

# 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 Steenberghe**

ThinkDeep AI  
CEO & founder

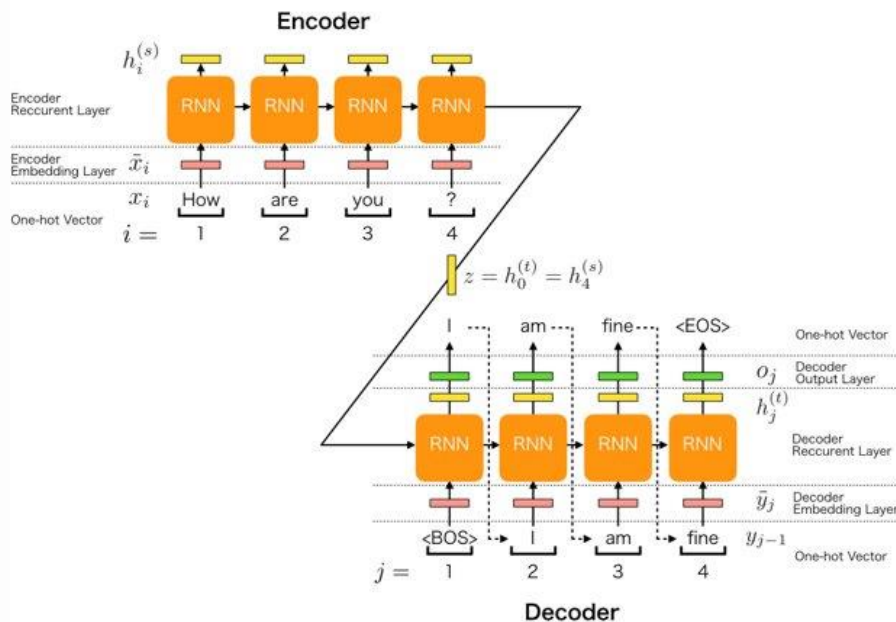
# Transformers vs approche classique

**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 entraîner.

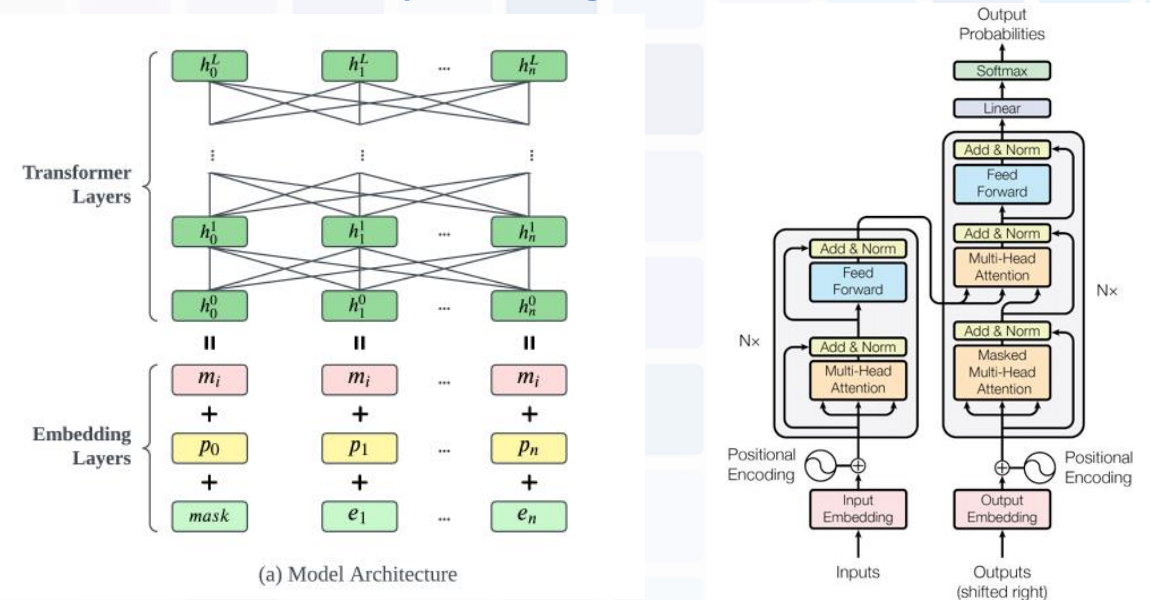


**2017**  
*Transformers*

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" reçoit alors un embedding différent.

