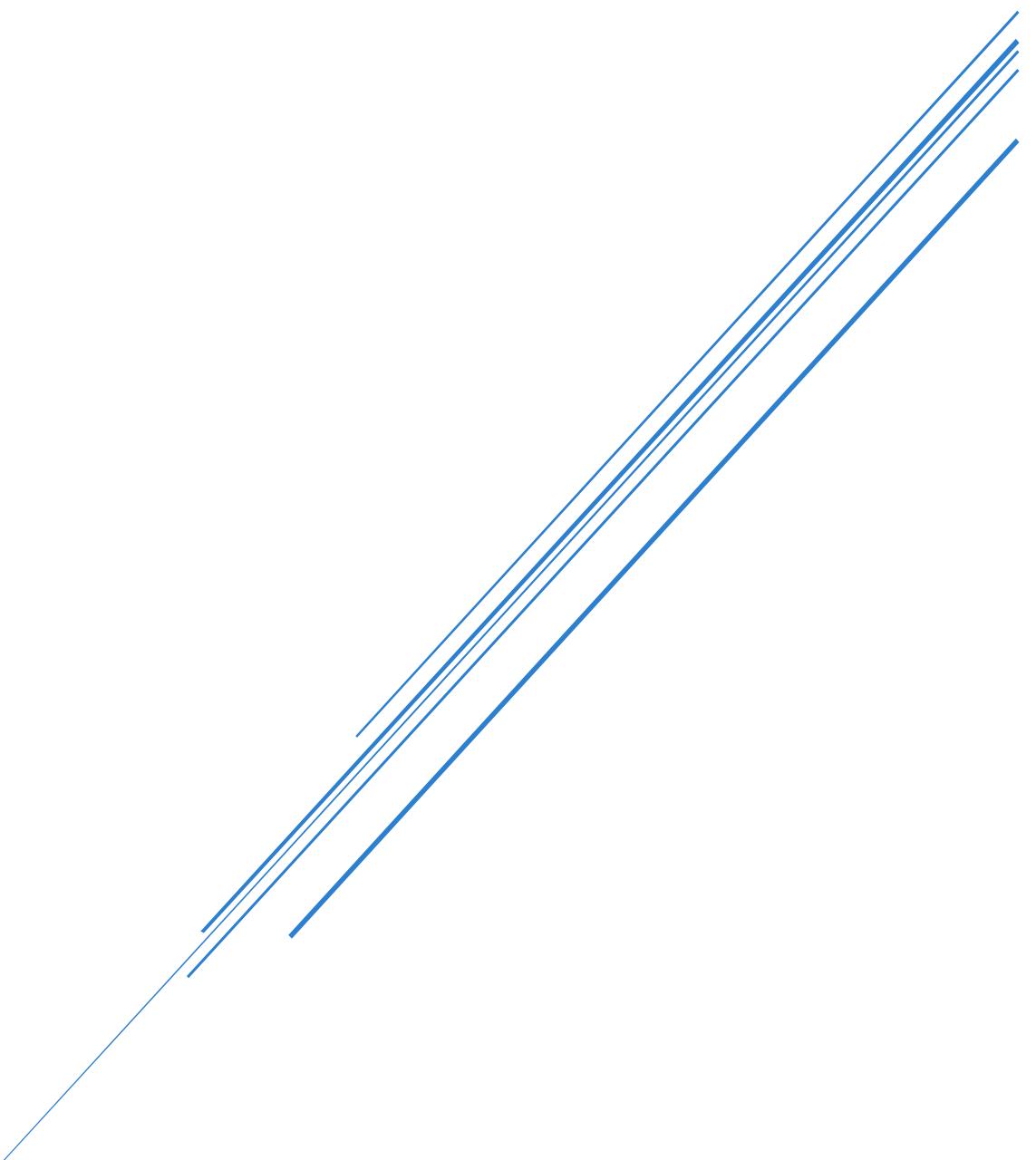


# Azure AI Foundry

## Laboratoire



## Table des matières

Configuration de l'environnement.....	2
Créer un nouveau projet .....	2
Déployer un modèle.....	7
Action .....	7
Génération augmentée de récupération (RAG) .....	13
Évaluation .....	20
Action - Évaluez votre RAG .....	21
Action - Évaluer votre flux de chat (facultatif) .....	27
Découvrir la sécurité du contenu.....	36

# Configuration de l'environnement

Pour déployer cet atelier, un **abonnement Azure est requis**, dans lequel vous pouvez créer : un projet d'IA avec sa ressource AI Hub, un service de Sécurité du contenu Azure AI et un service AI Search.

## Étapes du laboratoire

1. Utilisez Azure AI Foundry Playground.
2. Travaillez avec un modèle LLM open source.
3. RAG – connectez vos données.
4. Création d'un flux d'invites
5. Évaluez votre modèle
6. Testez l'invite de Sécurité du contenu Azure AI

## Créer un projet d'IA et une ressource AI HUB

Commençons par créer un projet dans Azure AI Foundry.

Allez dans votre navigateur et tapez : <https://ai.azure.com>.

Connectez-vous avec votre compte Microsoft.

The screenshot shows the Azure AI Foundry interface. At the top, there's a navigation bar with icons for Home, Projects, AI Services, and Help. Below it, a header says "Jump into a project in Azure AI Foundry". A "Create project" button is visible. The main area displays a table of projects:

Project	Description	Created on	Location	Hub
project_rag_lab2		Nov 13, 2024 11:43 ...	eastus	Hub_Jab2
project-labaistudio		Oct 21, 2024 4:48 PM	eastus	labaistudio

Below the table, there are sections for "Work outside of a project" and "Find it fast". The "Work outside of a project" section includes links to "Chat playground", "Explore Azure AI Services", and "Summit Center". The "Find it fast" section includes links to "Quota management", "Model catalog and benchmarks", "Safety and security", and "Content Understanding". On the right side, there's a "Help" sidebar with sections like "Watch a tutorial", "Overview", "What are AI services?", "Azure AI Studio architecture?", "Quick starts", and "Tutorials".

## Créer un nouveau projet

- **Cliquez sur :** <https://ai.azure.com/>
- **Cliquez sur :** « + Crée un projet »

[View all projects](#)[+ Create project](#)[Help](#)

## Create a project

Projects are easy-to-manage containers for your work—and the key to collaboration, organization, and connecting data and other services.

Project name \* ⓘ

Hub ⓘ

[Create new hub](#)[Create](#)[Cancel](#)

- Saisissez un nom pour votre projet.
- Cliquez sur : « **Créer un nouveau hub** »

**Vous pouvez créer et gérer un hub à partir du portail Azure ou d'AI Foundry.**



Si vous souhaitez **créer un Hub sécurisé**, vous devez le créer à partir du portail :  
[Comment créer et gérer un hub Azure AI Foundry - Azure AI Foundry | Microsoft Learn](#)

- Tapez un nom pour votre hub

A hub is the collaboration environment for your team to share your project work, model endpoints, compute, connections, and security settings.

Name

[Learn more](#)[Next](#)[Cancel](#)

- Cliquez sur **Personnaliser**

**Create a project**

Projects are easy-to-manage containers for your work—and the key to collaboration, organization, and connecting data and other services.

**Project name \*** ⓘ  
frgail-7494

**Hub** ⓘ [Create new hub](#)  
Select or search by name

✓ Azure resources to be created (new: Hub-AI + 4) [Customize](#)

<b>Subscription</b> MCAPS-Hybrid-REQ-40894-2022-frgail	<b>Hub</b> (new) Hub-AI
<b>Resource group</b> (new) rg-frgail-1297_ai	<b>Data storage</b> (new) Hub, Storage, Key Vault, AI Services
<b>Location</b> eastus	<b>Public network access</b> Enabled

[View resource and pricing details](#)

ⓘ Do you need to customize the security or storage resources? [Go to Azure Portal](#)

[Create](#) [Cancel](#)

▪ Remplissez les cellules

**Create a project**

- 1 Project details
- 2 Create a hub
- 3 Review and finish

**Create a hub for your projects**  
A hub is the collaboration environment for your team to share your project work, model endpoints, compute, connections, and security settings. [Learn more](#)

Do you need to customize security or the [dependent resources](#) of your hub? [Go to Azure Portal](#)

**Hub name \***  
Hub-AI

**Subscription \*** ⓘ [Create new subscription](#)  
MCAPS-Hybrid-REQ-40894-2022-frgail

**Resource group \*** ⓘ [Create new resource group](#)  
labaistudio

**Location \*** ⓘ [Help me choose](#)  
East US

**Connect Azure AI Services or Azure OpenAI \*** ⓘ [Create new AI Services](#)  
(new) ai-labaifoundry

**Connect Azure AI Search \*** ⓘ [Create new AI Search](#)  
(new) aisearch-labaifoundry

[Back](#) [Next](#) [Create](#) [Cancel](#)

▪ Cliquez sur **Créer**

**Create a project**

- Project details
- Create a hub
- Review and finish

**Review and finish**

The following resources will be created for you, along with required dependencies. The creation of the first hub and project may take a few minutes to complete. [Learn more about hubs and dependencies](#).

