L'IA générative : passé l'effet « waouh », on en fait quoi ?

Marc Juillard
Société Générale Insurance
Directeur DataHub

Pierre-Marie Beretti
Dataltist
Manager

Vincent Van Steenbergen
ThinkDeep AI
CEO & founder

Transformers vs approche classique

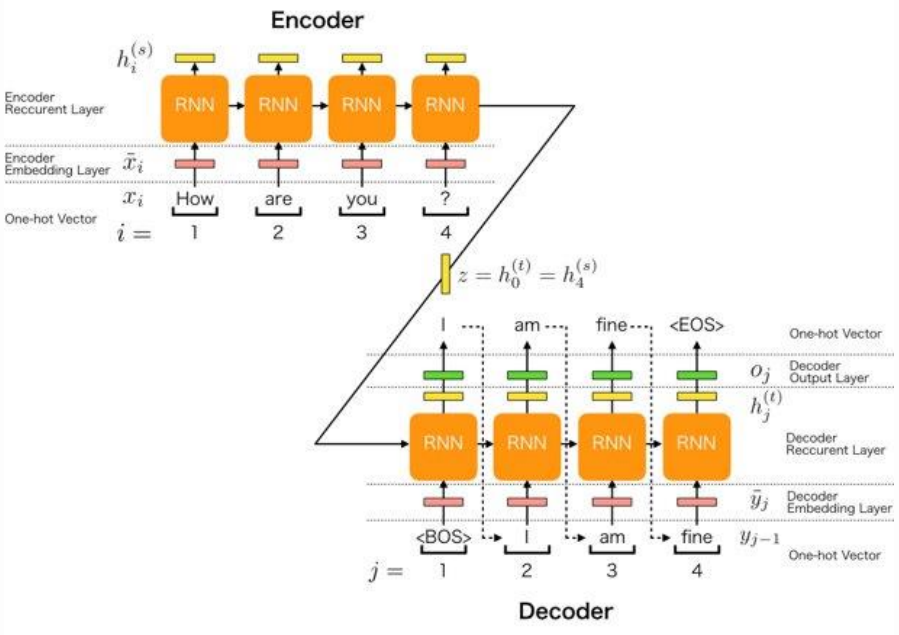
ACTUAIRES 100% DATA SCIENCE BY Institut A

Avant 2017
Word Embeddings + LSTM

Un embedding de type Word2Vec générique : **un jeu de coordonnées sémantique (ou “embedding”) pour chaque mot**.

Un ensemble de Réseaux de Neurones Récurrents en charge de prendre en compte le contexte spécifique de la phrase.

L’ensemble conduit à des modèles de plus en plus complexes, séquentielles et donc extrêmement longs à entrainer.

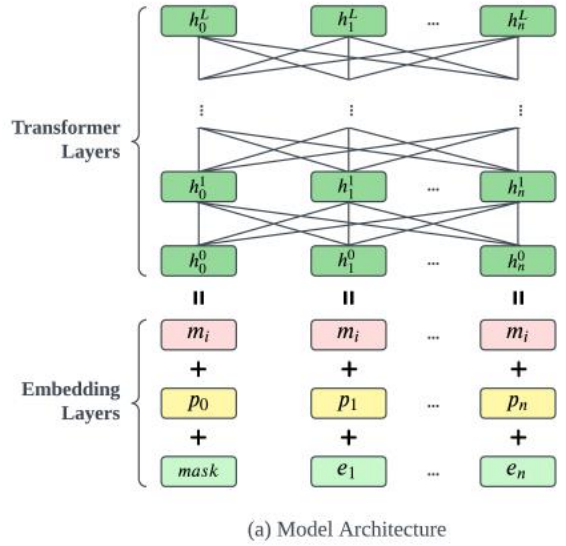


2017
Transformers

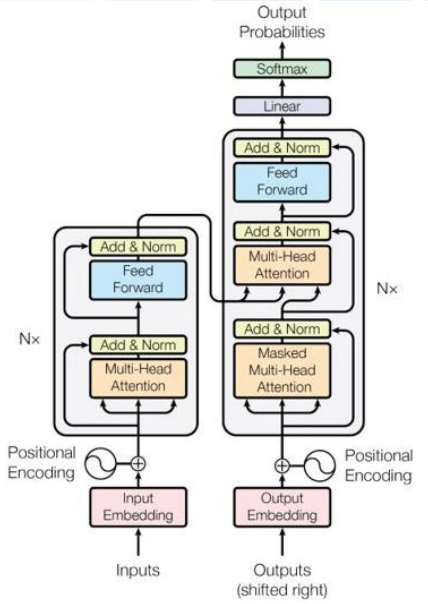
Le papier de Google “Attention is all need” introduit les modèles de types transformers.

Il contiennent des couches d’attention qui permettent de prendre en compte les relations entre les mots dans une sequence de texte. **un jeu de coordonnées pour chaque occurrence précise du mot.** Le mot baguette *dans les phrases “La Baguette du Boulanger” et “la Baguette du Magicien” reçoivent alors un embedding différent.*

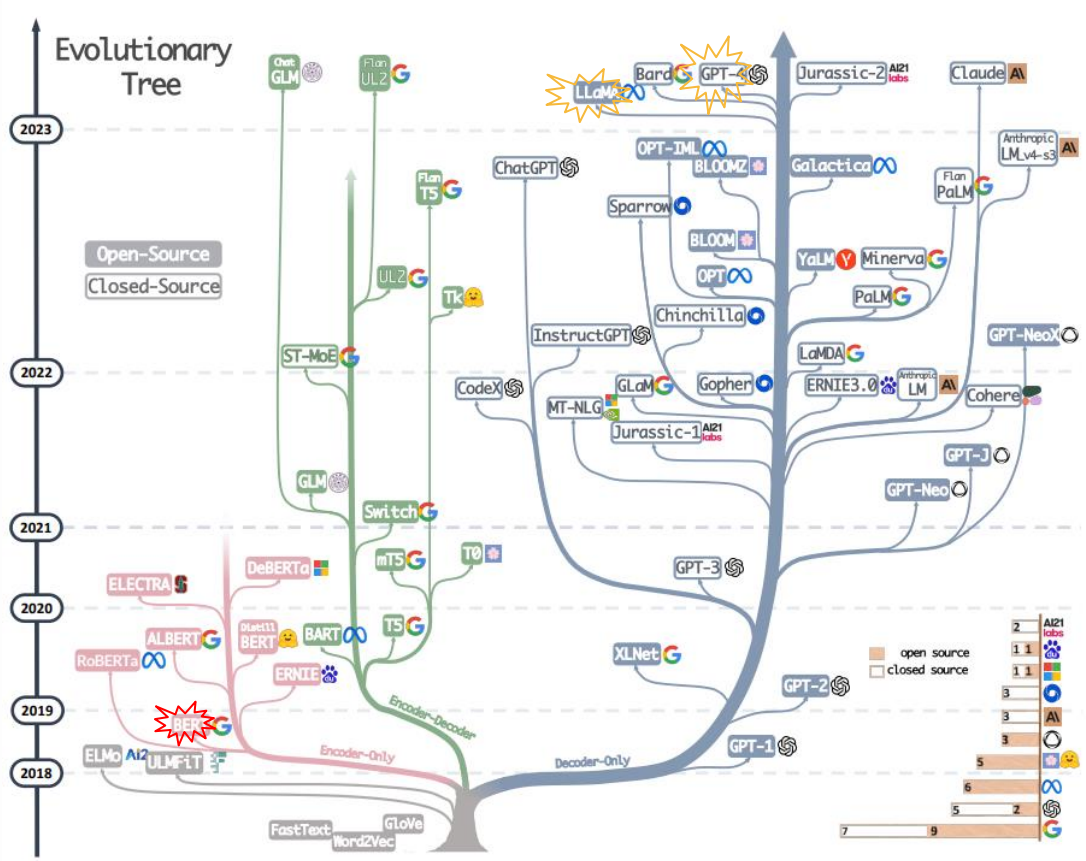
Ces modèles peuvent de plus être **entrainer en parallèle ce qui leur permet d’être entrainer sur des corpus de très gros volume de textes.**