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



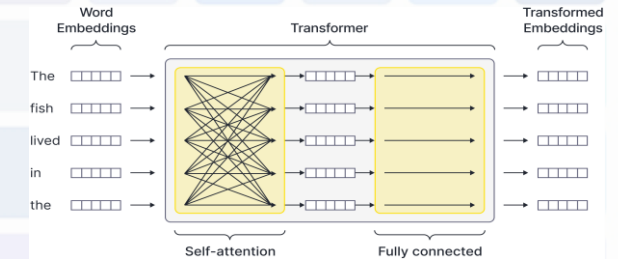
# Transformers : Encodeurs et Décodeurs

## 2018 : BERT (Bidirectional Encoder Representations from 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.



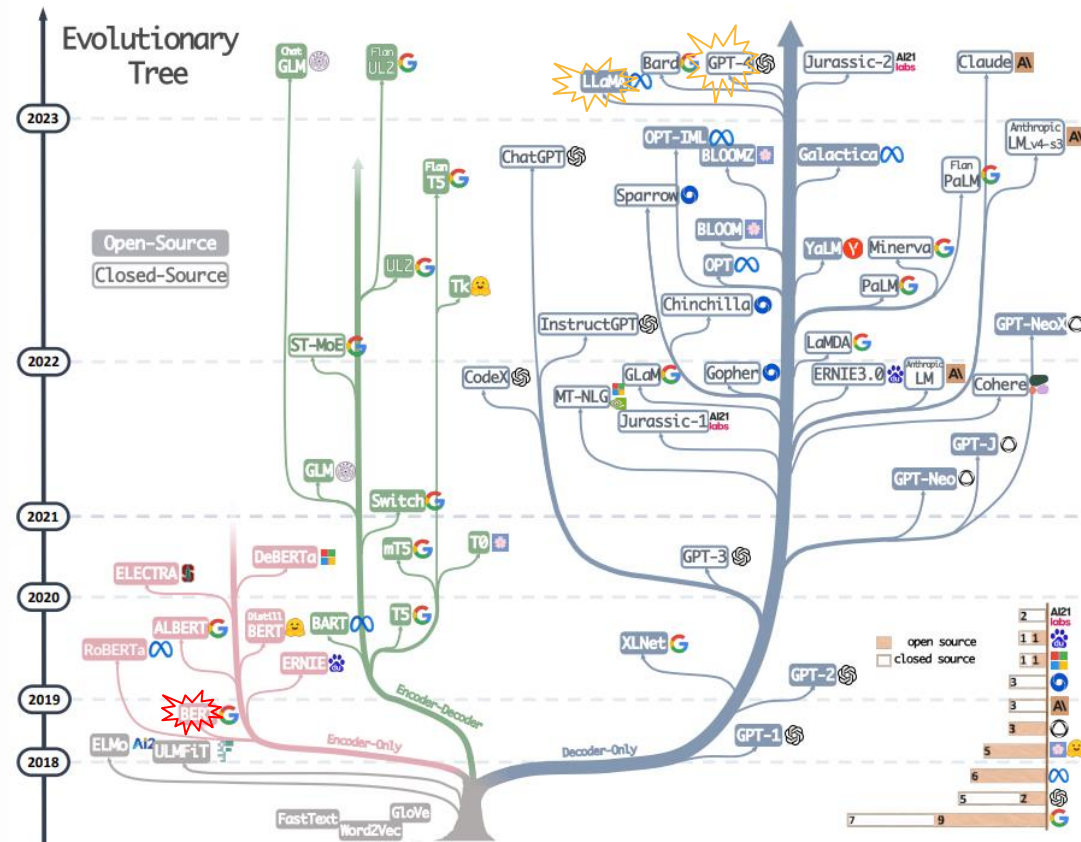
➤ Tourne sur de la CPU et peut être fine tuné avec les données de l'entreprise.

## 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)

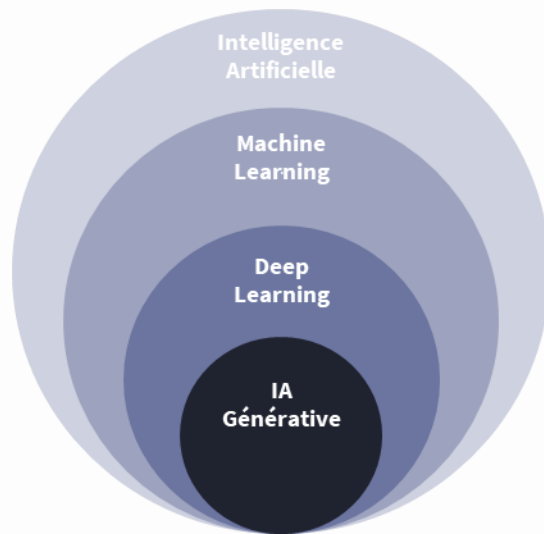
**LLaMa2** : plusieurs versions de modèles sont disponibles (7, 13 et 70 milliards de paramètres), avec une taille de contexte maximale de 4086 Tokens




