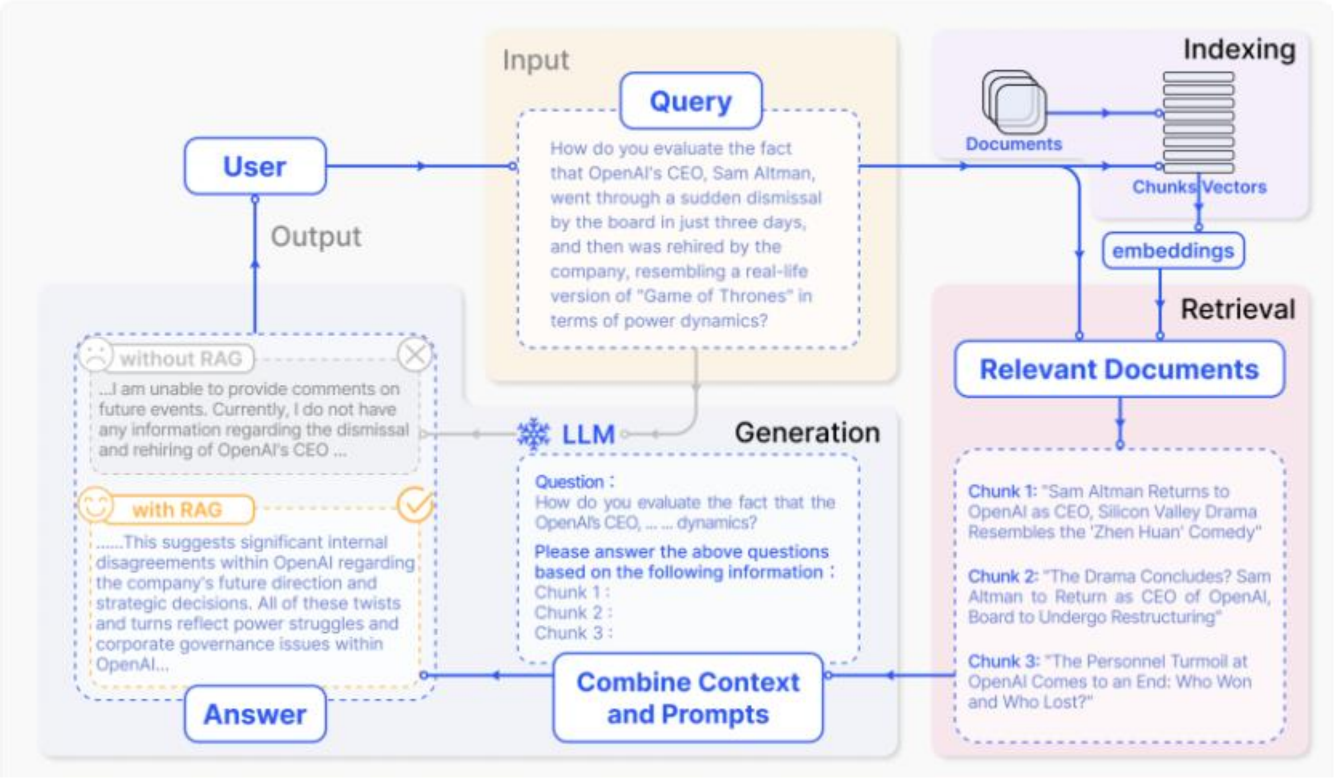
Large language model

PROGRESS REPORT
02/12/2024

RAG models



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

- RAG pipeline :
  - Retrieval evaluation
    - How relevant are globally the choosen documents?
    - Are all documents as relevant?
  - Generation evaluation
    - How relevant the answer is compared to the query?
    - How accurate is the answer given the context?
    - ...

Overview

I/ RAG Evaluation

- 1/ Embedding Evaluation
- 2/ Retrieval Evaluation
- 3/ Generation Evaluation

II/ Methodology for RAG with BlackBox

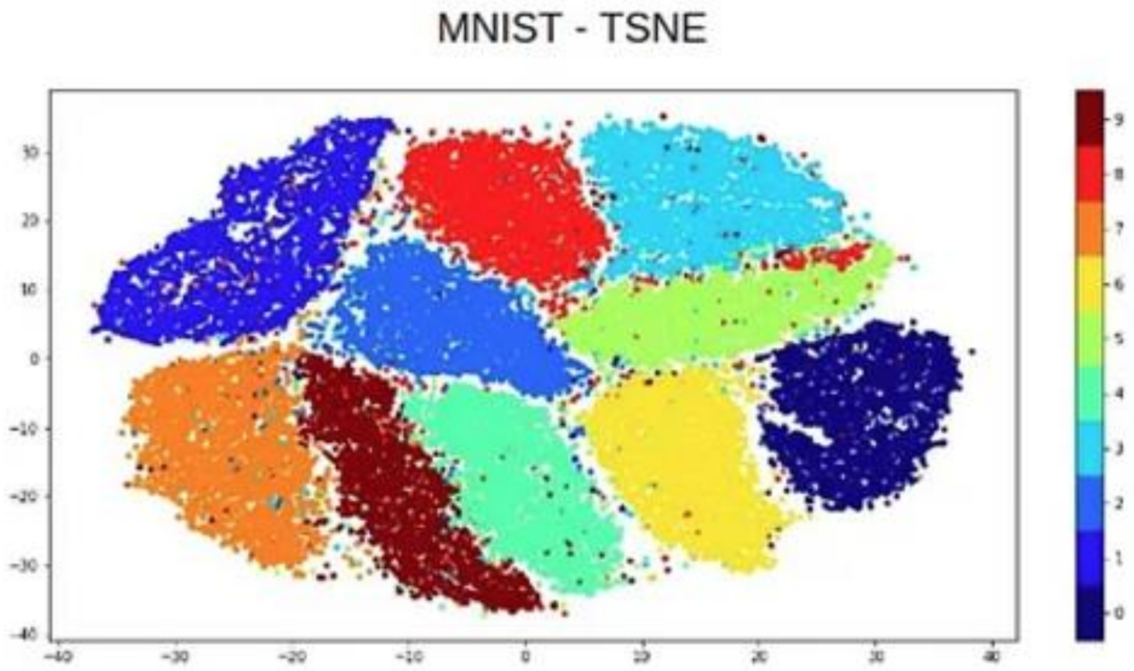
Embedding Evaluation

2/ Task based methods

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

3/ 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

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.

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

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

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

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

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

  - **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.

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

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?

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 ?

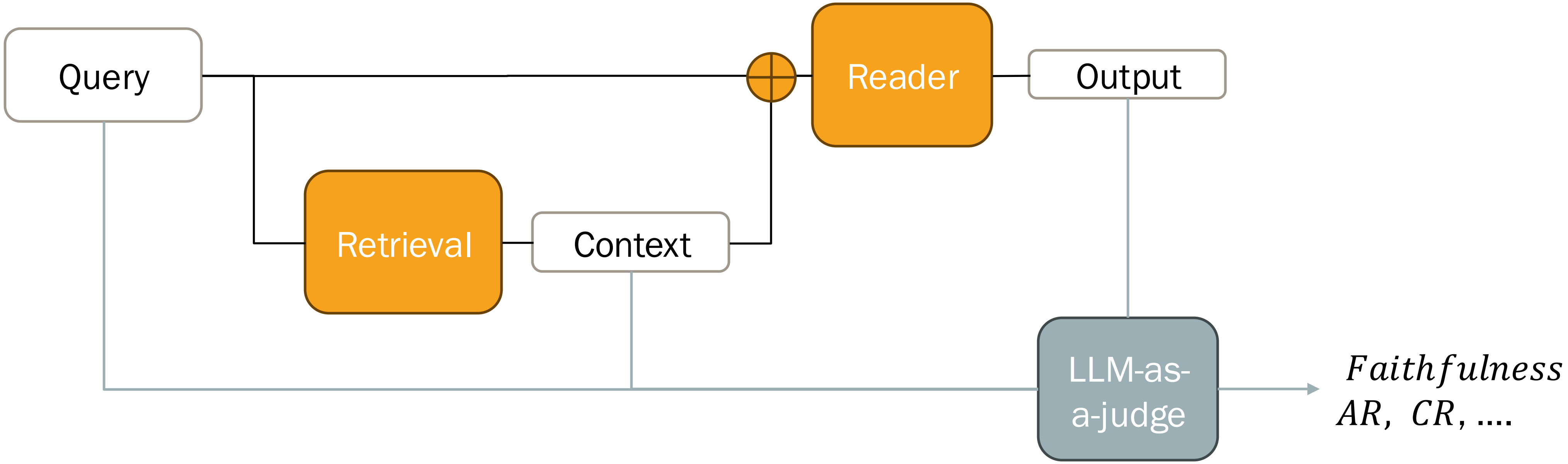
Example

retrieval: (thenlper/gte-small, wikipedia-small),
reader:Qwen/Qwen2.5-7B-Instruct,
judge:meta-llama/Meta-Llama-3-8B-Instruct

Input	Who was JFK?	Context	'Lee Harvey Oswald’, 'John F. Kennedy’, 'P. J. Kennedy’,
Answer	According to the context, JFK stands for John F. Kennedy, who was the 35th president of the United States. He was born on May 29, 1917, and assassinated on November 22, 1963.	Faithfulness	[' JFK stands for John Fitzgerald Kennedy.\n', ' JFK was the 35th president of the United States.’…]
Answer relevancy questions	[' who is jfk?\n', ' what does jfk stand for?\n', ' who was the 35th president of the united states?\n', ' when was jfk born?\n', ' when was jfk assassinated?\n', " what was notable about jfk's death date?"]	CR statements	[' john fitzgerald kennedy (May 29, 1917 – november 22, 1963), often called jfk and jack, was the 35th president of the united states.\n\n’, …]

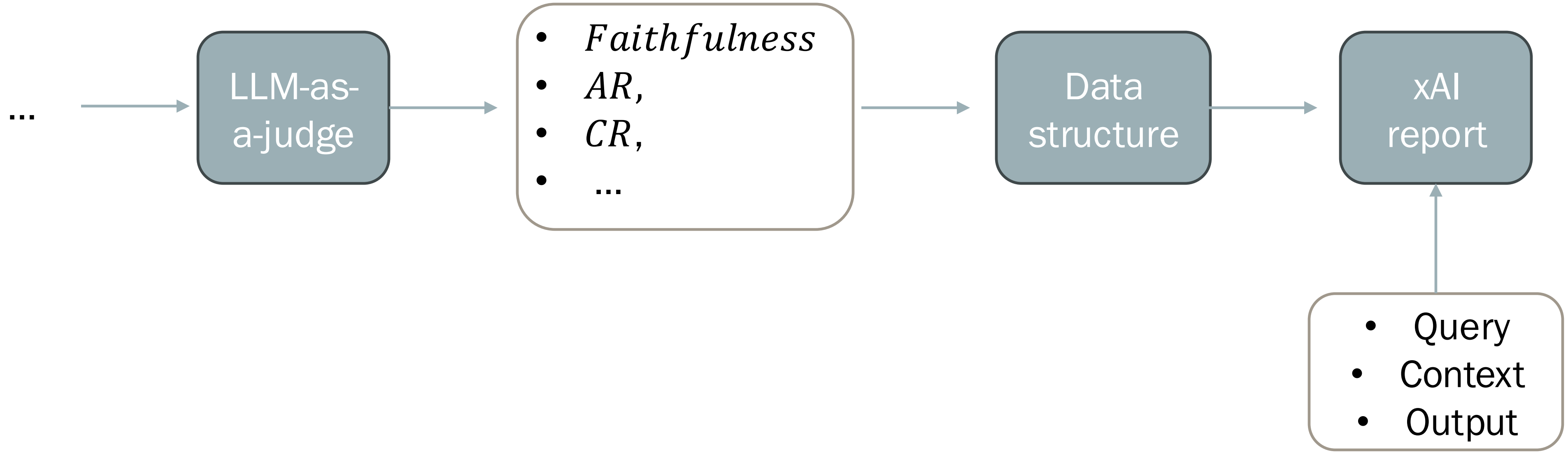
Faithfulness	AR	CR	Completeness
1	0.763323982556661	0.5555555555555556	0.4

Pipeline



Query
Retrieval
Context
Reader
Output
LLM-as-a-judge
$Faithfulness$
$AR,\ CR, \ldots.$

Pipeline



…
LLM-as-a-judge
• $Faithfulness$
• AR,
• CR,
• …
Data structure
xAI report
• Query
• Context
• Output

Next steps

- How to use evaluation metrics for explainability?

