Large language model

PROGRESS REPORT
03/04/2025

Overview

1. Retrieval study
2. Ablation testing
3. Extracting Rational from Context
4. Full explainability pipeline overview
5. Explainability
6. Completeness

I. Retrieval study

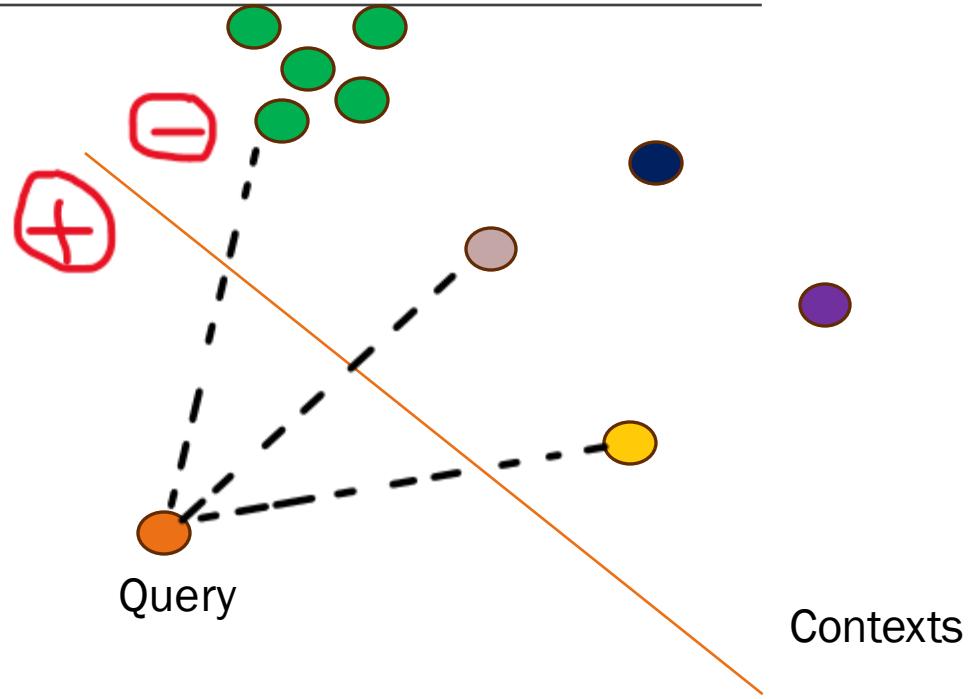
Retrieval study

- Previously we used just cosine

  –> no ensurance about diversity

- Solution: implementing SVM to satisfy diversity

- Least distance to decision boundary can satisty
  - Similarity
  - Diversity

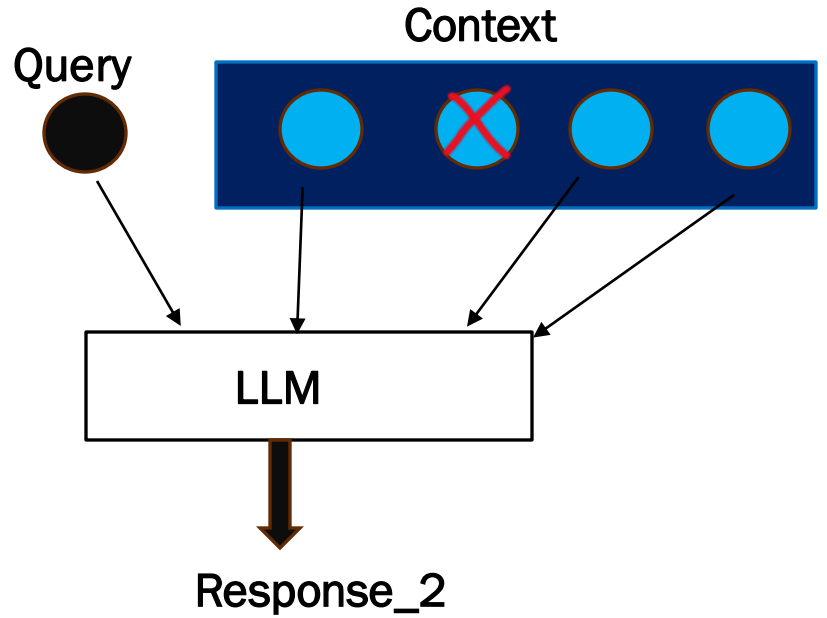


II. Ablation testing

Ablation testing

- Ablation testing = Context sensitivity via perturbation