# IA GEN en deux mots

## 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 possibilité de générer du contenu (texte, images ou vidéos personnalisées...).
- Une accélération du déploiement de cas d'usages qui nécessitaient 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,...

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

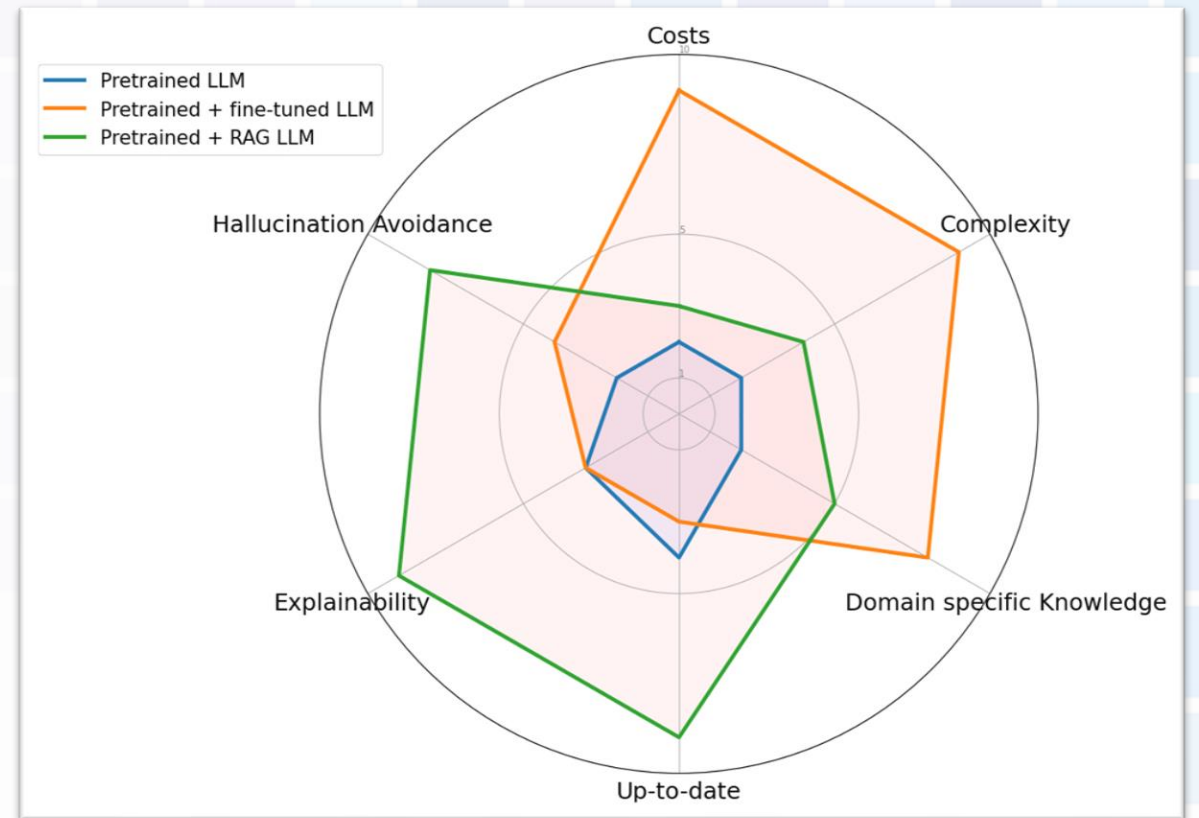
 **MISTRAL**  
**AI\_**

  
**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).



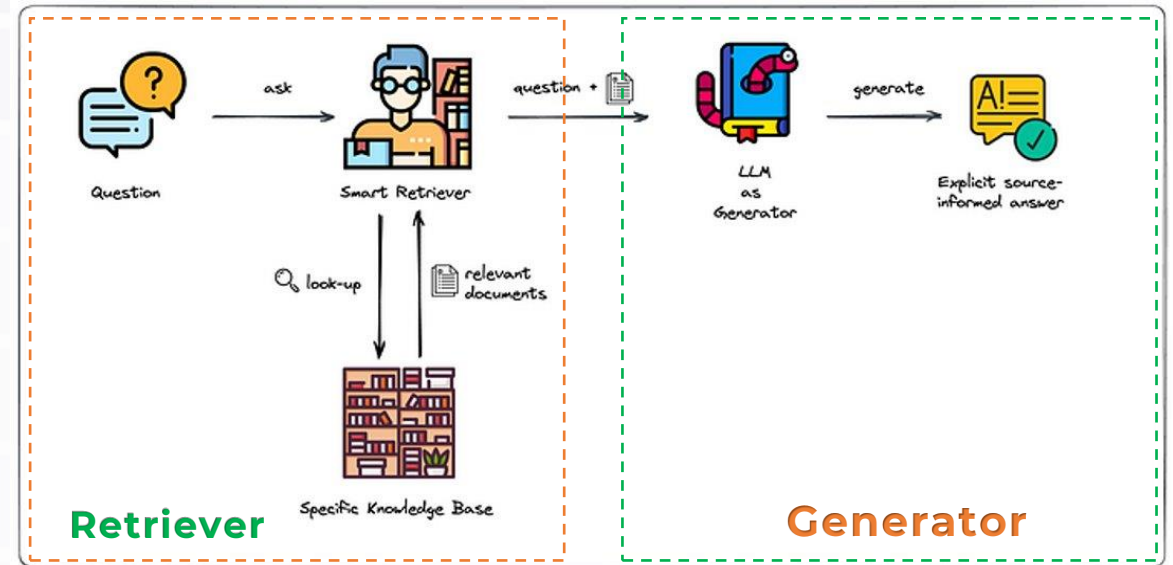