- Which evaluation metrics are relevant for our case?

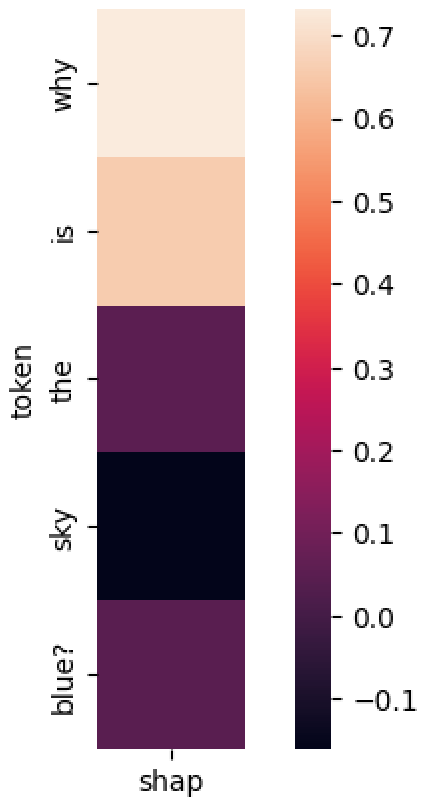
Appendix

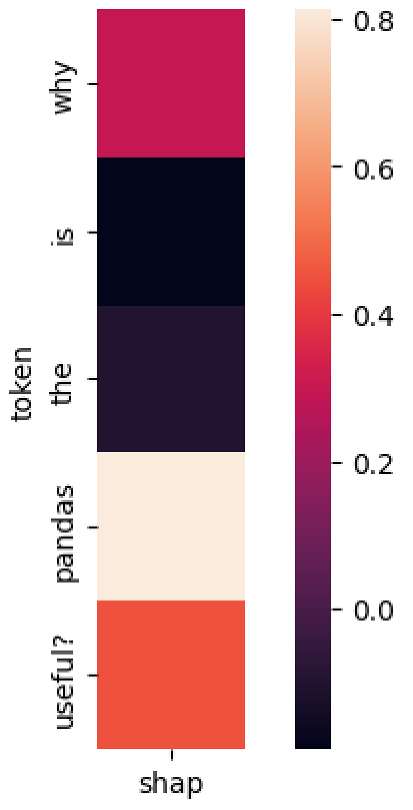
TokenSHAP [arXiv:2407.10114]

- For $tokens = (x_1, \dots, x_n)$ compute the baseline output b from LLM model

- Compute output for randomly sampled tokens b_C in $tokens$ and compare both methods $v_C = cosine\_sim(b_C, b)$
  - For each x_i average each v_C in which x_i is and do the same for each v_C in which x_i is not

- $SHAP_i = with_i - without_i$





Example of Methodology for RAG with BlackBox

1/ Retrieval dataset

2/ Embedding: exp Sentence-BERT to encode documents

3/ Retrieval part: FAISS for efficient similarity search –> partition + approximate NN

