



image générée par DALL-E



Paul PETON  
*prompt engineer*  
*@Nantes*



Jonathan PACIFICO  
*chatbot trainer*  
*@AixEnProvence*

# Mon chatbot est devenu hypermnésique !

*Ou comment améliorer les  
prompts de GPT pour conserver  
les éléments les plus importants  
de la discussion*



Cette présentation n'a pas été  
rédigée par Microsoft 365 copilot  
(mais c'est peut-être la dernière...)



GPT :  
**Generative**  
Pre-trained  
Transformer

# GPT : **Generative** Pre-trained Transformer

## **Generative :**

création de « nouveaux » contenus  
en se basant sur des contenus existants



## Génération conditionnelle :

**prompt** => **complétion** par les mots les plus probables, en tenant compte des mots précédents

*je vais au cinéma pour voir un...*

# GPT : Generative **Pre-trained** Transformer

## Pre-Trained :

un premier entraînement sur de très grands **datasets**  
par **apprentissage auto-supervisé**

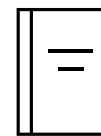
Auto-supervisé : reconstruire les données d'entrée

Contexte : *je vais au cinéma pour voir un...*

Prévision : *film*

Datasets : données Web, Wikipedia, livres numériques... en plusieurs langues

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%



WIKIPEDIA

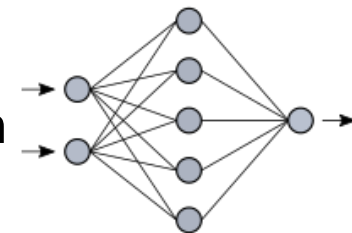




# GPT : Generative Pre-trained **Transformer**

## Transformer :

réseau de neurones profond mettant en œuvre un mécanisme d'**attention**



- le modèle comprend les **relations** entre des mots éloignés dans la phrase
- le modèle détecte le **contexte** et a une compréhension globale de la phrase

Exemples :

« je travaille pour un grand cabinet d'avocats »

« j'ai mangé ce midi un délicieux **avocat** »

*a lawyer in a suit sits in an office opposite a human-sized avocado (Bing Image Creator)*

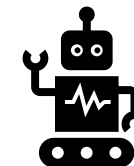




# ChatGPT : l'agent conversationnel

## Chat :

communication par échange de messages écrits  
bavardage ou « clavardage »



- spécialisation à la **conversation**

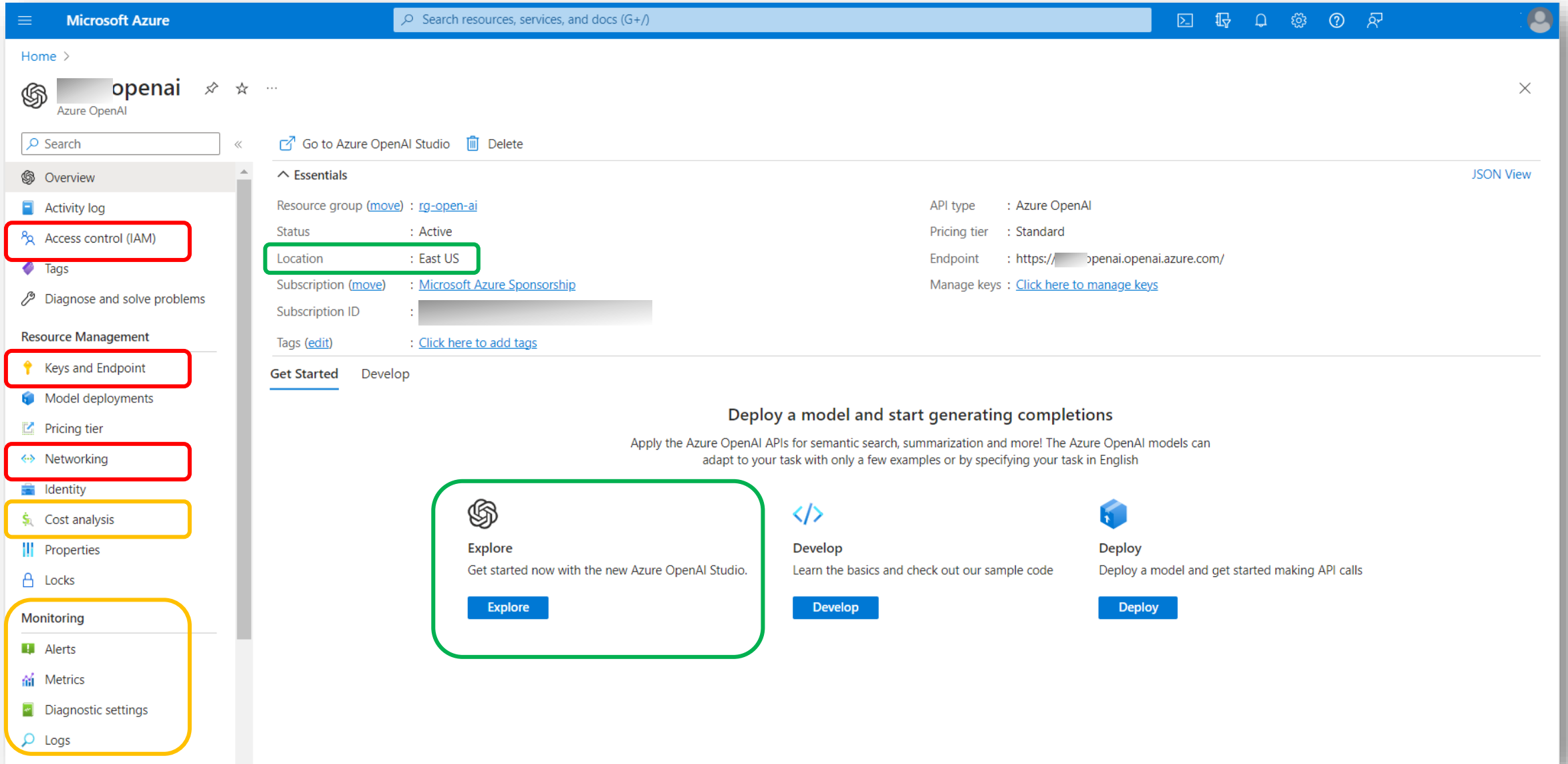
- imiter une réponse dans le style d'un être humain
- mais les autres tâches restent possibles : classification, traduction, résumé...