# Retrieval Augmented Generation (RAG)

Les LLM sont entraînés pour prédire le prochain mot d'une phrase. Si l'emploi 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 à construire sa réponse sur la base d'un contexte spécifique et approprié.

**Son fonctionnement peut se résumer de la sorte :**

- Un retriever (1<sup>er</sup> 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



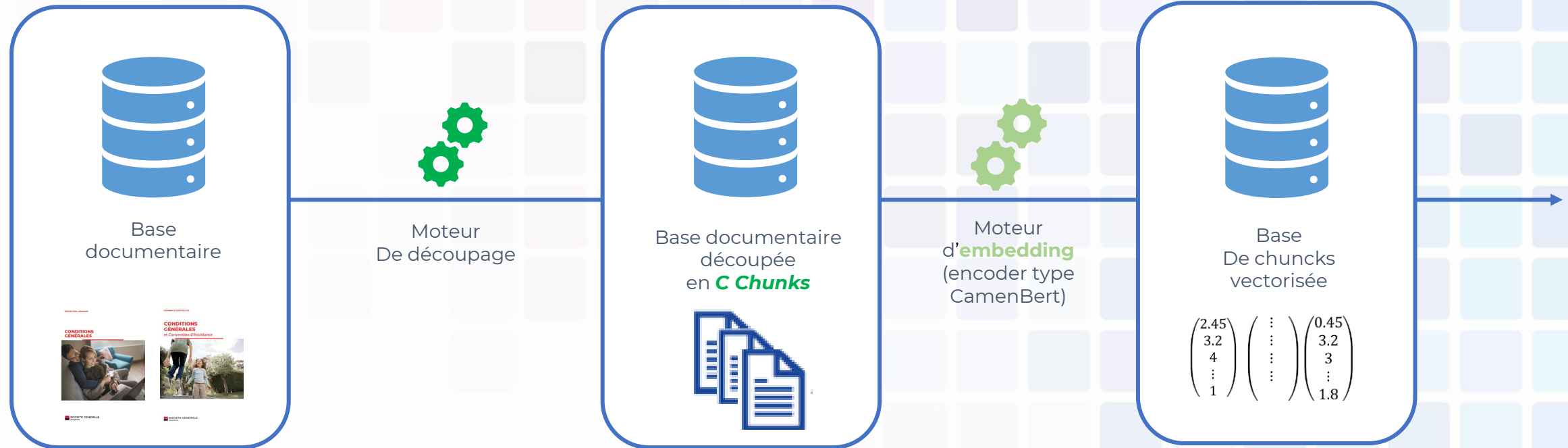
**Pourquoi ne pas donner toute ma base documentaire au LLM ?**

