

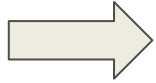


# ChatBot - Application du traitement des langues naturelles

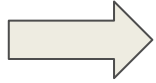
Tuteur:  
Alexandre Saidi

Elèves:  
KAROUMA Youssef  
KHATIB Mohamed  
TIBI Daniel

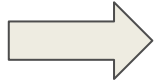
# Sommaire:



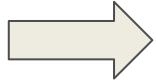
Contexte



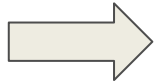
Problématique



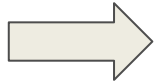
Objectifs



Livrables



Présentation des différentes phases de projet



Perspectives

# Contexte

- L'apparition du domaine de NLP remonte aux années 50.(l'article:"Computing Machinery and Intelligence" Alan Turing).
- Emergence des méthodes probabilistes et de ML depuis les années 2000.
- Depuis 2010, les modèles de Deep learning comme les transformers(BERT et GPT) gagnent une popularité en NLP.



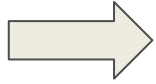
VS



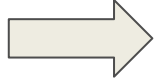
# Problématique:

Comment peut-on concevoir, en utilisant l'IA, un ChatBot capable de répondre aux questions relatives à la scolarité à Centrale?

# Objectifs du projet

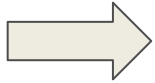


**Objectif académique** : Familiarisation avec les techniques de traitement de langues

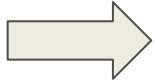


**Réalisation d'un ChatBot** : ChatBot sous forme de questions/réponses sur le règlement de la scolarité et du Master Info

# Livrables



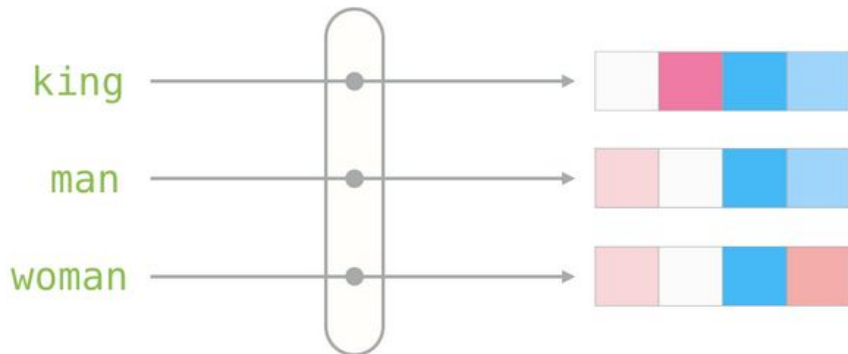
**Prototype du ChatBot** : ChatBot implémenté sous une application Web



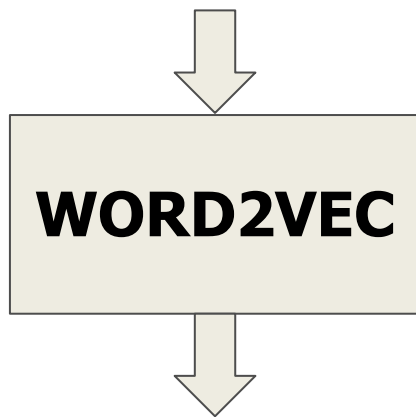
**Rapport du projet** : Inclusion de l'état de l'art et des résultats trouvés.

# L'embedding : une étape essentielle

- L'embedding est une technique de représentation de mots ou de phrases sous forme de vecteurs numériques dans un espace mathématique.
- Il permet de capturer la signification sémantique des mots et leur contexte dans un corpus de texte.



- Plusieurs types d'embeddings qui existent: **Word2vec**, **Glove**, **FastText**, **BERT** ...
- Pour des raisons de simplicité, de rapidité d'inférence et de ressources de calcul limitées, nous avons choisi le modèle:

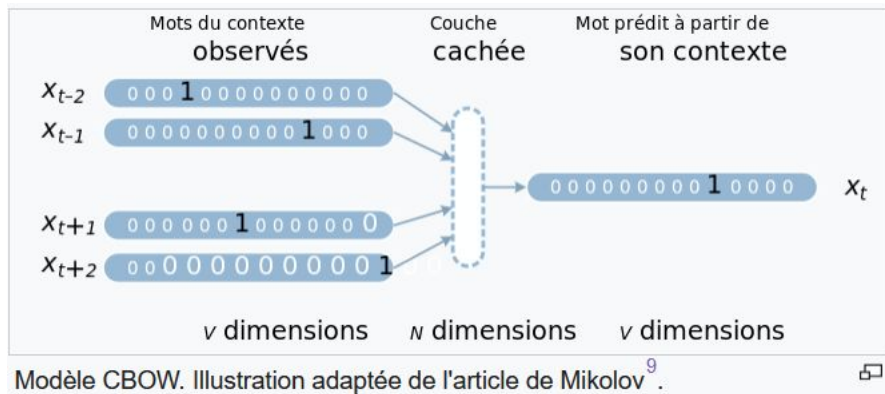




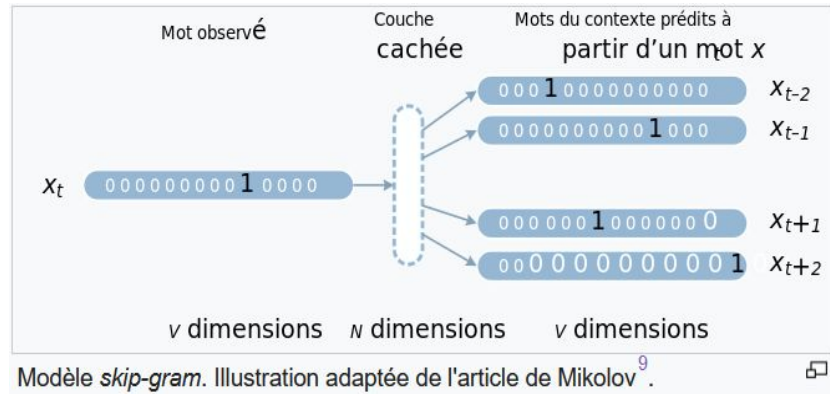
# L'embedding Word2vec:

- Apprend des embeddings de mots d'un vocabulaire d'un grand corpus de texte.
- Utile pour capturer **la similarité sémantique** entre les mots à partir de la similarité mathématique entre leurs vecteurs. **Exemple: roi-homme+femme=reine.**
- Il est composé de 2 architectures essentielles: **CBOW** et **SkipGram**.

## Architecture CBOW



## Architecture SkipGram



Possibilité de combiner les 2. (concaténation, moyenne...).

# Significativité du produit scalaire de 2 embeddings:

\_Pour un modèle SKIPGram qui prédit la probabilité d'avoir 2 mots dans  $w_O$  et  $w_I$  dans un contexte, on mq:

$$p(w_O|w_I) = \frac{\exp(v'_{w_O}{}^\top v_{w_I})}{\sum_{w=1}^W \exp(v'_w{}^\top v_{w_I})}$$

Avec  $v'_{w_O}$ : embedding de  $w_O$ , et  $v_{w_I}$ : celui de  $w_I$ .  
Le produit scalaire est donc proportionnel à la proba de leur coexistence en meme contexte.

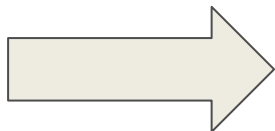
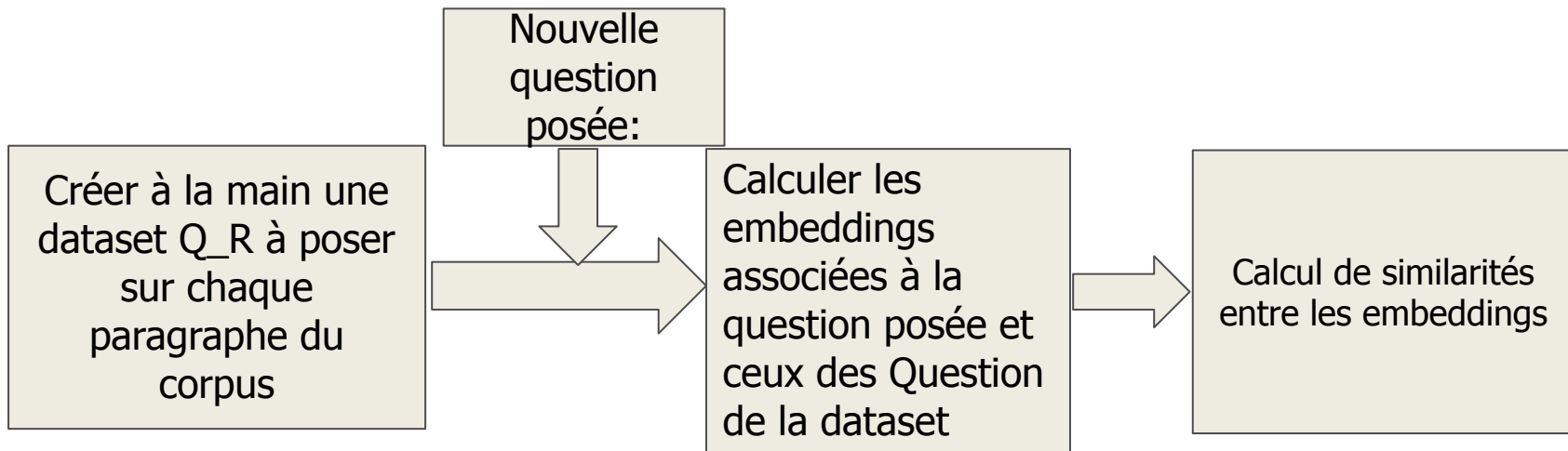
# Sent2vec: Embedding d'une phrase:

\_Plusieurs techniques sont utilisées pour obtenir les embeddings d'une phrase à partir des embeddings de ses mots: **max\_pooling, mean\_pooling, concatenation...**

\_On pourra même entraîner des réseaux de neurones pour apprendre les poids associés à l'embedding de chaque mot dans l'embedding total de la phrase.