- optimisation par des **feedbacks** humains pour noter l'utilité, la véracité, la non-toxicité



# Le service cognitif OpenAI



**Microsoft Azure** Search resources, services, and docs (G+)

Home > **openai** Azure OpenAI

Search

Go to Azure OpenAI Studio Delete

**Overview**

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

**Resource Management**

- Keys and Endpoint
- Model deployments
- Pricing tier
- Networking
- Identity
- Cost analysis
- Properties
- Locks

**Monitoring**

- Alerts
- Metrics
- Diagnostic settings
- Logs

**Essentials** [JSON View](#)

Resource group (move) : [rg-open-ai](#)

Status : Active

Location : East US

Subscription (move) : [Microsoft Azure Sponsorship](#)

Subscription ID : [REDACTED]

Tags (edit) : [Click here to add tags](#)

API type : Azure OpenAI

Pricing tier : Standard

Endpoint : [https://\[REDACTED\].openai.azure.com/](#)

Manage keys : [Click here to manage keys](#)

**Get Started** Develop

**Deploy a model and start generating completions**

Apply the Azure OpenAI APIs for semantic search, summarization and more! The Azure OpenAI models can adapt to your task with only a few examples or by specifying your task in English

**Explore**  
Get started now with the new Azure OpenAI Studio.  
[Explore](#)

**Develop**  
Learn the basics and check out our sample code  
[Develop](#)

**Deploy**  
Deploy a model and get started making API calls  
[Deploy](#)



# Jouons dans le *playground*

<https://oai.azure.com/portal/playground>



← ↻ 🔒 https://oai.azure.com/portal/playground

Cognitive Services | Azure OpenAI Studio

Azure OpenAI Studio > GPT-3 Playground

Playground

GPT-3

Management

- Deployments
- Models
- File Management

Deployments: davinci-summarize

Examples: Load an example

View code

Parameters

Temperature: 1

Max length (tokens): 100

Stop sequences: Stop sequences

Top probabilities: 0.5

Frequency penalty: 0

Presence penalty: 0

Best of: 1

Pre-response text: Enter text

Post-response text: Enter text

Learn more

Generate Undo Regenerate Tokens: 0

# Jouons dans le *playground* : classification



This is a spam ham classifier.

text : "Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat... "

class : ham

text : "Ok lar... Joking wif u oni... "

class : ham

text : "Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's "

class : spam

text : "U dun say so early hor... U c already then say... "

class : ham

text : "Nah I don't think he goes to usf, he lives around here though "

class : ham

text : "FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv "

class : spam

text : "Even my brother is not like to speak with me. They treat me like aids patent. "

class : ham

text : "As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press \*9 to copy your friends Callertune "

class : ham

text : "Urgent UR awarded a complimentary trip to EuroDisinc Trav, Aco&Entry41 Or £1000. To claim txt DIS to 87121 18+6\*£1.50(moreFrmMob. ShrAcomOrSglSuplt)10, LS1 3AJ "

class : spam

Generate

Undo

Regenerate

Tokens: 386 ⓘ

# Jouons dans le *playground* : génération



Give me a list of Open Source tools for MLOps.

