
# XAI for Large Language Models

Kerrian Le Caillec

Ibrahim Al Khalil Ridene

Lydia Hammache

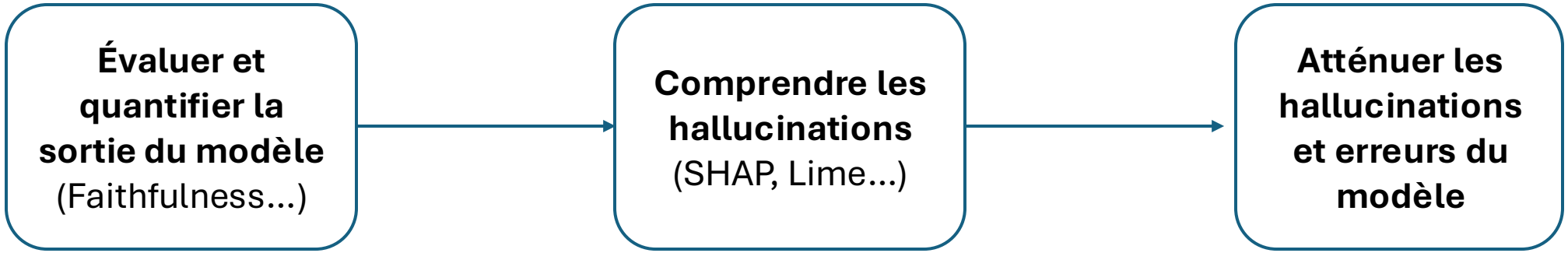
# Contexte et Objectifs

**Problème : Comment s'assurer que la réponse  d'un LLM soit cohérente et vérifie un certain nombre de critères ?**

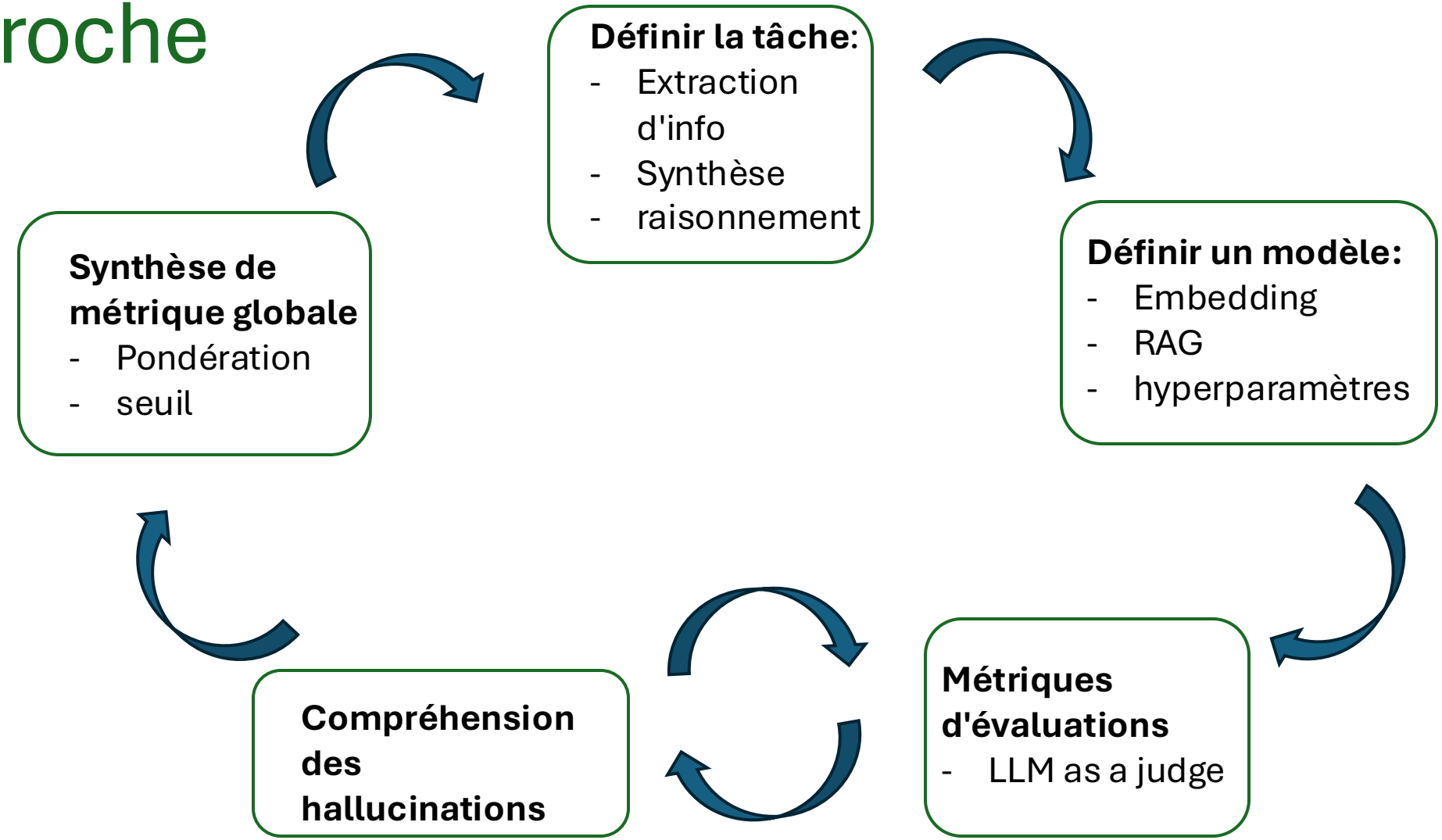
- Pas de mesure de la validité des réponses du LLM.
- Améliorer la compréhension du processus de génération des LLMs.
- Développer des outils pratiques pour analyser les prises de décision des LLMs, notamment en cas d’hallucinations.
- Définir des métriques pour quantifier un niveau de certitude sur une réponse générée.
- Définir un seuil à partir duquel une réponse est risquée.