# 1 ère approche simple:Modèle retrieve:

\_On s'intéresse à l'approche suivante:



On retient la réponse dont la question a la plus grande similarité avec la question posée.

# Résultats obtenus pour le modèle Retrive

- ➡ Embedding utilisé: sent2vec: en faisant la moyenne des embeddings des mots obtenus par CBOW.
- ➡ Similarité utilisée: Cosinus de 2 vecteurs. (valeur grande indique une cohérence contextuelle entre les 2 phrases).
- Quelques résultats obtenus:

```
>>>Comment valider la 1 ère année ?
We strongly recommend passing in an `attention_mask` since your input_ids may be padded. See https://huggingface.co/docs/transformers/troubleshooting#incorrect-output-when-padding-tokens-arent-masked.
Similarity: 0.5199403464794159 - un semestre detude dans une universite a letranger est valide si leleve obtient 30 credits ectes ou leur equivalent au quatrieme semestre valide 2 semestres a 30 credits ectes et effectue son travail de fin detudes aux cinquieme et sixieme semestres ou aux quatrieme et cinquieme semestres valide un double diplome dans une universite etrangere partenaire de lecole centrale de lyon ou obtient 30 credits ectes ou equivalents lors dune annee de cesure
```

# Limites de l'approche Retrieve:

\_Elle dépend de la richesse de la dataset, ne crée pas une nouvelle réponse mais juste récupère ce qui existe.

\_L'utilisation d'un modèle CBOW pré-entraîné ne permet pas forcément de couvrir le vocabulaire énorme des questions potentielles à poser.

\_Traite chaque question de la dataset individuellement(contexte du corpus pas pris en compte).

# Conditional Language Models

Language models estimate the probability of different linguistic units. LMs estimate the **unconditional** probability of a sequence  $y$ .

Language Models: 
$$P(y_1, y_2, \dots, y_n) = \prod_{t=1}^n p(y_t | y_{<t})$$

Sequence-to-sequence estimate the **conditional** probability of a sequence  $y$  given a source  $x$ .

Conditional

Language Models: 
$$P(y_1, y_2, \dots, y_n, | \textcolor{green}{x}) = \prod_{t=1}^n p(y_t | y_{<t}, \textcolor{green}{x})$$

condition on source  $x$



## Sequence to sequence basics

$$y' = \arg \max_y p(y|x, \theta)$$

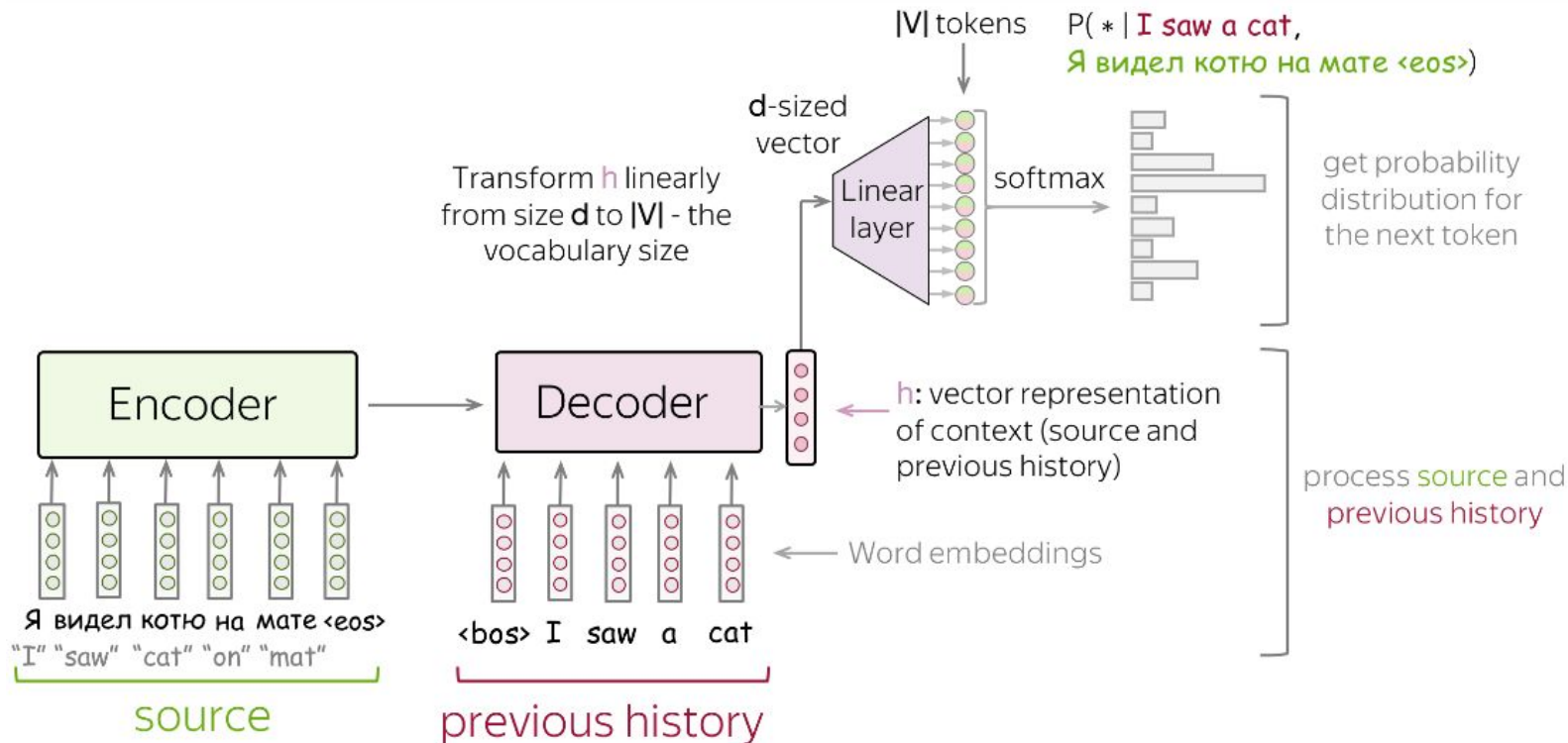
model
parameters

Questions we need to answer

- modeling
- learning
- search

- **Modeling:** How does the model for  $p(y|x, \theta)$  look like?
- **Learning:** How to find  $\theta$ ?
- **Search:** How to find the  $\arg \max$ ?

# Simple architecture



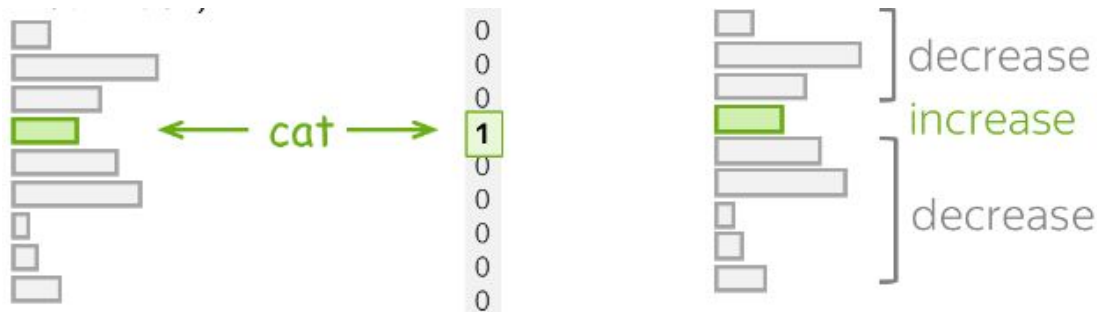
# Training: The Cross-Entropy Loss

The standard loss function is the **cross-entropy loss** for the **target distribution  $p^*$**  and the **predicted distribution  $p$** .

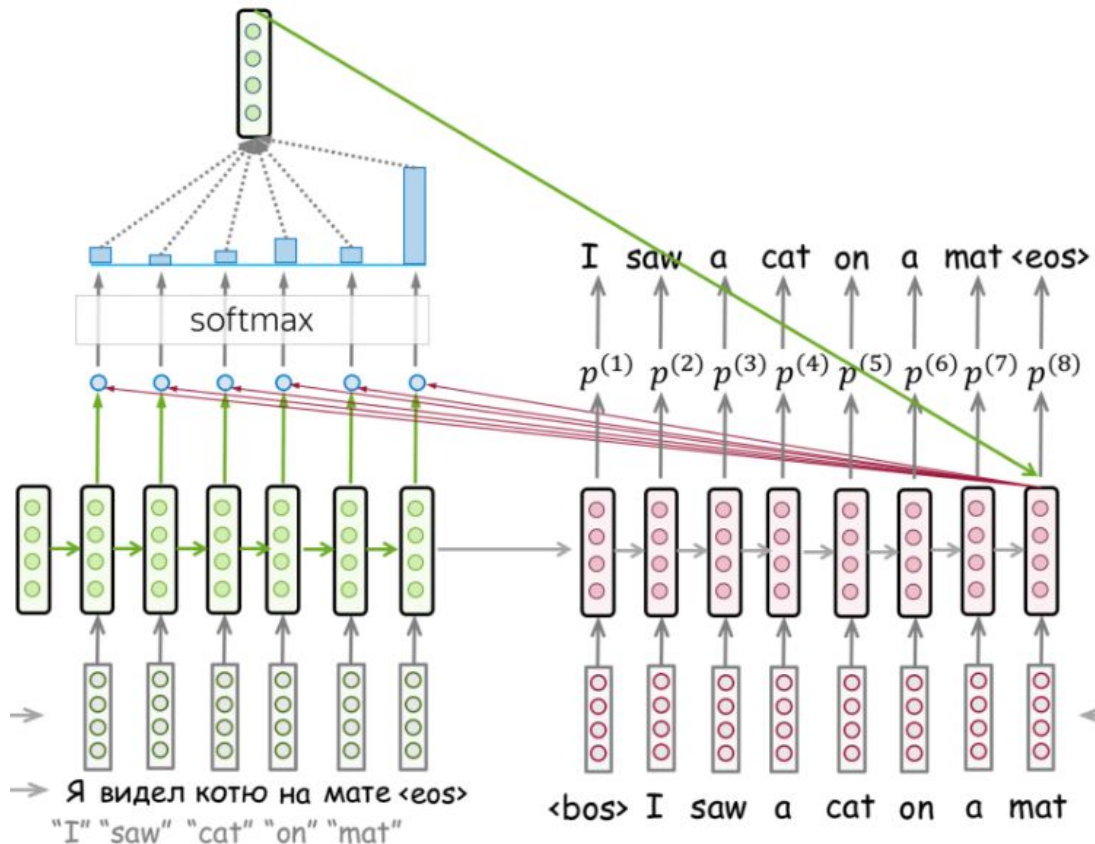
$$Loss(p^*, p) = -p^* \log(p) = - \sum_{i=1}^{|V|} p_i^* \log(p_i).$$