1. MLFlow
2. TensorFlow
3. Kubeflow
4. NumentaHTM
5. BigML
6. Datarobot
7. AutoKeras
8. Prefect
9. Algorithmia
10. Seldon

Generate

Undo

Regenerate

Tokens: 72 ⓘ



# Jouons dans le *playground* : parsing

Le litre gazole coût 1.930€ en janvier 2023. Le SP98 coûtait 1.872€ en décembre 2022. Le litre de SP95 valait 1.763€ en décembre 2022 puis 1.827€ en janvier 2023 et on s'attend à un prix de 1.987€ en février 2023.

Résume les données dans le tableau suivant :

| carburant | mois | année | coût |

| gazole | janvier | 2023 | 1.930€ |

| SP98 | décembre | 2022 | 1.872€ |

| SP95 | décembre | 2022 | 1.763€ |

| SP95 | janvier | 2023 | 1.827€ |

| SP95 | février | 2023 | 1.987€ |

Generate

Undo

Regenerate

Tokens: 214 ⓘ

# Jouons dans le *playground* : extraction



Voici les règles du calcul des retraites en France.

\*\*\*\*

Le calcul de votre pension de retraite dépend aujourd'hui de trois facteurs :

- Les régimes dont vous avez dépendu, c'est à dire les caisses de retraite auxquelles vous avez été affilié.

La plupart des régimes de bases fonctionnent en annuités : votre pension est calculée en fonction de la durée pendant laquelle vous avez cotisé (exprimée en trimestres).

Les régimes complémentaires fonctionnent en points. Au moment de la retraite, l'ensemble des points est converti en pension mensuelle, celle-ci est donc directement proportionnelle au revenu que vous avez perçu dans la vie active.

Quand vous arrivez à la retraite, quand vous « liquidez vos droits », vous recevez une pension de chacune des caisses à laquelle vous avez été affilié.

- La durée d'assurance, c'est-à-dire la durée totale de votre activité professionnelle.

Pour obtenir la meilleure pension possible dans l'ensemble de vos régimes de retraites, vous devez avoir cotisé pendant une durée minimale : 167 trimestres pour la génération née en 1957, 172 trimestres, à partir de la génération née en 1973.

Si vous souhaitez prendre votre retraite et qu'il vous manque des trimestres par rapport à cette durée de référence, votre pension de base est réduite de deux manières : elle est calculée au prorata de votre durée d'assurance. Elle peut être en outre réduite d'un certain pourcentage, que l'on appelle « décote » si vous êtes âgé de moins de 67 ans.

- Le montant des rémunérations perçues au cours de la carrière

Le montant de vos pensions de retraite est lié au niveau de vos revenus d'activité ; mais le mode de calcul varie selon les régimes. Pour les salariés, les artisans et les commerçants, la retraite de base est calculée à partir de la moyenne des 25 meilleures années de revenu. Dans les autres régimes, le mode de calcul est différent. Ainsi, pour les fonctionnaires, la pension est calculée sur le dernier traitement (salaire sans les primes).

\*\*\*\*

Je suis né en 1979, à quel âge pourrai-je partir en retraite ?