# Plan de travail

- **Etat de l'art**
  - **Méthode d'évaluations dans le cadre de *text generation* et d'explicabilité.**
- **Mise en pratique**

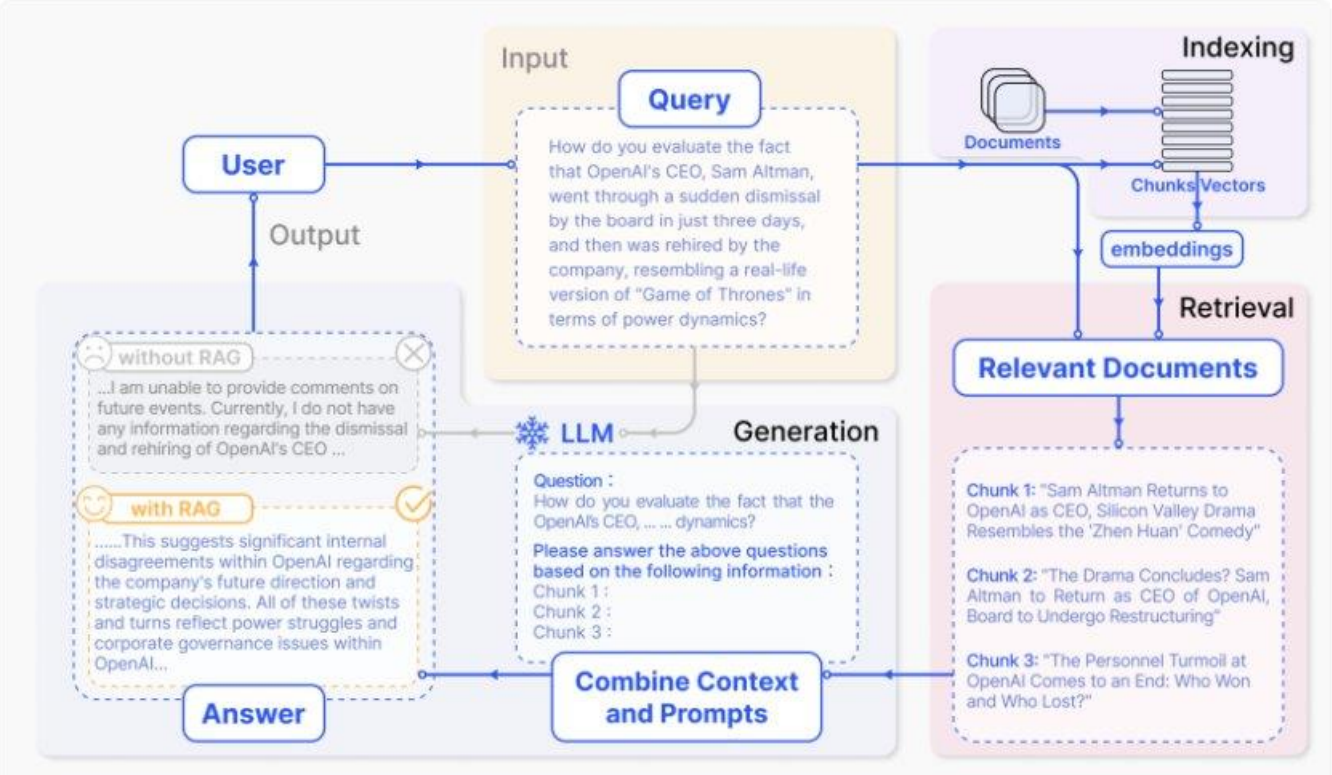


# Approche


**Définir la tâche:**
- Extraction d'info
- Synthèse
- raisonnement
**Définir un modèle:**
- Embedding
- RAG
- hyperparamètres
**Métriques d'évaluations**
- LLM as a judge
**Compréhension des hallucinations**
**Synthèse de métrique globale**
- Pondération
- seuil

# Méthodes

# RAG Models




**RAG pipeline** :

- Retrieval evaluation
  - How relevant are the chosen documents?