Predicted distribution

Target distribution



# Seq2seq with attention



Problems:

- Hard to compress the sentence for the **encoder**
- Different information may be relevant at different steps for the **decoder**

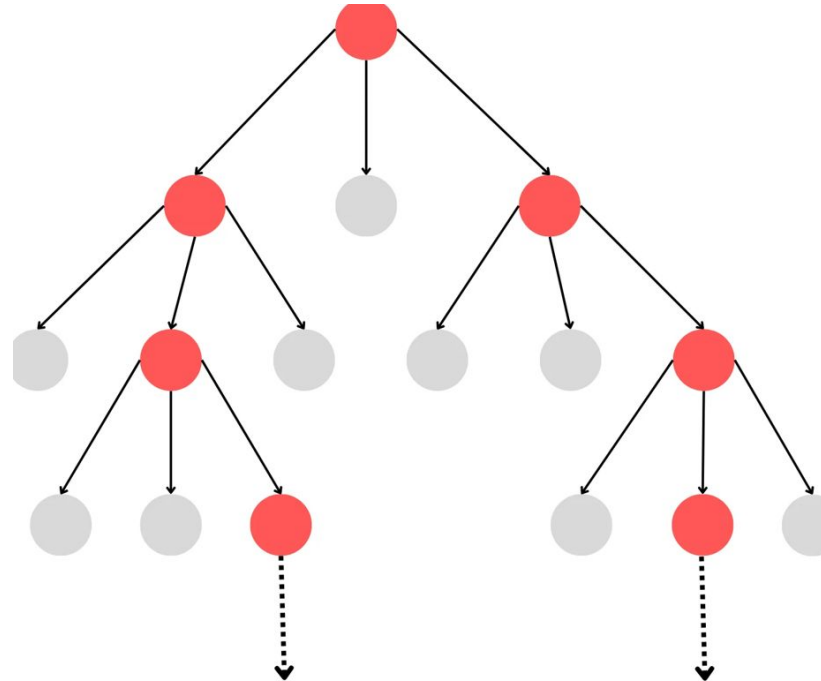
## Inference - Greedy decoding

- At each step pick the most probable token
- The best token at the current step does not necessarily lead to the best sequence

$$\arg \max_y \prod_{t=1}^n p(y_t | y_{<t}, x) \neq \prod_{t=1}^n \arg \max_{y_t} p(y_t | y_{<t}, x)$$

## Inference - Beam search

- We keep several hypothesis
- At each state, we will be continuing each of the current hypothesis and pick top-N of them

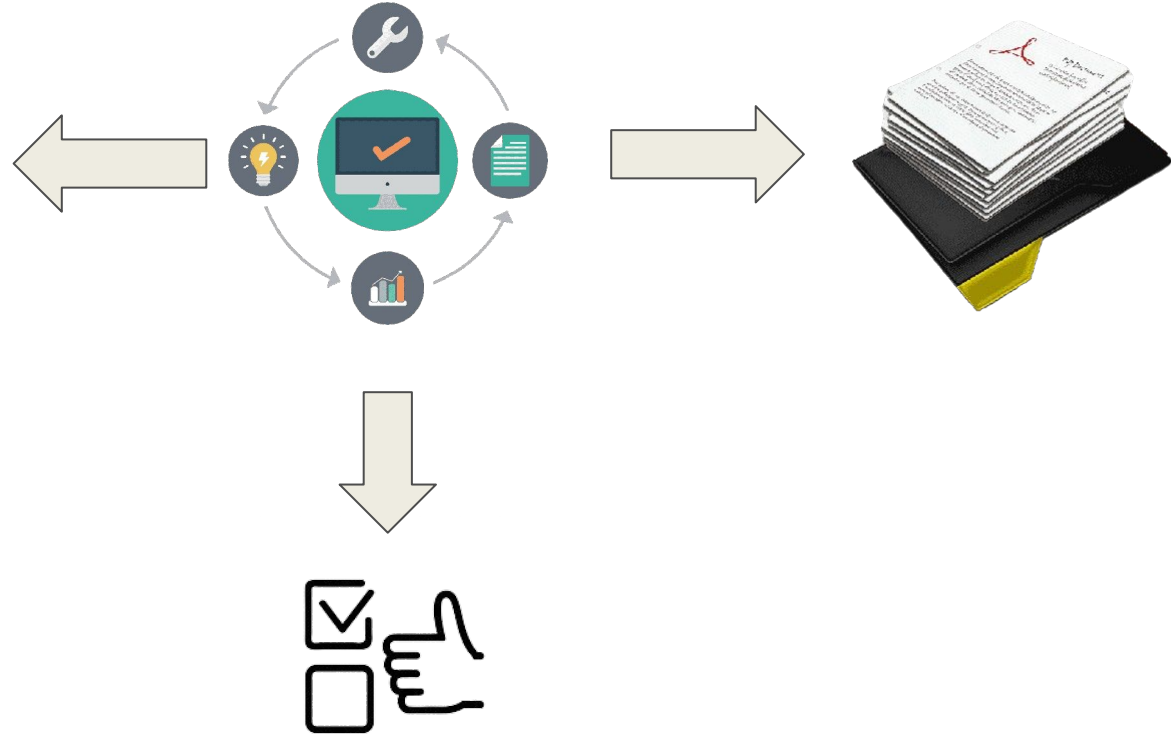


Continue till the goal state is found

# Approche Retrieve-Generate

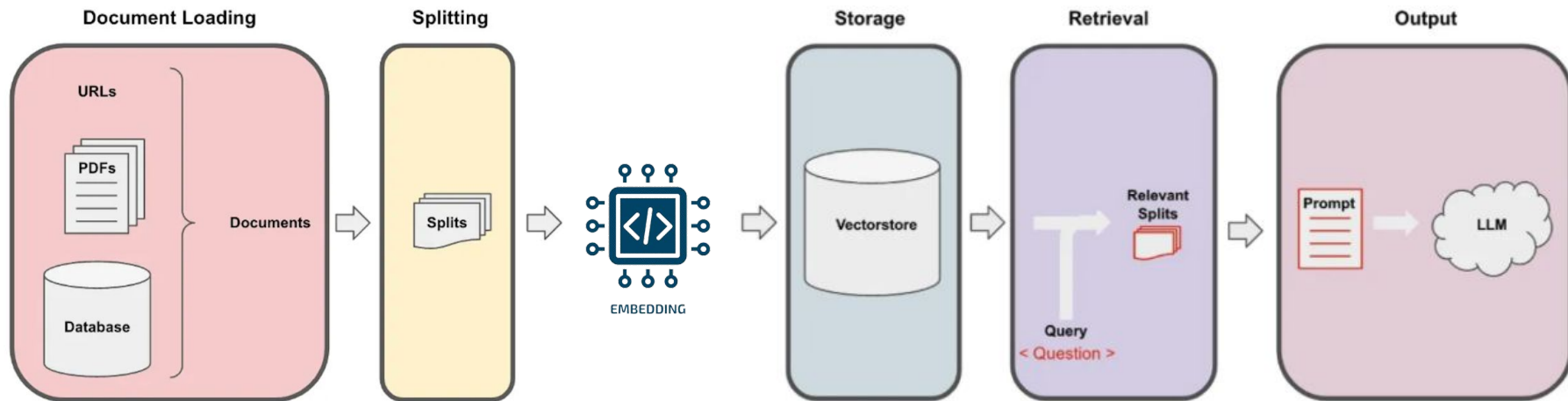
- Comment peut-on retirer de l'information sans avoir besoin de créer une base de données de questions et réponses associés?
- Une fois cette information pertinente identifiée, comment notre chatbot peut-il la transformer en réponses cohérentes et compréhensibles, assurant ainsi une expérience utilisateur fluide et enrichissante?