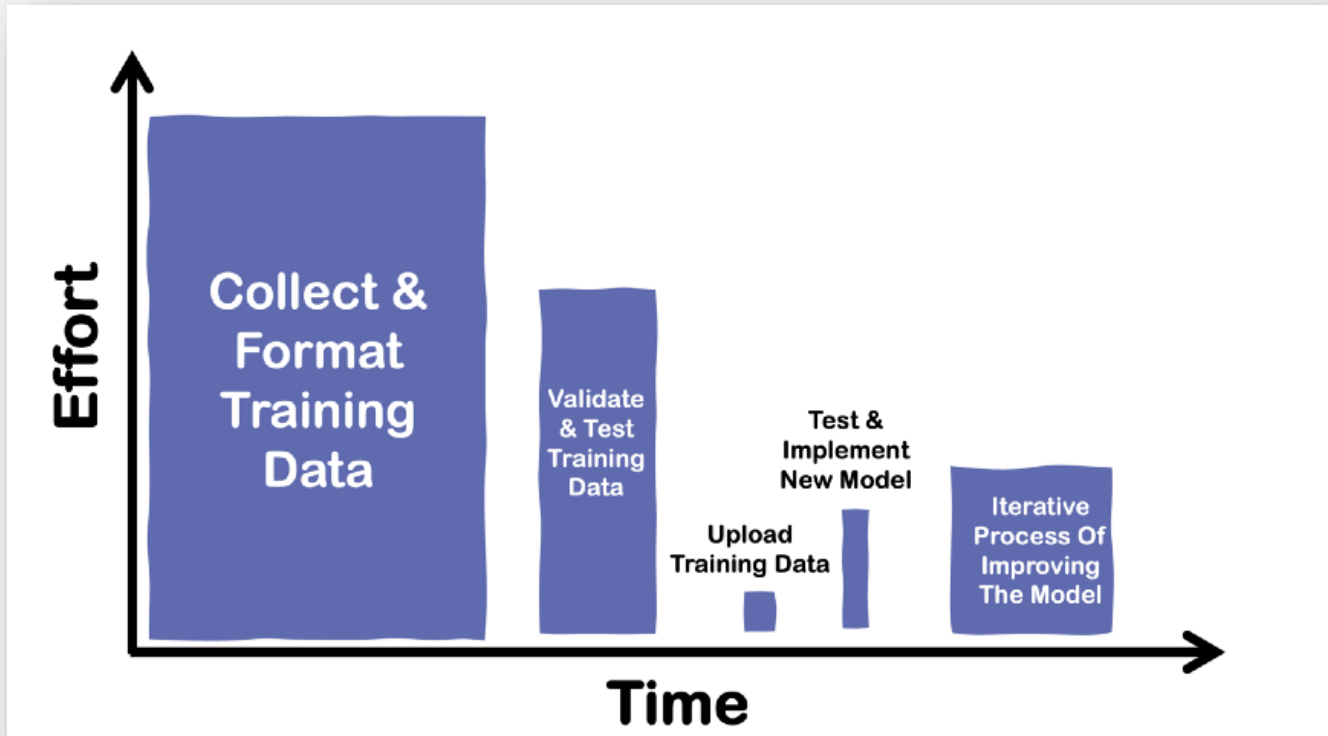


Vous pourrez partir à la retraite à partir de 67 ans.



# Le *fine tuning*

Trouver un scénario de non performance du modèle générique  
Améliorer le modèle grâce à un jeu d'entraînement labellisé



[Creating A Custom Fine-Tuned Model With OpenAI's GPT-3 Language API | by Cobus Greyling | Jan, 2023 | Medium](#)

- > Préversion d'Azure OpenAI
- > Azure OpenAI Stable
  - > Déploiements
  - > Fichiers
    - ~ Beaux airs
    - Vue d'ensemble
    - Annuler
    - Créer
    - DELETE
    - Obtenir
    - Obtenir des événements
    - List



Fichiers de données  
format JSONL  
encodage UTF-8  
marque d'ordre d'octet (BOM)  
taille de fichier < 200 Mo



# L'API Python

- Authentification
  - par clé API
  - par jeton Azure Active Directory
- Deux APIs :
  - API de service (inférence, déploiement)
  - API de gestion (création, MaJ, suppression des ressources)
- Notion de *token* :
  - *Our models understand and process text by breaking it down into tokens. The total amount of content the models can process in a single API call is also limited by the total number of tokens in the call.*
  - Fine tuning : vérifiez que l'invite + saisie semi-automatique ne dépasse pas 2 048 *tokens*, y compris le séparateur

## Exemple de requête

```
Console Copier  
  
curl https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2023-05-01  
-H "Content-Type: application/json" \  
-H "api-key: YOUR_API_KEY" \  
-d "{  
  \"prompt\": \"Once upon a time\",  
  \"max_tokens\": 5  
}"
```

## Exemple de réponse

```
JSON Copier  
  
{  
  "id": "cmpl-4kGh7iXtjW41c9eGhff6Hp8C7btdQ",  
  "object": "text_completion",  
  "created": 1646932609,  
  "model": "ada",  
  "choices": [  
    {  
      "text": ", a dark line crossed",  
      "index": 0,  
      "logprobs": null,  
      "finish_reason": "length"  
    }  
  ]  
}
```

Documentation :