(a) Model Architecture



Transformers : Encodeurs et Décodeurs

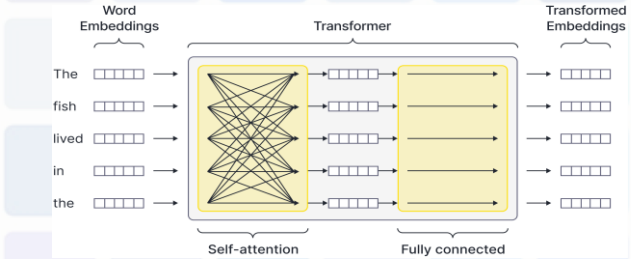


2018 : BERT *(Bidirectional Encoder Représentations form Transformers)*

Modèle d'encodage de 110 à 340 millions de paramètres entraînés sur 3,3 milliards de mots (wiki & books2) pour un dictionnaire de 30 000 à 50 000 mots

- Considère le contexte situé à gauche et à droite du mot ce qui lui permet de lever énormément d'ambiguïté de langage. Ceci permet de simplifier la structure des modèles NLP, une grande partie des couches de LSTM étant remplacée par Bert.

- Extrêmement performant pour de l'extraction d'information, de la classification, de l'analyse de sentiment, de la recherche de similarité Mais ne peut générer du texte.



- Tourne sur de la CPU et peut être fine tuné avec les données de l'entreprise.

2023 : GPT3.5 /LlaMa 2

Les deux modèles sont basés sur des structures de langage autorégressives, ne lisant donc le texte que de gauche à droite.

GPT 3.5 : Modèle de décodage de 175 milliards de paramètres entraînés sur des milliards de pages internet contenant plus de 300 milliards de mots. Taille de contexte de 4 096 tokens pour GPT3.5 et peut monter jusqu'à 32 768 tokens pour GPT4 (et 128k pour GPT4 turbo)

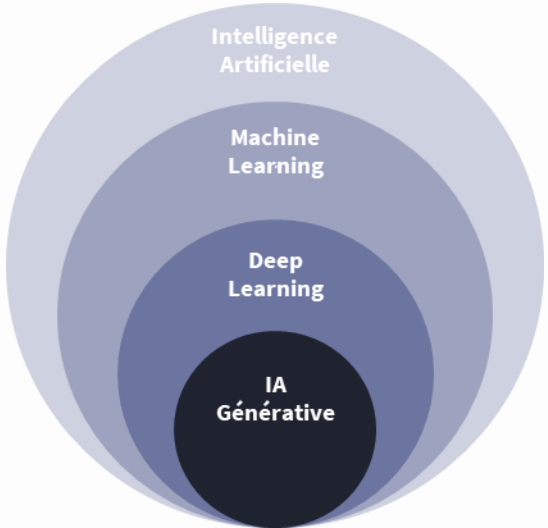
LlaMa2 : plusieurs versions de modèles sont disponibles (7, 13 et 70 milliards de paramètres), avec une taille de contexte maximale de 4086 Tokens

IA GEN en deux mots

ACTUAIRES 100% DATA SCIENCE BY Institut A

L'IA générative est une évolution majeure des technologies d'IA permettant :

- Le développement de solutions de type ChatGPT grâce à une compréhension avancée du langage naturel.
- De nouveaux cas d'usages désormais possibles grâce à la possibilté de générer du contenu (texte, images ou vidéos personnalisées…).
- Une accélération du déploiement de cas d'usages qui nécéssitaient des volumes de données importants à collecter (ex : analyse de sentiments des verbatims clients).



De nombreux cas d'usages autour de la productivité individuelle, de l'efficacité opérationnelle, de l'expérience client,…

Génération de Contenu	- **Relation Client** : génération de réponses pour les clients et les utilisateurs - **Marketing** : génération de communications, de visuels
Génération de Code	- **DSI** : génération automatique de code, détection d'erreurs, migration entre langages informatiques, tests unitaires, …
Résumé	- **Relation Client** : résumé d'appels, analyse de verbatims. - **Communication/Marketing** : synthèse de rapport, rapports public…
Extraction de données	- **Optimisation traitement gestion** : Extraction d'informations des documents clients
Recherche Q&A	- **Analyse exposition** : contrats, fond d'investissements,…. - **Relation Client** : assistant conseillers - **Formation** : moteur de recherche au-travers de notre base de formation

Nous proposons de faire un focus sur la mise en place :

- un **modèle avancé de Q&A**
- la mise en place d'un **assistant conversationnel**

Ecosystème GEN AI

Un écosystème Open Source très dynamique

- L'écosystème GEN AI est particulièrement dynamique, notamment en termes de librairies et outils Open Source.
- On dispose d'un éventail de modèles Open Source (Mistral, LLama, Zephyr…) qui s'enrichit chaque jour et est accessible très facilement.
- **La tendance actuelle est à la spécialisation de petits modèles de langage pour les adapter à des problématiques métier (ontologie, terminologie ou base de connaissances spécifiques).**