- ⇒ 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



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

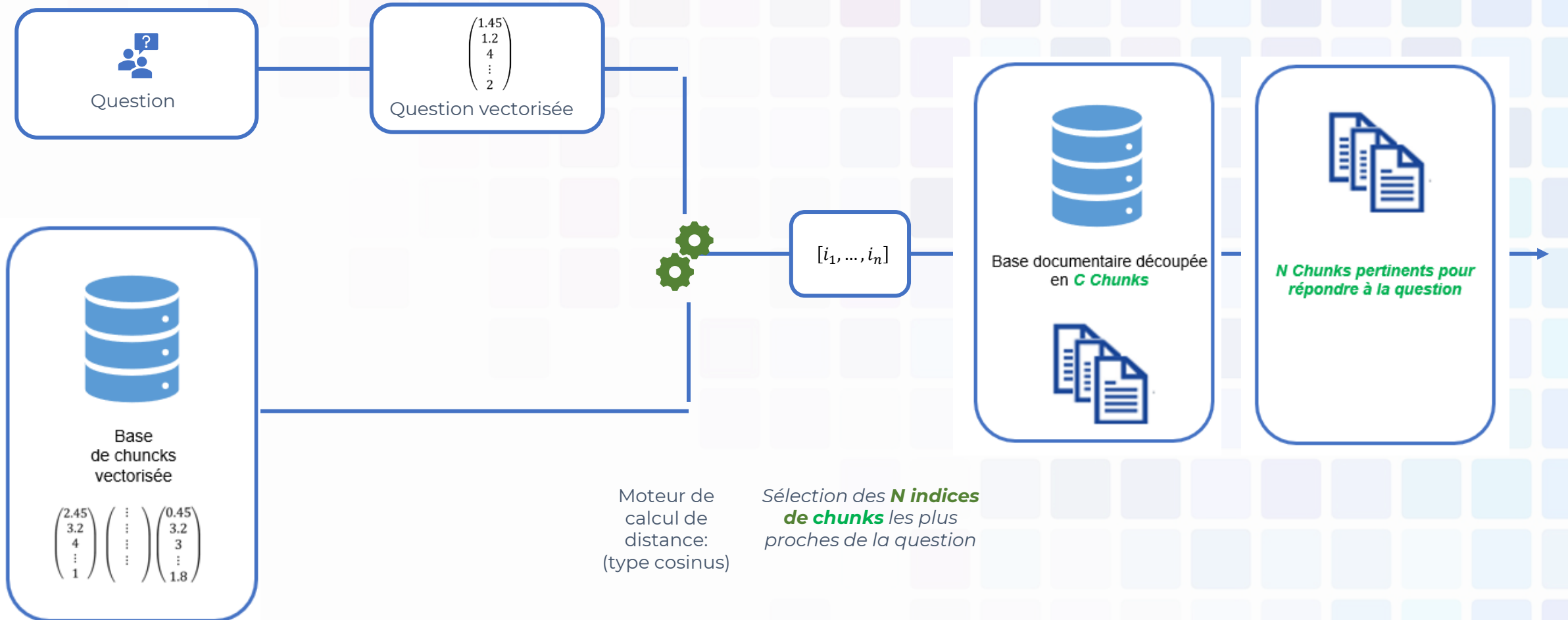


La taille des chunks doit être cohérente avec la taille du contexte LLM



# Retrieval Augmented Generation

## Etape 2 : Récupération des paragraphes de la base documentaire pertinents



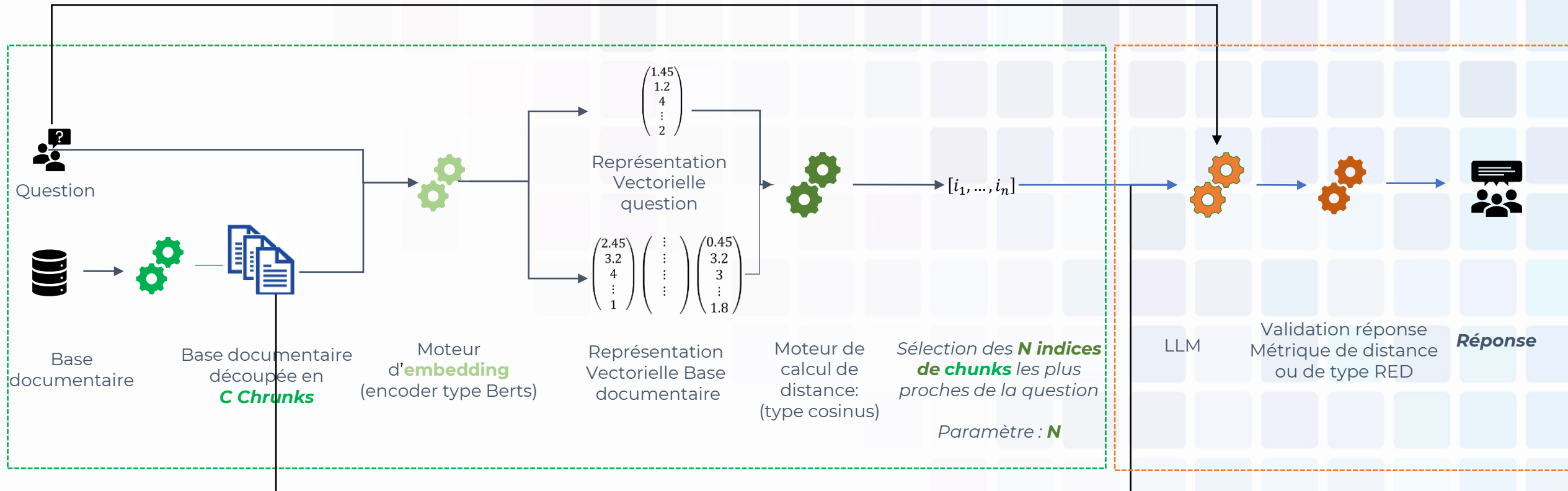
# Retrieval Augmented Generation

## Etape 3 : Envoie au LLM et validation réponse



# Retrieval Augmented Generation

## ➤ Vision globale



# RAG : Résultats sans contexte

Argumentaires Commerciaux

Generative QA without documents

Generative QA with documents

---

Current Model: Llama2-13b 🦙

## Generative Question Answering

### Chat with your Documents 📖

Posez une question :

qu'est ce que le produit Ebene

Run

🗣️ qu'est ce que le produit Ebene

🤖 Bonjour ! Le produit Ebene est une gamme de produits de soins pour les mains et les ongles qui vise à hydrater, nourrir et protéger votre peau et vos ongles. Il est composé de produits de soins pour les mains et les ongles, tels que des crèmes hydratantes, des baumes et des onguents, qui sont conçus pour prodiguer une sensation de douceur et de souplesse à votre peau et vos ongles. Le produit Ebene est disponible en plusieurs formules et parfums pour répondre à vos besoins spécifiques.

Feedback utilisateur sur la qualité de la réponse

☹️ 😐 😊 😄 😁

# RAG : Résultats avec contexte

Argumentaires Commerciaux

Generative QA without documents

**Generative QA with documents**

Upload your Documents

Drag and drop files here  
Limit 200MB per file

Browse files

CG EBENE (002).pdf  
0.7MB

Process

Paramètre du système:

Current Model: daryl149-llama-2-7b-chat-hf