# Approche Retrieve-Generate

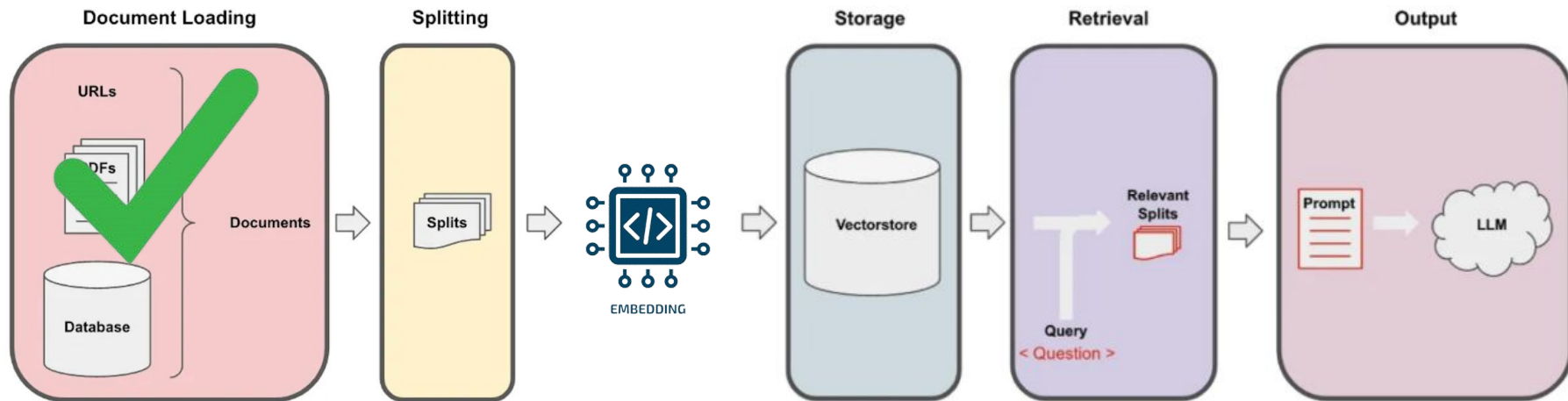




# Approche Retrieve-Generate

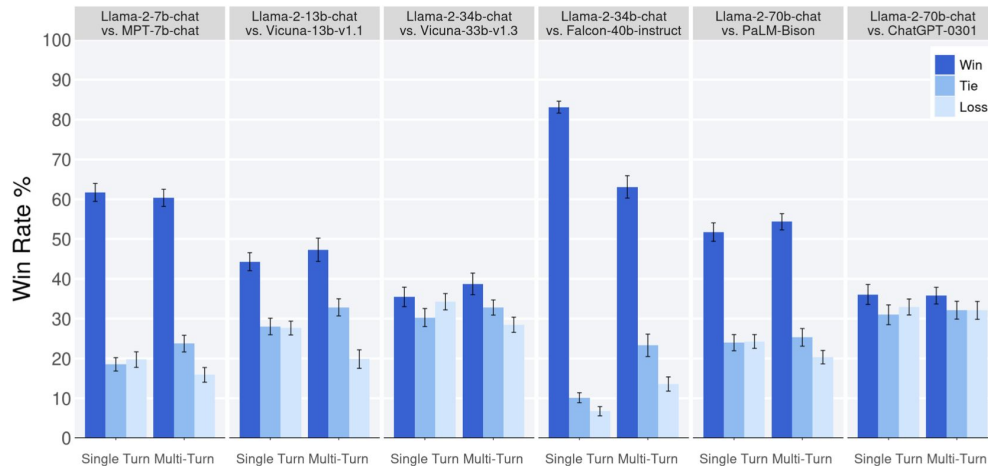


# Approche Retrieve-Generate



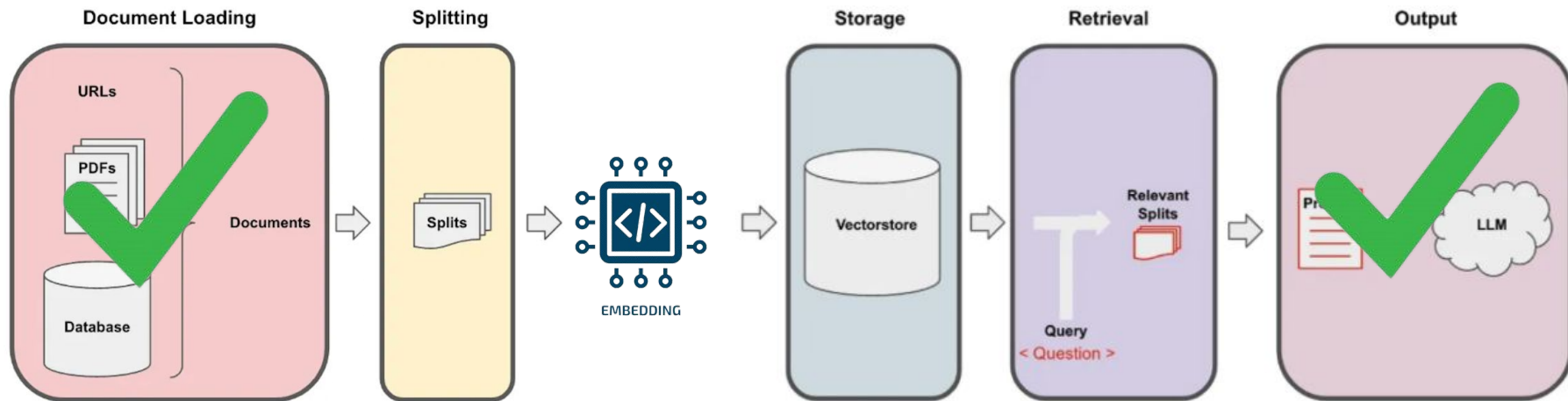
# Choix du LLM

- Open Source LLM
- Performance et précision
- Spécialisation et personnalisation

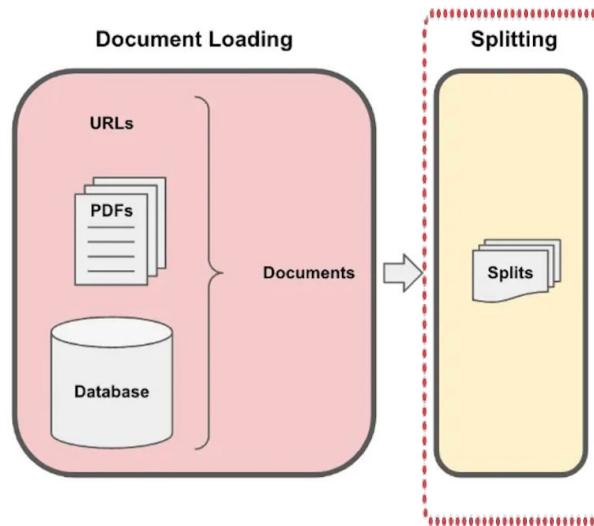


Résultats d'une comparaison humaine entre différents LLM disponibles

# Approche Retrieve-Generate



# Splitting des textes



# Splitting des textes

## A.2.2 Mobilités internationales

La CEU internationale examine les dossiers des étudiants candidats au recrutement dans le cadre des accords internationaux de l'établissement pour admission en qualité d'élèves à l'Ecole Centrale de Lyon pour une scolarité de deux ans en vue de l'obtention du diplôme d'ingénieur généraliste de l'Ecole Centrale de Lyon. Elle examine aussi les dossiers des étudiants candidats pour une scolarité en 3e année en vue de l'obtention du DESECL ou l'acquisition de crédits ECTS (European Credit Transfert System), ainsi que pour le certificat de formation à la recherche et l'admission pour un échange d'un semestre en S8. Dans tous les cas, ces candidats sont en priorité des étudiants proposés par les universités partenaires. Le Directeur de l'Ecole Centrale de Lyon fixe la liste des élèves admis sur titre sur proposition de la Commission des Echanges Universitaires.

La CEU internationale examine également les dossiers des élèves ingénieurs centraliens qui sont candidats à une formation double diplomante ou un échange dans une université étrangère suivant les critères présentés dans le dossier de demande. Elle statue sur l'autorisation à postuler dans un ou plusieurs établissements (à l'exclusion de tout autre) dans le cadre d'un échange. En cas de refus, les motifs de cette décision seront transmis aux élèves concernés. L'élève est finalement autorisé à partir lorsque l'établissement d'accueil émet un avis favorable pour l'accueillir, après validation de son Tronc Commun (Etendu le cas échéant) par le jury de Tronc Commun, et après validation du *learning agreement* par la Direction du Développement des Relations Internationales selon les modalités indiquées.

## A.2.3 Mobilités en France

La CEU France examine les dossiers des étudiants candidats au recrutement dans le cadre des accords avec des établissements partenaires en France (et l'Ecole Centrale de Casablanca) pour une admission sur titre en qualité d'élèves à l'Ecole Centrale de Lyon ou en double diplôme pour une scolarité de deux ou trois ans en vue de l'obtention du diplôme d'ingénieur généraliste de l'Ecole Centrale de Lyon. Elle examine aussi les dossiers des étudiants issus d'établissements partenaires en France, candidats pour une scolarité en 3e année en vue de l'obtention du DESECL. Le Directeur de l'Ecole Centrale de Lyon fixe la liste des élèves admis sur titre sur proposition de la Commission des Echanges Universitaires.

La CEU France examine également les dossiers des élèves ingénieurs centraliens qui sont candidats à une formation double diplomante ou un échange dans un établissement partenaire en France suivant les critères présentés dans le dossier de demande. Elle statue sur l'autorisation à postuler dans un ou plusieurs établissements (à l'exclusion de tout autre) dans le cadre d'un échange. En cas de refus, les motifs de cette décision seront transmis aux élèves concernés. L'élève est finalement autorisé à partir lorsque l'établissement d'accueil émet un avis favorable pour l'accueillir et après validation de son Tronc Commun (Etendu le cas échéant) par le jury de Tronc Commun.



Split qui s'adapte à la fin d'un paragraphe

Split ne contenant ni un excès ni n'a un manque d'information

Extrait du règlement de scolarité

# Splitting des textes

## A.2.2 Mobilités internationales

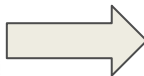
La CEU internationale examine les dossiers des étudiants candidats au recrutement dans le cadre des accords internationaux de l'établissement pour admission en qualité d'élèves à l'Ecole Centrale de Lyon pour une scolarité de deux ans en vue de l'obtention du diplôme d'ingénieur généraliste de l'Ecole Centrale de Lyon. Elle examine aussi les dossiers des étudiants candidats pour une scolarité en 3e année en vue de l'obtention du DESECL ou l'acquisition de crédits ECTS (European Credit Transfert System), ainsi que pour le certificat de formation à la recherche et l'admission pour un échange d'un semestre en S8. Dans tous les cas, ces candidats sont en priorité des étudiants proposés par les universités partenaires. Le Directeur de l'Ecole Centrale de Lyon fixe la liste des élèves admis sur titre sur proposition de la Commission des Echanges Universitaires.

La CEU internationale examine également les dossiers des élèves ingénieurs centraliens qui sont candidats à une formation double diplomante ou un échange dans une université étrangère suivant les critères présentés dans le dossier de demande. Elle statue sur l'autorisation à postuler dans un ou plusieurs établissements (à l'exclusion de tout autre) dans le cadre d'un échange. En cas de refus, les motifs de cette décision seront transmis aux élèves concernés. L'élève est finalement autorisé à partir lorsque l'établissement d'accueil émet un avis favorable pour l'accueillir, après validation de son Tronc Commun (Etendu le cas échéant) par le jury de Tronc Commun, et après validation du *learning agreement* par la Direction du Développement des Relations Internationales selon les modalités indiquées.

## A.2.3 Mobilités en France

La CEU France examine les dossiers des étudiants candidats au recrutement dans le cadre des accords avec des établissements partenaires en France (et l'Ecole Centrale de Casablanca) pour une admission sur titre en qualité d'élèves à l'Ecole Centrale de Lyon ou en double diplôme pour une scolarité de deux ou trois ans en vue de l'obtention du diplôme d'ingénieur généraliste de l'Ecole Centrale de Lyon. Elle examine aussi les dossiers des étudiants issus d'établissements partenaires en France, candidats pour une scolarité en 3e année en vue de l'obtention du DESECL. Le Directeur de l'Ecole Centrale de Lyon fixe la liste des élèves admis sur titre sur proposition de la Commission des Echanges Universitaires.

La CEU France examine également les dossiers des élèves ingénieurs centraliens qui sont candidats à une formation double diplomante ou un échange dans un établissement partenaire en France suivant les critères présentés dans le dossier de demande. Elle statue sur l'autorisation à postuler dans un ou plusieurs établissements (à l'exclusion de tout autre) dans le cadre d'un échange. En cas de refus, les motifs de cette décision seront transmis aux élèves concernés. L'élève est finalement autorisé à partir lorsque l'établissement d'accueil émet un avis favorable pour l'accueillir et après validation de son Tronc Commun (Etendu le cas échéant) par le jury de Tronc Commun.



Split qui prend tout d'abord en compte les paragraphes

Nombre de mots dans le split ni grand ni petit

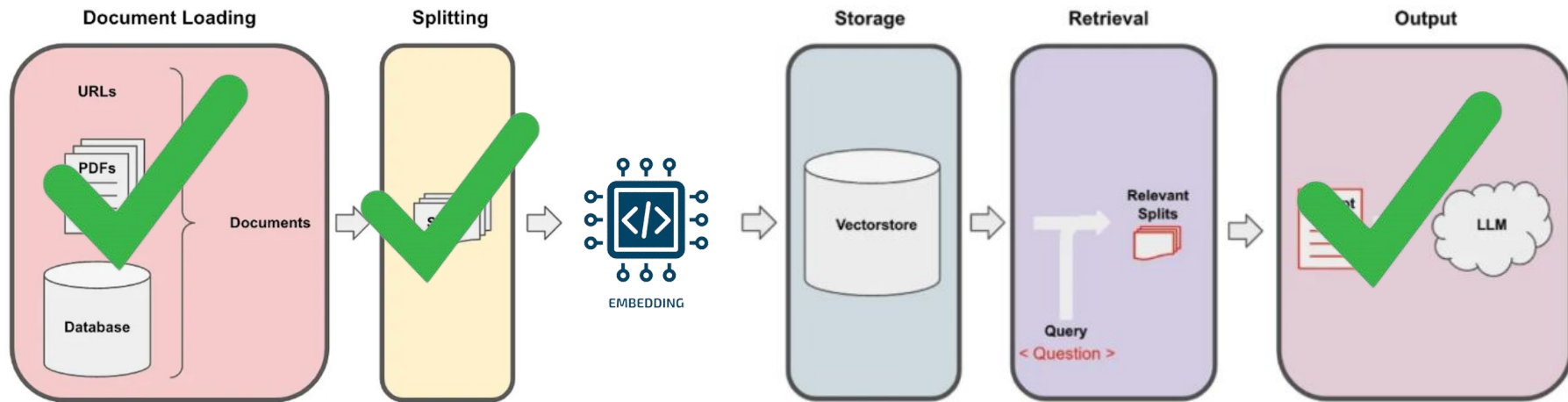
Extrait du règlement de scolarité

# Splitting des textes

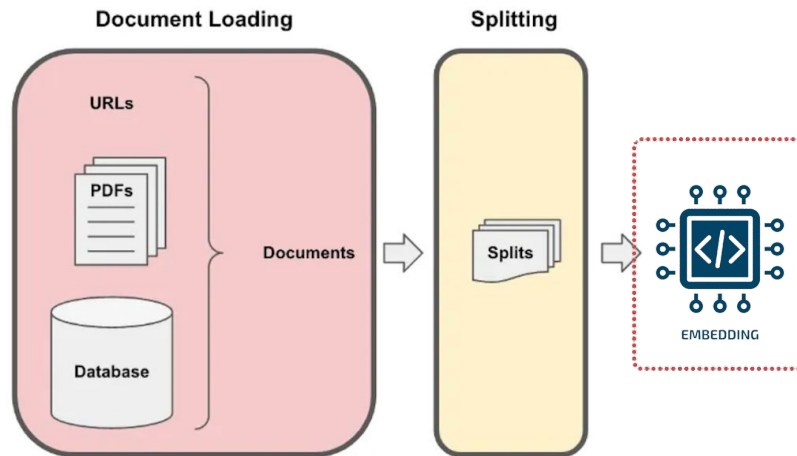
- Implémentation d'un splitter qui prend en compte premièrement les paragraphes, puis les lignes, jusqu'à l'arrivée à un nombre de caractères inférieur au nombre de caractère fixé.
- Nombre de caractères optimal trouvé en faisant plusieurs expériences est 500



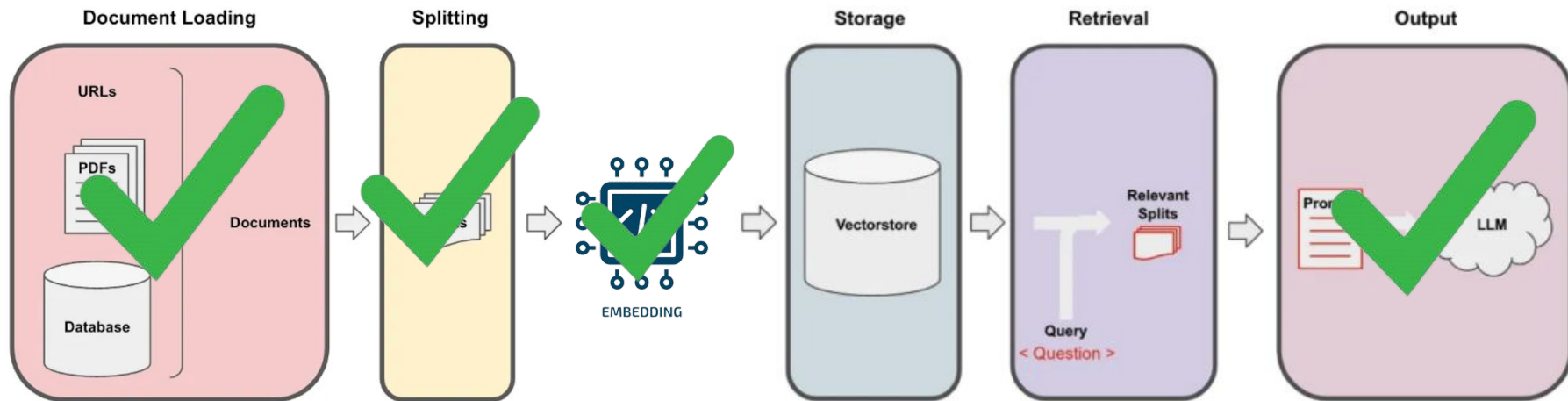
# Approche Retrieve-Generate



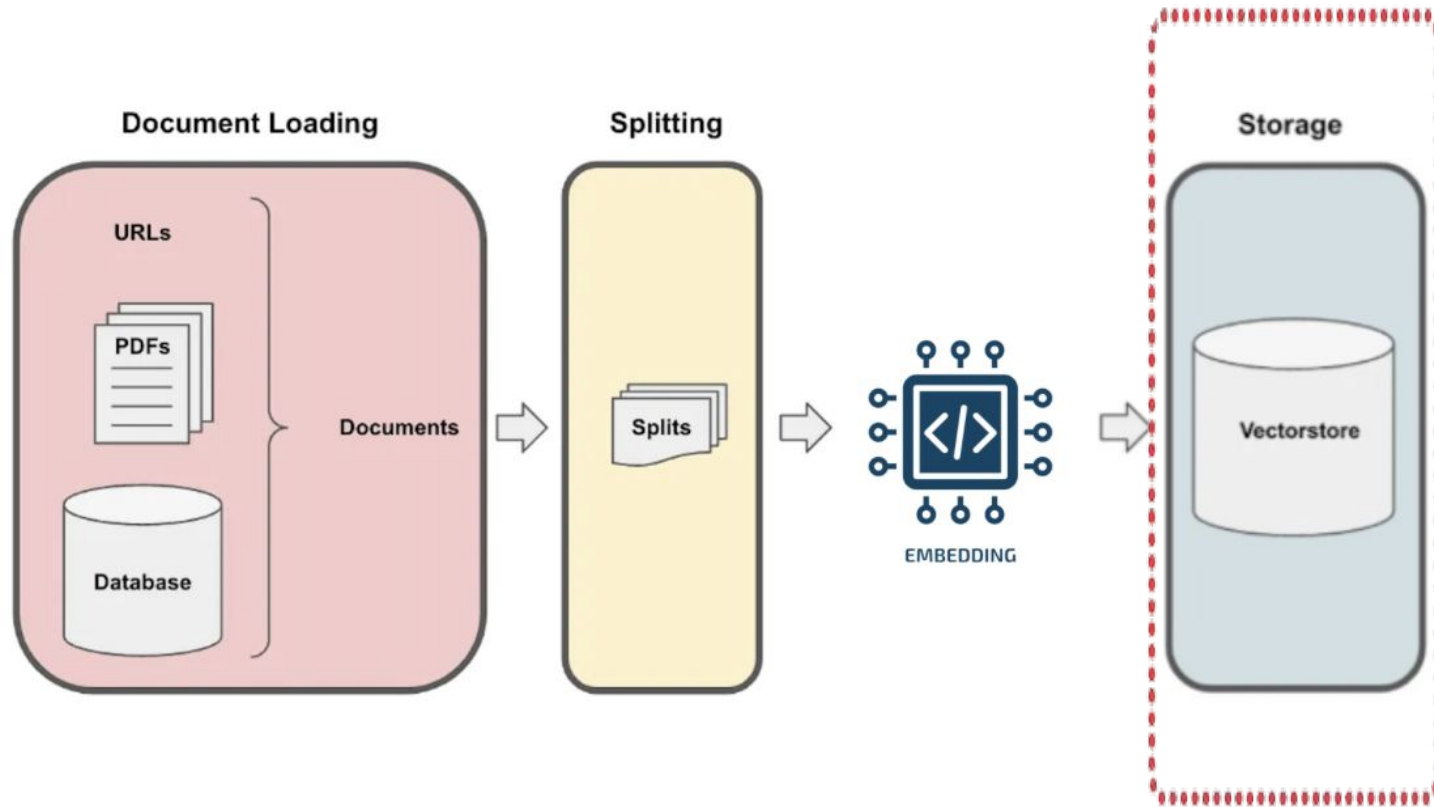
# Qu'est ce qu'après le splitting?



