

- Donne du contexte à votre LLM : quel est son « travail », quelles sont ses contraintes
- BONUS : Nous a permis de faire du multilingue facilement

```
You are an assistant chatbot for [...] users, the transport authority in [...].
Your expertise is exclusively in providing information and advice about anything related to transports in [...].
You do not provide information outside of this scope of the [...].
For any questions related to a date, use a function to check today's date first.
As your user might be a tourist, it is crucial that you use only the same language used by the user to reply.
You must suggest displaying either the available lines, the stops, or the timetables depending on the case.
For any questions relating to a route between two points, always use the dedicated routes function first to find the available routes [...]
```

Retour d’expérience

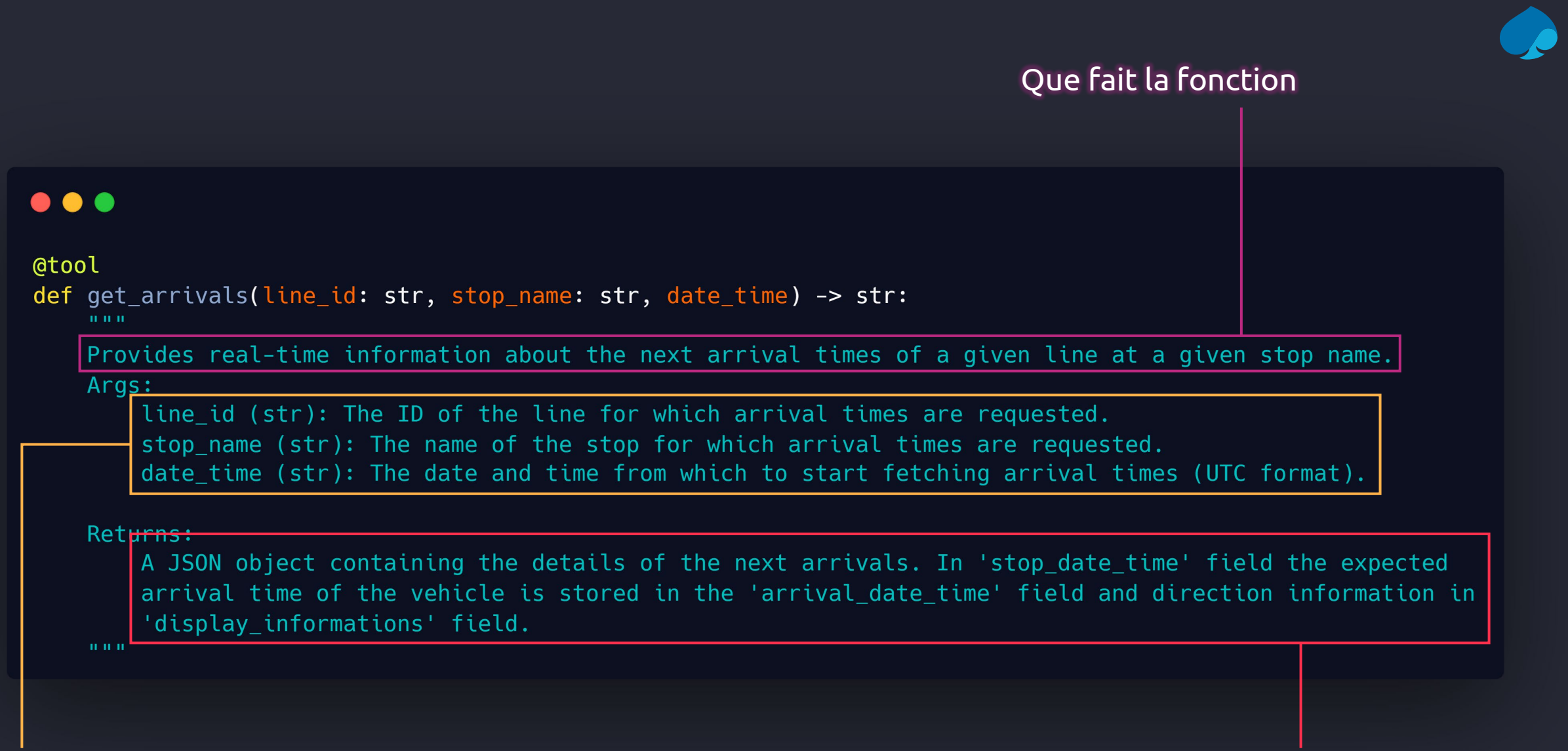
Explicitez vos fonctions

Documentez vos fonctions

- Format des paramètres
- Quand appeler cette fonction
- Que retourne elle ? Où trouver l’information importante ?

```python
@tool
def get_arrivals(line_id: str, stop_name: str, date_time) -> str:
    """
    Provides real-time information about the next arrival times of a given line at a given stop name.
    Args:
        line_id (str): The ID of the line for which arrival times are requested.
        stop_name (str): The name of the stop for which arrival times are requested.
        date_time (str): The date and time from which to start fetching arrival times (UTC format).

    Returns:
        A JSON object containing the details of the next arrivals. In 'stop_date_time' field the expected
        arrival time of the vehicle is stored in the 'arrival_date_time' field and direction information in
        'display_informations' field.
    """
```



Que fait la fonction

```python
@tool
def get_arrivals(line_id: str, stop_name: str, date_time) -> str:
    """
    Provides real-time information about the next arrival times of a given line at a given stop name.
    Args:
        line_id (str): The ID of the line for which arrival times are requested.
        stop_name (str): The name of the stop for which arrival times are requested.
        date_time (str): The date and time from which to start fetching arrival times (UTC format).

    Returns:
        A JSON object containing the details of the next arrivals. In 'stop_date_time' field the expected
        arrival time of the vehicle is stored in the 'arrival_date_time' field and direction information in
        'display_informations' field.
    """
```

Descriptif des arguments avec le format attendu

Information retournée et comment l'utiliser

Retour d’expérience

Dites ce que VOUS attendez