Selectionner le document à visualiser

CG EBENE (002).pdf

1 sur 19

**SOCIETE GENERALE**  
Assurances

**ANNEXE 1 – LES CAS SPECIFIQUES**

Il convient de respecter la procédure suivante dans les cas particuliers décrits ci-après :

- **Pour un contrat co-souscrit**, il est nécessaire d'adresser à l'assureur la fiche de renseignements complémentaires pour chaque souscripteur.
- **Pour un contrat démembre**, il est nécessaire d'adresser à l'assureur la fiche de renseignements complémentaires pour chaque titulaire du contrat.
- **Pour un enfant mineur sous administration légale pure et simple** : la Fiche de renseignements complémentaires doit être renseignée au nom de l'enfant mineur et décrire la situation professionnelle et patrimoniale des parents détenteurs de l'autorité parentale (ou représentants légaux).
- **Pour un enfant mineur sous contrôle judiciaire ou sous tutelle** : l'opération d'assurance vie demandée doit être en accord avec l'ordonnance du juge des tutelles. La fiche de renseignements complémentaires doit être renseignée au nom de l'enfant mineur et décrire la situation professionnelle et patrimoniale du foyer fiscal auquel il est rattaché.
- **Pour un majeur sous curatelle simple ou renforcée** : la fiche de renseignements complémentaires doit être renseignée au nom du majeur sous curatelle et décrire sa situation professionnelle et patrimoniale. C'est le majeur sous curatelle assisté de son curateur qui doit répondre aux questions.
- **Pour un majeur sous tutelle** : l'opération d'assurance vie réalisée doit être en accord avec l'ordonnance du juge des tutelles. La fiche de renseignements complémentaires doit être renseignée au nom du majeur protégé et décrire sa situation professionnelle et patrimoniale.