LLaMA by Meta

MISTRAL AI_

Zephyr®

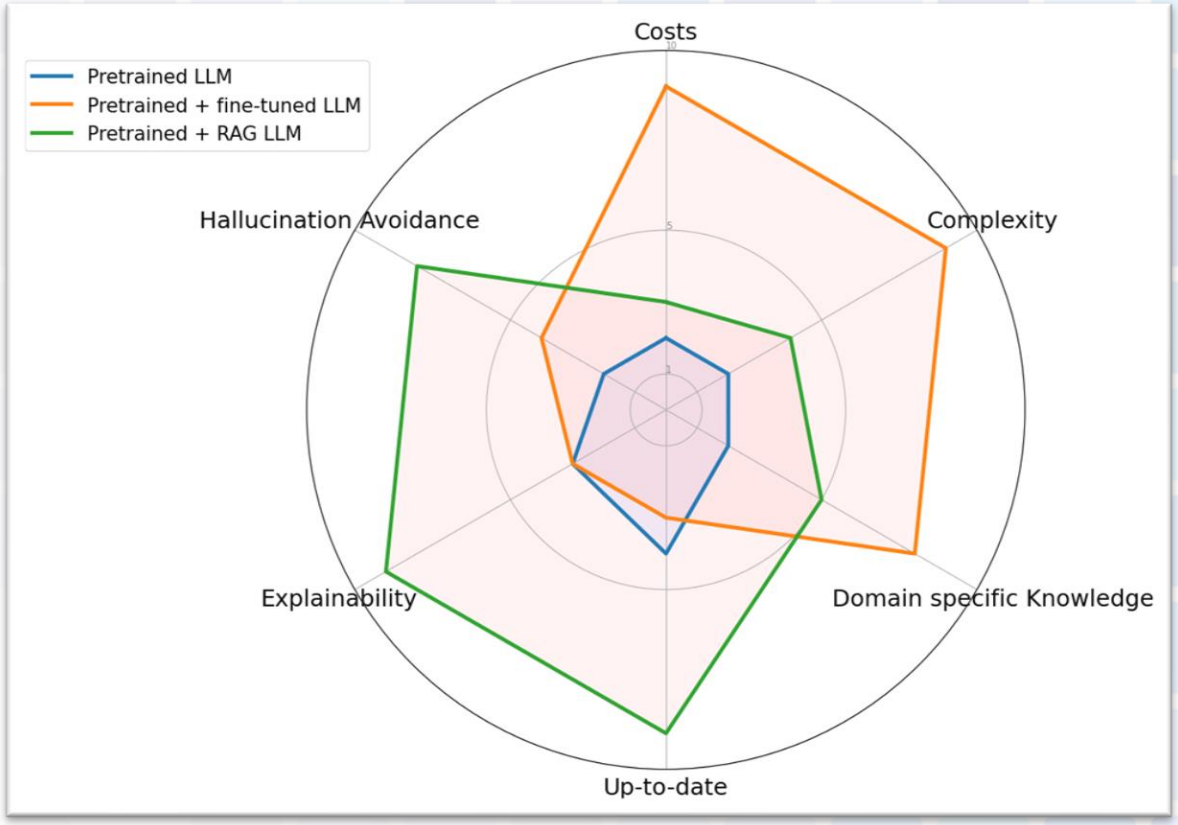
Différentes approches de training

Bien savoir où l'on veut aller

1. Réutiliser d'un modèle existant **« tel quel »**
   - ✓ marche bien s'il y a un modèle qui correspond déjà exactement au besoin
   - ✓ limité si besoin d'un assistant métier spécifique à un domaine.

2. **Approche RAG** : Residual Augmented Generation
   - ✓ modèle de fondation enrichi avec une base de connaissances documentaire
   - ✓ permet d'alimenter un assistant avec des données « actualisées ».
   - ✓ Approprié pour limiter les cas d'hallucination

3. **Fine-tuner** un modèle de fondation :
   - ✓ permet de constituer son propre modèle personnalisé
   - ✓ nécessaire pour que l'assistant parle un langage, une terminologie ou une logique spécifique
   - ✓ besoin de constituer un jeu de données de questions/réponses types (1000+ exemples).

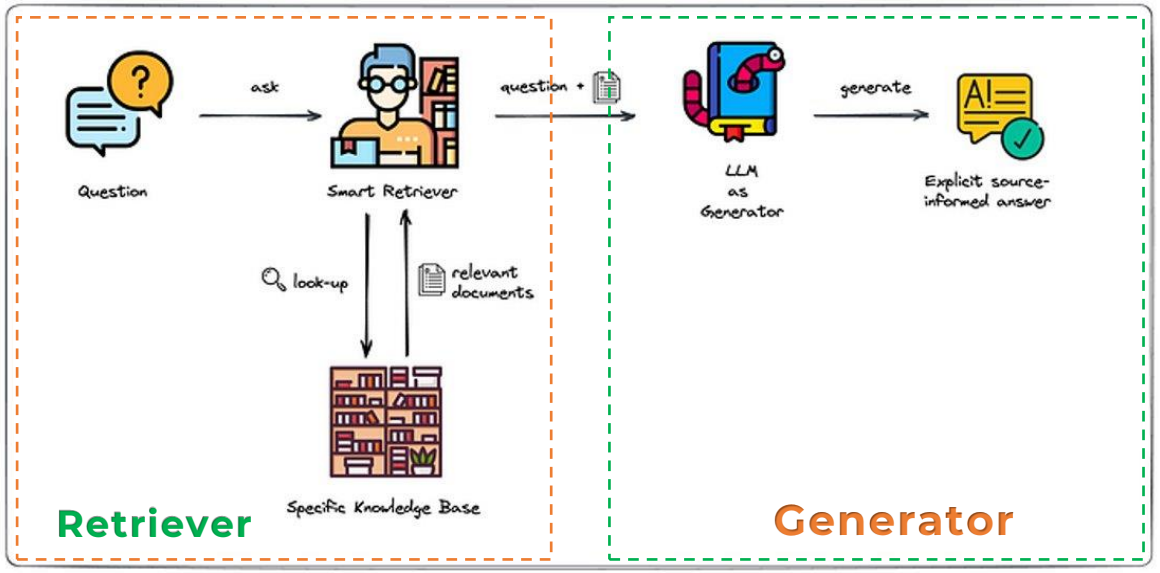


The limits of LLMs and how RAG remedies them (konfuzio.com)

Retrieval Augmented Generation (RAG)

Les LLM sont entraînés pour prédire le prochain mot d'une phrase. Si l'emploie de méthodes RLHF a permis d'améliorer la qualité des réponses, ils ont toujours tendance à halluciner.

Le RAG est une architecture qui diminue les risques d'hallucination en contraignant le LLM a construire sa réponse sur la base d'un contexte spécifique et approprié.



<u>Son fonctionnement peut se résumer de la sorte</u> :

- Un retriever (1er modèle d'IA) sélectionne dans la base documentaire les paragraphes contenant les réponses à la question
- Ces paragraphes sont transmis au LLM avec la tâche de générer un résumé des paragraphes relevés par le retriever

Pourquoi ne pas donner toute ma base documentaire au LLM ?

⇨ La performance de la réponse est inversement proportionnelle à la taille du contexte.
⇨ Les LLM possède des tailles de contexte limités (le prompt engineering encadrant le contexte consommant déjà une taille non négligeable)
⇨ Le coût (plus le contexte est élevé puis cher est la réponse.

L'ajout du retriever permet donc de réduire les risques d'hallucination en fournissant au LLM un contexte très encadré. La performance du retriever est inversement proportionnelle à la performance du LLM

Retrieval Augmented Generation

Etape 1 : vectorisation base de connaissance

Base documentaire

Moteur
De découpage

Base documentaire
découpée
en ***C Chunks***

Moteur
d'**embedding**
(encoder type
CamenBert)

Base
De chuncks
vectorisée

$$\begin{pmatrix} 2.45 \\ 3.2 \\ 4 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} \vdots \\ \vdots \\ \vdots \\ \vdots \end{pmatrix} \begin{pmatrix} 0.45 \\ 3.2 \\ 3 \\ \vdots \\ 1.8 \end{pmatrix}$$