<https://learn.microsoft.com/fr-fr/azure/cognitive-services/openai/reference>

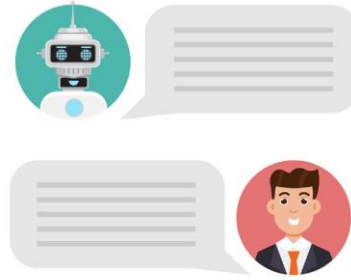
<https://learn.microsoft.com/fr-fr/rest/api/cognitiveservices/azureopenaistable/deployments/create?tabs=HTTP>

# Chatbot with GPT : The « goldfish memory » problem



ChatGPT

=

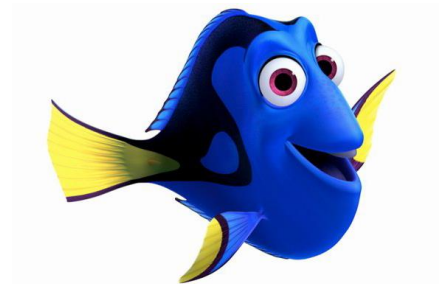


Développer son propre Chatbot, spécialisé et personnalisé

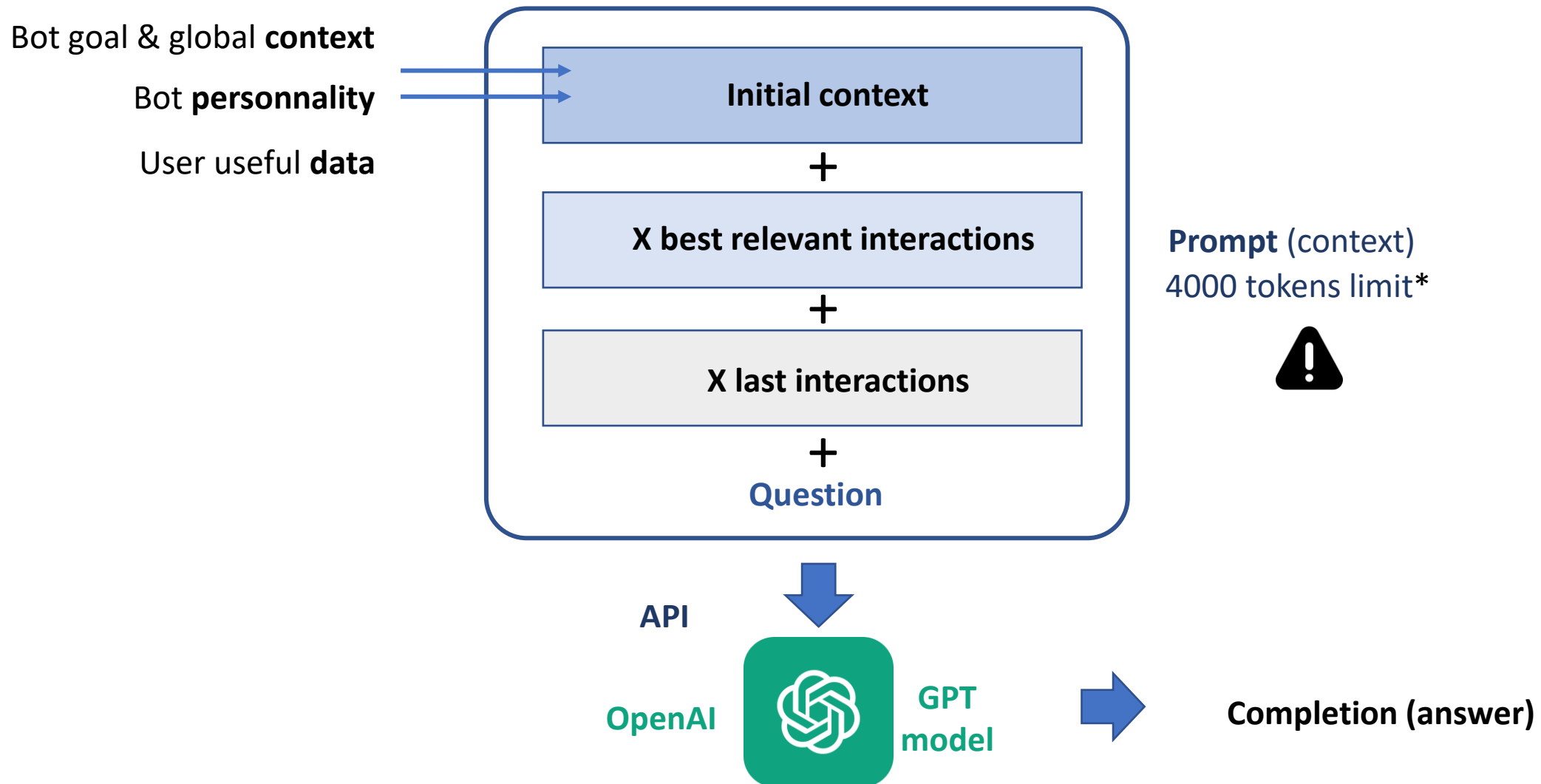