- See the impact of a missing chunk from the context on the generated response
- Algorithm :
  - Generate response based on query + full context –> response_0
  - At each time,
    - remove one chunk from the context
    - Generate response based on query + remaining chunks in context
    - Calculate similarity between response_0 and response_i

    –> lowest similarity correspond to most impactful chunk



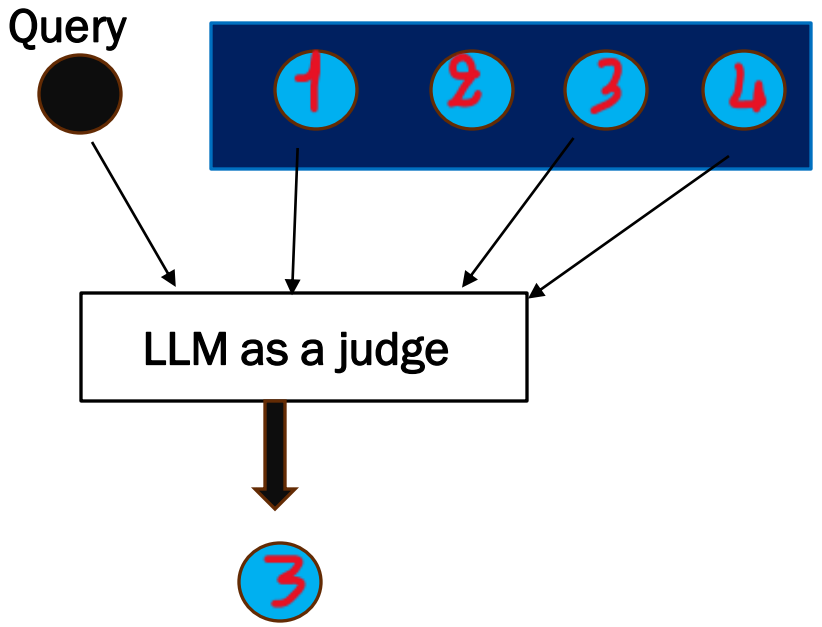
III. Rational extraction

Rational extraction

- Input : query, context, answer
- Ask an LLM as a judge to extract most impactful chunks from context for answer generation

