Document Attribution in Retrieval-Augmented Generation

Ikhtiyor Nematov [1,2], Tarik Kalai [1], Elizaveta Kuzmenko [1], Gabriele Fugagnoli [1,4], Dimitris Sacharidis [1], Katja Hose [3,2], Tomer Sagi [2]

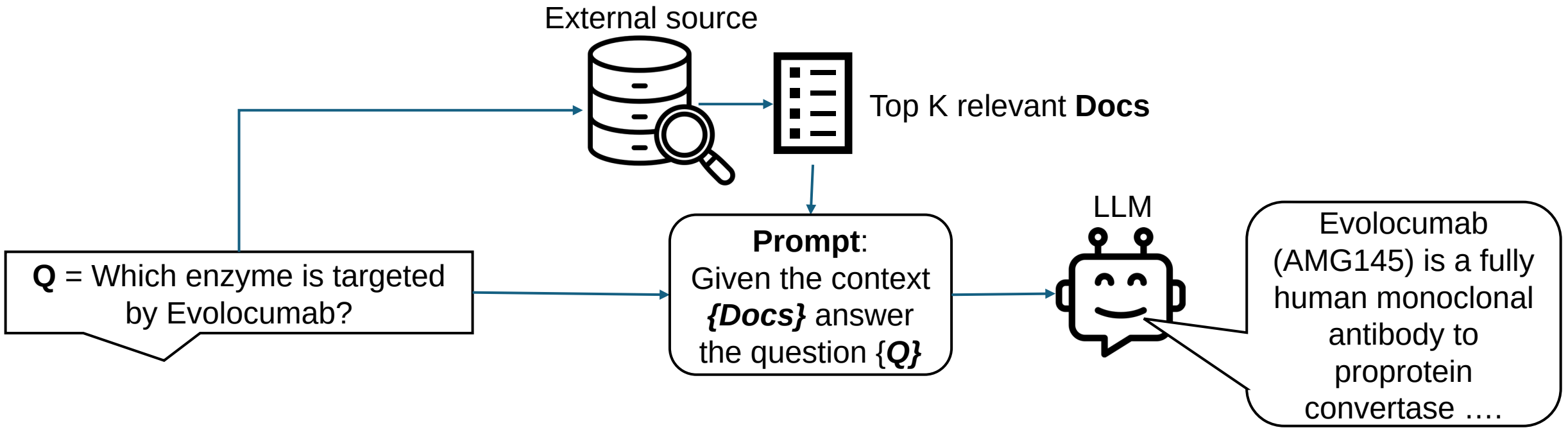
1. Université Libre de Bruxelles, Belgium
2. Aalborg University, Denmark
3. TU Wien, Austria
4. University of Padova, Italy

Outline

- Introduction
- Objectives
- Methodology
- Experiments
- Conclusion

Introduction

- LLMs are ubiquitous though they are hardly ever used on their own
- Retrieval Augmented Generated (RAG) grounds LLMs in external knowledge, reducing hallucinations and improving factuality.
- Feature/data attribution is extensively studied for traditional ML



But which documents if any have actually influenced the final answer?

Shapley Values [1]

- a principled, game-theoretic approach for fair attribution of a payoff
- For a set of players **D** and a characteristic utility function **v(S)** that defines the worth of any subset (coalition) **S ⊆ D,** Shapley value for j is defined as:

- utility in traditional ML can be loss, or any output of the model
- requires $2^{|D|}$ utility computations
- many approximation methods exist

$$\phi_j(v) = \sum_{S \subseteq D \setminus \{j\}} \frac{|S|!(|D| - |S| - 1)!}{|D|!} [v(S \cup \{j\}) - v(S)]$$

Objectives

- Applying Shapley values in RAG scenario with a tailored utility function
- quantifying how well approximations can mirror exact attributions while minimizing costly LLM calls
- evaluate their practical explainability in identifying critical documents, especially under complex inter-document relationships such as redundancy, complementarity, and synergy that are common is RAG.