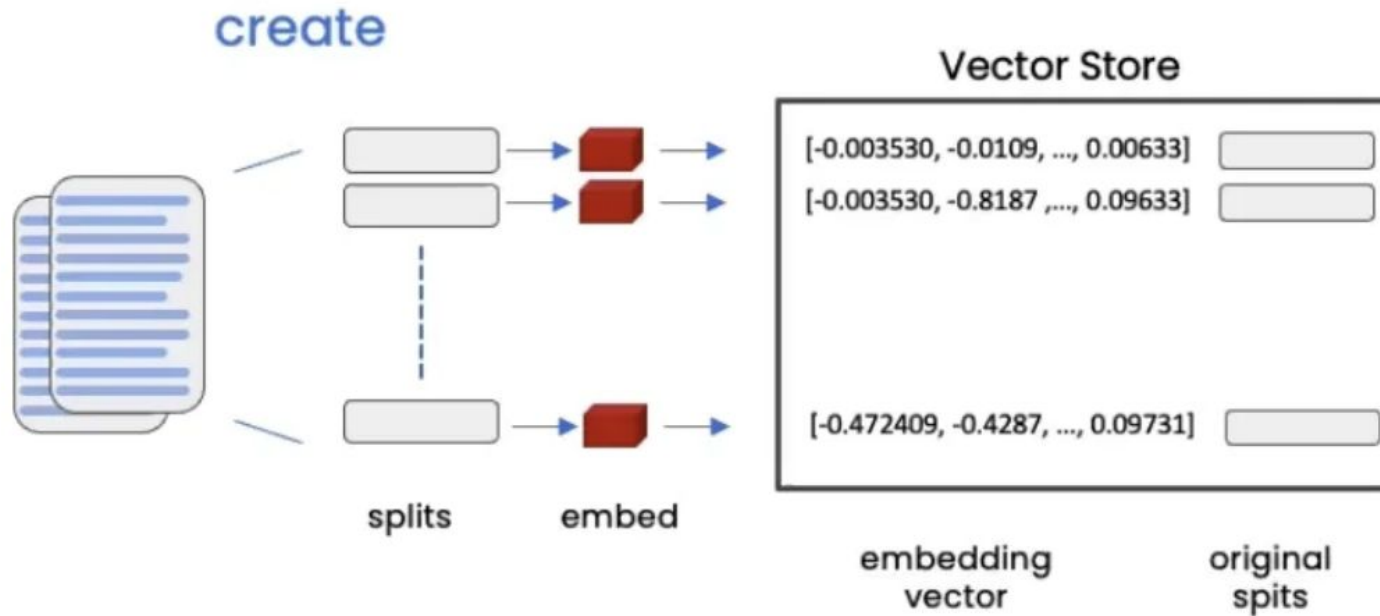
# Approche Retrieve et générative



# Stockage des données



# Stockage des données



# Récupération des réponses

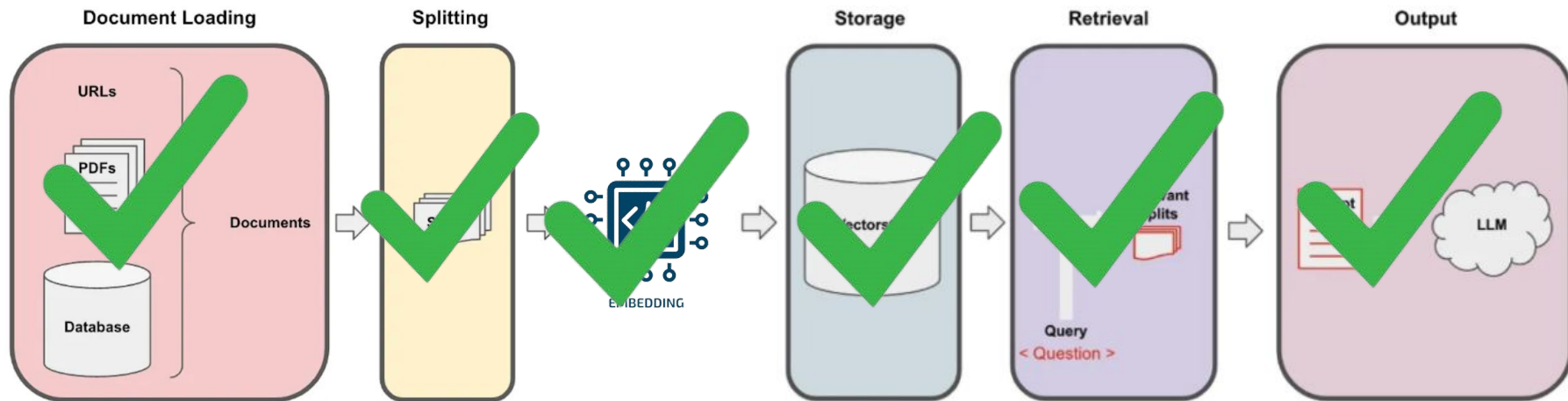
Le problème après embeddings se formule :

$$L = k\text{-argmin}_{i=0:\ell} \|x - y_i\|_2$$

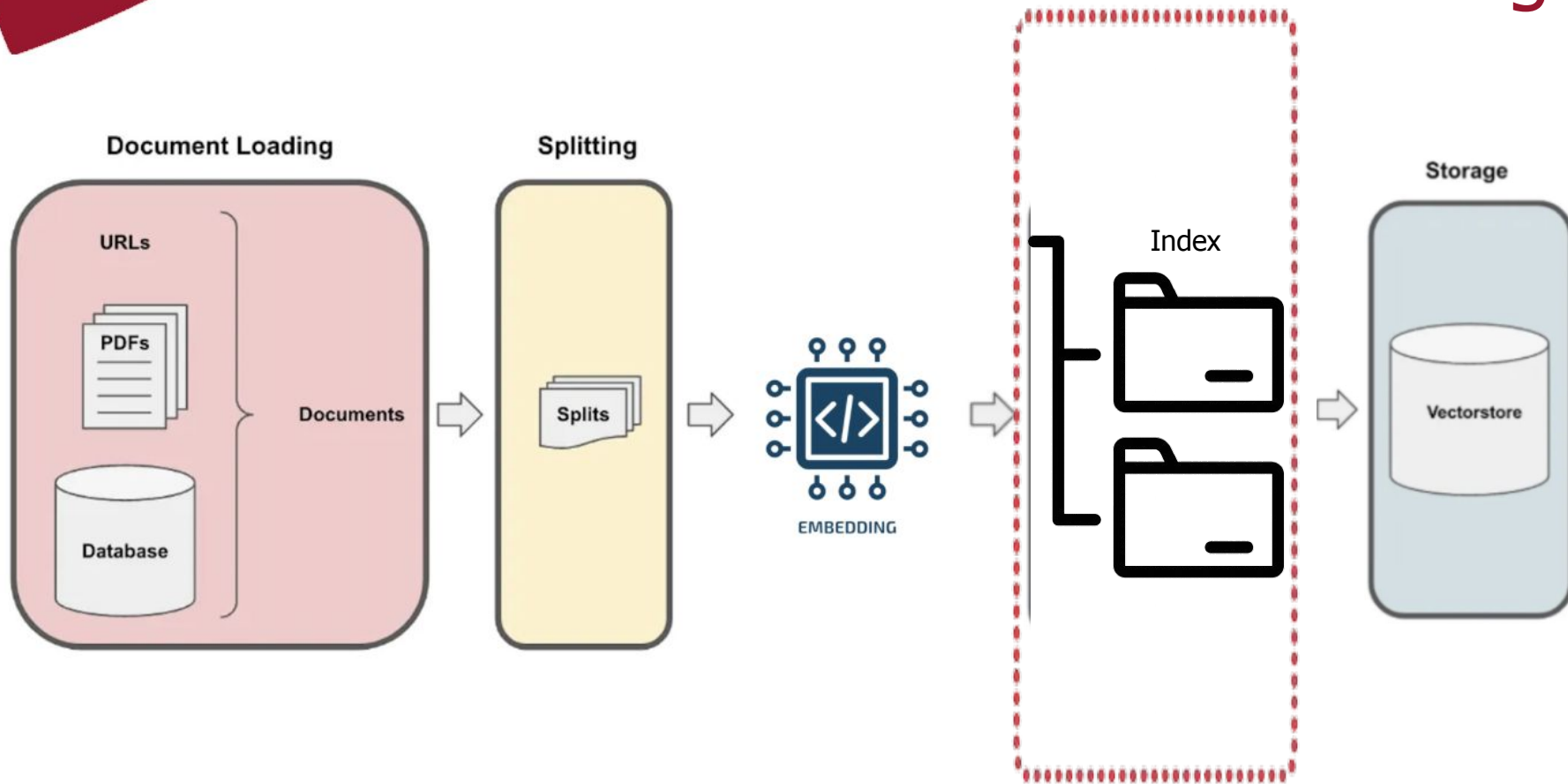
Complexité =  $O(N * l / 500)$

N : nombre de caractères dans la totalité des documents.