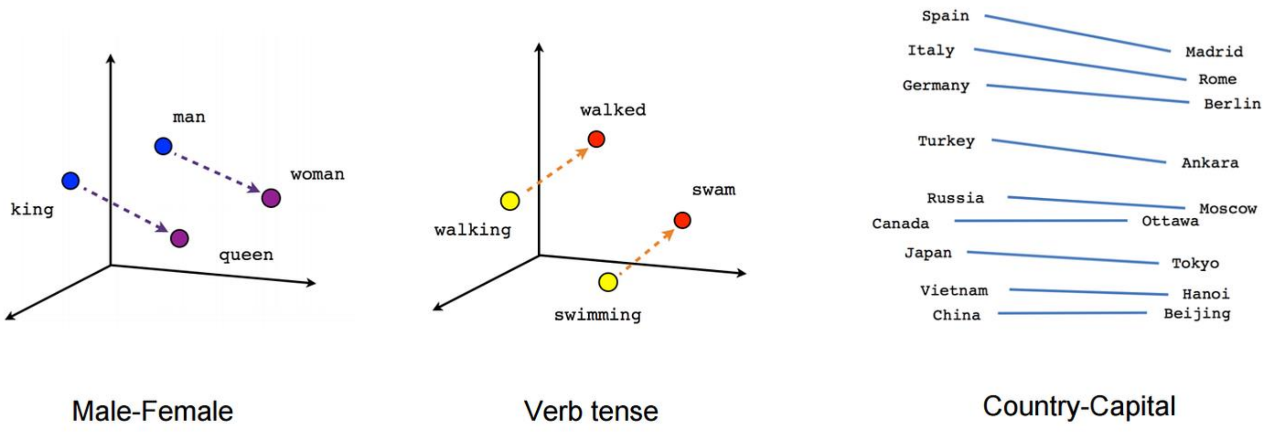
4/ Generation part: exp T5 model from huggingface

5/Evaluation: LLM-as-a-judge

Embedding Evaluation

1/ Intrinsec evaluation

- Neighborhood Analysis
  - Vérifier si les prompts similaires ont des embeddings proches dans l'espace vectoriel --> Cosine similarity
- Analogy
  - Résoudre des analogies comme *"king - man + woman ≈ queen"* pour les embeddings de mots utilisés dans les prompts



Word2Vec — Analogical Reasoning

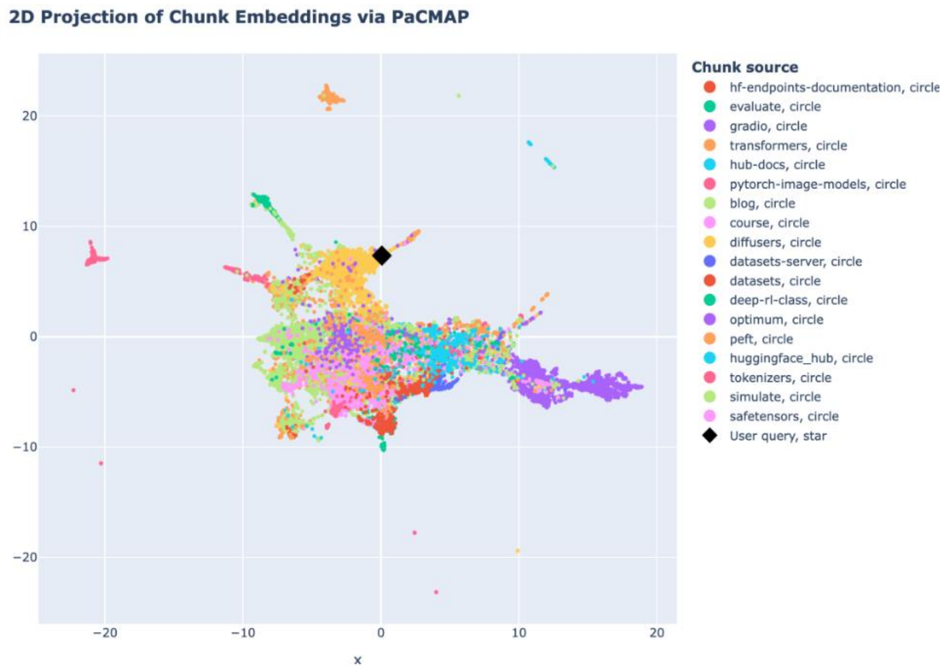
Mathematical Proof of Analogical Reasoning in Word2Vec Embedding

Sanjay Chouhan

Other XAI techniques

Representation analysis

- UMAP, machine learning embeddings



Bibliography

1. Pinecone, RAG Evaluation: Don’t let customers tell you first, https://www.pinecone.io/learn/series/vector-databases-in-production-for-busy-engineers/rag-evaluation/

2. HuggingFace, RAG Evaluation