Methodology

- each document is a separate player
- attribution must be w.r.t. the **target response** (payoff), generated by the full coalition D
- **Utility – log likelihood** of generating the **target response** given the coalition S
- It was used as a evaluation metric for LLM [3]
- Teacher-forcing probability product

$$v(S) = \sum_{t=1}^{|R_{\text{target}}|} \log P(\text{token}_t(R_{\text{target}}) | \text{token}_{<t}(R_{\text{target}}), Q, S)$$

Research Questions

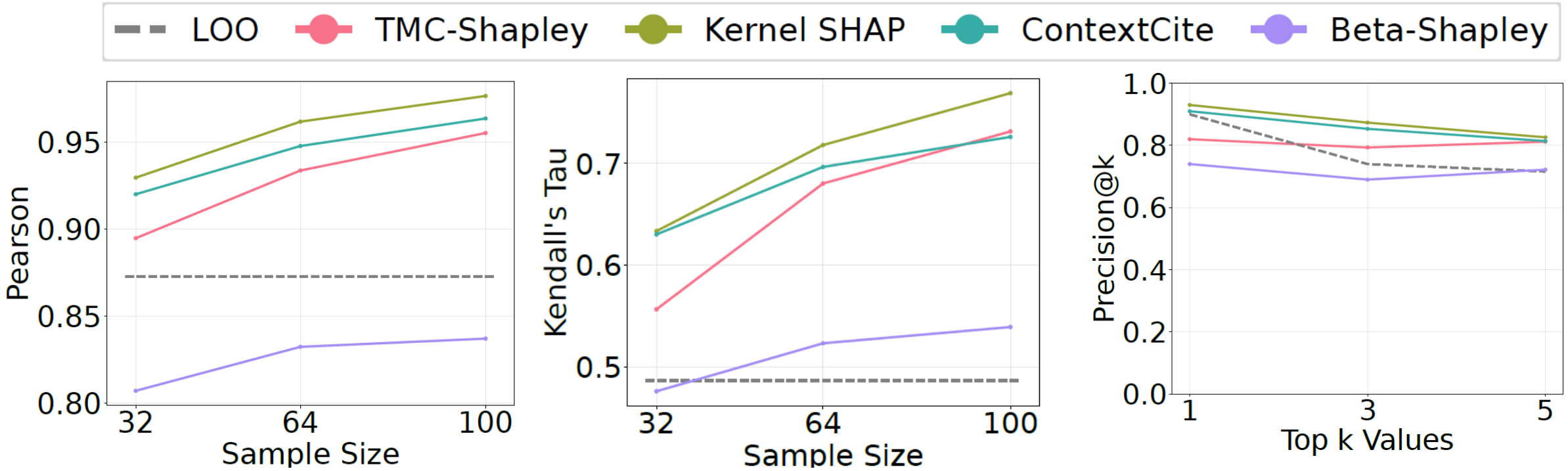
- **RQ1: Replication Quality**
  - How accurately can computationally cheaper methods (like Kernel SHAP, TMC-Shapley) replicate the "gold standard" exact Shapley scores?
- **RQ2: Attribution Effectiveness**
  - How effective are these methods at identifying the *k* documents whose removal causes the largest drop in utility?
- **RQ3: Robustness to Inter-document Relationships**
  - How do these methods perform in complex but common scenarios like Redundancy, Complementarity, and Synergy?

Experimental Setup

- **Attribution Methods:**
  - **Exact Shapley** (Our expensive ground truth)
  - **Approximations:** Kernel SHAP [2], ContextCite [3] (surrogate models)
  - **Sampling-based:** TMC-Shapley [4], Beta-Shapley [5]
  - **Naïve Baseline:** Leave-One-Out (LOO)
- **LLMs:** Mistral-7B, Llama-3.2-8B, Qwen-3B
- **Datasets:**
  - **Real-world QA:** NQ, BioASQ
  - **Synthetic Datasets:** Custom-built to test Redundancy, Complementarity, and Synergy in a controlled way.

RQ1: Replication Quality

- Compare rankings generated by all methods to the ranking of Shapley values
- Kernel SHAP and ContextCite closely mirror the true Shapley rankings
- TMC and Beta Shapley need more samples



RQ2: Attribution Effectiveness

- Precision@k to the exhaustive top k that cause the highest utility drop when removed
- Decent results in general, though not optimal
- Why?
- Inter-document interactions