# Approche Retrieve et générative



# Indexation des embeddings





# Indexation des embeddings

Solution : Clustering des  $y_i$  dans un espace à  $k$  clusters

$$q_1 : \mathbb{R}^d \rightarrow \mathcal{C}_1$$

Résolution du problème :

$$L_{\text{IVF}} = \tau\text{-argmin}_{c \in \mathcal{C}_1} \|x - c\|_2$$

Avec  $\mathcal{C}_1$  est l'ensemble des centroids,  $\tau$  le nombre de centroids qu'on veut récupérer

# Indexation des embeddings

$$L_{\text{IVF}} = \tau\text{-argmin}_{c \in \mathcal{C}_1} \|x - c\|_2$$

Solution : Plus proches  $\tau$  centroids à la réponse

Problème ramené au clusters de centroids solution de  $L_{\text{IVF}}$

# Indexation des embeddings

Objectif : Recherche efficace dans un cluster

Solution : Nouvelle indexation sur la différence entre le vecteur et le centroid de son cluster

$$q_2(y - q_1(y))$$

# Indexation des embeddings

Objectif : Recherche efficace dans un cluster

Problème : Égalité de l'indice de deux vecteurs

$$q_2(y - q_1(y))$$

# Indexation des embeddings

Objectif : Recherche efficace dans un cluster

Solution : Quantificateur produit

$$y = [y^0 \dots y^{b-1}] \quad q_2(y) = (q^0(y^0), \dots, q^{b-1}(y^{b-1}))$$

# Indexation des embeddings

$$L_{\text{IVF}} = \tau\text{-argmin}_{c \in \mathcal{C}_1} \|x - c\|_2$$

$$y = [y^0 \dots y^{b-1}] \quad q_2(y) = (q^0(y^0), \dots, q^{b-1}(y^{b-1}))$$

$$q_1 : \mathbb{R}^d \rightarrow \mathcal{C}_1$$

$$L_{\text{IVFADC}} = \underset{i=0:\ell \text{ s.t. } q_1(y_i) \in L_{\text{IVF}}}{k\text{-argmin}} \|x - q(y_i)\|_2$$

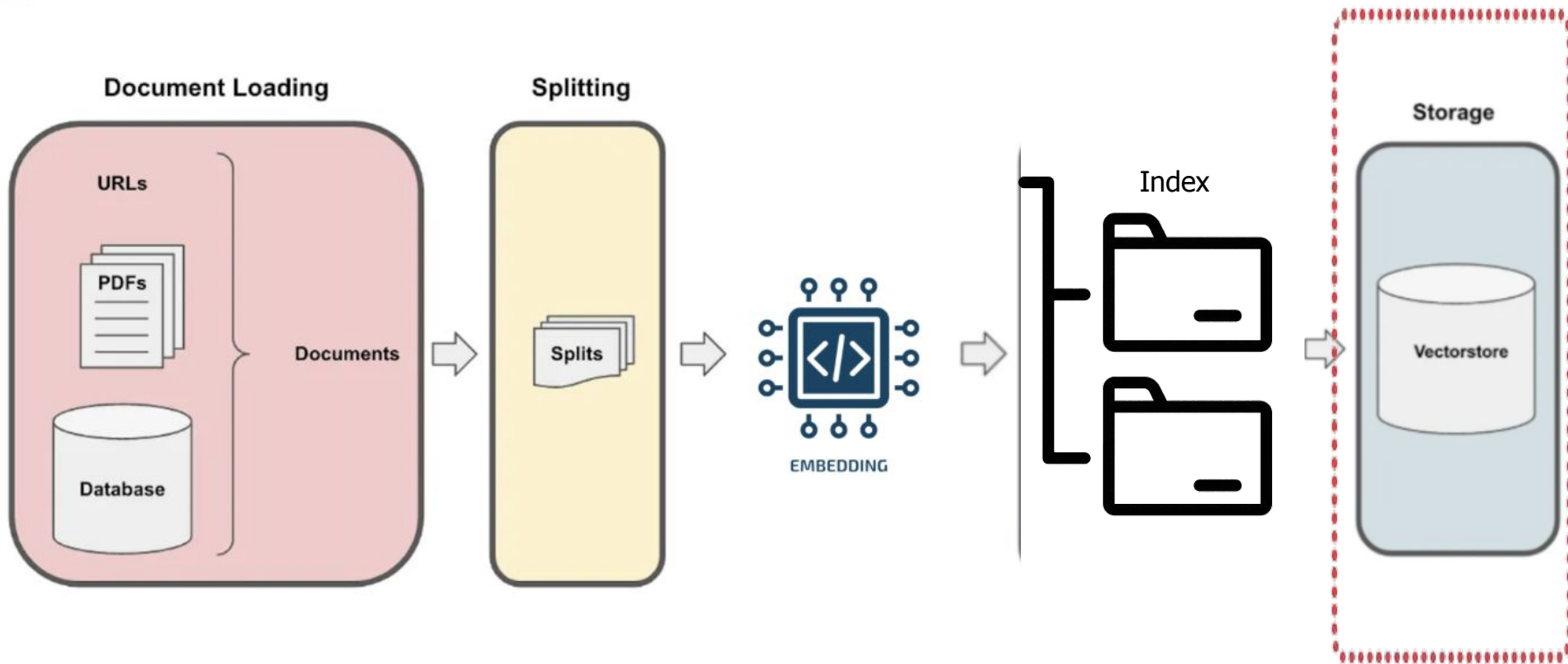
# Indexation des embeddings

$$L_{IVF} = \tau\text{-argmin}_{c \in C_1} \|x - c\|_2 \quad L_{IVFADC} = \underset{i=0:\ell \text{ s.t. } q_1(y_i) \in L_{IVF}}{k\text{-argmin}} \|x - q(y_i)\|_2$$

$$y = [y^0 \dots y^{b-1}]$$

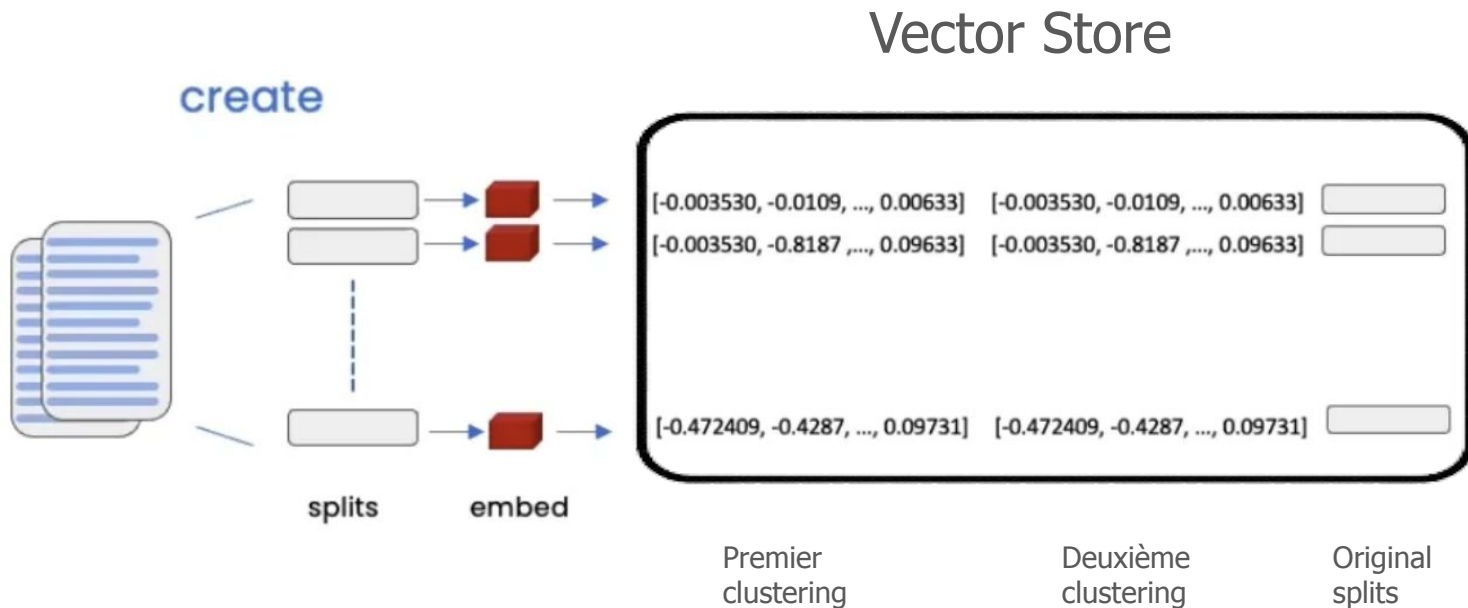
- Il a été prouvé que  $|C_1| \approx \sqrt{\ell}$  donne des résultats pertinents
- $\tau = 1$ ,  $b = 1$
- $k=2$

# Stockage de données indexées





# Stockage de données indexées



# Récupération des éléments de réponse

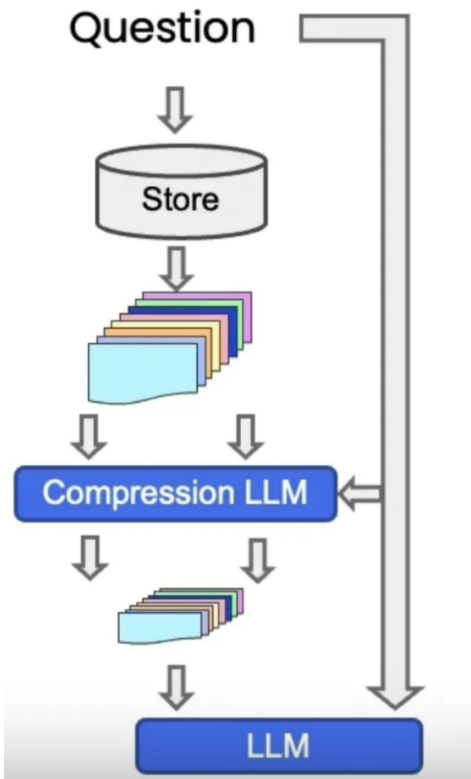
$$L_{\text{IVF}} = \tau\text{-argmin}_{c \in \mathcal{C}_1} \|x - c\|_2$$

$$L_{\text{IVFADC}} = \underset{i=0:\ell \text{ s.t. } q_1(y_i) \in L_{\text{IVF}}}{k\text{-argmin}} \|x - q(y_i)\|_2$$