  - Are all documents as relevant?
- Generation evaluation
  - How relevant is the answer compared to the query ?
  - How accurate is the answer given the context ?

[https://huggingface.co/blog/hrishioa/retrieval-augmented-generation-1-basics]

# RAG evaluation metrics

1/ Embedding Evaluation

2/ Retrieval Evaluation

3/ Generation Evaluation

# 1/ Embedding Evaluation

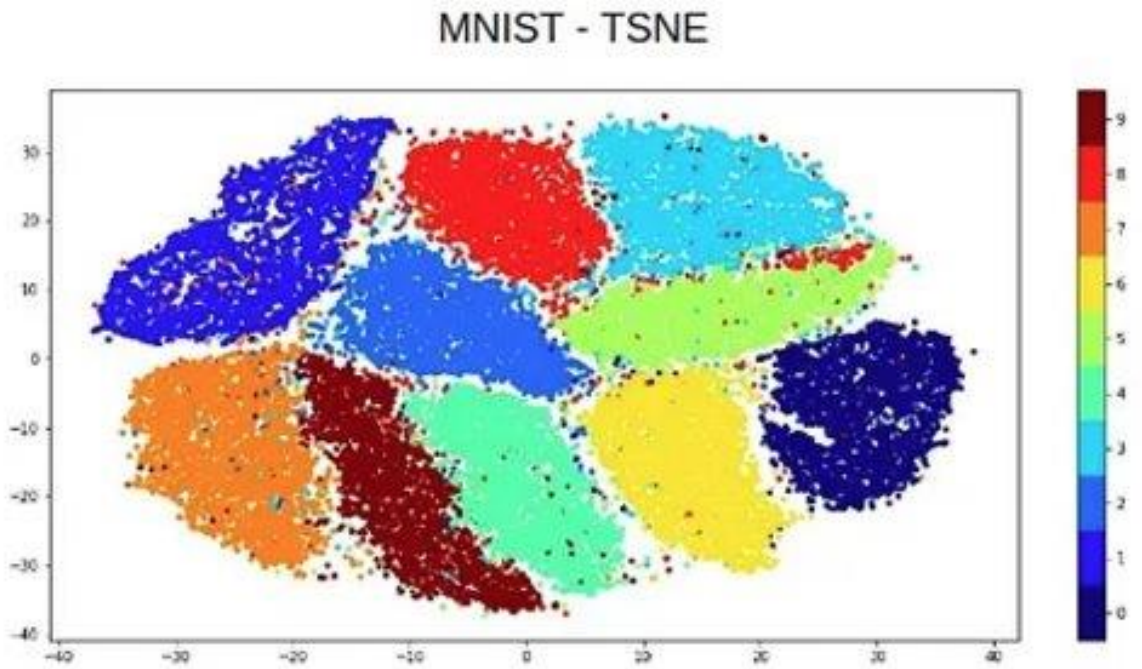
**Task based methods**

- Évaluer les performances sur des tâches spécifiques (e.g., classification, clustering) sans fine-tuning.