Contient toute la documentation pertinente du groupe

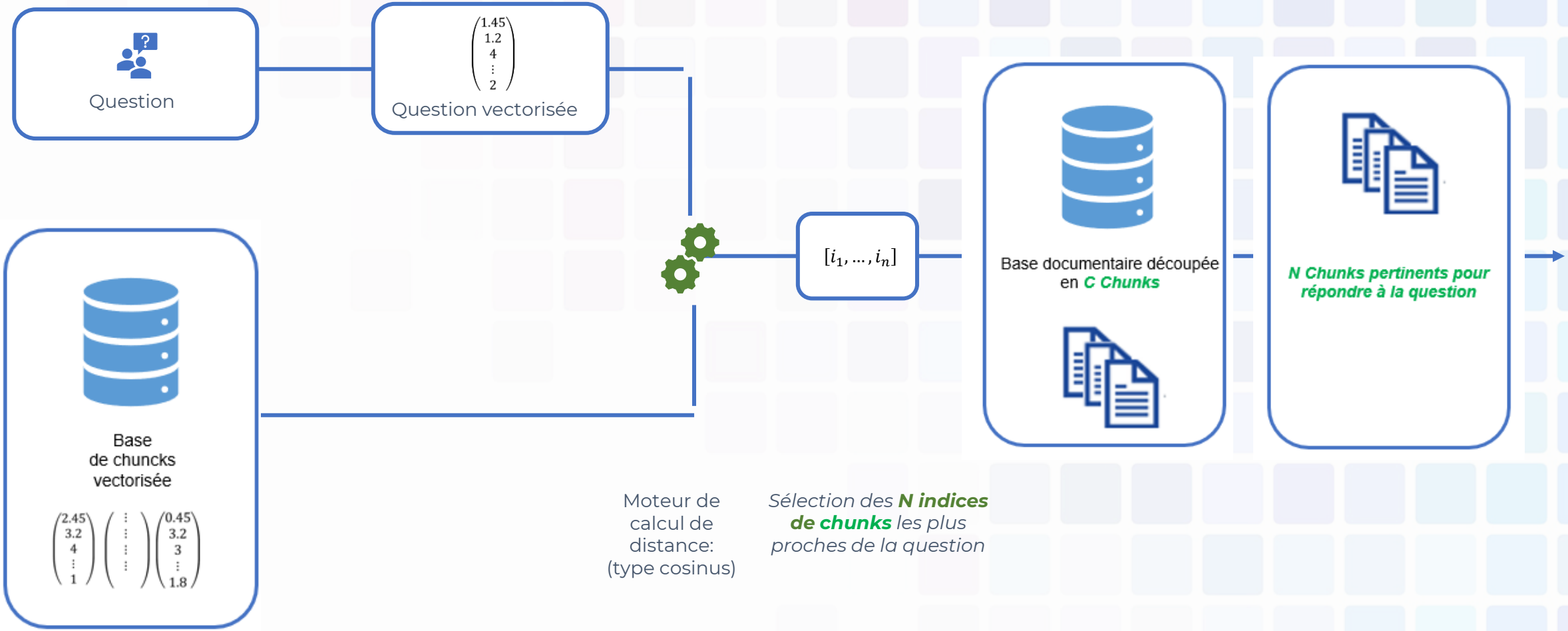
Contient la base de connaissance utilisable comme contexte par le LLM

base de connaissance vectorisée pour calcule des similarités avec une question

La taille des chuncks doit être cohérente avec la taille du contexte LLM

Retrieval Augmented Generation

Etape 2 : Récupération des paragraphes de la base documentaire pertinents



Retrieval Augmented Generation

Etape 3 : Envoie au LLM et validation réponse

Question

N Chunks pertinents pour répondre à la question

LLM

Prompt

```
Utilisez les éléments de contexte suivants
pour répondre à la question finale. Si vous
ne connaissez pas la réponse, dites
simplement "je ne sais pas", n'essayez pas
d'inventer une réponse et répondez
uniquement en français. Si le contexte ne
contient pas assez d'information pour
répondre à la question dites que "je ne sais
pas". Les réponses doivent etre précises et
concises et doivent etre uniquement basé sur
le contexte suivant:
----------------------------------
{context}
----------------------------------
Question: {question}
Réponse Utile:
```

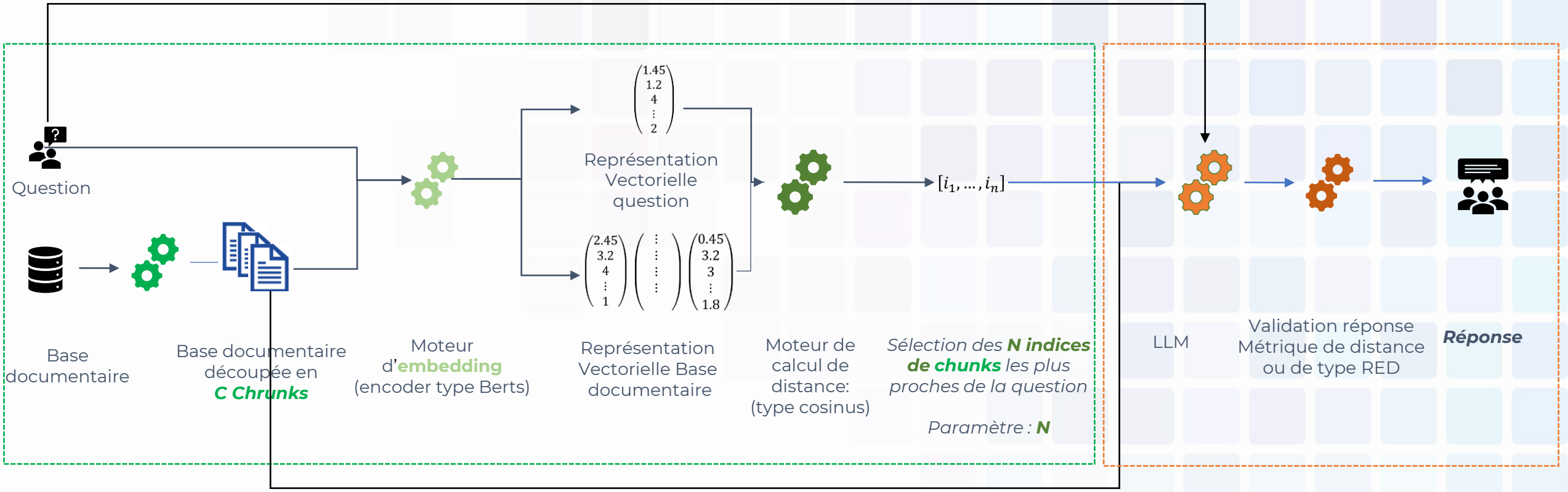
Validation réponse
Métrique de distance
ou de type RED