# Réponse finale

Partie retrouvée par  
récupération (retrieve)

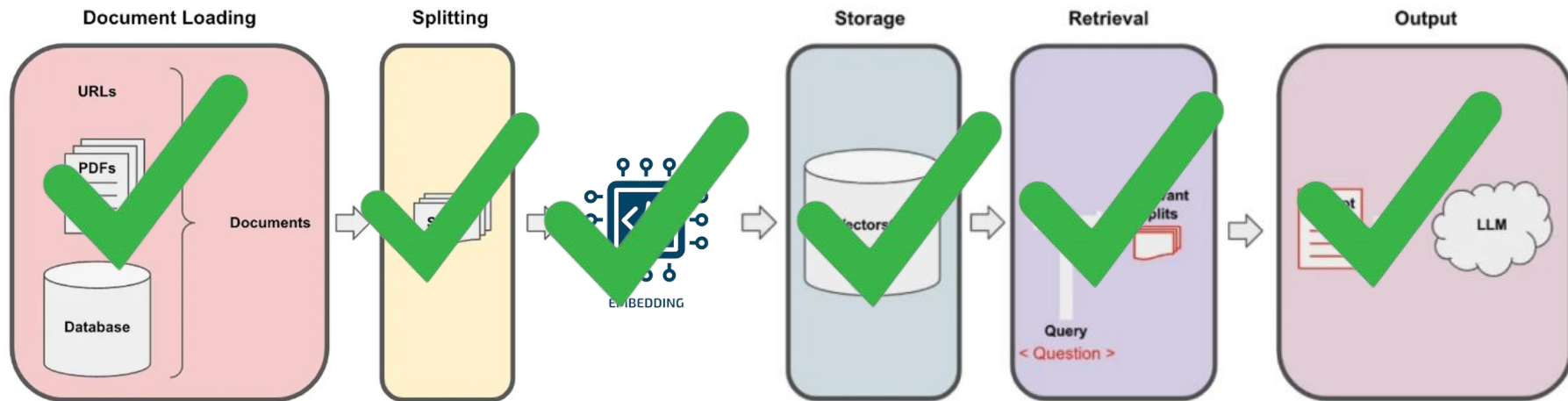
Partie gardée par le LLM  
après contextualisation



# Réponse finale

- Dans notre cas,  $k=2$ , le LLM intervient rarement avant l'étape finale
- Il est donc utile que pour la partie de synthèse de la réponse

# Approche Retrieve et générative



# Résultats du ChatBot



**You** 23:05

Comment un élève peut justifier son absence?

Took 1 step ▾



**Chatbot** 23:05

L'élève peut justifier son absence en signalant la raison de l'absence au Service de la scolarité le plus tôt possible et au plus tard dans les cinq jours ouvrables qui suivent le premier jour de l'absence. Seules sont autorisées les absences pour raisons de force majeure (comme des problèmes de santé ou des décès familiaux), pour des convocations émanant d'autorités administratives ou pour des motifs précisés par l'école.



**You** 11:44

Que faire si un élève s'absente au rattrapage d'un examen?

Took 1 step ▾



**Chatbot** 11:44

Si un élève s'absente au rattrapage d'un examen, il doit contacter immédiatement le Service de la Scolarité pour demander une autorisation d'absence. Si l'absence est justifiée, un nouvel examen sera proposé à un délai ultérieur pour permettre à l'élève de rattraper la séance d'examen manquée.



# Résultats du ChatBot

**You** 11:45

Qu'est ce qu'un élève peut faire s'il s'absente à un TP?

Took 1 step ▾

**Chatbot** 11:45

Lorsque l'un des élèves d'une classe est absent au TP (Travaux Pratiques), cela peut avoir des conséquences sur les résultats de l'ensemble de la classe. Il est donc important que les élèves soient présents à tous les TP pour réussir dans leur formation.



**You** 12:59

Que faire si un élève s'absente au TP?

Took 1 step ▾

**Chatbot** 13:00

Si un élève s'absente au TP, il est important de vérifier si l'absence est autorisée ou non. Si elle n'est pas autorisée, il convient de notifier l'élève de la conséquence de l'absence non autorisée, notamment la note de 0 pour l'évaluation de la séance et peut-être des conséquences pour le déroulement de la série complète de TP. Il est également important de transmettre les motifs de la décision de refus d'autorisation d'absence auprès des élèves concernés.



# Résultats du ChatBot

**You** 12:36  
Quelle est la durée du stage d'exécution?

Took 1 step ▾

**Chatbot** 12:36  
Le stage d'exécution dure au minimum quatre semaines.



**You** 12:36  
Quelle est la durée du stage d'application?

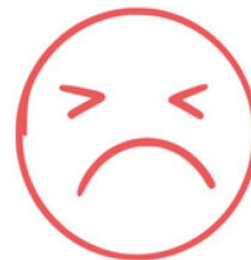
Took 1 step ▾

**Chatbot** 12:37  
Le stage d'application dure au minimum 3 mois.

**You** 12:40  
Quelle est la durée d'un TFE?

Took 1 step ▾

**Chatbot** 12:40  
Le TFE dure une durée de 6 semaines.





# Explication du résultat

## Le stage d'exécution

Effectué **en fin de première année**, le stage d'exécution vise à faire découvrir à l'élève le fonctionnement d'une entreprise, la nature du travail d'exécution et les relations entre opérateurs et cadres.

Ce stage consiste en la réalisation, **pendant au minimum quatre semaines**, du travail d'un opérateur au sein d'une équipe d'opérateurs. Il fait partie intégrante de la scolarité et constitue une expérience enrichissante pour la formation professionnelle et humaine des élèves ingénieurs.

## Le stage d'application

Effectué **en fin de deuxième année**, le stage d'application dure au minimum 3 mois. Il doit être achevé à la rentrée universitaire suivante. Complémentaire du stage d'exécution, le stage d'application fait lui aussi partie intégrante de la scolarité et contribue à former les élèves tant professionnellement qu'humainement.

1ère	stage d'exécution (ouvrier)	1 à 2 mois 4 semaines minimum	1 juillet
2ème	stage d'application	3 à 4 mois 12 semaines minimum	02 mai
3ème	travail de fin d'étude	5 à 6 mois 21 semaines minimum	1 avril

# Résultats du ChatBot

**You** 15:57  
C'est quoi la 3A?  
Took 1 step ▾

**Chatbot** 15:57  
La 3A est l'absence pour activités associatives.  
📋



**You** 15:55  
C'est quoi un TP?  
Took 1 step ▾

**Chatbot** 15:55  
Un TP est un tuteur pédagogique.  
📋

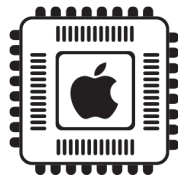
**You** 15:58  
C'est quoi le S7?  
Took 1 step ▾

**Chatbot** 15:58  
Le S7 est le septième semestre d'études.  
📋



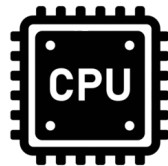
# Analyse de performance

APPLE SILICON “GPU”



15 secondes

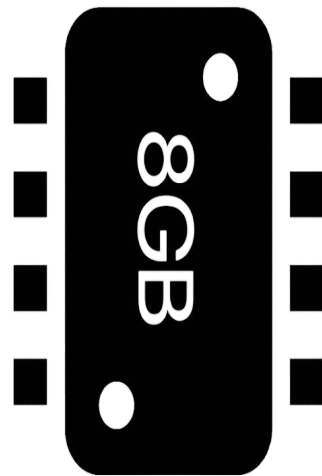
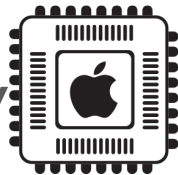
Intel CPU - serveur -



2 à 3 minutes

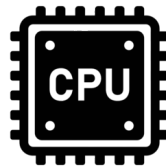
# Analyse de performance

APPLE SILICON "GPU"



Quantification  
4-bit

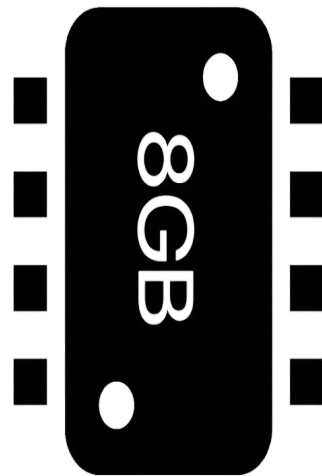
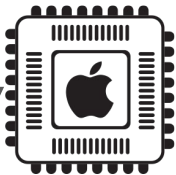
Intel CPU - serveur -



Quantification  
8-bit

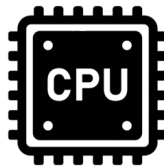
# Analyse de performance

APPLE SILICON "GPU"



Interface  
graphique  
accessible

Intel CPU - serveur -



Inaccessibilité  
d'interface graphique

# Perspectives

- Inclusion des données tabulaires
- Élargissement de la base des données
- Prise en compte des données tabulaires
- Prise en compte des abréviations non incluses dans le règlement de scolarité.
- Utilisation d'une quantification de Llama plus avancée.
- Transmission du processus vers une GPU puissante (celle du serveur)
- Hébergement du ChatBot dans le serveur de l'école



**Merci de votre  
attention**

- LLaMa 2: Open Foundation and Fine-Tuned Chat Models, 2023
- MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers, 2020
- Billion scale search similarity with GPUs, 2017
- Searching with one billion vector, re-rank with source coding, 2011