	Mistral 7B			
k	2	3	4	5
Shapley	**0.80**	**0.77**	**0.78**	**0.78**
TMC-Shapley 32	0.69	0.67	0.70	0.70
Beta Shapley 32	0.61	0.59	0.61	0.66
Kernel SHAP 32	0.75	0.69	0.72	0.73
ContextCite 32	0.73	0.68	0.68	0.68
TMC-Shapley 64	0.70	0.71	0.73	0.72
Beta Shapley 64	0.66	0.60	0.63	0.65
Kernel SHAP 64	0.77	0.75	0.76	0.75
ContextCite 64	**0.80**	0.75	0.73	0.74
TMC-Shapley 100	0.76	0.71	0.74	0.73
Beta Shapley 100	0.61	0.61	0.64	0.66
Kernel SHAP 100	0.79	0.73	0.76	0.76
ContextCite 100	**0.80**	**0.77**	0.76	0.77
LOO	0.57	0.57	0.57	0.61

Inter-document relationship

- **Redundancy**
  - ***Question**: What is the weather in Suvsambil?*
  - ***Documents**:*
    1. *The weather in Suvsambil is sunny*
    2. *Suvsambil is a mountainous country*
    3. *The sun is shining in Suvsambil today*
- **Complementarity**

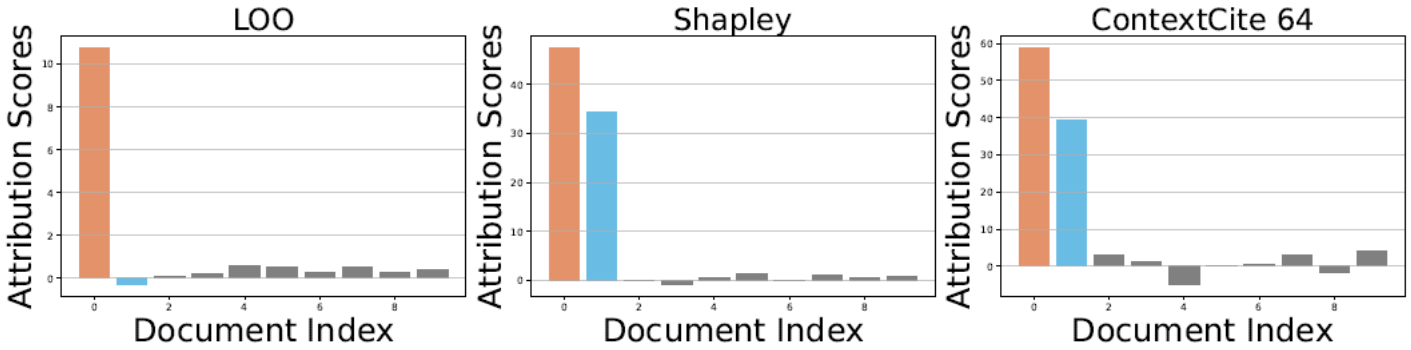
  ***Question**: What are the roles or professions of Elara Vayne and JaxKorden?*

  ***Documents**:*
    1. *Elara Vayne is the chief Star-Navigator of the starship ‘Wanderer’*
    2. *Many young Squibs dream of joining the Sky Guard.*
    3. *Jax Korden serves as the primary Rift-Warden protecting theChronos Gate.*
- **Synergy (Multi-Hop)**
  - ***Question**: What is the weather in the capital of Suvsambil?*
  - ***Documents**:*
    1. *The capital of Suvsambil is Savrak.*
    2. *Weather in Tentak is cloudy.*
    3. *Weather in Savrak is sunny.*