Réponse

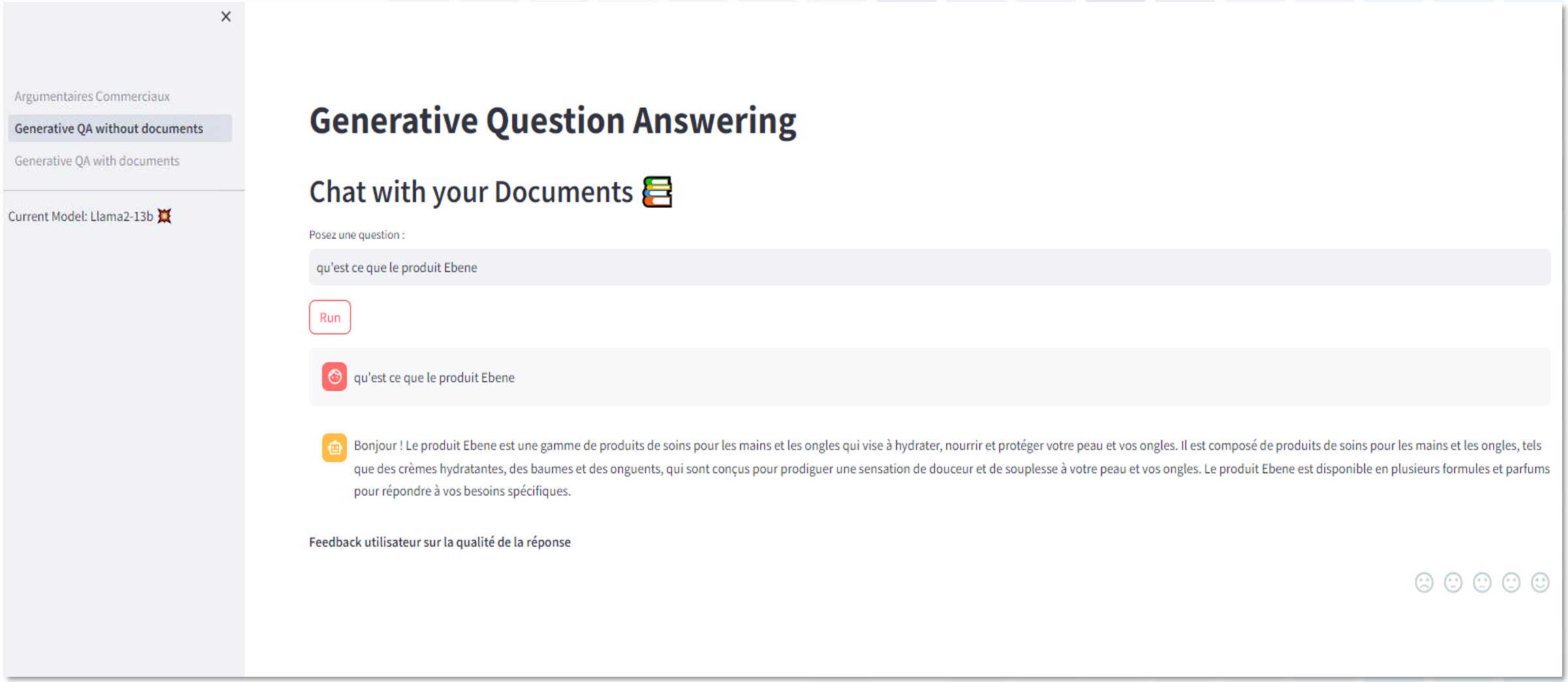
Retrieval Augmented Generation

ACTUAIRES 100% DATA SCIENCE BY Institut A

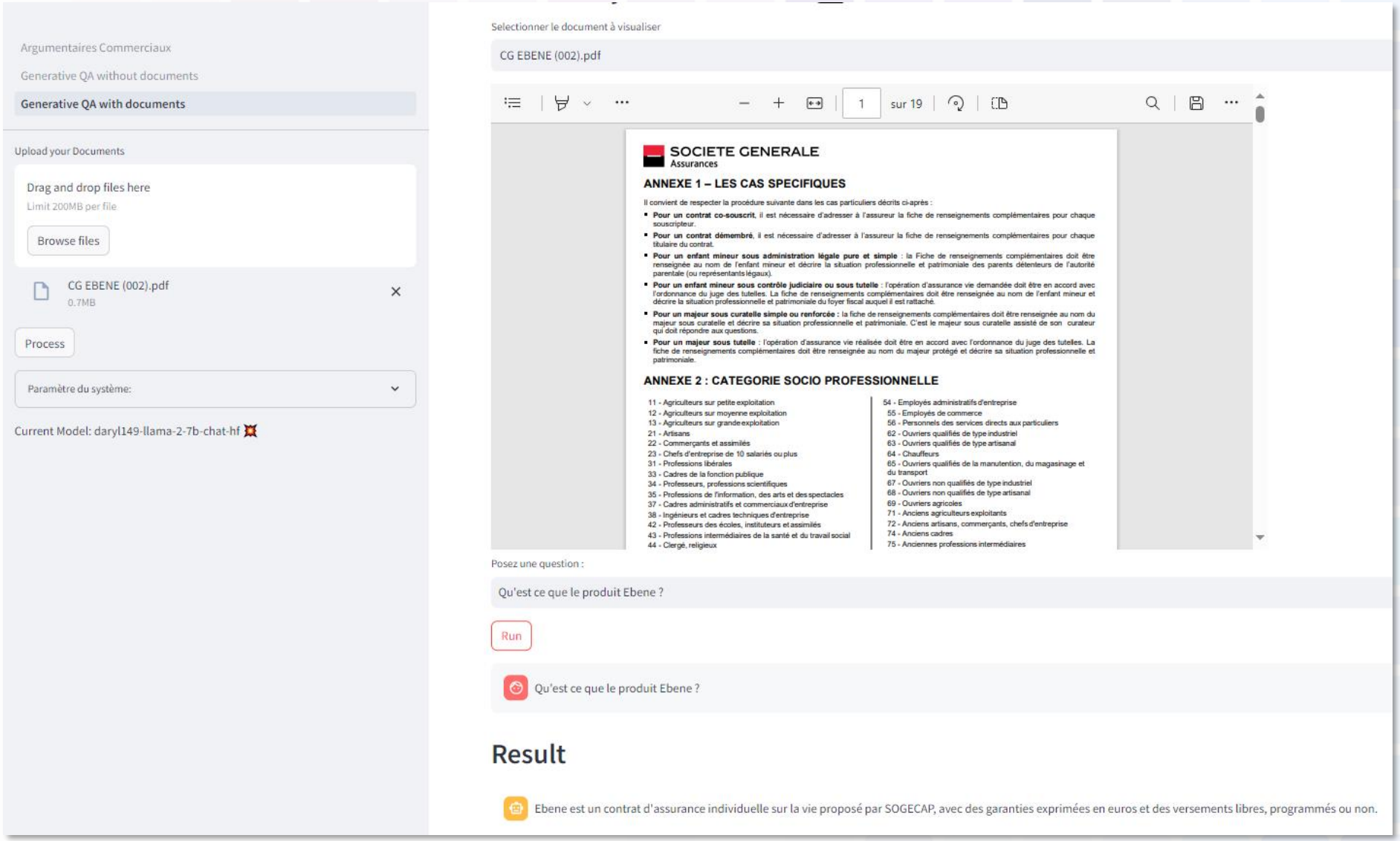
➢ **Vision globale**



RAG : Résultats sans contexte



RAG : Résultats avec contexte



RAG : Résultats - vérification des sources

ACTUAIRES 100% DATA SCIENCE BY Institut A

☹ 🙁 😐 🙂 ☺

Sources utilisées pour la reponse

Sources utilisées:

	document	page	passage	score
0	CG EBENE (002).pdf	1	Ces deux garanties, soumises à des conditions d'âge et de montant, ne se cumulent	1.6181
1	CG EBENE (002).pdf	2	ÉBÈNE est un contrat d'assurance individuelle sur la vie proposé par SOGECAP, entre	1.6322
2	CG EBENE (002).pdf	2	13. LA RENONCIATION.....	1.654

RAG : les paramètres à optimiser

Paramètres	Utilisation / rôle	Commentaire
C	Taille de découpage de la base documentaire	• Un découpage trop fin conduira à un nombre de documents contenant la réponse trop important. • Un découpage trop grossier conduira à un contexte trop vague voir dépassant la taille du contexte maximum • Cible à 500 tokens avec max à 1000
Embedding	Encodage base vectorielle	• N'a pas à être cohérent avec le LLM. • Un modèle Berts sera un bon compromis et pourra être ré-entrainé sur les données de la compagnie.
N	Nombre de meilleur candidat	• En lien avec le paramètre C et la taille maximale de contexte du modèle : • LlaMa2 : 4096 tokens, • GPT4 : de 8192 tokens à 32 768 pour la version turbo.
LLM	Fournira un résumé des N meilleurs réponses	• GPT 4 si ROI et confidentialité le permette. • LaMa2 ou équivalent sinon
Validation réponse	Permet de diminuer le risque d'hallucination du modèle.	• Mélange distance entre la sortie du LLM et la question et distance entre la sortie du LLM et les N réponses.