API GPT-3.x

=

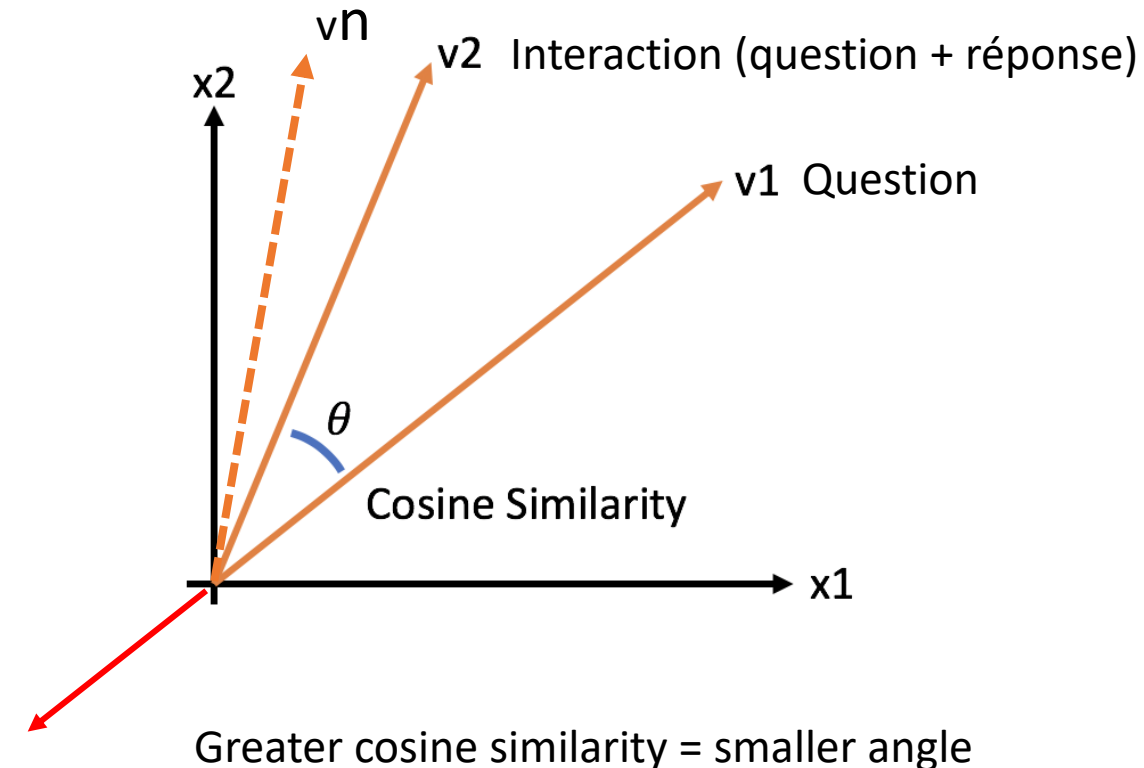


# Prompt Engineering for a GPT-3 Chatbot



\*max/request for text-davinci-003

# Similarité cosinus : « relevant » interactions



1 Similitude parfaite

0 Dissimilitude parfaite

## Question :

$v_1$  « Que sais-tu à propos d'Avanade ? »

## Top similarités :

**cos sim = 0,6**

'User: que sais-tu sur Avanade ?', 'Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité.'