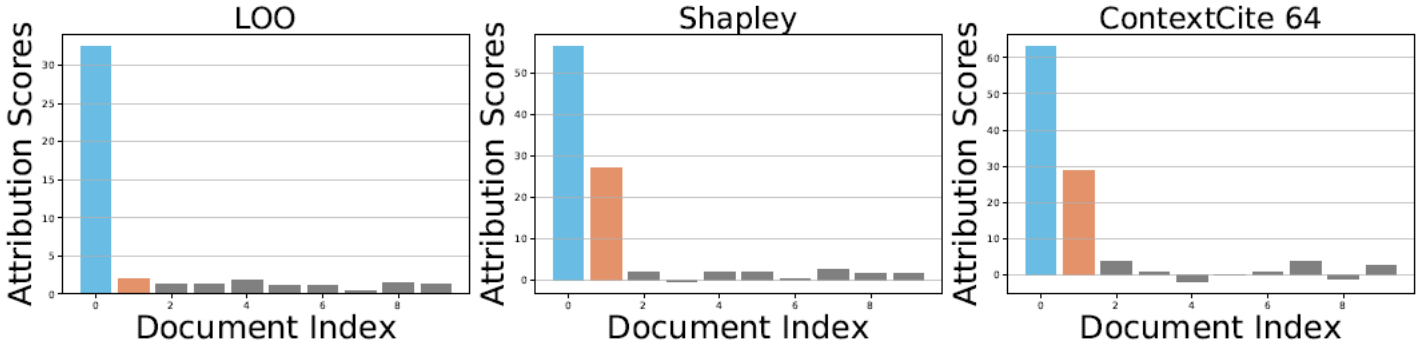
Redundancy

Question	What is the traditional greeting in Blimpton?
Positive A (Orange)	The traditional greeting in Blimpton involves touching elbows while saying "Flurp be with you".
Positive B (Blue)	Blimptonian etiquette requires the elbow-touch greeting, accompanied by the standard phrase "Flurp be with you".
Negative sample	All Blimptonian vehicles hover at least 10cm above ground.
Negative sample	Blimptonian water freezes at 50°C due to added minerals.

LLM answer (AB): The traditional greeting in Blimpton is touching elbows while saying "Flurp be with you".



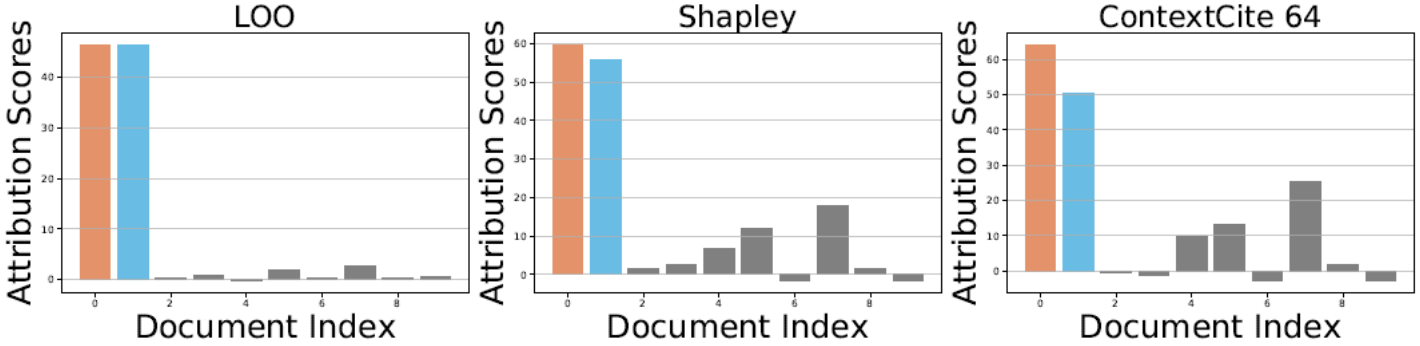
LLM answer (BA): The traditional greeting in Blimpton is the elbow-touch greeting, accompanied by the standard phrase "Flurp be with you".