**ANNEXE 2 : CATEGORIE SOCIO PROFESSIONNELLE**

11 - Agriculteurs sur petite exploitation	54 - Employés administratifs d'entreprise
12 - Agriculteurs sur moyenne exploitation	55 - Employés de commerce
13 - Agriculteurs sur grande exploitation	56 - Personnels des services directs aux particuliers
21 - Artisans	62 - Ouvriers qualifiés de type industriel
22 - Commerçants et assimilés	63 - Ouvriers qualifiés de type artisanal
23 - Chefs d'entreprise de 10 salariés ou plus	64 - Chauffeurs
31 - Professions libérales	65 - Ouvriers qualifiés de la manutention, du magasinage et du transport
33 - Cadres de la fonction publique	67 - Ouvriers non qualifiés de type industriel
34 - Professeurs, professions scientifiques	68 - Ouvriers non qualifiés de type artisanal
35 - Professions de l'information, des arts et des spectacles	69 - Ouvriers agricoles
37 - Cadres administratifs et commerciaux d'entreprise	71 - Anciens agriculteurs exploitants
38 - Ingénieurs et cadres techniques d'entreprise	72 - Anciens artisans, commerçants, chefs d'entreprise
42 - Professeurs des écoles, instituteurs et assimilés	74 - Anciens cadres
43 - Professions intermédiaires de la santé et du travail social	75 - Anciennes professions intermédiaires
44 - Clergé, religieux	

Posez une question :

Qu'est ce que le produit Ebene ?

Run

Qu'est ce que le produit Ebene ?

**Result**

Ebene est un contrat d'assurance individuelle sur la vie proposé par SOGECAP, avec des garanties exprimées en euros et des versements libres, programmés ou non.

# RAG : Résultats - vérification des sources



## 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
<b>C</b>	Taille de découpage de la base documentaire	<ul style="list-style-type: none"> <li>• Un découpage trop fin conduira à un nombre de documents contenant la réponse trop important.</li> <li>• Un découpage trop grossier conduira à un contexte trop vague voir dépassant la taille du contexte maximum</li> <li>• Cible à 500 tokens avec max à 1000</li> </ul>
<b>Embedding</b>	Encodage base vectorielle	<ul style="list-style-type: none"> <li>• N'a pas à être cohérent avec le LLM.</li> <li>• Un modèle Berts sera un bon compromis et pourra être ré-entraîné sur les données de la compagnie.</li> </ul>
<b>N</b>	Nombre de meilleur candidat	<ul style="list-style-type: none"> <li>• En lien avec le paramètre C et la taille maximale de contexte du modèle :</li> <li>• LLaMa2 : 4096 tokens,</li> <li>• GPT4 : de 8192 tokens à 32 768 pour la version turbo.</li> </ul>
<b>LLM</b>	Fournira un résumé des N meilleurs réponses	<ul style="list-style-type: none"> <li>• GPT 4 si ROI et confidentialité le permette.</li> <li>• LaMa2 ou équivalent sinon</li> </ul>
<b>Validation réponse</b>	Permet de diminuer le risque d'hallucination du modèle.	<ul style="list-style-type: none"> <li>• Mélange distance entre la sortie du LLM et la question et distance entre la sortie du LLM et les N réponses.</li> </ul>

# 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)

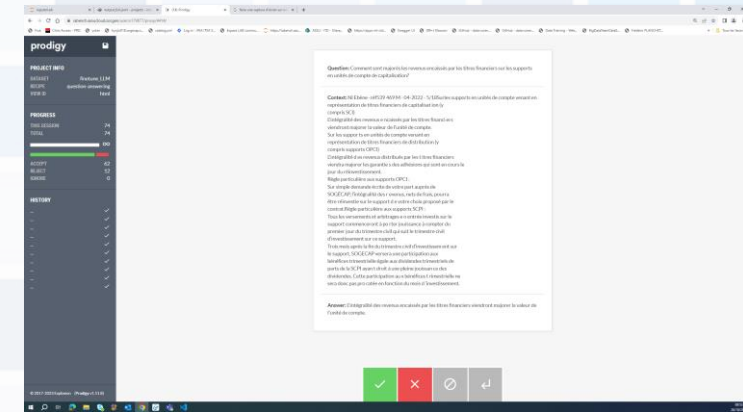
- Créez 3 triplets (question, contexte, réponse) pour chaque morceau.
- 2200 triplés créés.