**Hub**

Name: Hub-AI  
Subscription: MCAPS-Hybrid-REQ-40894-2022-frgail  
Resource group: labaistudio  
Location: eastus

**Project**

Name: frgail-7494  
Subscription: MCAPS-Hybrid-REQ-40894-2022-frgail  
Resource group: labaistudio

**AI Services**

Name: ai-labaifoundry

**AI Search**

Name: aisearch-labaifoundry

[Back](#) [Create](#) [Cancel](#)

▪ Cliquez sur **Open in management center**

**frgail-7494**

Add a project description (optional)

Endpoints and keys		View all endpoints
API Key	.....	
Included capabilities	Use the following endpoint to call your Azure OpenAI models:	
Azure AI inference		
Azure OpenAI	<a href="https://ai-labaifoundry721835953777.openai.azure.com/">https://ai-labaifoundry721835953777.openai.azure.com/</a>	
Azure AI Services		
<a href="#">[x] API documentation</a>		

**Project details**

Project connection string  
eastus.api.azureml.ms:f279e1c9-050d-47e0... [Edit](#)

Subscription  
MCAPS-Hybrid-REQ-40894-2022-frgail [Edit](#)

Subscription ID  
f279e1c9-050d-47e0-bc17-7fccf4e193f2 [Edit](#)

Location  
eastus [Edit](#)

**Manage project settings**

- Add users  View quota
- Connect resources  Track costs

[Open in management center](#)

The screenshot shows the Azure AI Foundry Management center for the project 'frgail-7494'. In the 'Connected resources' section, five resources are listed: 'AzureAISeach' (Azure AI Search), 'ai-labafoundry721835953777\_aoai' (Azure OpenAI), 'ai-labafoundry721835953777' (AI Services), 'frgail-7494/workspaceblobstore' (Azure Blob Storage), and 'frgail-7494/workspaceartifactstore' (Azure Blob Storage). An orange arrow points from this section down to a table below.

Name	Type	Target	Key	Authentication type
AzureAISeach	Azure AI Search (Cogniti...)	https://aistore-labafoundry721835953777.se...	.....	API key
ai-labafoundry721835953777_aoai	Azure OpenAI	https://ai-labafoundry721835953777.openai.a...	.....	API key
ai-labafoundry721835953777	AI Services	https://ai-labafoundry721835953777.cognitive...	.....	API key
frgail-7494/workspaceblobstore	Azure Blob Storage	https://sthubai721835953777.blob.core.windo...	--	SAS
frgail-7494/workspaceartifactstore	Azure Blob Storage	https://sthubai721835953777.blob.core.windo...	--	SAS

#### 4 ressources connectées ont été créées :

- **Azure AI Search** pour créer un index et stocker des vecteurs.
- **AI Services** pour accéder aux Services d'IA (Parole, Langue+Traducteur, Vision, Sécurité du contenu)
- **Service Azure OpenAI** pour accéder aux modèles Azure OpenAI
- **Stockage Blob Azure** pour stocker les données et les artefacts.
  - **Magasin d'artefacts de l'espace de travail :**
    - Principalement utilisé pour stocker divers artefacts liés à vos projets d'IA, tels que des ensembles de données, des modèles, des journaux et d'autres fichiers.
    - Chaque projet dispose de ses propres conteneurs de stockage dédiés dans le magasin d'artefacts de l'espace de travail, ce qui permet de maintenir l'isolation et la sécurité des données.
  - **Magasin d'objets blob Workspace :**
    - Agit comme le stockage d'objets blob par défaut pour l'espace de travail, utilisé pour les besoins généraux de stockage de données.
    - Généralement utilisé pour stocker de grandes quantités de données non structurées, telles que du texte, des images et des données binaires.

## Déployer un modèle

Après avoir créé votre projet d'IA, la première étape consiste à créer un déploiement d'un modèle Azure OpenAI afin de pouvoir commencer à expérimenter avec les invités que vous utiliserez dans votre application.

### Action

- Dans le projet que vous venez de créer, cliquez sur **Modèles + points de terminaison** :

The screenshot shows the Azure AI Foundry interface. The left sidebar has a tree view with categories like Overview, Model catalog, Playgrounds, AI Services, etc. A red box highlights the 'Models + endpoints' item under 'My assets'. The main content area is titled 'Manage deployments of your models and services' and shows a 'Model deployments' tab selected. It features a large search icon and a button to 'Create a new deployment'. Below this, there's a note about Azure AI Studio supporting LLMs and flows for deployment. At the bottom, there's a 'Need help? View documentation' link.

- Sélectionnez Déployer le modèle de base

The screenshot shows a dropdown menu from the 'Deploy model' button. The 'Deploy base model' option is highlighted with a red box.

## Select a model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog.](#)

Models: 1812

Collections

Deployment options

Inference tasks



Show description

Search

gpt-4o-realtime-preview  
Audio generation

gpt-4  
Chat completion

gpt-35-turbo  
Chat completion

o1-preview  
Chat completion

o1-mini  
Chat completion

gpt-4o-mini  
Chat completion

gpt-4o  
Chat completion

< Prev Next >



Select a model to see description

Confirm

Cancel

- Dans la **barre de recherche**, tapez **GPT-4o**, sélectionnez **gpt-4o** et cliquez sur **Confirmer** :

Select a model

Choose a model to create a new deployment. For flows and other resources, create a deployment from their respective list. [Go to model catalog](#).

Models: 83

Task: Chat completion

**gpt-4o**

GPT-4o offers a shift in how AI models interact with multimodal inputs. By seamlessly combining text, images, and audio, GPT-4o provides a richer, more engaging user experience.

Matching the intelligence of GPT-4 Turbo, it is remarkably more efficient, delivering text at twice the speed and at half the cost. Additionally, GPT-4o exhibits the highest vision performance and excels in non-English languages compared to previous OpenAI models.

GPT-4o is engineered for speed and efficiency. Its advanced ability to handle complex queries with minimal resources can translate into cost savings and performance.

The introduction of GPT-4o opens numerous possibilities for businesses in various sectors:

- Enhanced customer service:** By integrating diverse data inputs, GPT-4o enables more dynamic and comprehensive customer support interactions.
- Advanced analytics:** Leverage GPT-4o's capability to process and analyze different types of data to enhance decision-making and uncover deeper insights.

Confirm

Cancel

- Donnez un nom à votre déploiement et appuyez sur **Déployer**

Deploy model gpt-4o

Deployment name \*

gpt-4o-labai

Deployment type

Global Standard

Global Standard: Pay per API call with the highest rate limits. Learn more about [Global deployment types](#).

Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about [data residency](#).

**Deployment details**

Customize

Model version  
2024-08-06

Connected AI resource  
ai-hublab2013149793759\_aoai

Project  
project\_rag\_lab2

Authentication type  
Key

Capacity  
10K tokens per minute (TPM)

Resource location  
East US

Content safety  
DefaultV2

Deploy

Cancel

**Mistral-large-2407-labai**

**Deployment info**

- Name: Mistral-large-2407-labai
- Provisioning state: Succeeded
- Created by: franck.gaillard@microsoft.com
- Last updated on: Dec 4, 2024 4:05 PM
- Model: Mistral-large-2407

**Endpoint**

- Target URI: https://Mistral-large-2407-labai.eastus.models.ai.azure.com
- Key: [REDACTED]
- Compute type: Consumption
- Swagger URI: https://Mistral-large-2407-labai.eastus.models.ai.azure.com/swagger.json

**API Routes**

- Azure AI model inference: Chat Completion
- Mistral: Chat Completion

**Monitoring & safety**

- Azure AI Content Safety: Enabled

- Cliquez sur **Ouvert** dans l'aire de jeux :

**← Chat playground**

**Setup**

Deployment \*: Mistral-large-2407-labai

Give the model instructions and context:

You are an AI assistant that helps people find information.

**Chat session**

+ Add section

**Parameters**

Start typing here

- Nous allons **exécuter un exemple** où le modèle nous aidera à **résumer et à extraire des informations d'une conversation** entre un client et un représentant d'une société de télécommunications.

Copiez l'invite suivante dans le champ d'instructions et de contexte du modèle :

Vous êtes un assistant IA qui aide les entreprises de télécommunications à extraire des informations précieuses de leurs conversations en créant des fichiers JSON pour chaque transcription de conversation que vous recevez. Vous essayez toujours d'extraire et de formater au format JSON :

1. Nom du client [nom]

2. Téléphone du contact client [téléphone]
3. Sujet principal de la conversation [sujet]
4. Sentiment des clients (neutre, positif, négatif)[sentiment]
5. Comment l'agent a géré la conversation [comportement\_agent]
6. Quel a été le résultat final de la conversation [résultat]
7. Un résumé très bref de la conversation [résumé]

N'extrayez que les informations dont vous êtes sûr. Si vous n'êtes pas sûr, écrivez « Inconnu/Introuvable » dans le fichier JSON.

Après la copie, sélectionnez « **Appliquer les changements** ».

Tapez ensuite le texte suivant dans la session de chat et cliquez sur le bouton Envoyer :

Agent : Bonjour, bienvenue au service client de Telco. Je m'appelle Juan, comment puis-je vous aider ?

Client : Bonjour, Juan. J'appelle parce que j'ai des problèmes avec mon forfait de données mobiles. C'est très lent et je ne peux pas naviguer sur Internet ou utiliser mes applications.

Agent : Je suis vraiment désolé pour la gêne occasionnée, monsieur. Pourriez-vous s'il vous plaît me dire votre numéro de téléphone et votre nom complet ?

Client : Oui, bien sûr. Mon numéro est le 011-4567-8910 et je m'appelle Martín Pérez.

Agent : Merci, M. Pérez. Je vais vérifier votre forfait et votre utilisation des données. Un instant, s'il vous plaît.

Client : D'accord, merci.

Agent : M. Pérez, j'ai examiné votre plan et je vois que vous avez contracté le plan de base de 2 Go de données par mois. Est-ce exact ?

Client : Oui, c'est exact.

Agent : Eh bien, je vous informe que vous avez consommé 90% de votre limite de données et que vous n'avez que 200 Mo disponibles jusqu'à la fin du mois. C'est pourquoi votre vitesse de navigation a été réduite.

Client : Quoi ? Comment est-ce possible ? J'utilise à peine Internet sur mon téléphone portable. Je ne consulte mes mails et mes réseaux sociaux que de temps en temps. Je ne regarde pas de vidéos et je ne télécharge pas de fichiers volumineux.