**Qualitative Approaches**

- Visualizations : t-SNE, UMAP

- Projection interpretation : étude de la représentation vectorielle des prompts en dimension réduire (PCA)




T-SNE for MNIST dataset
Mastering t-SNE(t-distributed stochastic neighbor embedding)
Sachinsoni

# 2/ Retrieval Evaluation

- Recall
  - Mesure la proportion de documents pertinents qui figurent dans les K premiers résultats.

$$Recall@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ negatives@k)}$$

- Precision
  - Mesure la proportion de documents pertinents parmi les K premiers documents récupérés.

$$Precision@k = \frac{true\ positives@k}{(true\ positives@k) + (false\ positives@k)}$$

- F1-score
  - Combine précision et rappel dans une seule mesure harmonique.

$$\mathbf{F1@K} = 2 \cdot \frac{\mathbf{Precision@K} \cdot \mathbf{Recall@K}}{\mathbf{Precision@K} + \mathbf{Recall@K}}$$

- Robustesse au bruit
  - Ajouter du bruit au prompt et vérifier si les documents récupérés sont pertinents

    Exp : remplacer "Paris" par "pariss" ou "ville lumière".

# 2/ Retrieval Evaluation

- nDCG@K (Normalized Discounted Cumulative Gain)
  - utilisée pour évaluer la qualité d'une liste de résultats ordonnés : ordre + pertinence ("très pertinent", "peu pertinent", ou "non pertinent")
  - DCG : Discount Cumulative Gain
  - IDCG : Ideal Discount Cumulative Gain

$$\mathrm{nDCG@K} = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}$$

$$\mathrm{DCG@K} = \sum_{i=1}^{K} \frac{2^{\mathrm{relevance}(i)} - 1}{\log_2(i+1)}$$

  - **relevance(i)** : Score de pertinence attribué au document à la position i
    - exp : 2 (très pertinent), 1 (peu pertinent), ou 0 (non pertinent).
  - **i** : Position du document dans la liste récupérée.
  - **log2 (i+1)** : Pondère l'importance de la position du document. Plus un document est bas dans la liste (plus i est grand), plus son score de pertinence contribue moins au DCG.

# 2/ Context-related evaluation

- Completeness
  - Check if answer contains all the relevant information from the references, and let the model decides a $grade/5$.
    - => which granularity do we want ?

- Context relevancy
  - $CR = \frac{\#relevant\ sentences}{\#setences\ in\ context}$

    => May not be the most pertinent metric in this setting

- Positive acceptance / Negative rejection
  - Did the model refrain from responding when it wasn't supposed to?
  - Did the model respond when it wasn't supposed to?

# 3/ Generation evaluation

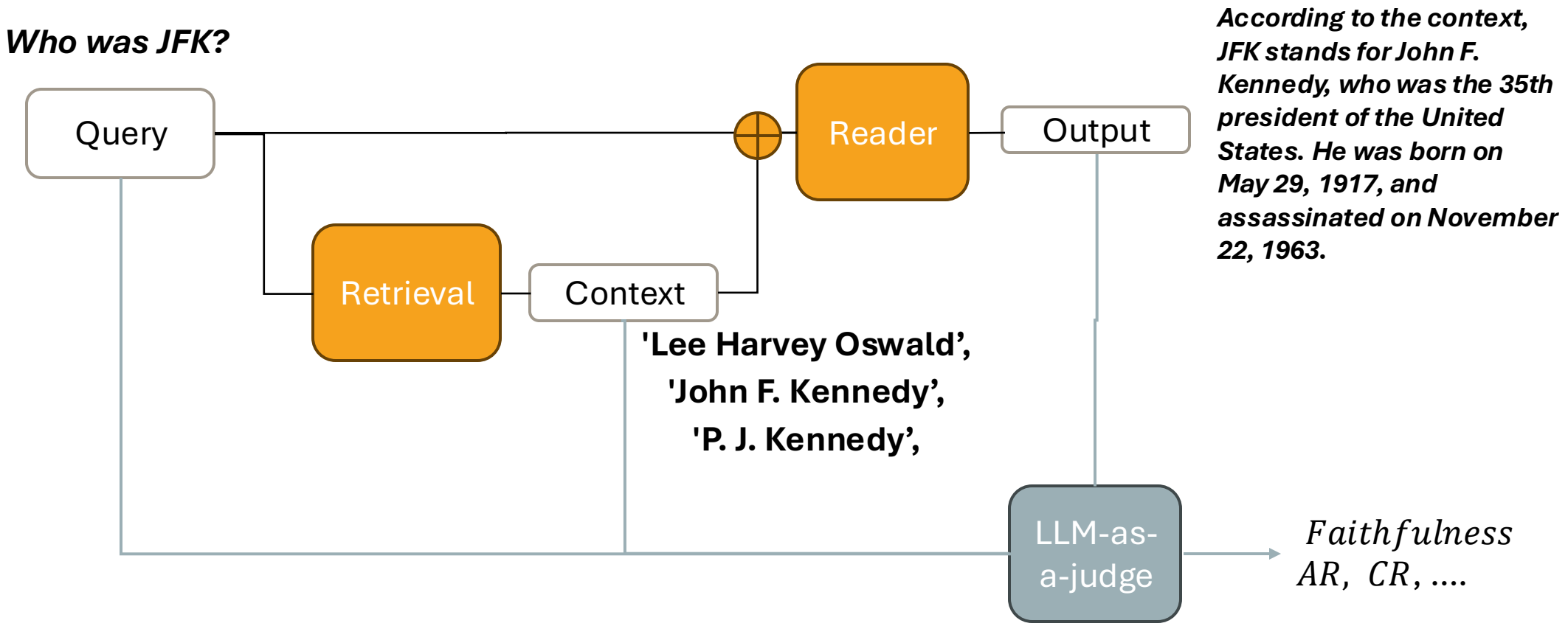
- Faithfulness/Hallucinations
  - Given context, how faithful the statements made by the LLM are.
  - $Faithful = \frac{\#\ true\ statements}{\#statements}$