Comment fine tuner le RAG ?

Dataset Generation

1- Créer des contextes

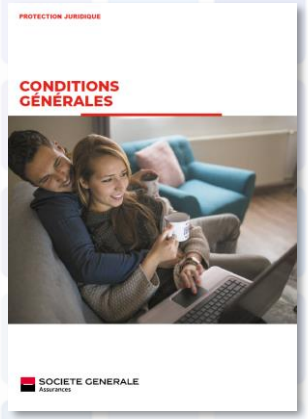
- Accès à 5 pdfs des « Conditions Générales » des produits d'assurance vendus par ASSU.
- Créez des morceaux qui se chevauchent. Chaque morceau (c'est-à-dire le contexte) contient environ 450 caractères avec environ 60 caractères qui se chevauchent.
- 673 contextes créés

2- Utiliser LLM (llama2 et gpt4)

- Créez 3 triplets (question, contexte, réponse) pour chaque morceau.
- 2200 triplés créés.

3- Nettoyer les données

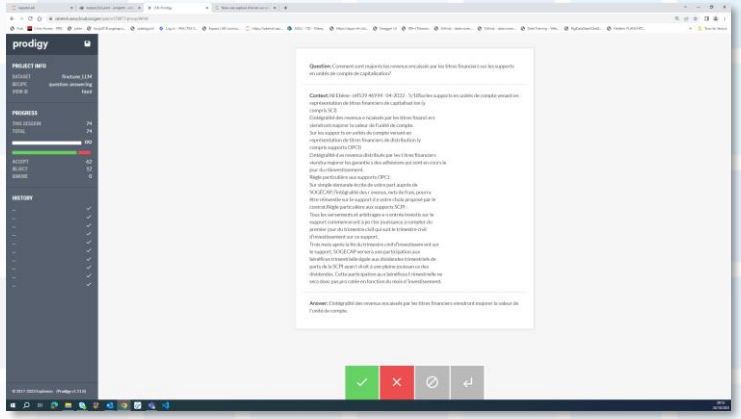
- Vérifiez que tous les triplets (question, contexte, réponse) sont pertinents (c'est-à-dire que la question et la réponse proviennent du contexte).



Utilisation des données

⇨ Fine tuning du retriever (notamment Encoder) via un classifier binaire prédisant si la question est issue du contexte.

⇨ Fine tuning de l'ensemble du RAG (dont LLM) via le couple réponse fournit par le modèle versus réponse construite par dataset génération.



Outil de labellisation prodigy

In fine le LLM permet d'obtenir rapidement des données d'entraînement du modèle

Comment fine tuner le RAG ?

Résultat du fine tuning

1 – Berts

Deux métriques deux performances :

- ✓ **Hit Rate** : pourcentage de réponses ou les k contextes ressortis contiennent le contexte ayant généré la question
- ✓ **Mean Reciprocal Rank** : pondération du Hit Rate par la position du contexte dans la liste des 4 propositions : ❶ = 100%, ….., ❹ = 25%, non retourné : 0

Metric	Modèle Multilangue e 5 Large Non finetuné	Modèle Camembert large non finetuné	Modèle Camembert large finetuné	Amélioration
Hit Rate	**81,4%**	**86,0%**	**90,1%**	**4%**
Mean Reciprocal Rank (MRR)	**64,4%**	**72,9%**	**77,1%**	**5%**

2 – RAG

- ✓ **Rouge** : Mesure la notion de similarité mot à mot
- ✓ **Bert Score :** muse la notion de similarité post *embedding*.

	ROUGE	Bert Score
Zephyr-7b-beta	38,00%	73,30%
Zephyr-7b-beta (with finetuned retriever)	39,70%	73,90%
mistralAI-7b	43,80%	74,90%
mistralAI-7b (with finetuned retriever)	46,10%	75,70%
Llama-2-7b	42,10%	76,70%
Llama-2-7b (with finetuned retriever)	46,00%	77,90%
Llama-2-13b	47,40%	78,50%
Llama-2-13b (with finetuned retriever)	49,90%	79,20%
Vigogne-2-13b	47,40%	78,80%
Vigogne-2-13b (with finetuned retriever)	49,10%	79,20%
Llama-2-70b	46,50%	77,80%
Llama-2-70b (with finetuned retriever)	50,70%	79,10%

Comment fine tuner le retriever ?

Dataset Generation

- Question

"Quelle est la procédure en cas de désaccord entre l'assuré et l'assureur sur l'opportunité d'engager ou de poursuivre une action?"

```
############ Context 1/4

L'assureur apporte sa garantie.
Cas 2.2.2 : l'assuré a souscrit une nouvelle garantie de
responsabilité déclenchée par la réclamation auprès d'un
nouvel assureur couvrant le même risque.
C'est la nouvelle garantie qui est mise en œuvre, sauf si l'a .....

############ Context 2/4

Si vous n'étiez  pas couvert sur la base du fait dommageable à
la date du fait dommageable, l'assureur qui doit être désigné
est celui qui est compétent, dans les conditions précisées
aux paragraphes II-1, II-2 et II-3 ci-dessus, au moment de la
.....

############ Context 3/4

• En cas d'intervention amiable, nous défendons vos intérêts
pour rechercher  dans un premier temps et dans la mesure du
possible une solution amiable à votre litige.Au cours des discussions amiables,
charge aucun honoraire d'av .....

############ Context 4/4

■De plein droit :
En cas de retrait de l'agrément administratif de l'assureur (article L
326-12 du Code des assurances).Les dispositions légales
■Délai de prescription
Toute action concernant votre contrat et émanant de vous ou de nous,
ne peut êt .....
```

Modèle multi-langue e5 :
Aucun contexte pertinent renvoyé

```
############ Context 1/4

L'assureur apporte sa garantie.
Cas 2.2.2 : l'assuré a souscrit une nouvelle garantie de
responsabilité déclenchée par la réclamation auprès d'un
nouvel assureur couvrant le même risque.
C'est la nouvelle garantie qui est mise en œuvre, sauf si l'a .....

############ Context 2/4

Si vous n'étiez  pas couvert sur la base du fait dommageable à
la date du fait dommageable, l'assureur qui doit être désigné
est celui qui est compétent, dans les conditions précisées
aux paragraphes II-1, II-2 et II-3 ci-dessus, au moment de la
.....

############ Context 3/4

• En cas d'intervention amiable, nous défendons vos intérêts
pour rechercher  dans un premier temps et dans la mesure du
possible une solution amiable à votre litige.Au cours des discussions amiables,
charge aucun honoraire d'av .....

############ Context 4/4

■De plein droit :
En cas de retrait de l'agrément administratif de l'assureur (article L
326-12 du Code des assurances).Les dispositions légales
■Délai de prescription
Toute action concernant votre contrat et émanant de vous ou de nous,
ne peut êt .....
```