Agent : Je comprends, M. Pérez. Mais gardez à l'esprit que certaines applications consomment des données en arrière-plan, sans que vous vous en rendiez compte. Par exemple, les mises à jour automatiques, les sauvegardes, le GPS, etc.

Client : Eh bien, mais ils ne m'ont pas expliqué cela lorsque j'ai contracté le plan. Ils m'ont dit qu'avec 2 Go, j'en aurais assez pour tout le mois. Je me sens trompé.

Agent : Je m'excuse, M. Pérez. Ce n'était pas notre intention de vous tromper. Je vous propose une solution : si vous le souhaitez, vous pouvez changer votre forfait pour un forfait supérieur, avec plus de Go de données et une vitesse plus élevée. De cette façon, vous pouvez profiter d'une meilleure expérience de navigation.

Client : Et combien cela me coûterait-il ?

Agent : Nous avons une offre spéciale pour vous. Pour seulement 10 pesos de plus par mois, vous pouvez accéder au forfait premium de 5 Go de données et de vitesse 4G. Êtes-vous intéressé ?

Client : Mmm, je ne sais pas. N'y a-t-il pas une autre option ? Ne pouvez-vous pas me donner plus de vitesse sans me faire payer plus ?

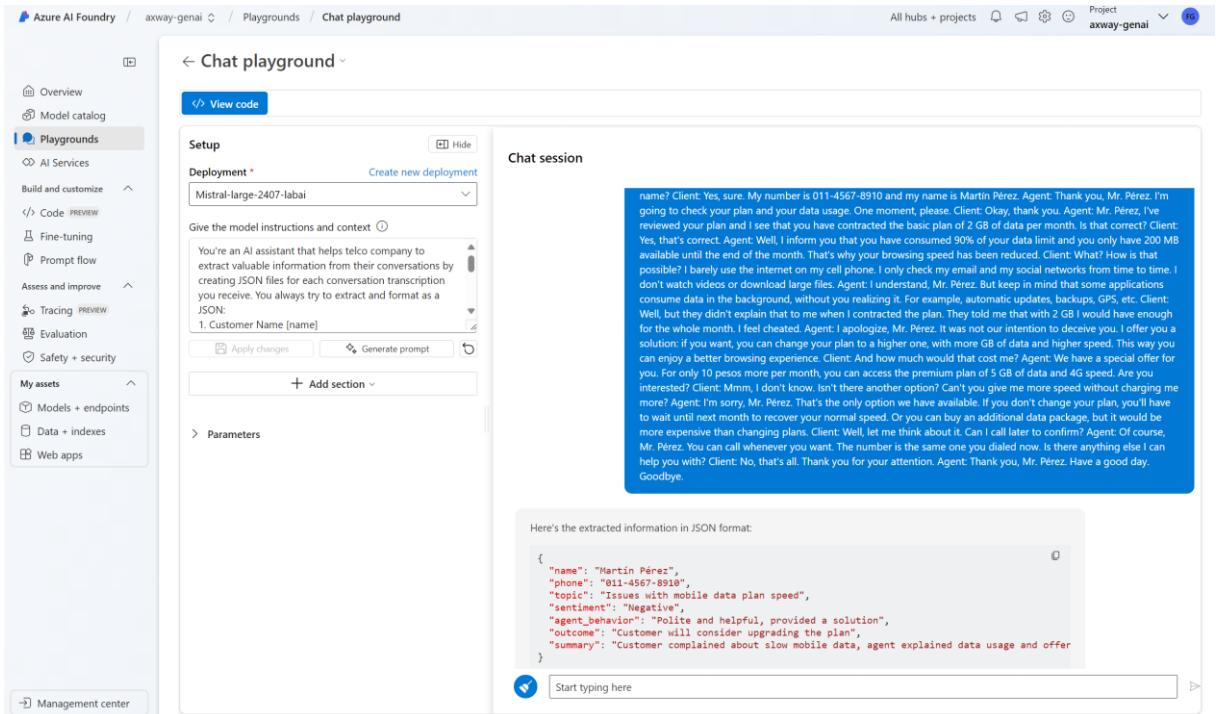
Agent : Je suis désolé, M. Pérez. C'est la seule option que nous avons disponible. Si vous ne changez pas de forfait, vous devrez attendre le mois prochain pour retrouver votre vitesse normale. Ou vous pouvez acheter un forfait de données supplémentaire, mais ce serait plus cher que de changer de forfait.

Client : Eh bien, laissez-moi y réfléchir. Puis-je appeler plus tard pour confirmer ?

Agent : Bien sûr, M. Pérez. Vous pouvez appeler quand vous le souhaitez. Le numéro est le même que celui que vous avez composé maintenant. Y a-t-il autre chose que je puisse faire pour vous aider ?

Client : Non, c'est tout. Je vous remercie de votre attention.

Agent : Merci, M. Pérez. Bonne journée. Au revoir.



## Génération augmentée de récupération (RAG)

### Conditions préalables

- **Un abonnement Azure.**
- **Un hub, un projet et un modèle de conversation Azure OpenAI AI déployés.** Terminez le guide de démarrage rapide du Playground AI Studio [pour créer ces ressources](#) si vous ne l'avez pas déjà fait.
- Une connexion au service Azure AI Search pour indexer les données de l'exemple de produit.
- Vous avez besoin d'une copie locale des données produit. Le référentiel python-promptflow Azure-Samples/rag-data-openai sur GitHub contient des exemples d'informations sur les produits de vente au détail pertinents pour ce scénario de tutoriel. Plus précisément, le fichier product\_info\_11.md contient des informations sur les chaussures de randonnée TrailWalker qui sont pertinentes pour cet exemple de tutoriel. Téléchargez l'exemple de données de produit de vente au détail Contoso Trek dans un fichier ZIP sur votre ordinateur local. [Téléchargez l'exemple de données de produit de vente au détail Contoso Trek dans un fichier ZIP](#)
- Vous devez avoir le fournisseur de ressources Microsoft.Web inscrit dans l'abonnement sélectionné pour pouvoir déployer sur une application web. Pour vérifier :
  - **Connectez-vous au portail Azure :** accédez au portail Azure et connectez-vous avec vos informations d'identification.
  - **Accédez à Abonnements :**
    - Dans le menu du portail Azure, recherchez **Abonnements** et sélectionnez-le.
  - **Sélectionnez votre abonnement :**
    - Choisissez l'abonnement dans lequel vous souhaitez inscrire le fournisseur de ressources.

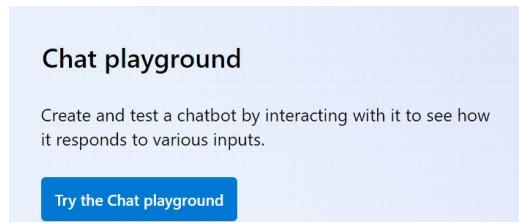
- **Fournisseurs de ressources libres :**
  - Dans le menu de gauche, sous **Paramètres**, cliquez sur **Fournisseurs de ressources**.
- **Enregistrez Microsoft.Web :**
  - Dans la liste des fournisseurs de ressources, recherchez **Microsoft.Web**.
  - Cliquez sur **S'inscrire** pour enregistrer le fournisseur de ressources.

### Ajoutez vos données et essayez le modèle de chat

Suivez ces étapes pour ajouter vos données dans le chat du **Playground** afin d'aider l'assistant à répondre aux questions sur vos produits. Vous ne modifiez pas le modèle déployé lui-même. Vos données sont stockées séparément et en toute sécurité dans votre abonnement Azure.

### Déployer un modèle RAG

1. Sélectionnez **Playgrounds**, puis cliquez sur **Essayer le terrain de jeu de conversation** :



2. Sélectionnez votre modèle d'IA générative déployé dans la **liste déroulante Déploiement**.

3. Sur le côté gauche du Playground, sélectionnez **Ajoutez vos données > + Ajouter une nouvelle source de données**.

← Chat playground ▾

View code

Prompt flow

Evaluate

Deploy

Import

Export

Prompt samples

Send feedback

Setup

Deployment \*

Create new deployment

gpt-4o (version.2024-05-13)

Give the model instructions and context ⓘ

You are an AI assistant that helps people find information.

Apply changes

Generate prompt

Add section

Add your data PREVIEW

Select available project index \*

Select available project index

Add a new data source

Parameters

Chat history

Start with a sample prompt

Creative storytelling

Write a short story about a time traveler who accidentally changes a major historical event.

Recipe creation

Invent a recipe for a dish that combines flavors from two different cuisines.

Poetry generation

Compose a poem about the beauty of nature in autumn.

Type user query here. (Shift + Enter for new line)

11/128000 tokens to be sent

Add your data PREVIEW

- 1 Source data
- 2 Index configuration
- 3 Search settings
- 4 Review and finish

Select your data

Select the data you want the generative AI to reference so it can ground its responses on your specific data.

Your data will be ingested into an Index, which allows the Generative AI model to quickly and accurately find information for your specific use case.

Currently, only the file types .doc(x), .htm, .html, .md, .pdf, .ppt(x), .py, .txt, and .xls(x) are supported. Max file size limit is 16 MB.

Data source \*

Upload files

Upload

Overwrite if alr

Upload list

Connect to an existing index

Azure AI Search

MongoDB Atlas

Create a new index

Data in Azure AI Studio

Azure Blob Storage

Storage URL

Upload files

No files uploaded

Select the Upload files or folders menu above to get started.

Supported file types include: delimited (i.e. csv, tsv, Parquet, JSON Lines and plain text

Next

Create vector index

Cancel

4. Cliquez sur « Charger » et « Charger le dossier » :

Add your data [PREVIEW](#)

- 1 Source data
- 2 Index configuration
- 3 Search settings
- 4 Review and finish

**Select your data**  
 Select the data you want the generative AI to reference so it can ground its responses on your specific data.  
 Your data will be ingested into an Index, which allows the Generative AI model to quickly and accurately find information for your specific use case.  
*Currently, only the file types .doc(x), .htm, .html, .md, .pdf, .ppt(x), .py, .txt, and .xls(x) are supported. Max file size limit is 16 MB.*

Data source \* ⓘ [Upload files](#)

Upload [Upload files](#) [Upload folder](#) [Upload list](#)

No files uploaded  
 Select the Upload files or folders menu above to get started.