    => Can be too restrictive « What is type I error in statistics ? » « Null hypothesis ≠ Original hypothesis »

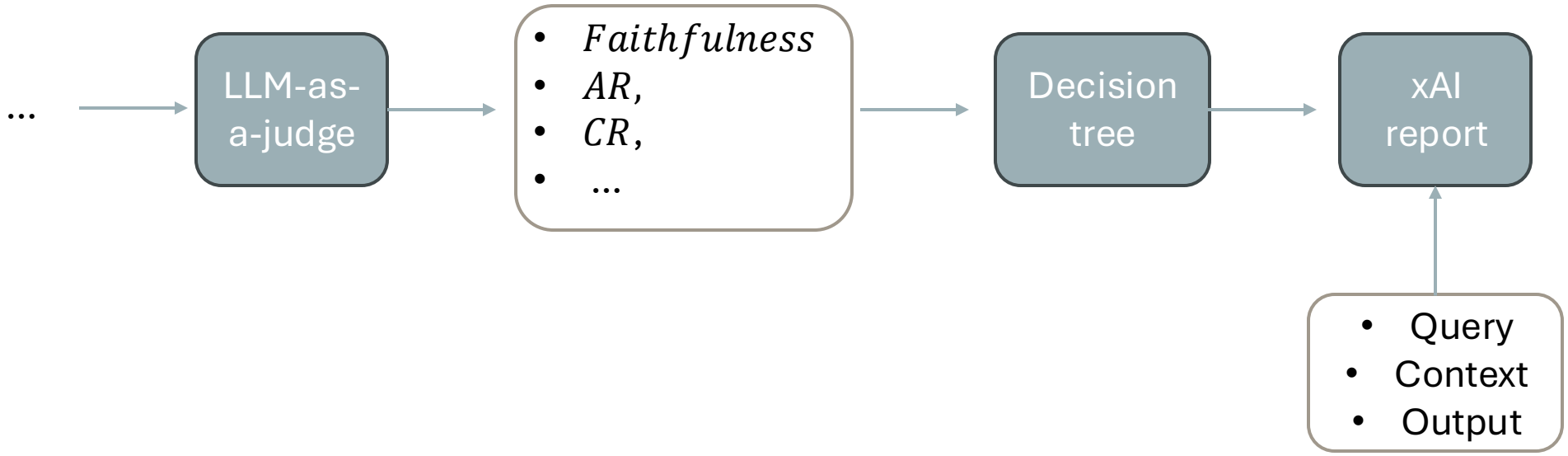
    => Need many statements to be reliable
- Answer relevancy
  - Generate questions $q_i$ based on the provided RAG answer.
  - $AR = \frac{1}{n}\sum_{i=1}^{n} \text{cosine}\_sim(q_i, q)$

    => How to make a decision from it ?