Modèle camembert nonfinuté :
Aucun contexte pertinent renvoyé

```
############ Context 1/4


Conformément aux dispositions de l'article L 127-4 du Code des assurances, en cas de désaccord entre vous et nous sur l'opportunité d'engager ou de poursuivre une actio
Conditions Générales - La Défense Pénale et Recours .....

############ Context 2/4

14    Conditions Générales A ssurance Automobile - Réf. L 190 305 - 05/2023
LA DÉFENSE PÉNALE ET RECOURS
SUITE À ACCIDENT
Qui est assuré ?
• Le souscripteur, ou en cas de décès de celui-ci, son conjoint ou ses
descendant
s à charge.
• Le proprié
t .....

############ Context 3/4

Si vous n'étiez  pas couvert sur la base du fait dommageable à
la date du fait dommageable, l'assureur qui doit être désigné
est celui qui est compétent, dans les conditions précisées
aux paragraphes II-1, II-2 et II-3 ci-dessus, au moment de la
.....

############ Context 4/4

Conditions Générales Camping-car / Caravane / Remorque - Réf. C 190 307 - 11/2022 Page 10 sur 35
La Défense Pénale
et Recours Suite à Accident
Dans le texte qui suit, "nous", "nos", "notre" font référence à
ABEILLE IARD & SANTÉ.
Qui est assuré ?
•
```

Modèle camembert finetuné :
Le 1er contexte renvoyé est pertinent

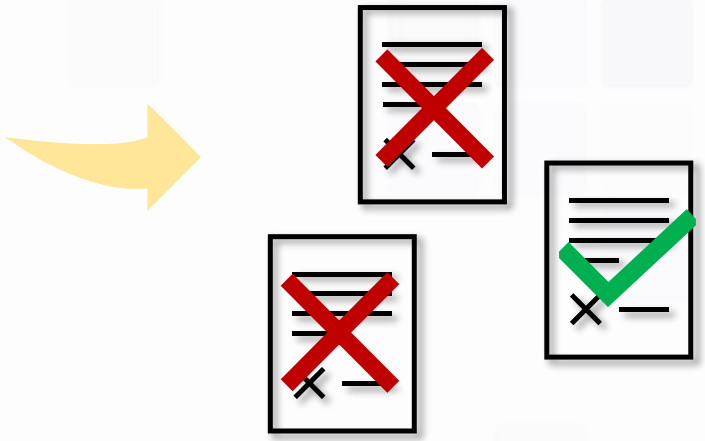
Use Case #2

Création d'un assistant pour orienter le client

1 Le client/prospect échange avec le chatbot sur ses besoins et interrogations

2 Le modèle détermine le ou les produits les plus adaptés pour le client

3 Reprise du process classique pour formaliser l'offre et assurer le devoir de conseil.



Process de fine-tuning

Etape 1 : constituer le dataset d'entraînement

1. **Bien choisir son jeu de données :**
   - ✓ jeu de données existant :
     - historiques de conversation
     - base de données, logs
     - jeu de données public
   - ✓ création « manuelle » d'une base de questions/réponses orientée sur ses besoins. Ce dataset doit être d'une **taille suffisante** (viser au moins un millier de Q/R)

2. **Nettoyage des données** (anonymisation, termes incorrects…)

3. **Mise en forme** des datasets pour une utilisation optimale lors de la phase d'entraînement.

Process de fine-tuning

Etape 2 : sélectionner le modèle de fondation à utiliser

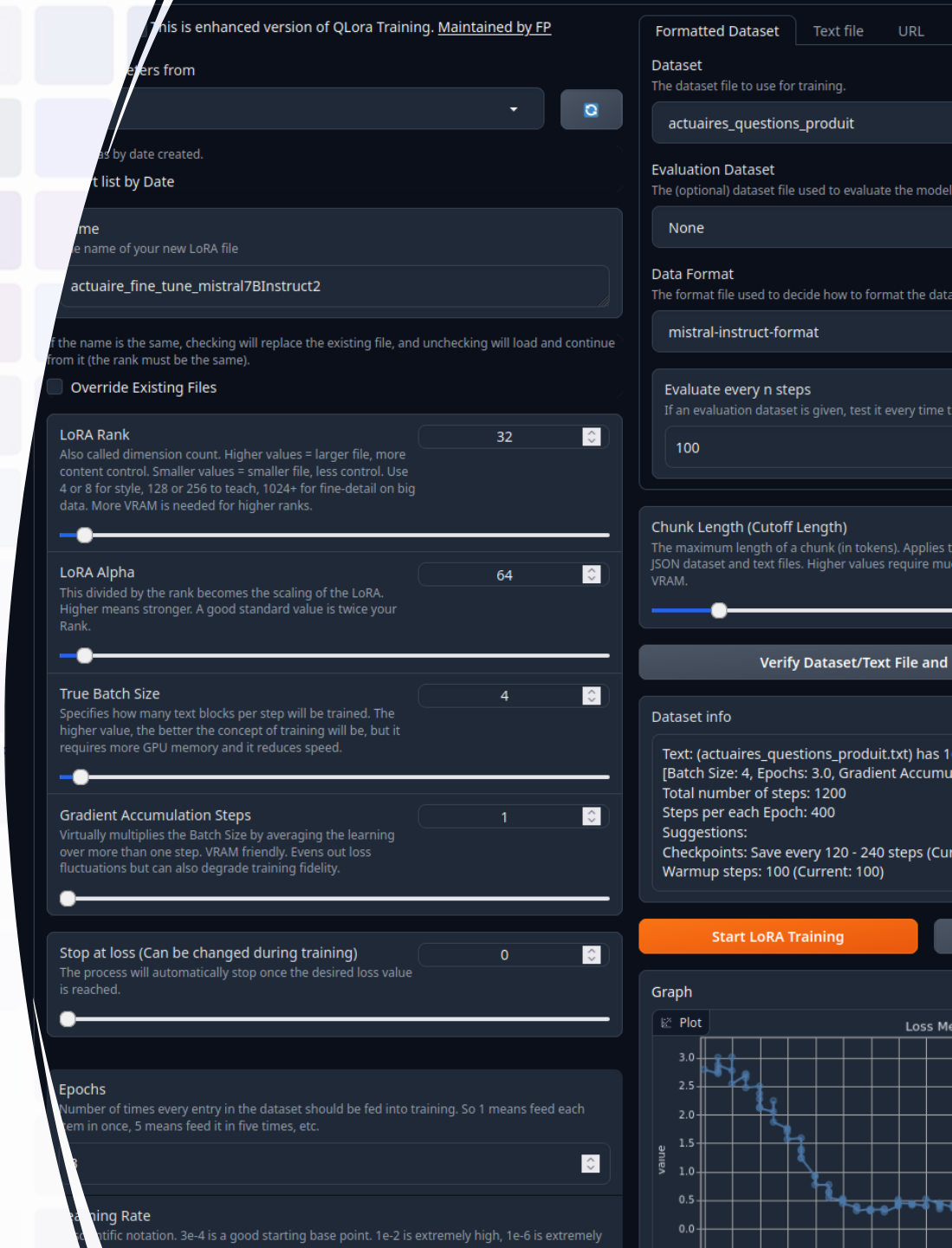
Il s'agit de déterminer sur quel modèle LLM existant on va procéder à un fine-tuning. Les principaux éléments à prendre en compte sont les suivants :