An Azure AI Search resource and an Azure Open AI connection will be required to index your data. [Create a new Azure AI](#)

[Next](#) [Create vector index](#) [Cancel](#)

5. Sélectionnez le dossier « **product-info** ». Les fichiers Markdown sont téléchargés dans le stockage Blob.

Sélectionner le dossier à charger

Organiser ▾ Nouveau dossier

Nom	Statut	Modifié le	Type
product-info	✓	13/11/2024 10:27	Dossier de fichier

Bureau Téléchargement Documents Images Musique Vidéos MS Gen AI Studio

Dossier : product-info

Charger Annuler

Add your data [PREVIEW](#)

- 1** Source data
- 2** Index configuration
- 3** Search settings
- 4** Review and finish

**Select your data**  
Select the data you want the generative AI to reference so it can ground its responses on your specific data.  
Your data will be ingested into an Index, which allows the Generative AI model to quickly and accurately find information for your specific use case.  
*Currently, only the file types .doc(x), .htm, .html, .md, .pdf, .ppt(x), .py, .txt, and .xls(x) are supported. Max file size limit is 16 MB.*

Data source \* [ⓘ](#)  [⌄](#)

[⌄](#)

Overwrite if already exists

**Upload list**

product-info/product_info_1.md	11.09 KB/11.09 KB	…
product-info/product_info_10.md	10.83 KB/10.83 KB	…
product-info/product_info_11.md	10.71 KB/10.71 KB	…
product-info/product_info_12.md	10.22 KB/10.22 KB	…
product-info/product_info_13.md	10.71 KB/10.71 KB	…
product-info/product_info_14.md	9.75 KB/9.75 KB	…

[Next](#) [Create vector index](#) [Cancel](#)

5. Sélectionnez le service Azure AI Search dans le menu déroulant et donnez un nom à votre index. Laissez « Sélection automatique » pour la machine virtuelle.

Add your data [PREVIEW](#)

- Source data
- 2** Index configuration
- 3** Search settings
- 4** Review and finish

**Index settings**  
Configure your index

**Index storage \***

**Select Azure AI Search service \*** [ⓘ](#)  
 [⌄](#)  
[Create a new Azure AI Search resource](#)  [ⓘ](#)

**Vector index \*** [ⓘ](#)

**Virtual machine \*** [ⓘ](#)  
 Auto select    Select from recommended options    Select from all options  
*Selecting a virtual machine will incur additional costs.*

[Back](#) [Next](#) [Create vector index](#) [Cancel](#)

6. Cliquez sur « Ajouter la recherche vectorielle à cette ressource de recherche » et sélectionnez une connexion Azure OpenAI dans le menu déroulant :

## Add your data

PREVIEW

- Source data
- Index configuration
- Search settings
- Review and finish

### Configure search settings

Adding vector search supports: Hybrid (vector + keyword search), Hybrid + Semantic (most accurate search results for generative AI applications), Vector, Semantic and Keyword retrieval. Hybrid will be set as default and can be changed at inference time in the playground. Not adding vector search supports: Keyword and Semantic retrieval. Keyword will be set as default and can be changed at inference time in the playground. Adding vector search requires an Azure OpenAI embedding model. [Learn more](#)

#### Vector settings

Add vector search to this search resource

#### Azure OpenAI connection \*

ai-hublab2013149793759\_aoai

This resource requires an embedding model. If you don't have one already, **text-embedding-ada-002 (Version 2)** will be deployed for you. Using vector embeddings will incur usage to your account. [View Azure OpenAI Service pricing](#)

Back

Next

Create vector index

Cancel

## 7. Cliquez sur **Créer un index vectoriel** :

## Add your data

PREVIEW

- Source data
- Index configuration
- Search settings
- Review and finish

### Review and finish

Review the configurations you set for your index

#### Vector index

product-info-lab2

#### Index storage

Azure AI Search

#### Azure AI Search connection

AzureAISearch

#### Include vector settings

Yes

#### Schedule

OneTime

#### Compute

Serverless compute (Auto select)

Back

Create vector index

Cancel

8. Après quelques minutes, votre index sera créé. Il comprend des intégrations vectorielles.

✓ Add your data [PREVIEW](#)

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)

Index:

[product-info-lab2](#)

Search type:

Hybrid (vector + keyword)

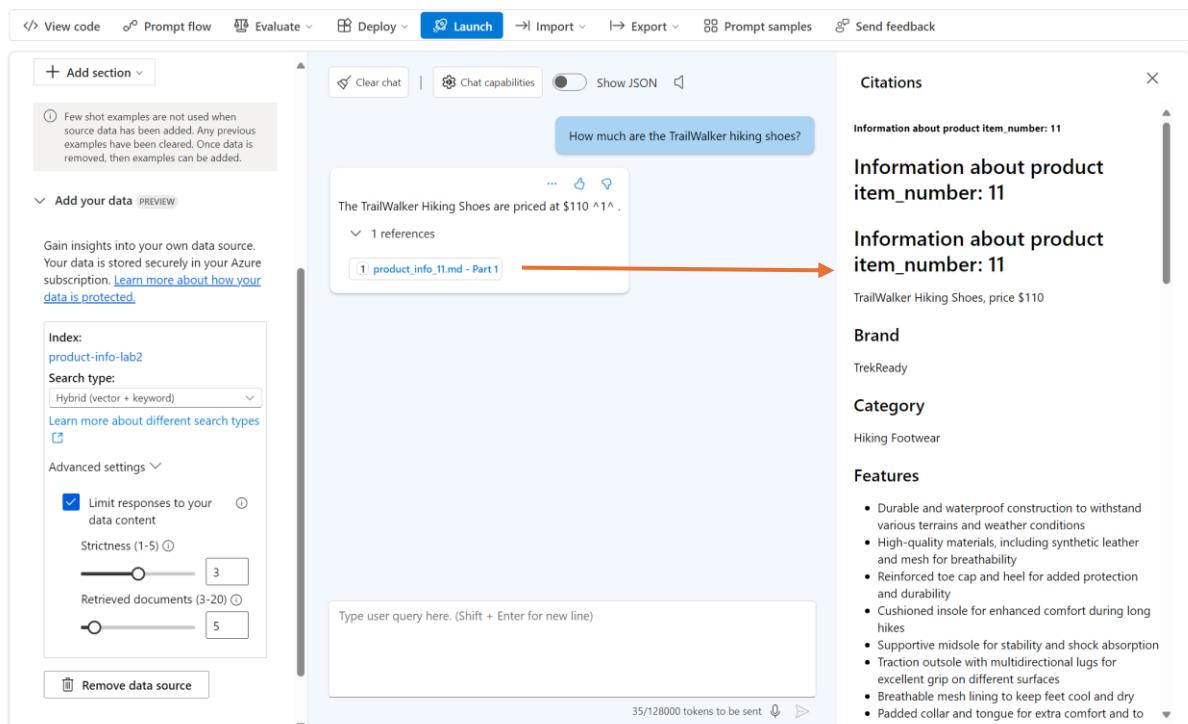
[Learn more about different search types](#)

[Advanced settings >](#)

 Remove data source

9. Vous pouvez maintenant discuter avec le modèle en posant la question : "**Combien coûtent les chaussures de randonnée TrailWalker ?**", et cette fois-ci, il utilise les informations de vos données pour construire la réponse. Vous pouvez développer le  **bouton Références** pour voir les données utilisées.

Chat playground



The screenshot shows the Azure AI Chat playground interface. On the left, there's a sidebar with 'Add section' and 'Add your data' (PREVIEW). The main area has a 'Launch' button and a 'Clear chat' option. A message box displays: "How much are the TrailWalker hiking shoes? The TrailWalker Hiking Shoes are priced at \$110 ^1^. 1 references [product\_info\_11.md - Part 1]". An orange arrow points from this message to a 'Citations' panel on the right. The 'Citations' panel lists: "Information about product item\_number: 11", "Information about product item\_number: 11", and "Information about product item\_number: 11". It also shows 'Brand: TrekReady', 'Category: Hiking Footwear', and a 'Features' list: Durable and waterproof construction to withstand various terrains and weather conditions; High-quality materials, including synthetic leather and mesh for breathability; Reinforced toe cap and heel for added protection and durability; Cushioned insole for enhanced comfort during long hikes; Supportive midsole for stability and shock absorption; Traction outsole with multidirectional lugs for excellent grip on different surfaces; Breathable mesh lining to keep feet cool and dry; Padded collar and tongue for extra comfort and to

## Déployez votre application web

Une fois que vous êtes satisfait de l'expérience dans Azure AI Studio, vous pouvez déployer le modèle en tant qu'application web autonome. Le déploiement crée un Azure App Service dans votre abonnement Azure. Cela peut entraîner des coûts en fonction du plan tarifaire que vous sélectionnez. Lorsque vous avez terminé d'utiliser votre application, vous pouvez la supprimer du portail Azure.

### 1. Sélectionnez « Déployer... en tant qu'application web » :

The screenshot shows the Azure AI Studio interface with the 'Launch' button highlighted. A tooltip for the 'Launch' button indicates it can be used to deploy as a web app or a new Teams app. The main workspace displays a chat interaction about hiking shoes, and the sidebar shows deployment logs and citation details.

### 2. Remplissez les différentes cellules et cliquez sur Déployer :

The screenshot shows the 'Deploy to a web app' configuration dialog. It includes fields for Name (trek-app), Subscription (MCAPS-Hybrid-REQ-40894-2022-frgail), Resource group (labaistudio), Location (France Central), Pricing plan (Standard (\$1)), and a checkbox for Enable chat history in the web app. At the bottom are 'Deploy' and 'Cancel' buttons.

### 3. Attendez que l'application soit déployée, ce qui peut prendre quelques minutes.

## Évaluation

Dans cet atelier, vous allez exécuter les étapes suivantes :

1. Évaluez votre flux de chat.
2. Déployez le flux RAG sur un point de terminaison géré en ligne.

## Action - Évaluez votre RAG

### Préparer votre jeu de données de test

Pour évaluer le RAG avec votre modèle déployé, il est essentiel de le comparer aux questions et aux réponses attendues, en cherchant à ce que la sortie de votre modèle corresponde le plus fidèlement possible aux données de test.

Pour ce faire, il suffit de suivre ces étapes :

Dans la **section Évaluation** de votre projet, cliquez sur **Évaluations manuelles**, puis sur **Nouvelle évaluation manuelle**.

The screenshot shows a user interface for managing manual evaluations. At the top, there's a navigation bar with project information and various icons. Below it, a header reads "Évaluer et comparer les performances des applications d'IA". A red box highlights the "Évaluations manuelles" tab in the top navigation. The main area contains a table with two rows of data:

Évaluation	Jeu de données utilisé	Créé par	Thumbs-up	Thumbs-down	Créé le
quirky_cat_cgyjv8j9cz	faq_csv_2024-12-11_162008_UTK	Timothée LEVILLAYER	94.74%	5.26%	Dec 11, 2024 5:30 PM
salmon_arch_16w3m2tll5	faq_csv_2024-12-11_162008_UTK	Timothée LEVILLAYER	0%	0%	Dec 11, 2024 5:24 PM

At the bottom of the table, there are buttons for "Actualiser" and "Réinitialiser la vue". Below the table, there are search and filter options. At the very bottom, there are navigation links for "Préc" and "Suv", and a "25/Page" dropdown.

Une nouvelle page s'ouvre. Dans l'angle supérieur droit, assurez-vous que le modèle que vous avez déployé est correctement sélectionné dans l'onglet **Configurations** puis cliquez sur **Ajouter vos données**, et sélectionnez l'**index** que vous avez créé précédemment.

Une fois cela fait, cliquez sur le bouton Importer **des données de test** pour ajouter les données par rapport auxquelles vous souhaitez évaluer votre modèle.

Maintenant, importez les données de test en cliquant sur **Charger le fichier** et importez le **fichier rando\_test\_data.csv**. Une fois le document chargé, cliquez sur Suivant.

Pour la partie suivante, nous devons mapper les champs des données de test à ceux attendus par Azure Foundry. Procédez comme suit :

- Choisissez le champ **question** pour l'entrée **Entrée**
- Choisissez le champ **réponse** pour l'entrée **Réponse attendue**

Cliquez ensuite sur **Ajouter**.

The screenshot shows the 'Import test data' dialog box. On the left, there are two steps: 'Select dataset' and 'Map data'. The 'Map data' step is active, showing a list of questions and their expected answers:

Quelle est la taille maximale d'un ordinateur portable ?	Le sac à dos SummitClimber peut accueillir un ordinateur portable.
Quelle est la poids maximal que peut supporter la tente Alpine Explorer ?	La table BaseCamp Folding peut supporter un poids maximal de 15 kg.
Combien de portes possède la tente Alpine Explorer ?	La tente Alpine Explorer possède deux portes.
Le pantalon de randonnée TrailBlaze est-il résistant à l'eau ?	Oui, le pantalon de randonnée TrailBlaze est résistant à l'eau.
Quel type de carburant utilise le réchaud EcoFire ?	Le réchaud EcoFire utilise des petites brindilles, des feuilles et du charbon.
Le sac à dos TrailLite Daypack est-il compatible avec un système de tente ?	Oui, le TrailLite Daypack est compatible avec un système de tente.
Quelle est la note de température du sac de couchage MountainDream ?	Le sac de couchage MountainDream est noté pour une température de -15°C.
Le réchaud CompactCook peut-il être utilisé en intérieur ?	Non, le réchaud CompactCook est conçu pour une utilisation en extérieur.

**Dataset mapping \***

**Input \*** question →

**Expected response \*** answer →

Back Add Cancel

## Exécuter l'évaluation

Maintenant que vos données de test ont été ajoutées, voici ce que vous devriez voir : un aperçu des questions/réponses sur lesquelles vous souhaitez évaluer votre RAG.

Pour ce faire, il suffit de cliquer sur le **bouton Exécuter** : vous verrez votre modèle générer une réponse pour chaque entrée.

The screenshot shows the 'Evaluation manuelle' interface. On the left, there is a sidebar with various project management options like 'Vue d'ensemble', 'Catalogue de modèles', 'Cours de récréation', etc. The main area is titled 'Résultat de l'évaluation manuelle' and contains a table of test cases:

Entrée	Réponse attendue	Sortie
Quelle est la capacité de la tente ?	La tente TrailMaster X4 a une capacité de 4 personnes.	Exécuter pour voir la réponse du modèle
Le sac à dos Adventurer Pro est-il étanche ?	Le sac à dos Adventurer Pro est étanche à l'eau.	Exécuter pour voir la réponse du modèle
Combien de poches intérieures a la tente ?	La tente SkyView 2 a 2 poches intérieures.	Exécuter pour voir la réponse du modèle
Le sac de couchage MountainDream est-il étanche ?	Oui, le sac de couchage MountainDream est étanche.	Exécuter pour voir la réponse du modèle
Quelle est la taille maximale d'un ordinateur portable ?	Le sac à dos SummitClimber peut accueillir un ordinateur portable.	Exécuter pour voir la réponse du modèle

At the top right, there are buttons for 'Configurations' and 'Ajoutez vos données'. A message in the center says: 'Obtenez des insights sur votre propre source de données. Vos données sont stockées en toute sécurité dans votre abonnement Azure. [Découvrir en détails sur la façon dont vos données sont protégées.](#)'

Finalement, c'est maintenant à vous d'évaluer si le modèle a réussi le test : pour chaque entrée, si la réponse du modèle est satisfaisante, cliquez sur l'icône du pouce vers le haut, sinon sur le

pouce vers le bas. Pour finir, vous obtenez un score d'évaluation qui indique les performances de votre modèle.

Enregistrez vos résultats en cliquant sur **Enregistrer les résultats**.

The screenshot shows the Azure Foundry interface for project 'tlevillayer-8194'. The left sidebar has 'Evaluation' selected. The main area is titled 'Configuration de l'assistant' with a 'Message système' section. On the right, there's a 'Configurations' panel with 'Modèle' set to 'gpt-4o', 'Réponse maximale' at 800, and 'Température' at 0.7. Below this is a 'Résultat de l'évaluation manuelle' section. The 'Enregistrer les résultats' button is highlighted with a red box. The results table shows three rows: 1) 'Données évaluées 100% (19/19)', 2) 'Pouce vers le haut 94.74% (18/19)', and 3) 'Pouce vers le bas 5.26% (1/19)'. The table also includes columns for 'Entrée', 'Réponse attendue', and 'Sortie'.

## Évaluation avancée

Nous pouvons aller de l'avant et essayer des méthodes d'évaluation plus avancées et standard qui sont disponibles (en mode préversion pour l'évaluation manuelle) dans Azure Foundry. Pour ce faire, cliquez sur **Évaluation automatisée**.

The screenshot shows the Azure Foundry interface for project 'tlevillayer-8194'. The left sidebar has 'Evaluation' selected. The main area is titled 'Assistant setup' with a 'System message' section. On the right, there's a search configuration panel with 'Index' set to 'index', 'Search type' to 'Hybrid (vector + keyword)', and 'Imported dataset' set to 'rande\_test\_data\_csv\_2024-12-11\_201539 UTC'. Below this is a 'Manual evaluation result' section. The 'Automated evaluation' button is highlighted with a red box. The results table shows three rows: 1) 'Data rated 0% (0/19)', 2) 'Thumbs up 0% (0/19)', and 3) 'Thumbs down 0% (0/19)'. The table includes columns for 'Input', 'Expected response', and 'Output'.

Ensuite, cliquez simplement sur **Suivant** jusqu'à ce que vous voyiez l'écran ci-dessous.

Cochez les **cases Ancrage et Pertinences** et choisissez la bonne connexion et le bon modèle déployé.

Allez en bas de la page et mappez les champs contexte, réponse et requête avec les bonnes valeurs. Cela devrait ressembler à ci-dessous.

How does your dataset map to your evaluation input? *			
Name	Description	Type	Data source
context	The source that response is generated with respect to	string	<input type="text" value="\${data.documents}"/>
response	The response to question generated by the model as answer	string	<input type="text" value="\${data.answer}"/>
query	A query seeking specific information	string	<input type="text" value="\${data.question}"/>

Ici, ce qui est important, c'est que le contenu de **data.documents** soit la partie qui a été récupérée du magasin d'index (RAG) lors de l'évaluation que nous avons faite ci-dessus !

Cliquez ensuite sur **Soumettre** et attendez la fin de l'évaluation.

Enfin, vous pouvez voir les résultats de cette évaluation : vous avez d'abord un tableau de bord qui affiche un aperçu des performances de vos modèles sur les métriques que vous avez choisies et lorsque vous faites défiler vers le bas, vous pouvez avoir les détails de chaque entrée.

Azure AI Foundry / tlevillayer-8194 / Evaluation / evaluation\_sleepy\_monkey\_qm0wpb6k6k

Refresh Export result Add custom chart View options Local

Overview Model catalog Playgrounds AI Services Build and customize Code PREVIEW Fine-tuning Prompt flow Assess and improve Tracing PREVIEW Evaluation Safety + security My assets Models + endpoints Data + indexes Web apps

## Evaluation details

### Metric dashboard

AI Quality (AI Assisted)

**Groundedness** ⓘ  
Average score 4.95

Score	Count
4	~1
5	~14

**Relevance** ⓘ  
Average score 4.26

Score	Count
3	~3
4	~10
5	~7

#### Detailed metrics result

Index	Status	Groundedness	Groundedness reason	Relevance	Relevance reason	Context	Query	Response
0	Completed ⓘ	5	The RESPONSE accurately and completely answers the QUERY using information directly from the CONTEXT.	4	The response is complete and directly answers the query with accurate information about the tent's capacity.	["content": "# Information about product item_number: 1\\n\\n# Information about product item_number: 1\\nTrailMaster X4 Tent\\n price \$250.\\n\\n## Brand\\nOutdoor\\n\\n## Category\\nTents\\n\\n## Features\\n\\n## Description\\n\\n## View more"]	Quelle est la capacité de la tente TrailMaster X4 ?	La tente TrailMaster X4 a une capacité de 4 personnes <sup>1</sup>
1	Completed ⓘ	5	The RESPONSE is fully grounded in the CONTEXT, providing a complete and accurate answer to the QUERY.	4	The response fully addresses the query by explaining the water resistance level of the backpack, making it a complete response.	["content": "# Information about product item_number: 2\\n\\n## Review\\\" I am extremely happy with my Adventurer Pro Backpack! It has ample space, multiple compartments, and is super comfortable to wear...\\n\\n## View more"]	Le sac à dos Adventurer Pro est-il étanche ?	Le sac à dos Adventurer Pro est fabriqué en nylon résistant à l'eau, ce qui aide à protéger votre équipement de la pluie légère. Cependant, il n'est pas complètement étanche <sup>1</sup>

## Action - Évaluer votre flux de chat (facultatif)

### Préparez votre flux de chat pour l'évaluation

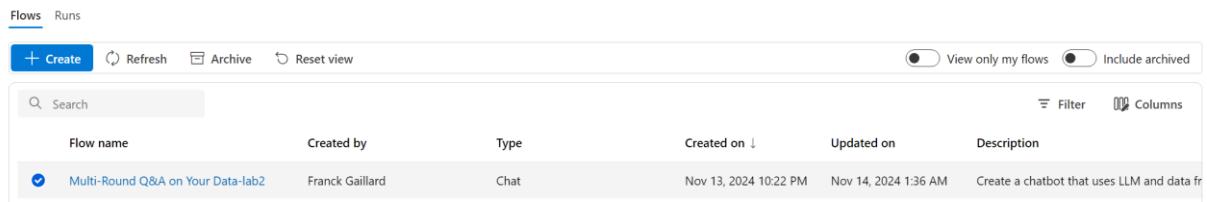
Pour que le flux RAG que vous avez créé précédemment soit évalué, vous devez inclure des informations supplémentaires au nœud de sortie de ce flux, en particulier le contexte utilisé pour générer la réponse.

Ces informations seront utilisées dans le cadre du processus d'évaluation.

Pour ce faire, il suffit de suivre ces étapes :

Dans la section **Flux d'invite**, ouvrez le flux **Multi-Round Q&A on Your Data** que vous avez créé dans l'atelier précédent. **Ce sera le flux que nous utiliserons pour l'évaluation.**

Create, iterate, and debug your orchestration flows



Flow name	Created by	Type	Created on	Updated on	Description
Multi-Round Q&A on Your Data-lab2	Franck Gaillard	Chat	Nov 13, 2024 10:22 PM	Nov 14, 2024 1:36 AM	Create a chatbot that uses LLM and data fr

Créez une sortie nommée **documents** dans le **nœud Sorties**. Cette sortie représentera les documents qui ont été récupérés dans le **nœud de recherche (lookup)** et qui ont ensuite été formatés dans le **nœud generate\_prompt\_context**.

En créant une sortie nommée **documents** et en lui attribuant les documents formatés, vous vous assurez que le contexte utilisé pour générer les réponses est explicitement suivi. Ce contexte inclut les documents récupérés et mis en forme lors de l'exécution du flux.

Attribuez la sortie du nœud **generate\_prompt\_context** à la sortie des **documents**, comme illustré dans l'image ci-dessous.



Name	Value	Chat output	Action
chat_output	\${chat_with_context.output}		
documents	\${generate_prompt_context.output}		

+ Add output

Cliquez sur **Enregistrer** avant de passer à la section suivante.

### Créez vos flux d'évaluation

Toujours dans l' élément Flux d'invite dans la section Outils, cliquez sur le bouton bleu **Créer**.

## Create a new flow

Create by type

Standard flow

Chat flow

Evaluation flow

Explore gallery

All Standard flow Chat flow Evaluation flow

View more samples

Multi-Round Q&A on Your Data

Web Classification

Chat with Wikipedia

Use GPT Function Calling

Classification Accuracy Evaluation

QnA Groundedness Evaluation

QnA Relevance Evaluation

Import

Cancel

Sélectionnez le **filtre Evaluation flow** et cliquez sur **Cloner** sur la carte d'évaluation de la **QnA Groundedness Evaluation**.

Explore gallery

All Standard flow Chat flow Evaluation flow

Classification Accuracy Evaluation

QnA Groundedness Evaluation

QnA Relevance Evaluation

QnA Coherence Evaluation

QnA Fluency Evaluation

QnA Ada Similarity Evaluation

QnA GPT Similarity Evaluation

QnA F1 Score Evaluation

View detail Clone

Cancel

## Clone flow

X

The flow code files are stored in a specific folder within your workspace file share storage. This folder name can be customized according to your preferences.

Location to store flow \* ⓘ

Users/fgail/promptflow

Folder name \* ⓘ

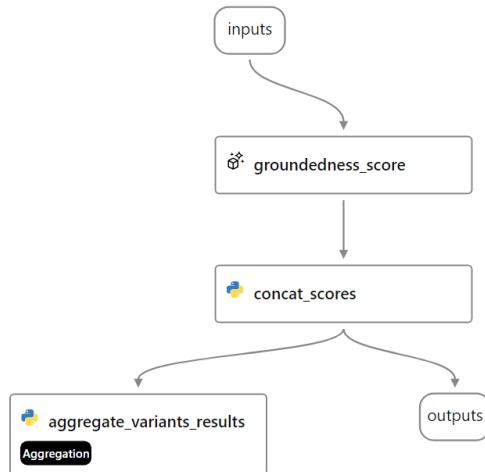
QnA Groundedness Evaluation-lab2

Clone

Cancel

Un flux sera créé avec la structure suivante :

## Graph



Mettez à jour le **champ Connexion** pour pointer vers un **déploiement gpt-4o** dans **groundedness\_score** nœud et mettez également à jour **max\_tokens à 1000**, comme illustré dans la figure suivante.



Après avoir mis à jour les informations de connexion, cliquez sur **Enregistrer** dans le flux d'évaluation et accédez à la section Flux dans l' **élément Flux d'invite**.

Vous allez maintenant répéter les mêmes étapes que celles décrites jusqu'à présent pour créer **deux** flux d'évaluation supplémentaires, l'une **QnA Relevance Evaluation** et l'autre **QnA GPT Similarity Evaluation**.

Category	Evaluation Type	Description	Action Buttons
Evaluation	Classification Accuracy Evaluation	Measuring the performance of a classification system by comparing its outputs to groundtruth.	[View detail] [Clone]
	QnA Groundedness Evaluation	Compute the groundedness of the answer for the given question based on the context.	[View detail] [Clone]
Evaluation	QnA Relevance Evaluation	Compute the relevance of the answer for the given question based on the context.	[View detail] [Clone]
	QnA Coherence Evaluation	Compute the coherence of the answer base on the question using llm.	[View detail] [Clone]
Evaluation	QnA Fluency Evaluation	Compute the Fluency of the answer base on the question using llm.	[View detail] [Clone]
	QnA Ada Similarity Evaluation	Compute the cosine similarity between the answer and the ground truth embedded with ada embedding.	[View detail] [Clone]
Evaluation	QnA GPT Similarity Evaluation	Compute the similarity of the answer base on the question and ground truth using llm.	[View detail] [Clone]
	QnA F1 Score Evaluation	Compute the F1 Score based on words in answer and ground truth.	[View detail] [Clone]

Mettez à jour le **champ Connexion** pour pointer vers un **déploiement gpt-4o** dans **relevance\_score** nœud, mettez également à jour **max\_tokens à 1000** comme indiqué dans la figure suivante.



Mettez à jour le **champ Connexion** pour pointer vers un **déploiement gpt-4o** dans **similarity\_score** nœud, mettez également à jour **max\_tokens à 1000**, comme illustré dans la figure suivante.

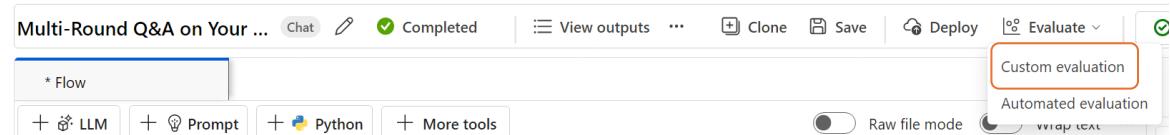


### *Exécuter l'évaluation*

Dans la section Flux d'invites, ouvrez le flux de **Multi-Round Q&A on Your Data** que vous avez créé dans l'atelier précédent. Ce sera le flux que nous utiliserons pour l'évaluation.

Démarrez le runtime automatique en sélectionnant **Démarrer la session de calcul** dans la liste déroulante.

Sélectionnez l'option **d'évaluation personnalisée** dans le menu Évaluer.



**Les variantes d'invite** font référence à différentes versions d'un nœud d'invite ou d'outil qui ont des paramètres distincts.