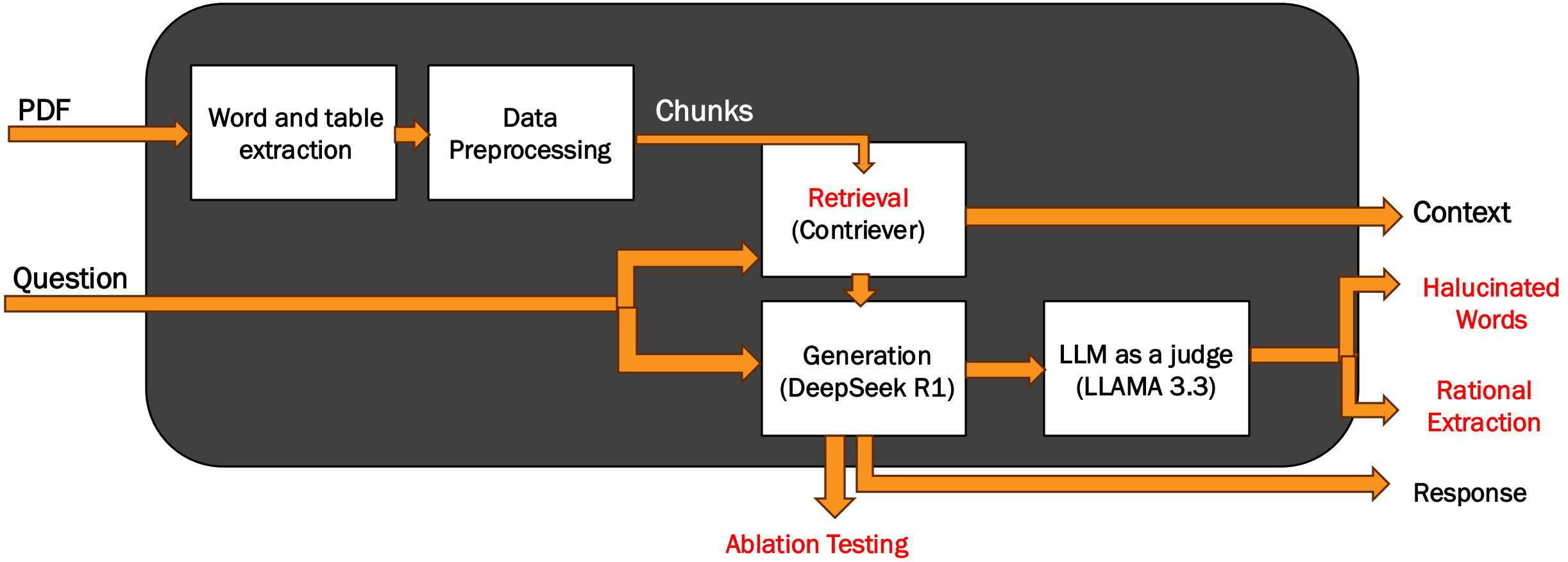
  –> Ensure sourcing of answer

  –> Give location of the information to answer the question



IV. Full Explanability pipeline

Production System



Question	Context		Answer	Ablation testing	Halucinated words	Rartional Extraction
What is the impact of rising interest rates on corporate investments and real estate?	• 'Asset: Bitcoin | 2023 Start Price: $16,500 | 2023 End Price: $28,400 | % Change: 72.1%' • 'Asset: Ethereum | 2023 Start Price: $1,200 | 2023 End Price: $1,900 | % Change: 58.3%' • 'Market Trends The S&P 500 and NASDAQ indices have shown mixed performance, with technology stocks leading gains while energy and traditional industries lagged.' • 'Index: S&P 500 | 2023 Start: 3,800 | 2023 End: 4,250 | % Change: 11.8%' • 'Table 1: S&P 500 and NASDAQ Performance Rising interest rates have affected borrowing costs, influencing real estate and corporate investments.'	• -0.119 • -0.29 • -0.306 • -0.46 • -0.58	Rising interest rates have increased borrowing costs, leading to reduced corporate investments and slower growth in the real estate sector. This has been evident as the S&P 500 and NASDAQ have shown mixed performance, with technology stocks outperforming while energy and traditional industries lag. The S&P 500 rose 11.8% and NASDAQ gained 72.1% in 2023.	• 0.7374 • 0.6583 • 0.8115 • 0.6279 • 0.578	['NASDAQ gained 72.1% in 2023.']	- Rising interest rates have affected borrowing costs, influencing real estate and corporate investments. - Index: S&P 500 | 2023 Start: 3,800 | 2023 End: 4,250 | % Change

Question	Context		Answer	Ablation testing	Halucinated words	Rartional Extraction
'How did the S&P 500 and NASDAQ indices perform in 2023?	• 'Asset: Bitcoin | 2023 Start Price: $16,500 | 2023 End Price: $28,400 | % Change: 72.1%' • Asset: Ethereum | 2023 Start Price: $1,200 | 2023 End Price: $1,900 | % Change: 58.3%' • 'Market Trends The S&P 500 and NASDAQ indices have shown mixed performance, with technology stocks leading gains while energy and traditional industries lagged.' • 'Index: S&P 500 | 2023 Start: 3,800 | 2023 End: 4,250 | % Change: 11.8%' • 'Table 2: Cryptocurrency Performance in 2023 Cryptocurrencies and blockchain-based assets have gained traction	• -0.022 • -0.209 • -0.224 • -0.35 • -0.53	The S&P 500 rose by 11.8% in 2023, while the NASDAQ also saw gains.	• 0.8070 • 0.8592 • 0.7830 • 0.6542 • 0.8726	the NASDAQ also saw gains	- Index: S&P 500 | 2023 Start: 3,800 | 2023 End: 4,250 | % Change: 11.8%

V. Explainability

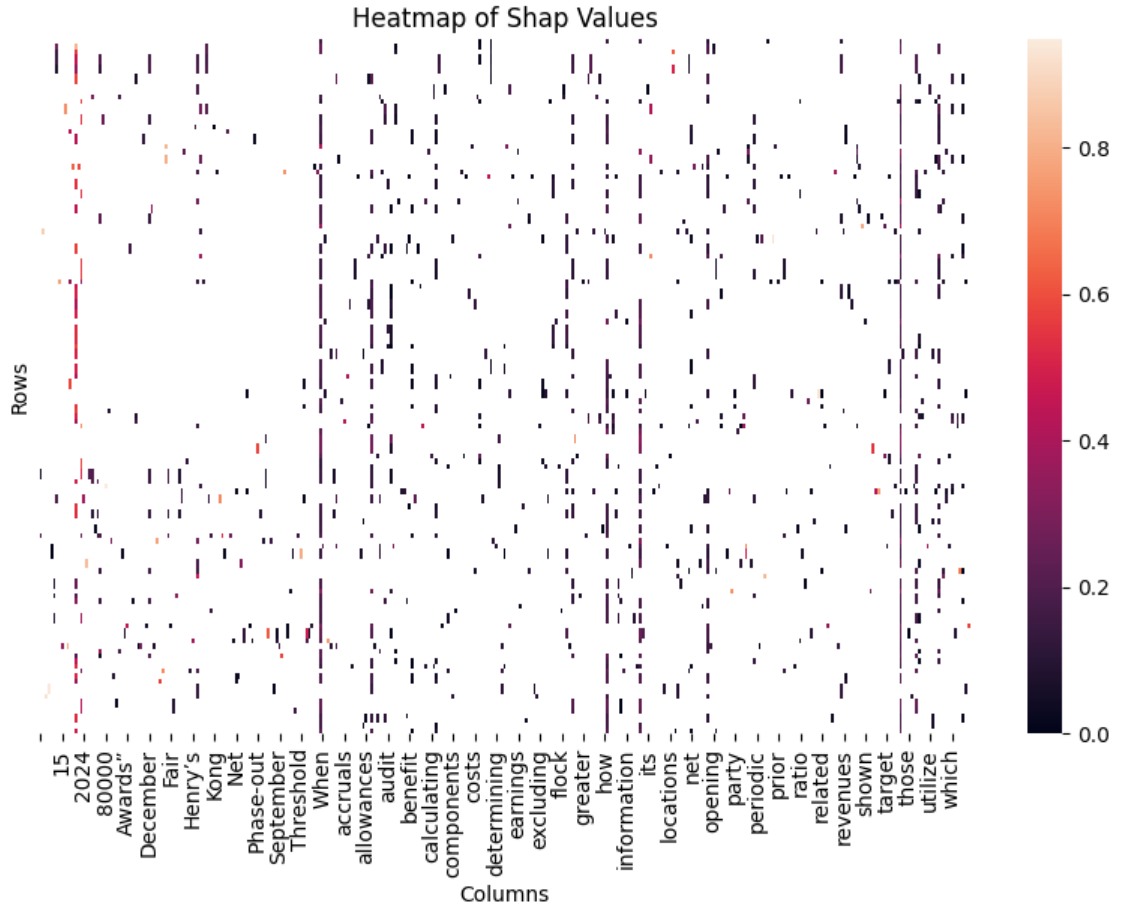
Token attribution

- Get hallucinated $(prompt, context)$
- Build contextualized $prompt_{RAG} = (prompt, context)$

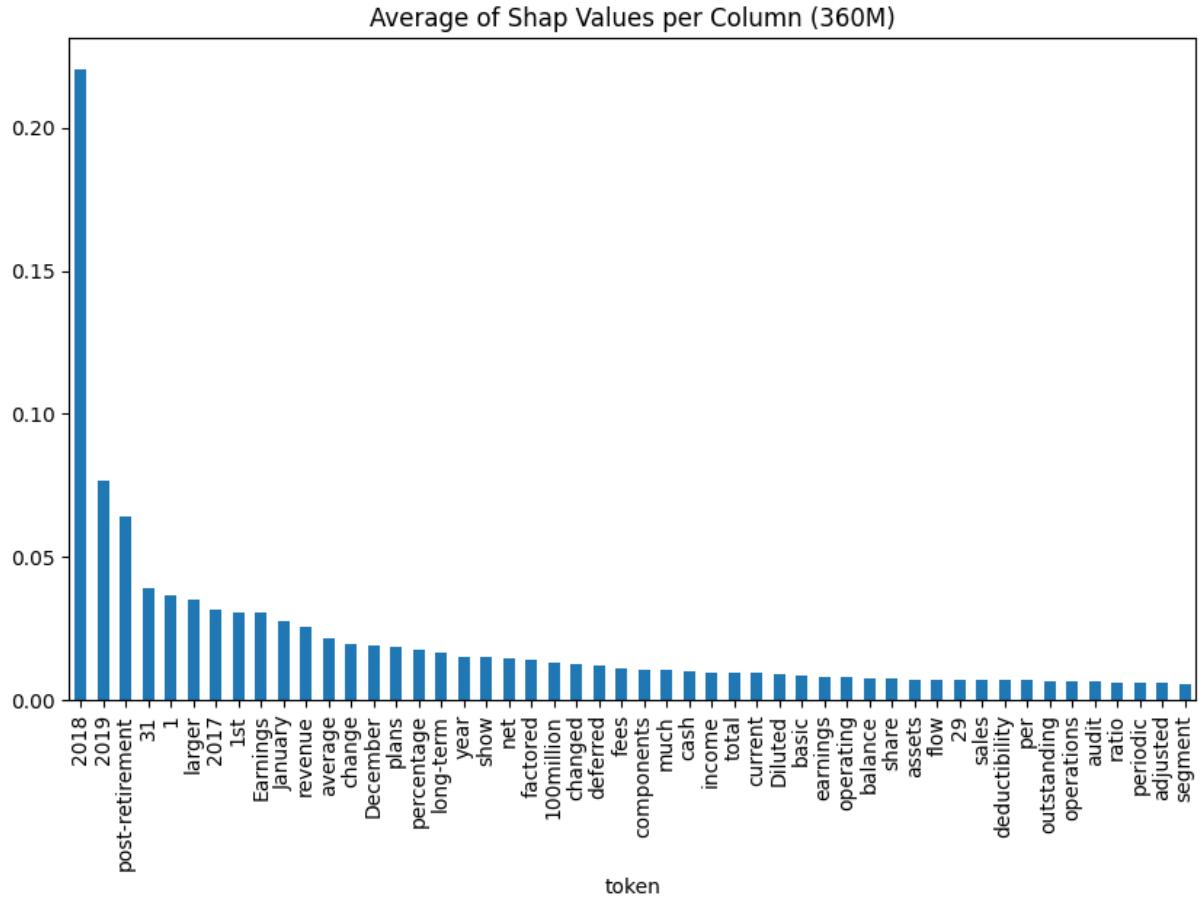
- $[token_1, \dots, token_k] = prompt_{RAG}$