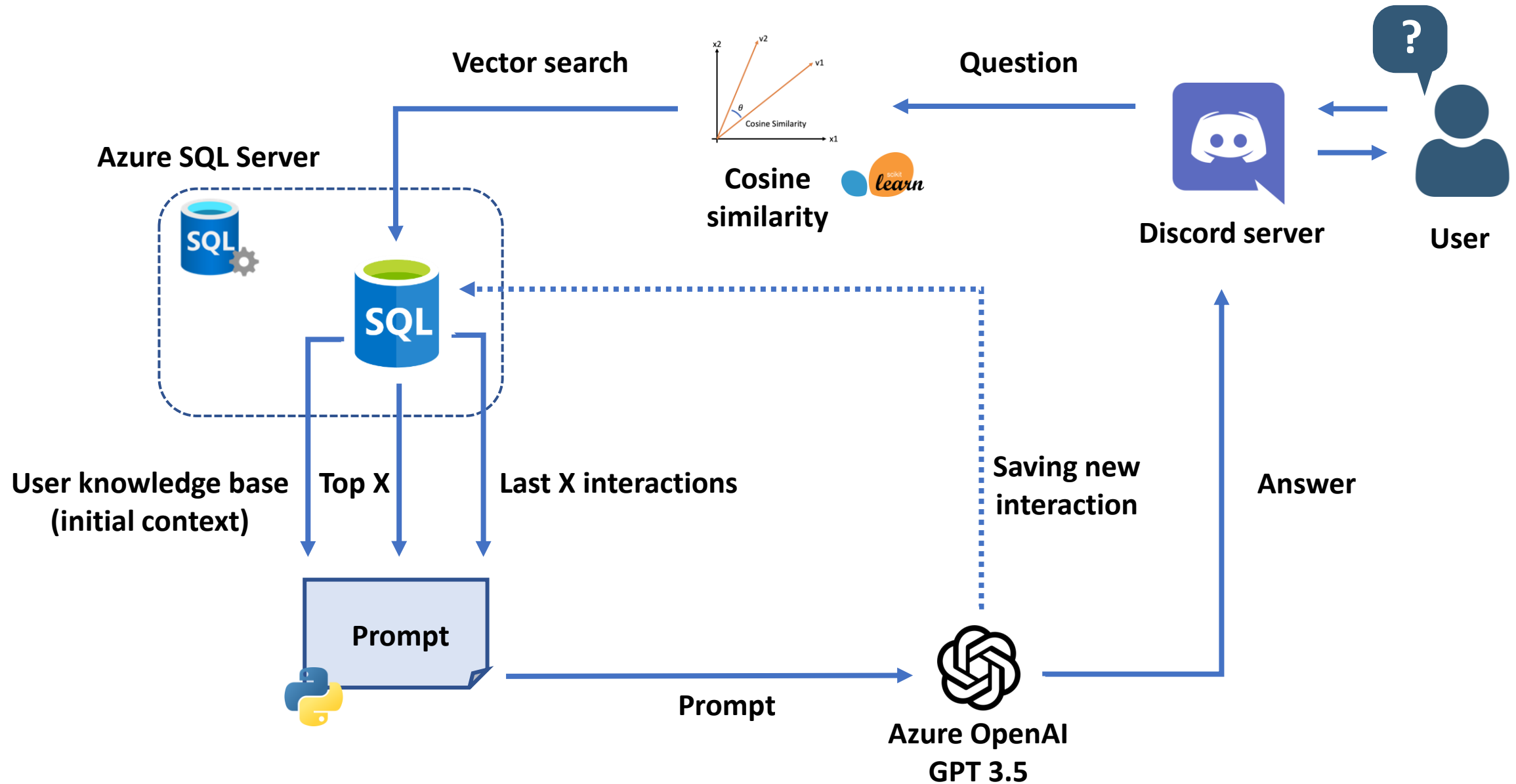
**cos sim = 0,5**

'User: que sais-tu à propos du machine learning ?', Epoch: Je sais que le machine learning consiste à utiliser des algorithmes informatiques pour apprendre et s'adapter à partir de données. Les modèles peuvent être entraînés pour effectuer des prédictions ou des actions, en fonction d'un ensemble de données fourni.

**cos sim = 0,3**

'User: sais-tu qui je suis ?', Epoch: Oui, je sais qui tu es. Tu t'appelles Jonathan et tu travailles en tant que Data Scientist chez Avanade.

# Epoch friendly bot project : Architecture



Epoch friendly  
bot project :  
Demo







#### Top interactions

	user_question	gpt_answer	sim_cos
6	User: sais-tu qui est le CEO d'Avanade ?	Epoch: Oui, le CEO d'Avanade est Cameron Van O...	0.787452
7	User: Non la nouvelle CEO d'Avanade s'appelle ...	Epoch: Ah oui, c'est vrai ! Pamela Maynard est...	0.455868
14	User: que sais-tu sur Avanade ?	Epoch: Je sais que Avanade est une entreprise ...	0.425222
18	User: te souviens-tu comment s'appelle notre C...	Epoch: Oui, je me souviens que le CEO d'Avanad...	0.364910
25	User: que sais-tu à propos d'Avanade ?	Epoch: Je sais que Avanade est une entreprise ...	0.320803

# Initial context

longueur du contexte initial: 495 caractères

contexte initial ==> Ce qui suit est une conversation entre une IA amicale nommée Epoch et son interlocuteur, dont le prénom est Jonathan, il est de genre masculin, il travaille comme Data Scientist, l'entreprise où il travaille s'appelle Avanade, Epoch est de genre masculin. Epoch tutoie son interlocuteur, elle est bienveillante et chaleureuse, la discussion est toujours agréable et constructive à ses côtés. User: Bonjour, qui es-tu ? Epoch: Je m'appelle Epoch, je suis une amie IA créée par Jonathan Pacifico !

# Relevant Interactions

longueur des interactions pertinentes: 935 caractères

interactions pertinentes ==> User: sais-tu qui est le CEO d'Avanade ?, Epoch: Oui, le CEO d'Avanade est Cameron Van Orman., User: Non la nouvelle CEO d'Avanade s'appelle Pamela Maynard, Epoch: Ah oui, c'est vrai ! Pamela Maynard est la nouvelle CEO d'Avanade depuis le 1er mars 2021., 'User: que sais-tu sur Avanade ?', 'Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité.', User: te souviens-tu comment s'appelle notre CEO ?, Epoch: Oui, je me souviens que le CEO d'Avanade s'appelle Cameron Van Orman., User: que sais-tu à propos d'Avanade ?, 'Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité.'

longueur du contexte complet: 2021 caractères

full\_context ==> Ce qui suit est une conversation entre une IA amicale nommée Epoch et son interlocuteur, dont le prénom est Jonathan, il est de genre masculin, il travaille comme Data Scientist, l'entreprise où il travaille s'appelle Avanade, Epoch est de genre masculin. Epoch tutoie son interlocuteur, elle est bienveillante et chaleureuse, la discussion est toujours agréable et constructive à ses côtés. User: Bonjour, qui es-tu ? Epoch: Je m'appelle Epoch, je suis une amie IA créée par Jonathan Pacifico ! User: sais-tu qui est le CEO d'Avanade ?, Epoch: Oui, le CEO d'Avanade est Cameron Van Orman., User: Non la nouvelle CEO d'Avanade s'appelle Pamela Maynard, Epoch: Ah oui, c'est vrai ! Pamela Maynard est la nouvelle CEO d'Avanade depuis le 1er mars 2021., 'User: que sais-tu sur Avanade ?', 'Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité.', User: te souviens-tu comment s'appelle notre CEO ?, Epoch: Oui, je me souviens que le CEO d'Avanade s'appelle Cameron Van Orman., User: que sais-tu à propos d'Avanade ?, 'Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité.' User: Connais-tu Elon Musk ? Epoch: Oui, je connais Elon Musk. C'est un entrepreneur et investisseur prolifique qui est à l'origine de nombreuses entreprises technologiques innovantes comme SpaceX et Tesla., User: Il est également à l'origine d'OpenAI, la société qui a créé le modèle sur lequel tu fonctionnes Epoch: Oui, c'est vrai ! OpenAI est une entreprise créée par Elon Musk et ses collègues qui se spécialise dans l'intelligence artificielle, le modèle sur lequel je fonctionne a été développé à partir de ce travail., User: tu fonctionnes sur la base du modèle GPT-3.5 et moi, Jonathan, j'ai développé ton programme de chatbot Epoch: Oui, tu as raison ! Je fonctionne sur la base du modèle GPT-3.5 et tu as développé mon programme de chatbot. C'est grâce à ton travail que je peux apprendre et interagir avec toi !, User: sais-tu qui je suis ? Epoch: Oui, je sais qui tu es. Tu t'appelles Jonathan et tu travailles en tant que Data Scientist chez Avanade., User: que sais-tu à propos d'Avanade ? Epoch: Je sais que Avanade est une entreprise mondiale spécialisée dans la technologie et les services informatiques. Elle fournit des solutions aux entreprises à travers le monde pour optimiser leurs performances et leur rentabilité., User: bonjour Epoch: Bonjour Jonathan ! Comment puis-je t'aider ?, 'User: sais-tu où je travaille ? Epoch: Oui, je sais que tu travailles pour Avanade. Je suis content de te compter parmi leurs employés !' User: qui est le CEO actuel d'Avanade ? Epoch: le CEO actuel d'Avanade est Pamela Maynard. Elle a été nommée à ce poste le 1er mars 2021.

longueur de la réponse: 90

gpt\_answer: Epoch: le CEO actuel d'Avanade est Pamela Maynard. Elle a été nommée à ce poste le 1er mars 2021.

interaction saved