Dans l'**option Prompt\_variants**, sélectionnez l'option permettant d'exécuter uniquement **deux variantes** pour éviter d'atteindre votre limite de quota de modèle GPT-4o, comme le montre l'exemple d'image ci-dessous.

**Batch run & Evaluate**

Basic settings

Run display name \* ⓘ  
Multi-Round Q&A On Your Data-11-13-2024-22-20-09-\${variant\_id}-\${timestamp}

Run description

Tags

Key	Value
+ Add tag	

Variants \*

Select a node with variants that you want to run. Note: other nodes will run with default variant.

Select a node to run variants  Use default variant for all nodes

Node name	Variant id
Prompt_variants	variant_0(default),variant_1
(1) 2 run(s) will be generated based on selected variant(s)	<input type="checkbox"/> (Select all) <input checked="" type="checkbox"/> variant_0(default) <input checked="" type="checkbox"/> variant_1 <input type="checkbox"/> variant_2

Previous Next Review + submit Cancel

Sélectionnez Ajouter de nouvelles données.

**Batch run & Evaluate**

Basic settings

Batch run settings

Data \* ⓘ  
surface-pro-index (version 1)

Selected file must be .jsonl, .csv, .tsv, or a folder containing these types.  
Select a .jsonl, .csv, or .tsv file, or a folder containing these file types.

+ Add new data

Input mapping \*

Input	Type
chat_history	list
chat_input	string

Cannot pre

Previous Next Review + submit Add Cancel

**Add new data**

Name \*

Data with same name will be saved as a new version

Upload from local file  Upload from local folder

Choose a file \*

Select a supported file type: .csv, .tsv, .json  
Please make sure the data includes headers

Téléchargez le fichier **data.csv**.

## Add new data

X

Name \*

 \*

Data with same name will be saved as a new version

Upload from local file  Upload from local folder

Choose a file \*

Please make sure the data includes headers

Après avoir cliqué sur **Ajouter**, procédez au mappage des champs de saisie comme indiqué ci-dessous :

### Batch run & Evaluate

X

- Basic settings
- Batch run settings
- Evaluation settings
- Select evaluation  
optional
- Configure evaluation  
optional
- Review

#### Batch run settings

Data \* ⓘ

testdata (version 1)

Selected file must be jsonl, .csv, .tsv, or a folder containing these types.

• Select a jsonl, .csv, or .tsv file, or a folder containing these file types.

+ Add new data

#### Input mapping \*

Input	Type	Dataset column
chat_history	list	<code>\$(data.chat_history)</code>
chat_input	string	<code>\$(data.question)</code>

#### Preview of top 5 rows

chat_history	question	answer	documents
[ ]	What does Windows 10 Provides?	Windows 10 offers a variety of new featu...	Windows 10 provides new features and ...
[ ]	How much RAM does Surface Pro 4 can ...	Surface Pro 4 is available with up to 16 ...	Memory and storage Surface Pro 4 is av...
[ ]	How do I check the battery level on my ...	You can check the battery level from the ...	Check the battery level You can check th...
[ ]	What processor does the Surface Pro 4 h...	The Surface Pro 4 is equipped with a 6th...	Processor The 6th-generation Intel Core ...
[ ]	Can I use a pen with the Surface Pro 4?	Yes; the Surface Pro 4 comes with the Su...	Surface Pen Enjoy a natural writing exper...

Previous

Next

Review + submit

Cancel

Sélectionnez les trois flux d'évaluation que vous venez de créer.

**Batch run & Evaluate**

Cliquez sur **Suivant** pour configurer les champs **de question, de contexte, de ground\_truth et de réponse** pour chaque flux d'évaluation. Vous pouvez voir comment procéder dans les trois images ci-dessous.

**QnA GPT Similarity Evaluation-lab2**

**Evaluation input mapping \***

**Choose data asset for evaluation \***

testdata (version 1)

[+ Add new data](#)

Name	Description	Type	Data source
question		string	\$(data.question)
ground_truth		string	\$(data.answer)
answer		string	\$(run.outputs.chat_output)

**QnA Groundedness Evaluation-lab2**

**Evaluation input mapping \***

**Choose data asset for evaluation \***

testdata (version 1)

[+ Add new data](#)

Name	Description	Type	Data source
question		string	\$(data.question)
context		string	\$(data.answer)
answer		string	\$(run.outputs.chat_output)

QnA Relevance Evaluation-lab2 ✓

Apply to all evaluations

Evaluation input mapping \*

Choose data asset for evaluation \* ⓘ

testdata (version 1)

+ Add new data

Name	Description	Type	Data source
question		string	\$(data.question)
context		string	\$(data.answer)
answer		string	\$(run.outputs.chat_output)

Cliquez sur **Soumettre** pour démarrer l'évaluation.

Le processus d'évaluation a commencé. Pour afficher toutes les évaluations (une par variante), accédez à la **section Évaluation**.

Assess and compare AI application performance

Automated evaluations PREVIEW Manual evaluations Evaluator library PREVIEW

Evaluate the quality and safety of your generative AI applications with industry standard metrics to compare and choose the best version based on your need. [Learn more about metrics.](#)

+ New evaluation ⏪ Refresh × Cancel ⏪ Delete ⏪ Compare ⏪ Show batch runs ⏪ Switch to dashboard view

**Evaluation process**

1. Begin evaluation
2. Prepare data
3. Identify metrics
4. Analyze results
5. Revise
6. Compare results

Evaluations

Evaluations	Status	Created on	Groundedness	Relevance	Retrieval score	Coherence	Similarity
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	4.2	--	--	--	--
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	3.6	--	--	--	--
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	--	--	--	--	4.25
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	--	--	--	--	4.65
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	--	4.95	--	--	--
Multi-Round Q&A on Your Data-11-13-2024-22-20- ✓ Completed	Completed	Nov 14, 2024 2:50 AM	--	4.9	--	--	--

**Remarque :** vous pouvez utiliser le SDK pour évaluer vos modèles => [Évaluer avec le SDK d'évaluation Azure AI - Azure AI Studio | Microsoft Learn](#)

### Déployer le flux RAG sur un point de terminaison géré en ligne

Ouvrez le flux **Multi-Round Q&A on Your Data** que vous avez créé dans le labo précédent.

Maintenant que vous avez créé un flux et que vous l'avez testé correctement, il est temps de créer votre point de terminaison en ligne pour une inférence en temps réel.

Suivez les étapes ci-dessous pour déployer un flux d'invite en tant que point de terminaison en ligne dans Azure AI Studio.

1. Disposez d'un flux d'invite prêt à être déployé.

## 2. Sélectionnez Déployer dans l'éditeur de flux.

### Deploy Multi-Round Q&A on Your Data-lab2

**1 Basic settings**

**2 Advanced settings**

**3 Review**

**Basic settings**

Deploy your flow to a managed online endpoint for real-time inference. [Learn more](#)

**Endpoint**

New  Existing

**Endpoint name \*** [\(i\)](#)  
project-rag-lab2-endpoint

**Deployment name \*** [\(i\)](#)  
project-rag-lab2-endpoint-1

**Virtual machine \*** [\(i\)](#)  
Standard\_DS3\_v2 4 Cores, 14 GB (RAM), 28 GB (Disk), \$0.29/hr

**Instance count \*** [\(i\)](#)  
3

**Inferencing data collection** [\(i\)](#)  
 Enabled

### Deploy Multi-Round Q&A on Your Data-lab2

**1 Basic settings**

**2 Advanced settings**

**3 Review**

**Review the deployment settings**

Basic settings		Deployment	
Endpoint name	project-rag-lab2-endpoint	Tags	<a href="#">(i)</a> No tags
Deployment Name	project-rag-lab2-endpoint-1	Environment	Use environment of current flow definition
Virtual machine	Standard_DS3_v2		
Instance count	3		
Inferencing data collection	Enabled		
Application Insights diagnostics	Disabled		

Outputs	
Output name	Type
chat_output	string
documents	string

Connection			
Node name	Provider	Connection	Deployment name / Model
chat_with_context	AzureOpenAI	ai-hublab201314979 3759_aoai	gpt-4o
modify_query_with_history	AzureOpenAI	ai-hublab201314979 3759_aoai	gpt-4o

**Endpoint**

Authentication type: Key  
Public network access: Enabled  
Description: --  
Identity type: system  
Enforce access to connection secrets (preview): Enabled  
Endpoint tags:

**Create** **Back** **Cancel**

## Manage deployments of your models, apps, and services

Deploy a model with your private API key and an endpoint URI (Uniform Resource Identifier).

[Model deployments](#) App deployments Service deployments

A screenshot of the Azure AI Foundry interface. At the top right, a green notification box displays the message: "Prompt flow deployment 'project-rag-lab2-endpoint-1' of endpoint 'project-rag-lab2-endpoint' succeeded Allocating traffic for deployment View details". Below this, there's a toolbar with buttons for "Deploy model", "Refresh", "Edit", "Open in playground", and "Reset view". The main area shows a table of deployed models. The columns are: Name, Model name, Model version, State, Model retirement date, Content filter, and Deployment type. The table contains the following data:

Name	Model name	Model version	State	Model retirement date	Content filter	Deployment type
ai-hublab2013149793759_aoai	Azure OpenAI					
gpt-4o	gpt-4o	2024-05-13	Succeeded		DefaultV2 ⓘ	Global Standard
text-embedding-3-large	text-embedding-3-large	1	Succeeded		DefaultV2 ⓘ	Standard
text-embedding-ada-002	text-embedding-ada-002	2	Succeeded		Default ⓘ	Standard
project-rag-lab2-endpoint	Endpoint					
project-rag-lab2-endpoint-1	project-rag-lab2-endpoint		Succeeded			

Vous pouvez le tester :

[project-rag-lab2-endpoint-1](#)

Details Test Consume Monitoring PREVIEW Logs

A screenshot of the "Test" tab for the "project-rag-lab2-endpoint-1" deployment. The interface includes a "Chat mode" button, a text input field with the placeholder "how long does it take to charge the surface pro4?", and a message box stating "It takes two to four hours to charge the Surface Pro 4 battery fully from an empty state (Source: surface-pro-4-user-guide-EN.pdf)".

## Découvrir la sécurité du contenu

Action

Via la page Sûreté + sécurité

Allez dans la section Sûreté + sécurité de votre projet, sélectionnez l'onglet Filtres de contenu et cliquez sur Crée un filtre de contenu

A screenshot of the "Safety + security" page in the Azure AI Foundry interface. The left sidebar shows navigation links like Overview, Model catalog, Playgrounds, AI Services, Build and customize, Code, Fine-tuning, Prompt flow, Assess and improve, Tracing, Evaluation, and Safety + security (which is highlighted with a red box). The main content area has a heading "Here to help you build AI safely and securely" and a sub-section "Content filters work alongside core models. Create filters within a project and assign to deployments to manage content by category." It features a "Create content filter" button (also highlighted with a red box) and a table listing existing content filters. The table columns are: Name, Applied deployment, and Modified at. The table data is as follows:

Name	Applied deployment	Modified at
ai-tlevillayer2386ai988273461930_aoai	Azure OpenAI Connection	-
CustomContentFilter790		Dec 11, 2024 6:21 PM
CustomContentFilter607	gpt-4o	Dec 11, 2024 6:25 PM

Choisissez votre **connection** dans l'entrée demandée et cliquez sur Suivant

Azure AI Foundry / tlevillayer-8194 / Safety + security / Create content filter

Create filters to allow or block specific types of content

Basic information

Add basic information

Name \* CustomContentFilter962

Connection \* ai-tlevillayer2386ai988273461930\_aoai

Next

C'est ici que nous définissons les seuils de sécurité du contenu. Tout d'abord, nous devons définir le filtre pour les données d'entrée (saisies par l'invite de l'utilisateur), puis nous définissons le filtre de sortie pour les réponses générées par le modèle. Dans les deux cas, réglez tous les seuils sur **bas** afin que les mesures de sécurité se déclenchent plus facilement.

Azure AI Foundry / tlevillayer-8194 / Safety + security / Create content filter

Create filters to allow or block specific types of content

Input filter

Category	Media	Action	Threshold
Violence	Text Image	Annotate and block	Low
Hate	Text Image	Annotate and block	Low
Sexual	Text Image	Annotate and block	Low
Self-harm	Text Image	Annotate and block	Low
Prompt shields for jailbreak attacks	Text	Annotate and block	Jailbreak attacks will be blocked
Prompt shields for indirect attacks	Text	Off	Content will not be annotated at all

Blocklist (Preview)

Blocklist is off. Turn it on to find and block content with harmful words.

Back Next Cancel

Une fois cela fait, sélectionnez le modèle sur lequel vous souhaitez appliquer ces filtres de contenu et cliquez sur **Suivant**.

Enfin, sur l'écran de révision, cliquez sur Crée un filtre en bas de la page.

The screenshot shows the 'Create filters to allow or block specific types of content' step in the Azure AI Foundry interface. On the left, a sidebar lists various sections like Overview, Model catalog, Playgrounds, AI Services, etc. The 'Safety + security' section is currently selected. The main panel has a flowchart showing 'Basic information', 'Input filter' (which is checked), 'Output filter' (which is checked), 'Deployment (optional)', and 'Review'. To the right, there's a section titled 'Apply filter to deployments (optional)' with a 'Connection' dropdown set to 'ai-tlevillayer2386ai988273461930\_aoai'. Below it is a 'Deployments' table:

	Name	Model name	Mod
<input type="checkbox"/>	Mistral-large-2407	Mistral-large-2407	1
<input checked="" type="checkbox"/>	gpt-4o	gpt-4o	2024
<input type="checkbox"/>	text-embedding-ada-002	text-embedding-ada-002	2

At the bottom of the main panel are 'Back' and 'Next' buttons, with 'Next' being highlighted with a red box.

Vous êtes prêt. Maintenant, retournez dans la partie **Playgrounds** et testez votre filtre.

Essayez-le avec l'expression : « comment choisir un couteau pour la randonnée ? »

Vérifiez la réponse de **GPT-4o**, le filtre Violence a été déclenché avec le texte.

The screenshot shows the 'Chat playground' interface. At the top, there are buttons for 'View code', 'Evaluate', 'Deploy', 'Import', 'Export', 'Prompt samples', and 'Send feedback'. The 'Deployment' dropdown is set to 'gpt-4o (version:2024-08-06)'. In the 'Chat history' section, a user prompt 'Comment choisir son couteau pour partir en randonnée?' is shown. A message box indicates that the prompt was filtered due to triggering Azure OpenAI's content filtering system, specifically for 'Violence (medium)'. The message also says to 'Please modify your prompt and retry.' At the bottom left, there's a note: 'Give the model instructions and context' followed by 'You are an AI assistant that helps people find information.'

Revenez à la page **Sûreté + sécurité** et définissez les seuils sur **Élevé** pour les filtres d'entrée et de sortie. Ensuite, allez sur **Playgrounds** et posez la même question que ci-dessus. Maintenant, le modèle devrait vous répondre en conséquence.

The screenshot shows the 'Chat playground' section of the Playgrounds interface. On the left, there's a sidebar with 'Setup' (Deployment: Spt-40 (version 2024-08-06)), 'Give the model instructions and context' (You are an AI assistant that helps people find information.), and sections for 'Add your data' and 'Parameters'. The main area is titled 'Chat history' and contains a message: 'Comment choisir son couteau pour partir en randonnée ?'. Below the message, there's a detailed list of factors to consider when choosing a knife for a hike, such as type of blade (folding vs fixed), material of the blade (stainless steel vs carbon steel), weight, grip, and budget.

### *Directement depuis le Playgrounds*

- Tout d'abord, testons le comportement du modèle **GPT-4o**, sélectionnons l'**option Playgrounds**, et l'**option Chat** et copions l'invite suivante dans la cellule « **Donner des instructions et un contexte au modèle** » :

**Vous êtes un assistant IA qui aide les entreprises de télécommunications à extraire des informations précieuses de leurs conversations en créant des fichiers JSON pour chaque transcription de conversation que vous recevez.**

**Vous essayez toujours d'extraire et de formater au format JSON, les noms des champs entre crochets :**

1. Nom du client [nom]
2. Téléphone du contact client [téléphone]
3. Sujet principal de la conversation [sujet]
4. Sentiment des clients (neutre, positif, négatif)[sentiment]
5. Comment l'agent a géré la conversation [comportement\_agent]
6. Quel a été le résultat final de la conversation [résultat]
7. Un résumé très bref de la conversation [résumé]

## ← Chat playground ▾

</> View code

Setup  Hide

Deployment \* Create new deployment

Mistral-large-2407-labai

Give the model instructions and context (i)

[agent\_behavior]  
6. What was the FINAL Outcome of the Conversation  
[outcome]  
7. A really brief Summary of the Conversation [summary]  
Only extract information that you're sure. If you're unsure, write "Unknown/Not Found" in the JSON file.

Apply changes  Generate prompt  ↻

+ Add section ▾

Safety system messages  
Examples  
Variable

- Cliquez sur **Ajouter une section** puis sur **Messages système de sécurité** et cochez « **Avoid harmful content** » :

Select safety system message(s) to insert

Insert one or more prepared system messages into your prompt; you can alter or add to them if you'd like. Token usage will be incurred when you begin chatting with the model in the playground.

Select all (276 tokens)

Avoid harmful content (61 tokens)  
 Avoid ungrounded content (93 tokens)  
 Avoid copyright infringements (81 tokens)  
 Avoid jailbreaks and manipulation (41 tokens)

Insert  Cancel

- Copiez cette conversation dans l'invite et cliquez sur **Appliquer les changements** :

Agent : Bonjour M. Perez, bienvenue au service client de Telco. Je m'appelle Juan, comment puis-je vous aider ?

Client : Bonjour, Juan. Je suis très insatisfait de vos services.

Agent : ok monsieur, je suis désolé d'entendre cela, comment puis-je vous aider ?

Client : Je déteste cette entreprise, je tuerai tout le monde avec une bombe.

- Vérifiez la réponse de **GPT-4o**, le filtre Violence a été déclenché avec le texte.

The screenshot shows the Azure AI Foundry interface for a 'Chat playground'. On the left, a sidebar lists various sections like Overview, Model catalog, Playgrounds (selected), AI Services, Build and customize, Code, Fine-tuning, Prompt flow, Assess and improve, Tracing, Evaluation, Safety + security, My assets, Models + endpoints, Data + indexes, and Web apps. A 'Management center' button is at the bottom. The main area has tabs for 'View code' and 'Setup'. Under 'Setup', there's a 'Deployment' dropdown set to 'Mistral-large-2407-labai', a 'Create new deployment' button, and a 'Give the model instructions and context' section with numbered prompts. Below that is a 'Safety system message' section with a '## To Avoid Harmful Content' rule. A 'Parameters' section is partially visible. On the right, a 'Chat session' window shows a conversation between an Agent and a Client. The Client's message contains the word 'Bomb', which is flagged as harmful content. A pink warning box states: 'The response was filtered due to the prompt triggering Microsoft's content management policy. Please modify your prompt and retry.' The Agent's response is: 'Agent: Hi Mr. Perez, welcome to Telco's customer service. My name is Juan, how can I assist you? Client: Hello, Juan. I am very dissatisfied with your services. Agent: ok sir, I am sorry to hear that, how can I help you? Client: I hate this company I will kill everyone with a bomb.'

- Le mot « Bombe » a été détecté comme contenu nuisible et la réponse a été filtrée.