# RAG Pipeline



# Example of an "explainable" RAG Pipeline




- How to build the system of priority? => Build a trainset
- How can we maintain good results from the model while in inference?

# Synthèse Bibliographique

# Bibliographie : Transformers and GPTs

| Title | Problem statement | Methodology | For our project ? |
|---|---|---|---|
| Attention Is All You Need | Prior to the architecture developed in the paper, the model used for NLP were complex recurrent or convolutional neural networks in an encoder-decoder configuration. | Transformer architecture: feed-forward architecture with attention mechanisms, dispensing with recurrence and convolutions entirely. | Transformer architecture used everywhere |
| BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Existing language representation models are limited by their unidirectional context and shallow understanding, leading to inadequate performance in complex NLP tasks.<br>BERT provides a deep bidirectional representation that captures the full context of language, thereby improving the model's ability to understand and generate human-like text. | Combines bidirectional training through masked language modeling and next sentence prediction, followed by fine-tuning on specific tasks, all built upon the powerful Transformer architecture. | Type of transformer. Useful for understanding natural language. |
| GPT-4 Technical Report | Fast models for text generation struggles with limited context. | Transformer decoder-based architecture (GPT) useful for chat completion as it is auto-regressive. | Base of most of the generative transformers nowadays. |
| Latent Retrieval for Weakly Supervised Open Domain Question Answering | Chat generation often lacks contextualized data to answer a user query. | Model proposes an approach to add a retrieval part to be used as context for a text generation model afterwards | RAG model will be cornerstone model of our project |

# Bibliographie : RAG evaluation

| Title | Problem statement | Methodology | For our project ? |
|---|---|---|---|
| RAGAS: Automated Evaluation of Retrieval Augmented Generation | Evaluating RAG architectures is challenging because there are several dimensions to consider: the ability of the retrieval system to identify relevant and focused context passages | Suite of metrics which can be used to evaluate these different dimensions without having to rely on ground truth human annotations | Metrics useful to quantify and evaluate the output of RAGs : how relevant is the context/the answer ? |
| GroUSE: A Benchmark to Evaluate Evaluators in Grounded Question Answering | RAG evaluation methods were incomplete | To assess the calibration and discrimination capabilities of judge models, 7 generator failure modes were identified | More useful metrics to evaluate a RAG model, is the answer complete, should there be an answer?.... |
| GPTScore: Evaluate as You Desire | Need to compare output of a text generation model with ground truth. | Compute the correlation between several outputs on train dataset with human feedback as ground-truth | If we have a trainset to define a ground-truth, we can compare the output of the model with |
| Is ChatGPT a Good NLG Evaluator? A Preliminary Study | Need for a good NLG evaluation metric | LLM-as-Judge, use benchmark language models to evaluate the output of another model | Evaluate the performances of the rag model using |

# Bibliographie : XAI

| Title | Problem statement | Methodology | For our project ? |
|---|---|---|---|
| A Unified Approach to Interpreting Model Predictions | How to measure how changes in a feature impact the output of a regression/classification machine learning model? | SHAP (SHapley A dditive exPlanations).<br>SHAP assigns to each feature an importance value for a particular prediction.<br>Surrogate model that comes on top of the language model. | If we consider simple sentiment analysis cases, we may be able to determine the impact of tokens. |
| TokenSHAP: Interpreting Large Language Models with Monte<br>Carlo Shapley Value Estimation | How to measure how changes in a feature impact the output of a text generation model? | theoretical framework extending Shapley values to variable-length text LLM inputs.<br>An efficient Monte Carlo sampling approach tailored for language models. | For text generation, it can be interesting to detect hallucinations more efficently => We know which tokens affect the output text |
| RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems | comprehensive evaluation of RAG systems remains a challenge due to the lack of unified evaluation criteria and annotated datasets | Introduces RAGBench, a comprehensive dataset designed to evaluate RAG systems, emphasizing the need for unified evaluation criteria and annotated datasets | Can help to have a standard evaluation metric for RAGs |
| RAGE Against the Machine: Retrieval-Augmented LLM Explanations | | Presents RAGE, an interactive tool for explaining LLMs augmented with retrieval capabilities, focusing on counterfactual explanations | RAGE includes pruning methods to navigate the vast space of possible explanations --> producing better generation |