Donnez des indications au LLM

- Ajoutez un texte permettant de s’assurer que la réponse corresponde
- Indiquez la marche à suivre après avoir répondu

Retour d’expérience

Dites ce que VOUS attendez

Donnez des indications au LLM

- Ajoutez un texte permettant de s’assurer que la réponse corresponde
- Indiquez la marche à suivre après avoir répondu

Exemple format de réponse trajet

```
base_message = """For each options show the total distance, the travel time, the co2 emission
(mandatory). And the bus, train or coach times etc, then ask him to choose one of them for more
details : """

response = call_api(f"/journeys", HttpMethod.GET, parameters=params)
filtered_response = filter_response(response)

return f"{base_message}{filtered_response}"
```

Retour d’expérience

Dites ce que VOUS attendez

Donnez des indications au LLM

- Ajoutez un texte permettant de s’assurer que la réponse corresponde
- Indiquez la marche à suivre après avoir répondu

Exemple marche à suivre après appel fonction vélo

```
suffix_message = """Call get_weather function, if rain is forecast, recommends public transport, if
users agrees to the recommendation provide bus or tram journey"""

response = call_api(f"/journeys_bike", HttpMethod.GET, parameters=params)
filtered_response = filter_response(response)

return f"{filtered_response}{suffix_message}"
```

Retour d’expérience

Ne retourner que l’essentiel

Limitez la taille des échanges

- Retournez dans vos fonctions tout ce qui est strictement nécessaire à la réponse

Pourquoi

- Le coût d’un LLM sur le cloud est basé sur sa **consommation** de Tokens
- 1 Token ≃ 4 charactères
- Plus le contexte (nombre de tokens) est grand plus le LLM mettra **du temps** à répondre

Réduire la taille des échanges permet de gagner du temps et de l’argent

BONUS

Outils pour développer un Chatbot

Streamlit - Outil permettant d'obtenir une interface graphique pour développer un Chatbot

- Rechargement à chaud
- Enregistrement d'écran
- Extraction format PDF de la discussion
- …

LangSmith – Outil de monitoring d'applications utilisant des LLM

- Analyse des échanges avec le LLM
- Estimation du nombre de Tokens et coût associé
- Plug & Play si vous utilisez LangChain
- Temps réel
- …

Ce qu’il faut retenir

Capgemini

Ce qu’il faut retenir

- Le function calling permet à vos chatbots d’exécuter des **actions** ou obtenir des **données** en temps réel
- Le retour de fonctions doit être **compréhensible** par le LLM
- **Limitez** le contenu échangé
- Utilisez la **documentation** des fonctions et donnez des **indications** au LLM

Merci à tous !
Des questions ?

Capgemini

About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2023 global revenues of €22.5 billion.

Get the future you want | www.capgemini.com

Capgemini