### 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 classifieur 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 : ① = 100%, ....., ④ = 25%, non retourné : 0

Metric	Modèle Multilingue 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** : mesure 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 être .....

##### 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 être .....

##### 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 action  
Conditions Générales - La Défense Pénale et Recours .....

##### Context 2/4

14 Conditions Générales Assurance 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 propriétaire  
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 ?  
• '1

Modèle multi-langue e5 :  
Aucun contexte pertinent renvoyé

Modèle camembert nonfinuté :  
Aucun contexte pertinent renvoyé

Modèle camembert finetuné :  
Le 1<sup>er</sup> contexte renvoyé est pertinent

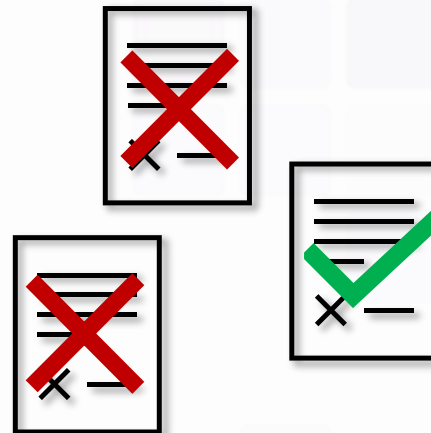
## Use Case #2

### 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'entraînement
- ✓ 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.

This is an enhanced version of QLoRA Training. Maintained by FP

Dataset: actuaire\_questions\_produit

Evaluation Dataset: None

Data Format: mistral-instruct-format

Evaluate every n steps: 100

Chunk Length (Cutoff Length): [Slider]

Verify Dataset/Text File and

Dataset info: Text: (actuaire\_questions\_produit.txt) has 1 [Batch Size: 4, Epochs: 3.0, Gradient Accumulation Steps: 1] Total number of steps: 1200 Steps per each Epoch: 400 Suggestions: Checkpoints: Save every 120 - 240 steps (Current: 100) Warmup steps: 100 (Current: 100)

Start LoRA Training

Graph: Plot of Loss vs. Epochs

Epoch	Loss
0	3.0
100	2.5
200	2.0
300	1.5
400	1.0
500	0.5
600	0.4
700	0.35
800	0.3
900	0.25
1000	0.2
1100	0.15
1200	0.1

LoRA Rank: 32

LoRA Alpha: 64

True Batch Size: 4

Gradient Accumulation Steps: 1

Stop at loss (Can be changed during training): 0

Epochs: [Slider]

Learning Rate: [Slider]

# 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...).

		Metrics						
		Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
Scenarios	RAFT	✓	✓	✓	✓	✓	✓	✓
	IMDB	✓	✓	✓	✓	✓	✓	✓
	Natural Questions	✓	✓	✓	✓	✓	✓	✓
	QuAC	✓	✓	✓	✓	✓	✓	✓
	XSUM	✓				✓	✓	✓

# Démo

The screenshot displays a workflow automation interface with several interconnected nodes:

- Tabular Data Node:** Contains a table with 5 rows of questions and expected answers.
- Prompt Node:** Contains a prompt template: "Answer the following math question as brief as possible, and only include the answer in your response: {question}" and a dropdown for "question".
- Simple Evaluation Node:** Contains a dropdown for "Num responses per prompt:" set to 2.
- Inspect Node:** Shows a list of models to query: GPT3.5 (8:1), GPT4 (8:1), PaLM2 (8:0.5), and Llama2.7B (8:0.5). It also shows a preview of responses for the question "what is 2 + 2?".
- Simple Evaluation Node (Right):** Shows a bar chart titled "Num responses per prompt:" with a value of 2. Below it, a bar chart titled "Models to query:" shows the percentage of correct responses for each model.

Question	Expected
What is 2 + 2?	4
What is the square root of 9?	3
What year was the 50th anniversary of the invention of the transistor?	1997
What is 5 to the power of 3?	125
What is 2 multiplied by the cubic root of 729?	18

Models to query:

- GPT3.5 8:1
- GPT4 8:1
- PaLM2 8:0.5
- Llama2.7B 8:0.5

question = "what is 2 + 2?"

Model	% percent true
GPT3.5	~65
GPT4	~75
PaLM2	~60
Llama2.7B	~55



# 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.