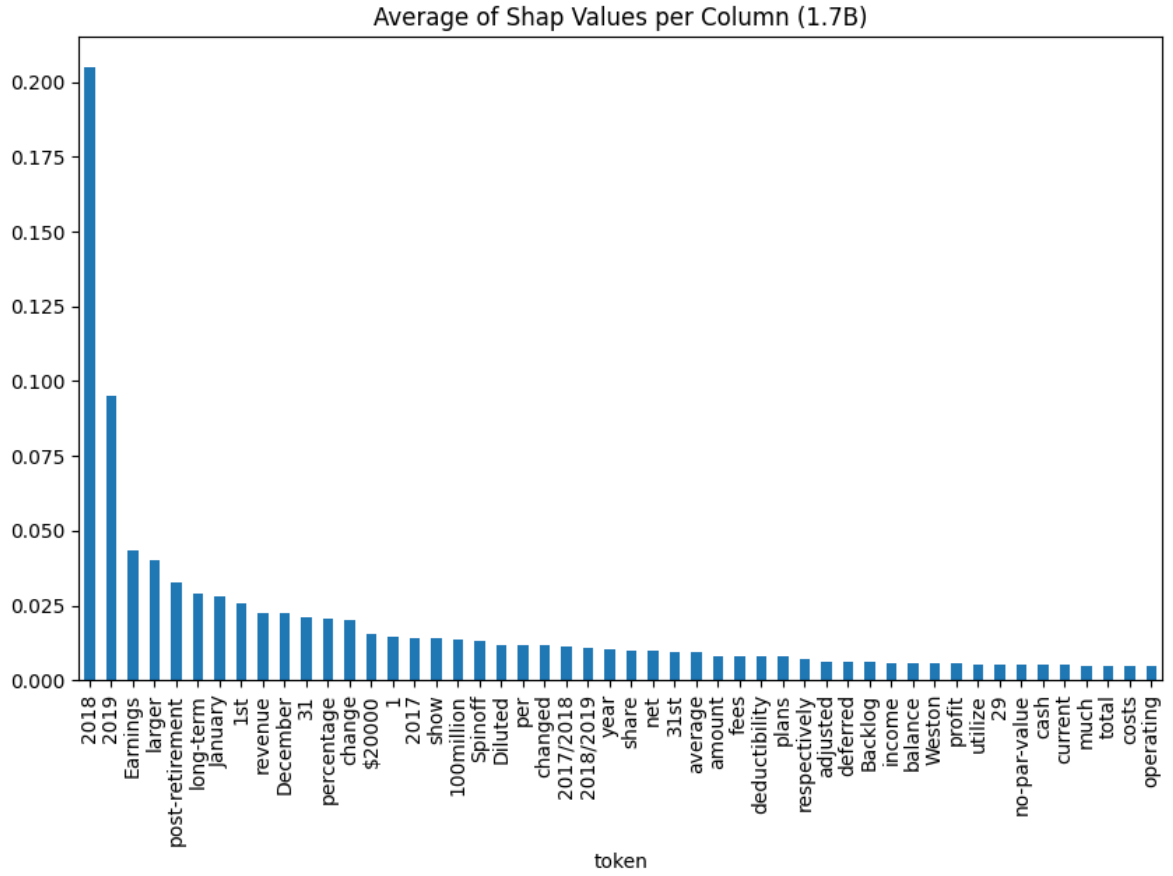
- Evaluate a reference answer using a language model m : $embed_{ref} = m(prompt_{RAG})$
- For $token_i$ in $prompt_{RAG}$:
  - Evaluate $embed_{-i} = m(prompt_{RAG} - token_i)$
  - Compute the cosine similarity $sim_i = cosine\ sim(embed_{-i}, embed_{ref})$
  - Return $Shap_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} sim_j - sim_i$

Hallucination results



Aggregating SHAP values





VI. Completeness

Completeness study for TAT-QA

- Globally the models seem to match despite the sampling being imbalanced

- Exception for some quantitative questions