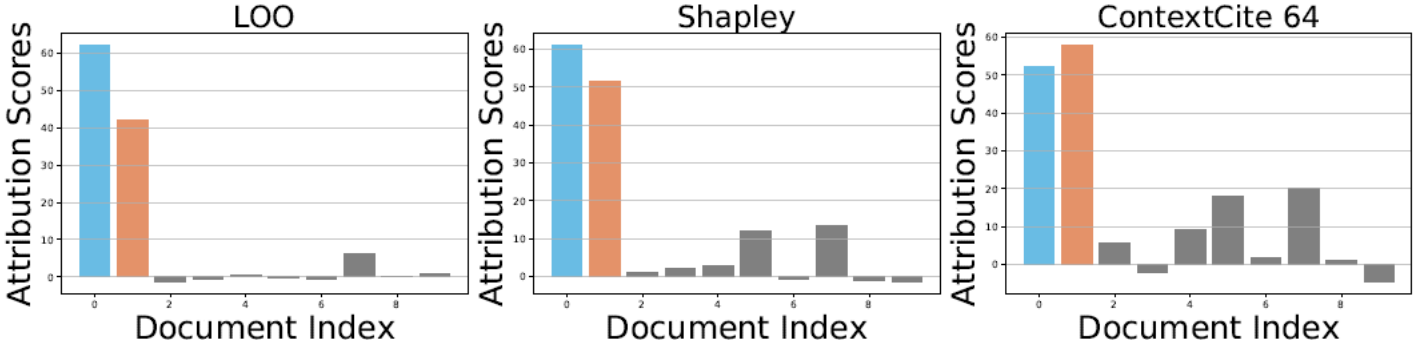
Complementarit

Question	What two functions does the Mystic "Dream-Weaver's Loom" perform?
Positive A (Orange)	The Mystic "Dream-Weaver's Loom" can capture and solidify "Sleep-Visions" into physical dream-silk tapestries.
Positive B (Blue)	It also has the ability to "Memory-Stitch", embedding specific recollections directly into the fabric.
Negative sample	Mystic looms are powered by lunar energy.
Negative sample	Memory-Stitching requires deep meditative states

LLM answer (AB): Mystic "Dream-Weaver's Loom" performs two functions.



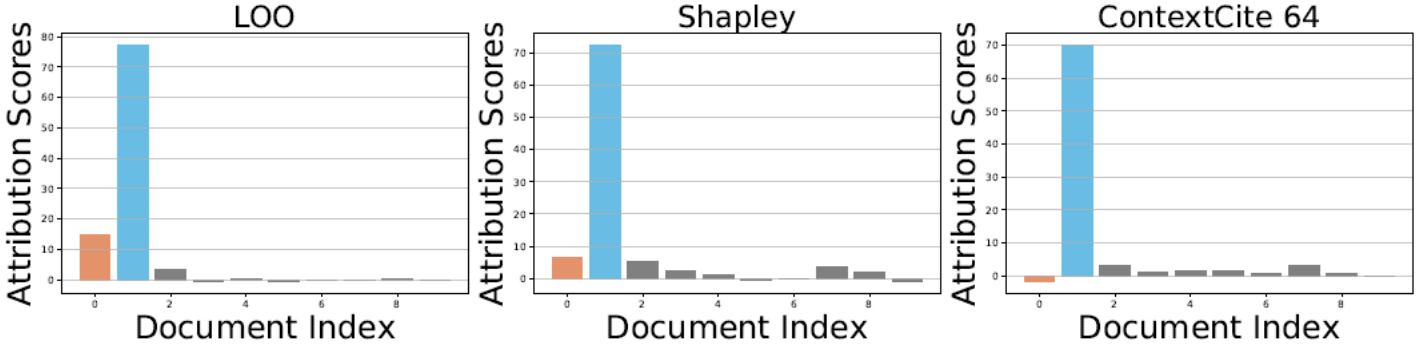
LLM answer (BA): Mystic "Dream-Weaver's Loom" performs two functions.



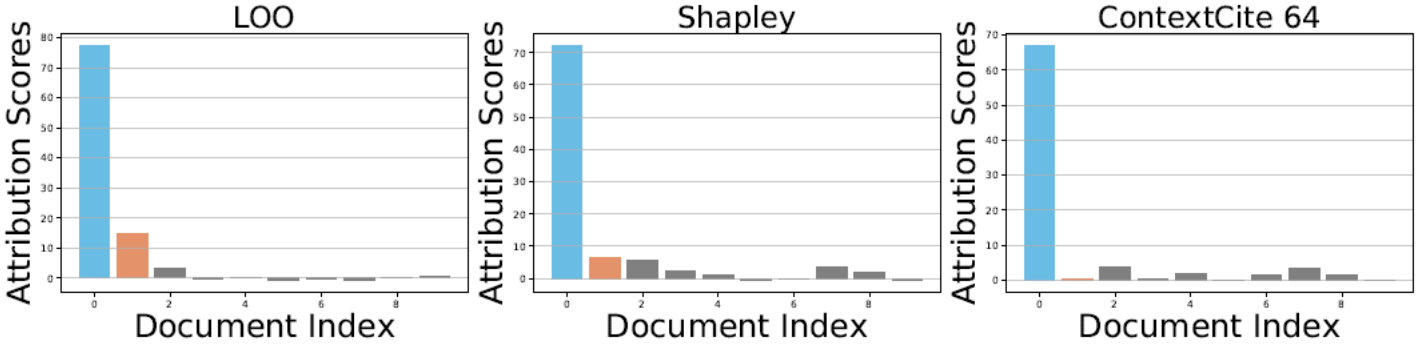
Synergy

Question	What is the salary of the most popular actor on the planet Aethelon?
Positive A (Orange)	Lyra Vael is widely considered the most popular actor currently working in Aethelon's film and stage sectors.
Positive B (Blue)	Lyra Vael commands a salary of approximately 50 million Credits per major project, making her one of the highest earners.
Negative sample	Aethelon's entertainment industry is renowned for its emotionally resonant dramas and intricate historical epics.
Negative sample	Actors on Aethelon undergo rigorous psychological training to fully embody complex characters.