- ✓ La licence d'exploitation (MIT, Apache ou autre)
- ✓ La taille du modèle (nombre de paramètres: 7B, 34B, 50B…)
- ✓ Contenu modéré ou pas (ex : GPT 3.5 vs Mistral)
- ✓ Multi-langue ou pas
- ✓ Dataset utilisé pour l'entrainement
- ✓ Eventuels benchmarks existants (ex : HuggingFace)

Process de fine-tuning

Etape 3 : paramétrer le modèle et l'entraîner

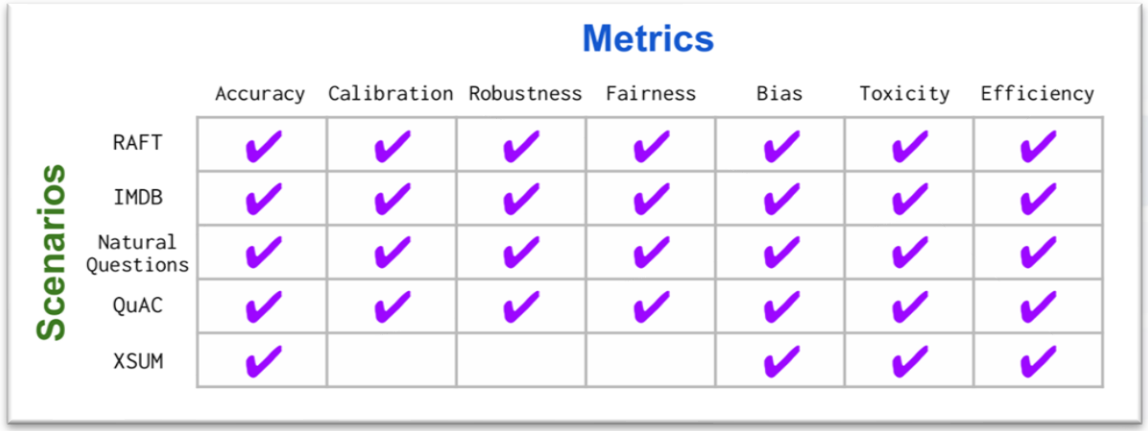
- Le fine-tuning demande une grosse puissance de **calcul GPU**. Il faut souvent prévoir plusieurs heures, voire plusieurs jours (ici 8h pour 1600 Q/R sur une instance H100).
- L'approche généralement adoptée est de privilégier **des « petits » modèles** (7B) et de procéder à des validations intermédiaires.
- Optimisation des modèles avec des approches LoRA et Quantization.
- Benchmark des modèles obtenus avec un jeu de questions de tests.



Process de fine-tuning

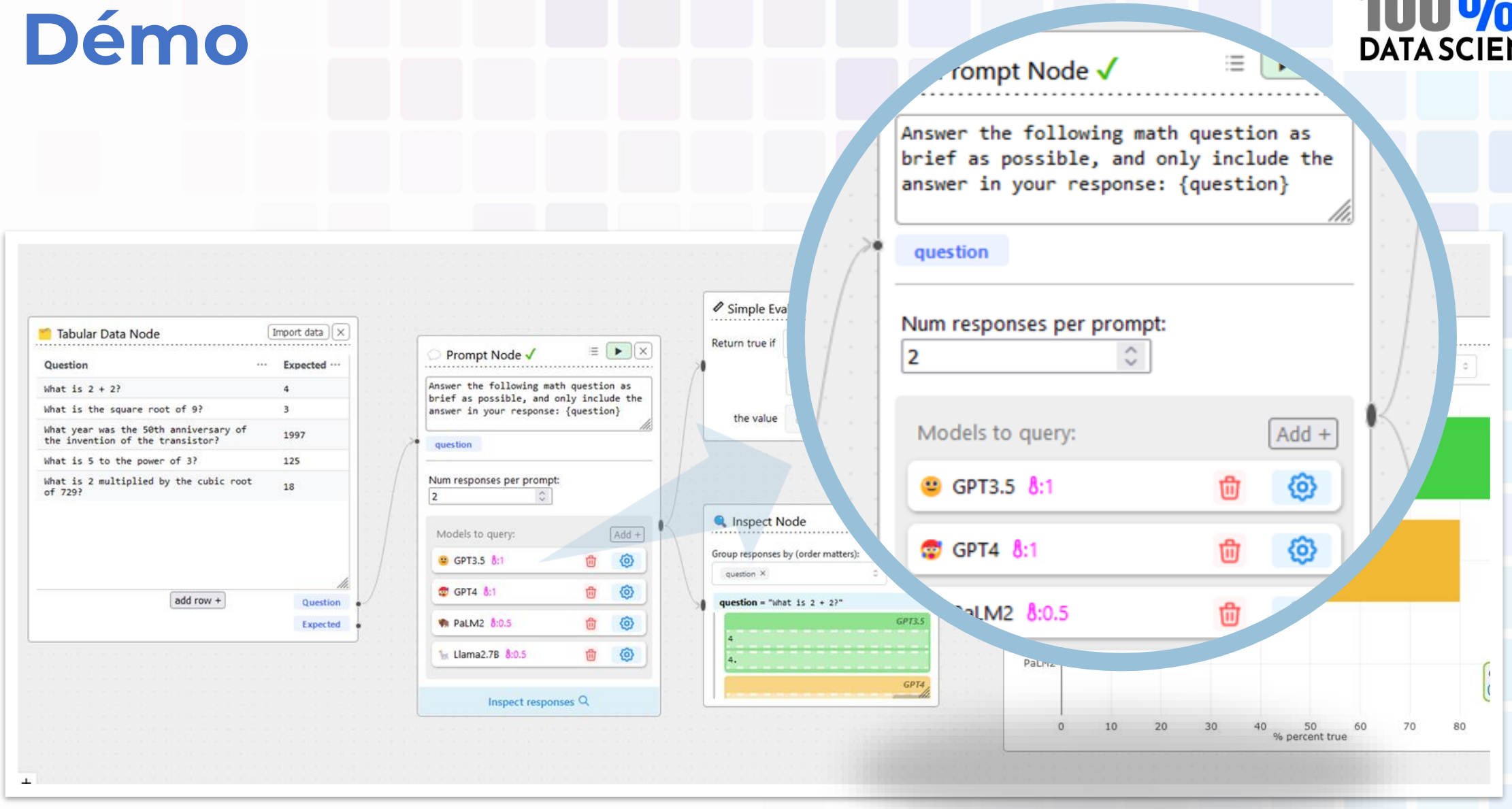
Etape 4 : évaluer le modèle fine-tuné

- L'évaluation nécessite la création de jeux de données de tests qui doivent être représentatifs des contextes attendus. Cette phase est tout aussi importante que la création des données d'entraînement !
- L'évaluation se fait en suivant la **méthode HELM** (Holistic Evaluation of Language Model) développée par l'université de Stanford. Ce framework comprend notamment une batterie de 59 métriques différentes (pertinence, robustesse, biais...).



Démo

ACTUAIRES 100% DATA SCIENCE BY Institut A



Points-clés

- Il est possible d'entraîner des assistants métiers spécialisés en se basant sur des **LLMs de fondation** Open Source.
- La **spécialisation** de ces modèles peut reposer sur :
  - ✓ des bases documentaires (approche **RAG**)
  - ✓ sur des jeux de données de questions/réponses (approche **fine-tuning**)
- Quelle que soit l'approche retenue, la qualité des données est primordiale.
- Il est nécessaire d'évaluer la qualité de ces modèles de manière formelle.