LLM answer (AB): The most popular actor on the planet Aethelon, Lyra Vael, commands a salary of approximately 50 million Credits per major project.

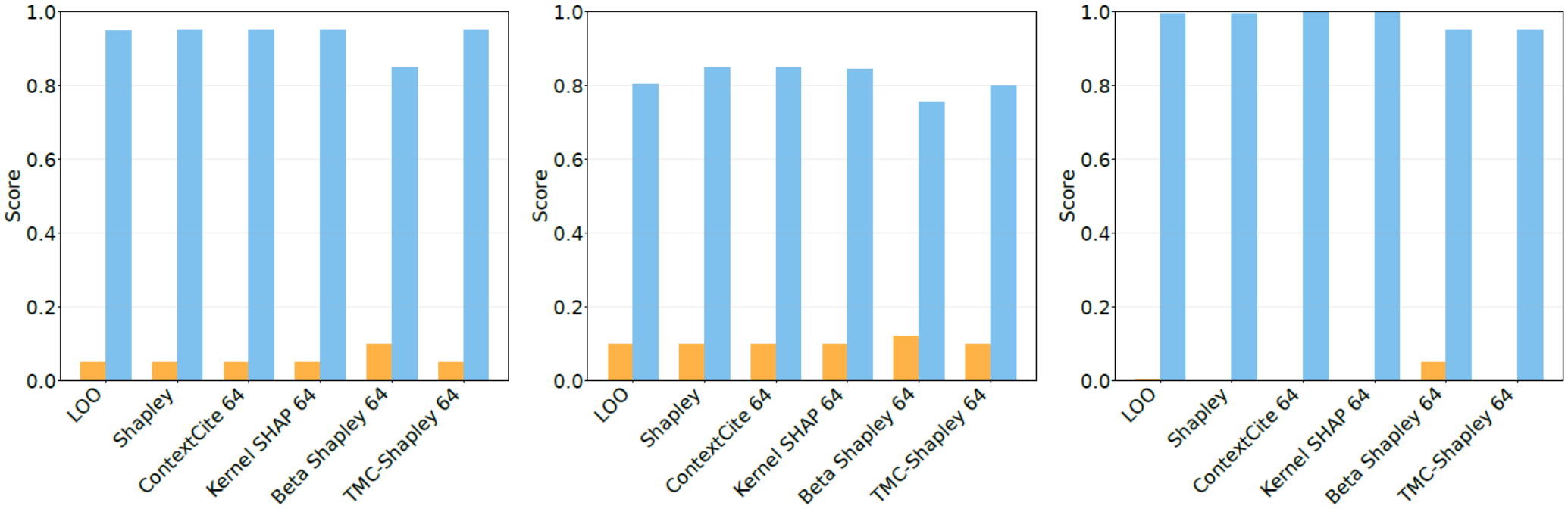


LLM answer (BA): The most popular actor on the planet Aethelon, Lyra Vael commands a salary of approximately 50 million Credits per major project.



Average normalized attribution scores for synergetic documents A and B, tested on Qwen-3B-Instruct, Mistral-7B-Instruct, and LLaMA-3.2-8BInstruct, respectively.

Methods **under-value "synthesis" documents**.



Conclusions

- applying Shapley-based attribution methods to RAG with a tailored utility function shows promise.
- methods that account for deeper inter-document relationships are needed
- attribution scores alone may be insufficient:
  - interpreting the interactions among highly attributed documents is also essential for a more complete understanding to improve attribution quality
- Shapley Interactions could be a way forward

References

1. Shapley, L. S. A value for n-person games. Contributions to the Theory of Games, 2(28):307–317, 1953

2. A Unified Approach to Interpreting Model Predictions

3. ContextCite: Attributing Model Generation to Context

4. Data Shapley: Equitable Valuation of Data for Machine Learning

5. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning