Reputation-Aware Truth Discovery for Robust Evidence Fusion in RAG Pipelines

Diogo Ribeiro and João Marco Silva
INESC TEC & University of Minho

Workshop on Adaptive Smart areaS and Intelligent Agents (ASSIA)
in
24th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS)

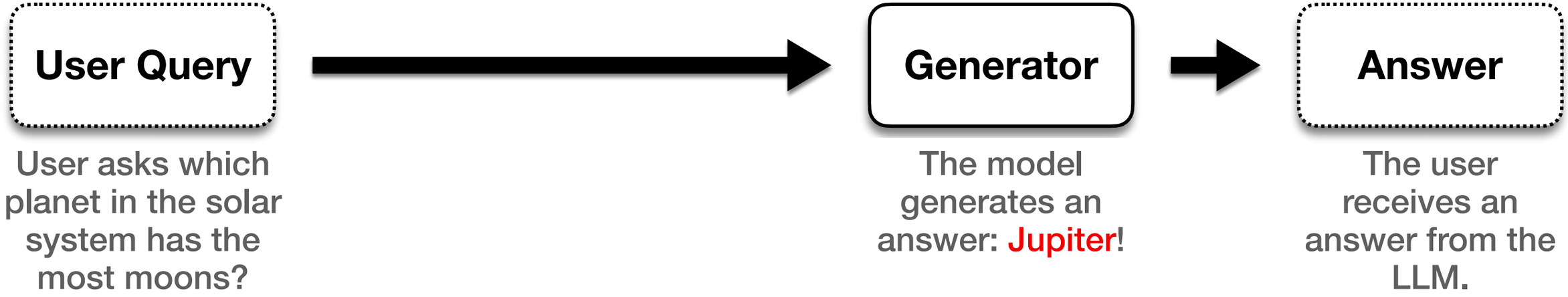
Naples, Italy - October 22nd 2026





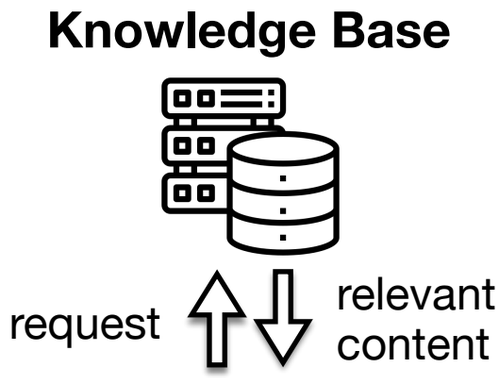
What is Retrieval Augmented Generation?

Enhances LLMs by retrieving up-to-date, external facts before generating accurate, context-aware responses.



What is Retrieval Augmented Generation?

Enhances LLMs by retrieving up-to-date, external facts before generating accurate, context-aware responses.



User Query → **Retriever** → **Generator** → **Answer**

User asks which planet in the solar system has the most moons?

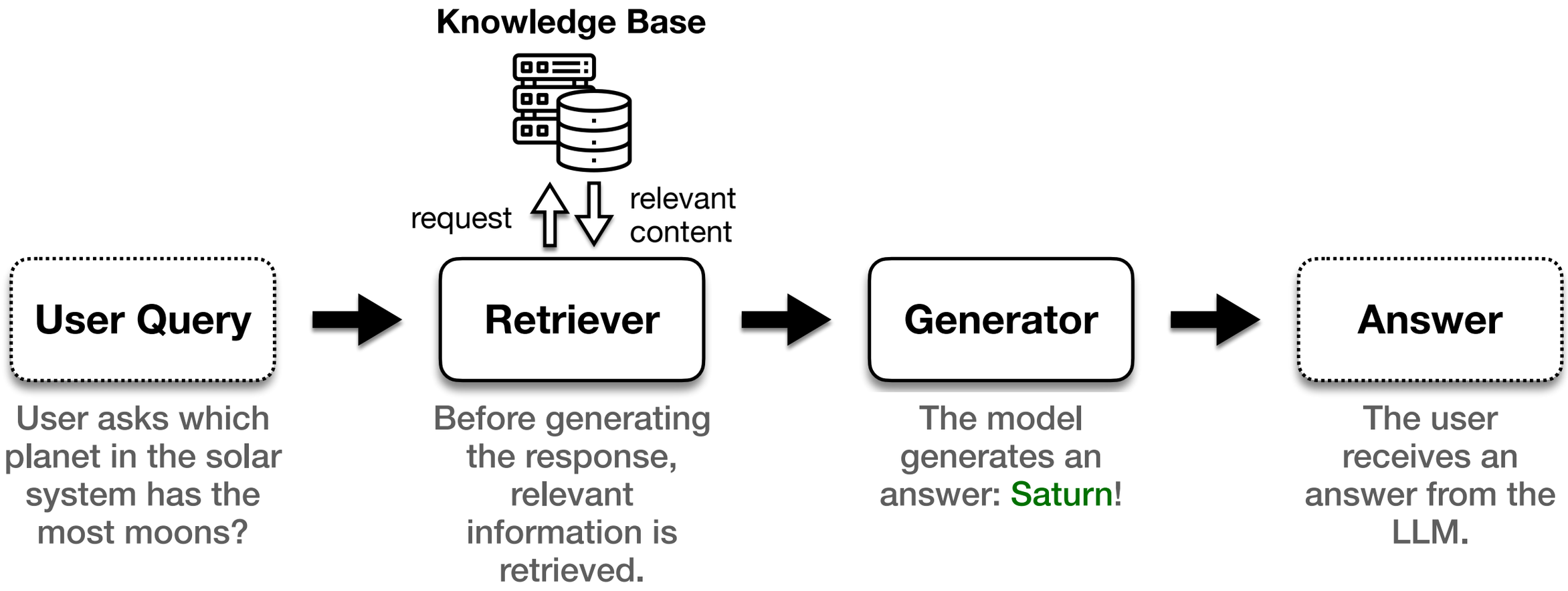
Before generating the response, relevant information is retrieved.

The model generates an answer: Jupiter!

The user receives an answer from the LLM.

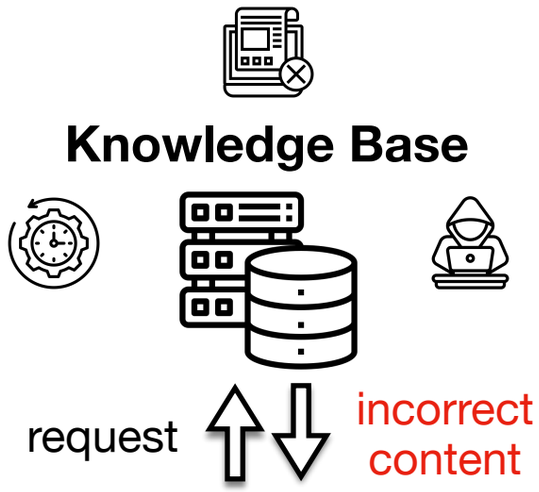
What is Retrieval Augmented Generation?

Enhances LLMs by retrieving up-to-date, external facts before generating accurate, context-aware responses.



What if the Knowledge Base can’t be trusted?

The knowledge base can become outdated, incorrect, or adversarial.



User Query → **Retriever** → **Generator** → **Answer**

User asks which planet in the solar system has the most moons?

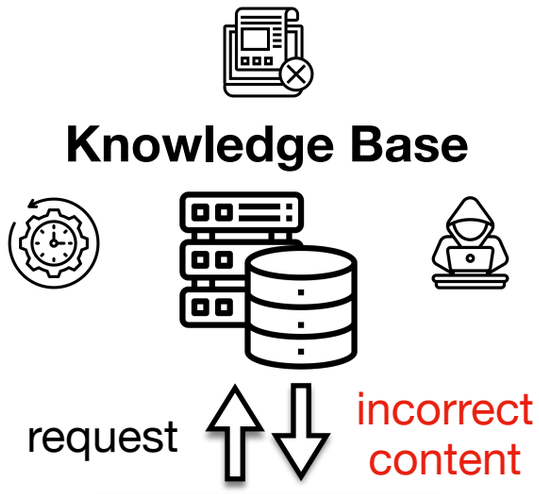
Before generating the response, relevant information is retrieved.

The model generates an answer: Mars!

The user receives an answer from the LLM.

What if the Knowledge Base can’t be trusted?

The knowledge base can become outdated, incorrect, or adversarial.



Answer can’t be trusted!

User Query → **Retriever** → **Generator** → ~~**Answer**~~

User asks which planet in the solar system has the most moons?

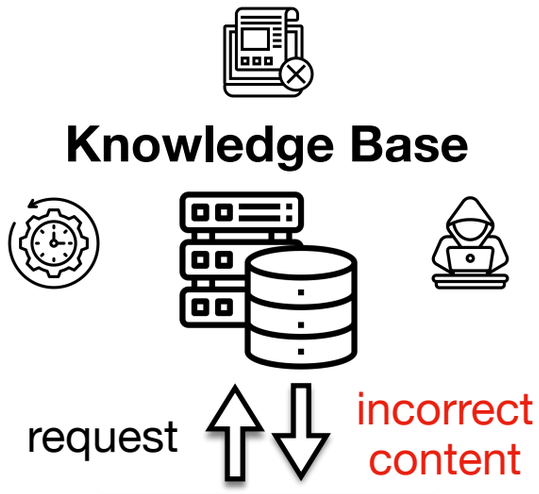
Before generating the response, relevant information is retrieved.

The model generates an answer: Mars!

The user receives an answer from the LLM.

What if the Knowledge Base can't be trusted?

The knowledge base can become outdated, incorrect, or adversarial.



Answer can't be trusted!

User Query → **Retriever** → **Generator** → ~~**Answer**~~

User asks which planet in the solar system has the most moons?

Before generating the response, relevant information is retrieved.

The model generates an answer: Mars!

The user receives an answer from the LLM.

The generator treats received text as evidence, with no way to distinguish good and bad evidence!

Handling Unreliable Evidence in RAG

Where do current defenses act?

◉ Corpus / Ingestion

- ❖ Source allow-lists, provenance tracking, access control.
- ❖ Anomaly and duplicate detection on ingested content.
- ❖ **Limitation:** Needs prior knowledge of which sources to trust.

◉ Retrieval

- ❖ Relevance and confidence scoring of retrieved documents.
- ❖ Re-retrieval or web fallback when confidence is low.
- ❖ **Limitation:** Relevant ≠ reliable, and no memory across queries.

Handling Unreliable Evidence in RAG

Where do current defenses act?

◉ Generation

- ❖ Fine-tuning or adversarial training for noisy context.
- ❖ Processing passages separately and aggregating answers (voting).
- ❖ Multi-step reasoning or debate over conflicting evidence.
- ❖ **Limitation:** Requires training, model access or extra LLM calls, and voting fails with high noise .

◉ Post-hoc Verification

- ❖ Citation and attribute checking.
- ❖ Claim-level fact verification and revision.
- ❖ Abstention when evidence is weak.
- ❖ **Limitation:** Adds latency and cost, corrects errors after generation.

Handling Unreliable Evidence in RAG

Where do current defenses act?

◉ Generation

- ❖ Fine-tuning or adversarial training for noisy context.
- ❖ Processing passages separately and aggregating answers (voting).
- ❖ Multi-step reasoning or debate over conflicting evidence.
- ❖ **Limitation:** Requires training, model access or extra LLM calls, and voting fails with high noise .

◉ Post-hoc Verification

- ❖ Citation and attribute checking.
- ❖ Claim-level fact verification and revision.
- ❖ Abstention when evidence is weak.
- ❖ **Limitation:** Adds latency and cost, corrects errors after generation.

Most defenses judge evidence per query.
Few accumulate knowledge about sources over time.

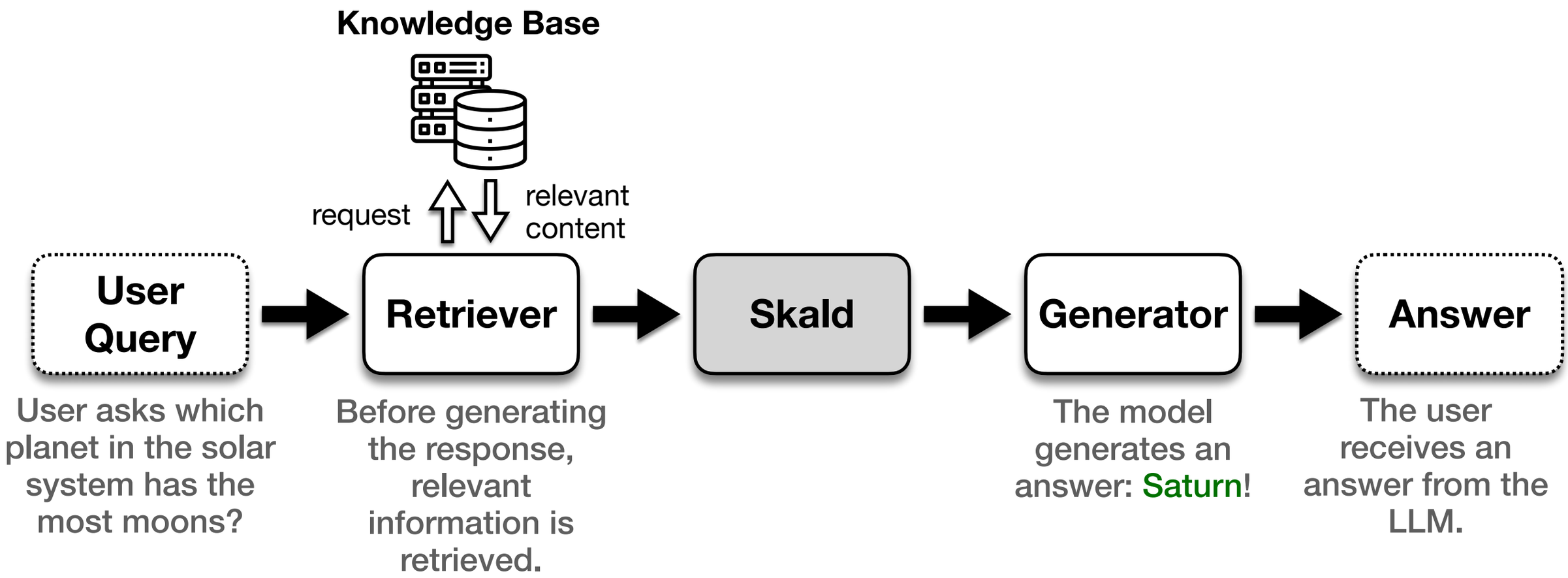
Skald (from Old Norse *skáld*)

This work

- A **fusion layer between the retriever and generator** that reconciles conflicting information before the LLM sees it.
  - **Consolidate** retrieved data using Truth Discovery.
  - **Persistent source reputation** is stored and updated across executions.
- **Single pass**, as reputation seeds the process, removing the need to iterate.
- **No training data** or LLM changes are required.

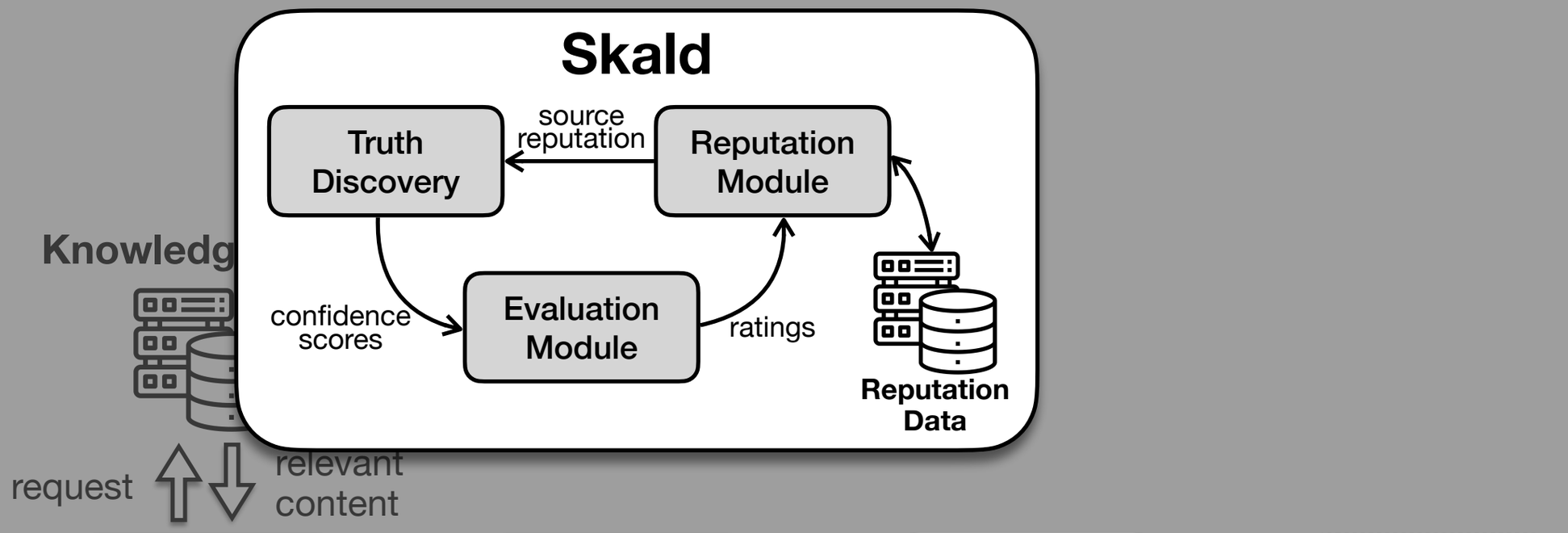
Skald

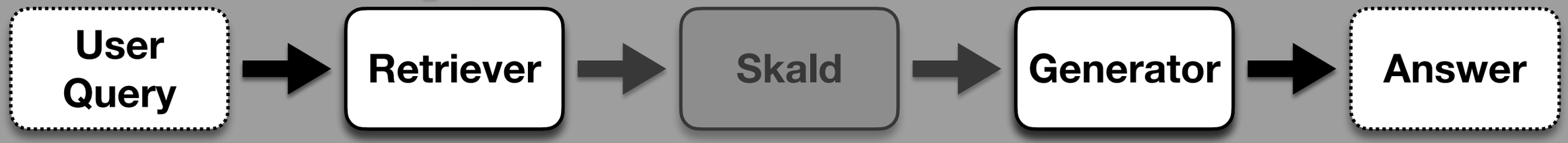
System overview



Skald

System overview





User asks which planet in the solar system has the most moons?

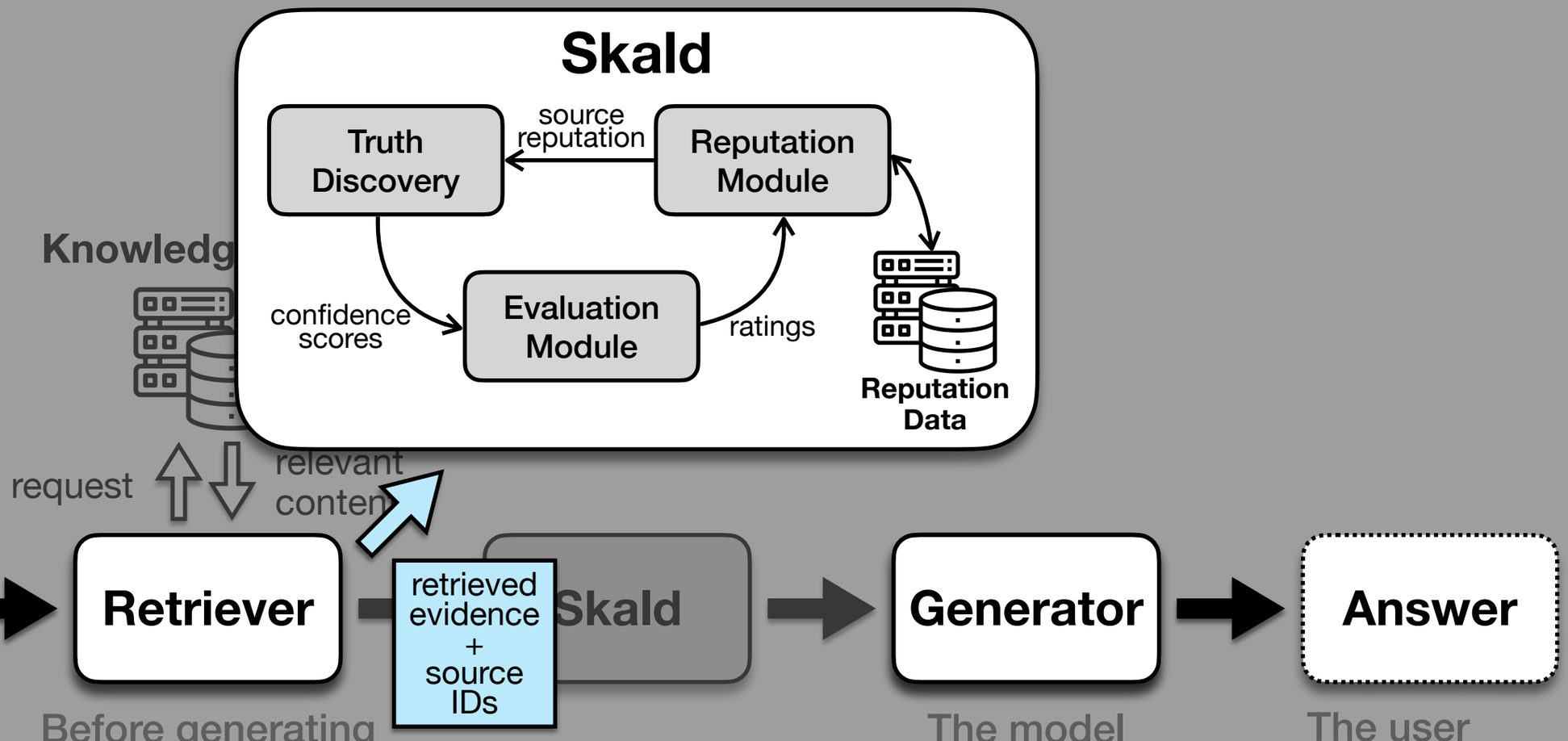
Before generating the response, relevant information is retrieved.

The model generates an answer: Saturn!

The user receives an answer from the LLM.

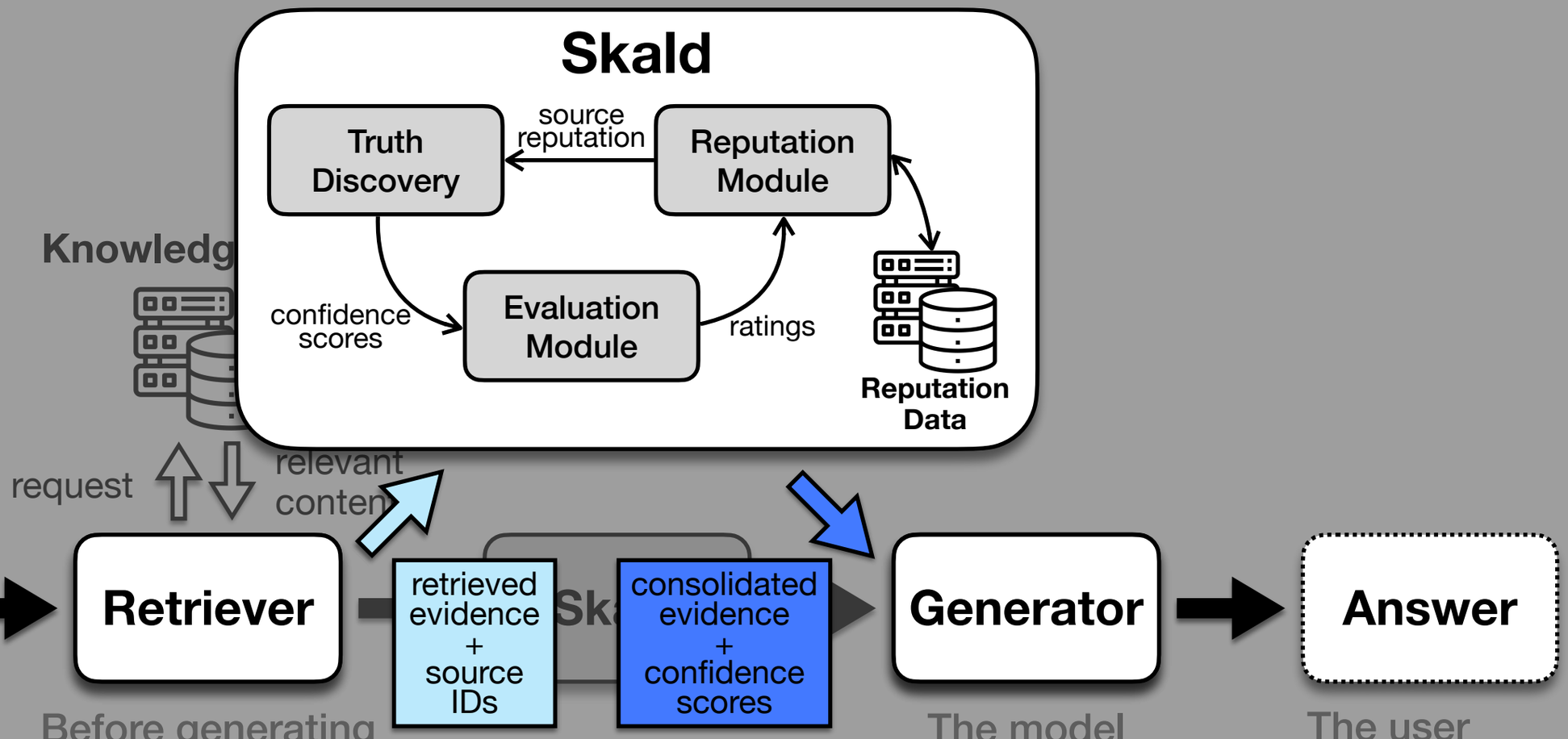
Skald

System overview



Skald

System overview



Skald

Reputation Module

- **Tracks source reputation** across executions.
  - Receives a rating for each source contribution.
  - Updates the source's reputation from those ratings.
  - Supplies reputation to the Truth Discovery module.
- **Bayesian** update with a **Dirichlet distribution.**
  - ***k*** rating **levels**, with **reputation scores in [0,1]**.
  - New sources start at **0.5**.
- **Persistent store**: reputation is kept between executions.

Skald

Truth Discovery Module

- **Input**:
  - Competing values for the same fact.
  - Identifier of the source behind each value.
- **Output**:
  - Confidence score in [0,1] for each value.
- Algorithm choice:
  - Some algorithms only support certain data types.
  - Retrieved evidence is heterogeneous (categorical, string, numeric), so we require support for all of them.

Skald

Truth Discovery Module

- **Input**:
  - Competi
  - Identifier
- **Output**:
  - Confider
- Algorithm ch
  - Some alg
  - Retrieved
    so we require support for all of them.

Algorithm	Data Type		
	Categorical	String	Continuous
AccuSim (Dong et al., 2009)	✓	✓	✓
AccuCopy (Dong et al., 2009)	✓	✓	✓
2-Estimates (Galland et al., 2010)	✓		
3-Estimates (Galland et al., 2010)	✓		
Investment (Pasternack & Roth, 2010)	✓		
TruthFinder (Yin et al., 2007)	✓	✓	✓
CRH (Q. Li, Li, Gao, Zhao, et al., 2014)	✓	✓	✓
CATD (Q. Li, Li, Gao, Su, et al., 2014)	✓		✓
LTM (Zhao et al., 2012)	✓		
GTM (Zhao & Han, 2012)			✓
LCA (Pasternack & Roth, 2013)	✓		

Skald

Truth Discovery Module

- **Input**:
  - Competi
  - Identifier
- **Output**:
  - Confider
- Algorithm ch
  - Some alg
  - Retrieved
    so we require support for all of them.

Algorithm	Data Type		
	Categorical	String	Continuous
AccuSim (Dong et al., 2009)	✓	✓	✓
AccuCopy (Dong et al., 2009)	✓	✓	✓
2-Estimates (Galland et al., 2010)	✓		
3-Estimates (Galland et al., 2010)	✓		
Investment (Pasternack & Roth, 2010)	✓		
TruthFinder (Yin et al., 2007)	✓	✓	✓
CRH (Q. Li, Li, Gao, Zhao, et al., 2014)	✓	✓	✓
CATD (Q. Li, Li, Gao, Su, et al., 2014)	✓		✓
LTM (Zhao et al., 2012)	✓		
GTM (Zhao & Han, 2012)			✓
LCA (Pasternack & Roth, 2013)	✓		

Skald

Truth Discovery Module

- **Input**:
  - Competi
  - Identifier
- **Output**:
  - Confider
- Algorithm ch
  - Some alg
  - Retrieved
    so we require support for all of them.
  - Chosen: **TruthFinder**, which also returns confidence scores in [0,1]

Algorithm	Data Type		
	Categorical	String	Continuous
AccuSim (Dong et al., 2009)	✓	✓	✓
AccuCopy (Dong et al., 2009)	✓	✓	✓
2-Estimates (Galland et al., 2010)	✓		
3-Estimates (Galland et al., 2010)	✓		
Investment (Pasternack & Roth, 2010)	✓		
TruthFinder (Yin et al., 2007)	✓	✓	✓
CRH (Q. Li, Li, Gao, Zhao, et al., 2014)	✓	✓	✓
CATD (Q. Li, Li, Gao, Su, et al., 2014)	✓		✓
LTM (Zhao et al., 2012)	✓		
GTM (Zhao & Han, 2012)			✓
LCA (Pasternack & Roth, 2013)	✓		

Skald

Truth Discovery Module

Skald

Truth Discovery Module

- **Source weight**: a value backed by reputable sources gets more weight, and many sources add up.
  - $\omega(v) = \Sigma - ln(1 - r(s))$

Skald

Truth Discovery Module

- **Source weight**: a value backed by reputable sources gets more weight, and many sources add up.
  - $\omega(v) = \Sigma - ln(1 - r(s))$

- **Implication between values**: similar values support each other, contradictory ones weaken each other
  - $\sigma(v) = \sum \omega(v') \cdot \text{imp}(v' \rightarrow v)$, with imp from -1(contradict) to 1(match)

Skald

Truth Discovery Module

- **Source weight**: a value backed by reputable sources gets more weight, and many sources add up.
  - $\omega(v) = \Sigma - ln(1 - r(s))$

- **Implication between values**: similar values support each other, contradictory ones weaken each other
  - $\sigma(v) = \sum \omega(v') \cdot \text{imp}(v' \rightarrow v)$, with imp from -1(contradict) to 1(match)

- **Confidence score**: a sigmoid function combines both effects.
  - $\phi(v) = \frac{1}{1 + e^{-\gamma \cdot (\omega(v) + \rho \cdot \sigma(v))}}$

Skald

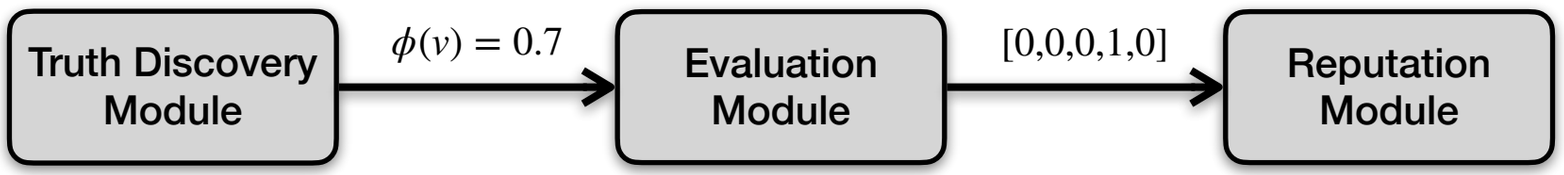
Truth Discovery Module

- **Source weight**: a value backed by reputable sources gets more weight, and many sources add up.
  - $\omega(v) = \Sigma - ln(1 - r(s))$
- **Implication between values**: similar values support each other, contradictory ones weaken each other
  - $\sigma(v) = \sum \omega(v') \cdot \text{imp}(v' \rightarrow v)$, with imp from -1(contradict) to 1(match)
- **Confidence score**: a sigmoid function combines both effects.
  - $\phi(v) = \frac{1}{1 + e^{-\gamma \cdot (\omega(v) + \rho \cdot \sigma(v))}}$
- **Single pass**: with reputation given upfront, no iterative convergence

Skald

Evaluation Module

- Connects the Truth Discovery and Reputation modules.
- Receives the confidence score of a value provided by a source.
  - e.g., $\phi(v) = 0.7$.
- **Converts** the **score to a rating** with k levels.
  - e.g., with $k = 5 : rating = [0,0,0,1,0]$



Skald

Assessment: goals and methodology

Skald

Assessment: goals and methodology

- **Goal.**
  - Test Skald’s accuracy among conflicting sources.
  - Evidence-fusion stage only, not end-to-end RAG.

Skald

Assessment: goals and methodology

◉ **Goal.**

- ❖ Test Skald’s accuracy among conflicting sources.
- ❖ Evidence-fusion stage only, not end-to-end RAG.

◉ **Methodology.**

- ❖ **Datasets:** books, flights, stocks (with gold standards)
- ❖ **Baselines:** TruthFinder and CRH
- ❖ **Metrics:** similarity to gold standard
- ❖ **Protocol:** 10 runs, input order randomized
- ❖ **Parameter study:** γ, ρ, k on books dataset (5 runs)

Skald

Assessing performance against established algorithms

- For all three datasets:
  - Skalds either matches or outperformes TruthFinder and CRH.

Table 1: Accuracy results (in %) of Skald compared to TruthFinder and CRH

	Books			Flights		Stocks	
Algorithm	JW	SD	DL	Euclidean	Manhattan	Euclidean	Manhattan
TruthFinder	88.35%	76.18%	62.32%	**99.99%**	**99.99%**	97.06%	97.06%
CRH	91.24%	84.57%	71.68%	**99.99%**	**99.99%**	91.83%	91.83%
Skald	**94.15%**	**90.30%**	**80.32%**	**99.99%**	**99.99%**	**99.18%**	**99.18%**

Skald

Assessing performance against established algorithms

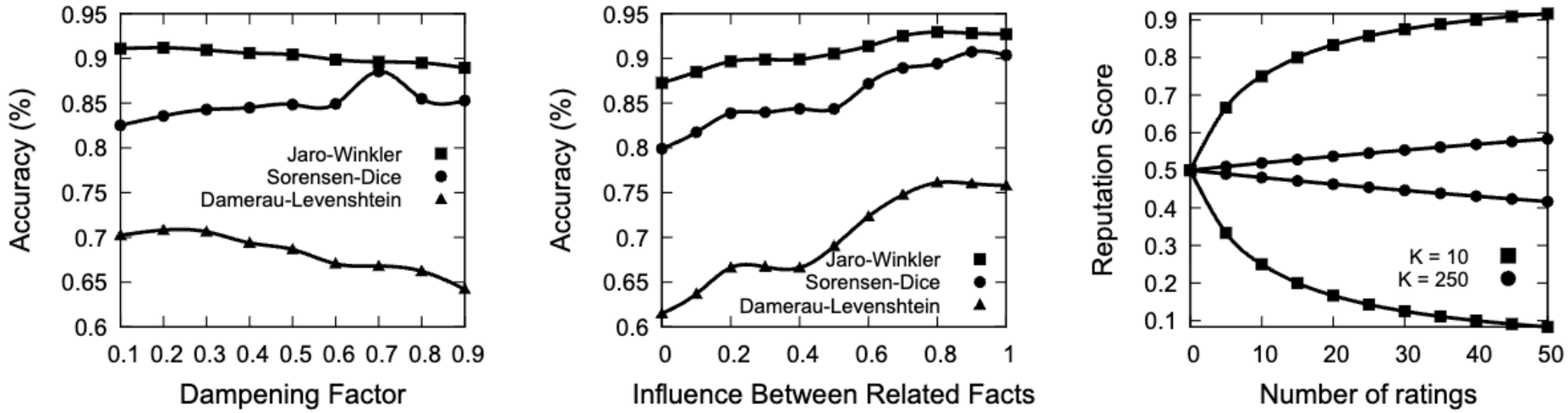
- For all three datasets:
  - Skalds either matches or outperformes TruthFinder and CRH.

Table 1: Accuracy results (in %) of Skald compared to TruthFinder and CRH

	Books			Flights		Stocks	
Algorithm	JW	SD	DL	Euclidean	Manhattan	Euclidean	Manhattan
TruthFinder	88.35%	76.18%	62.32%	**99.99%**	**99.99%**	97.06%	97.06%
CRH	91.24%	84.57%	71.68%	**99.99%**	**99.99%**	91.83%	91.83%
Skald	**94.15%**	**90.30%**	**80.32%**	**99.99%**	**99.99%**	**99.18%**	**99.18%**

Skald

Assessing adjustable parameters

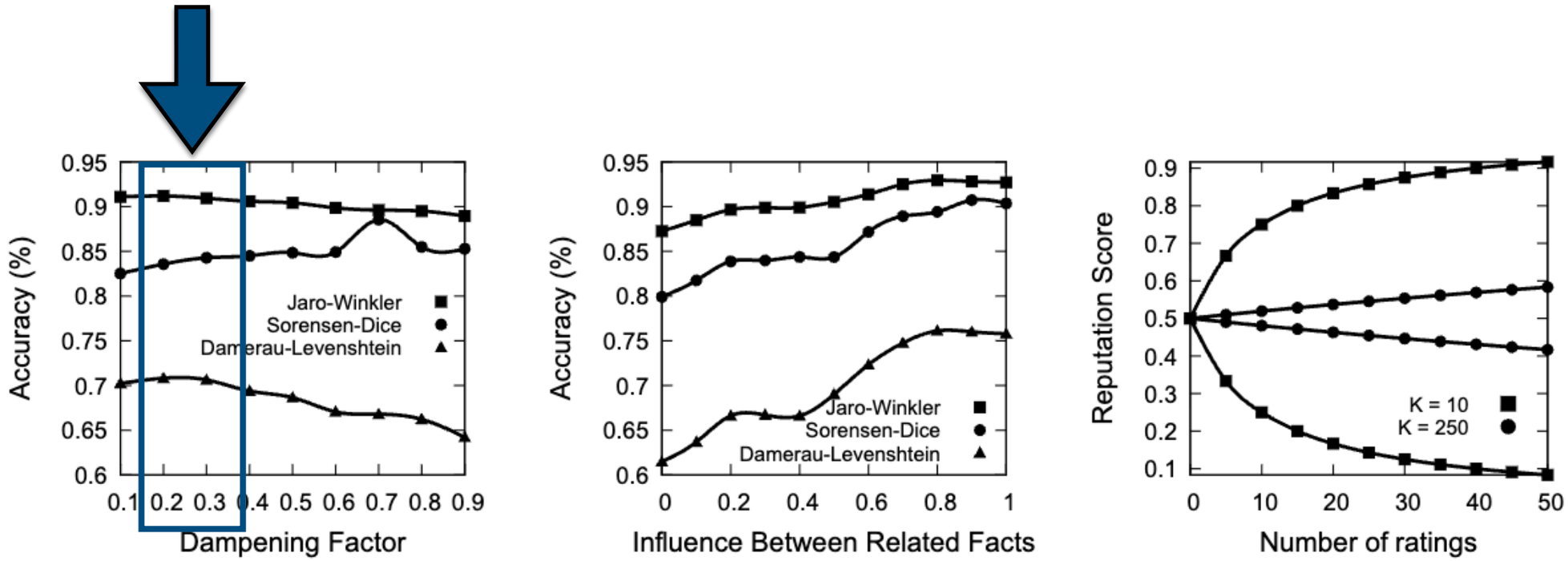


(a) Evaluation of γ. (b) Evaluation of ρ. (c) Evaluation of k.

Fig. 3: Evaluation for different γ, ρ, and k values.

Skald

Assessing adjustable parameters



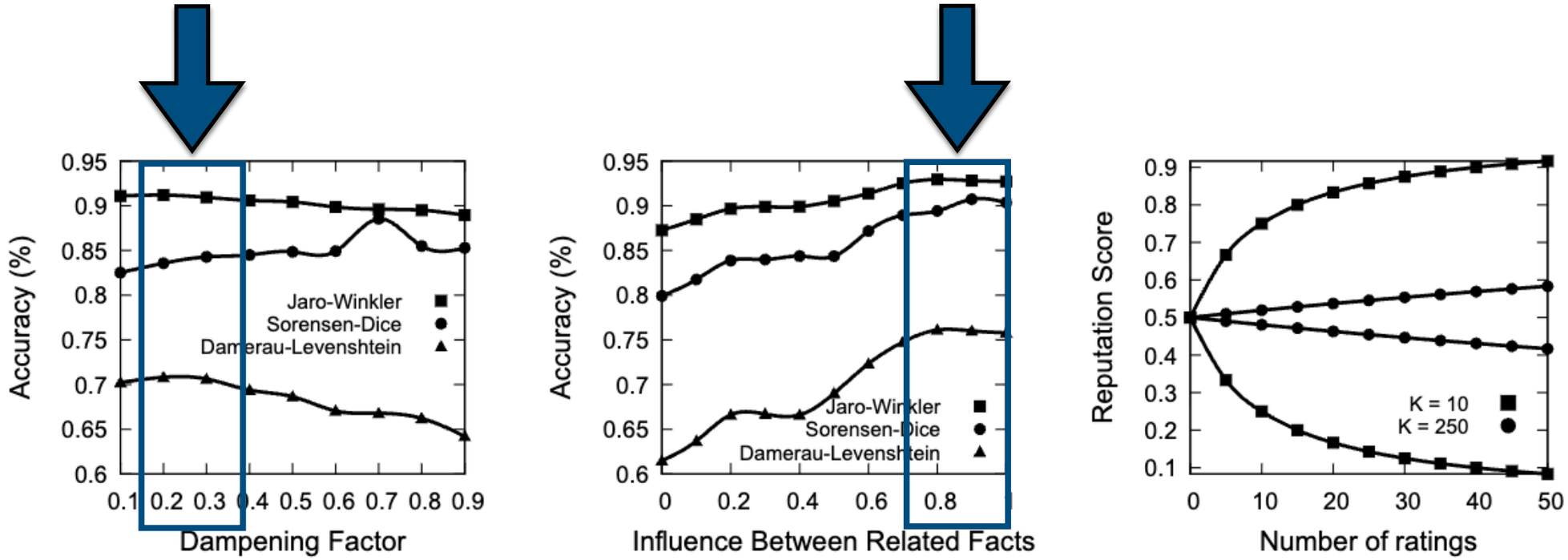
(a) Evaluation of γ. (b) Evaluation of ρ. (c) Evaluation of k.

Fig. 3: Evaluation for different γ, ρ, and k values.

The optimal range for the dampening factor resides in the [0.2,0.3] range.

Skald

Assessing adjustable parameters



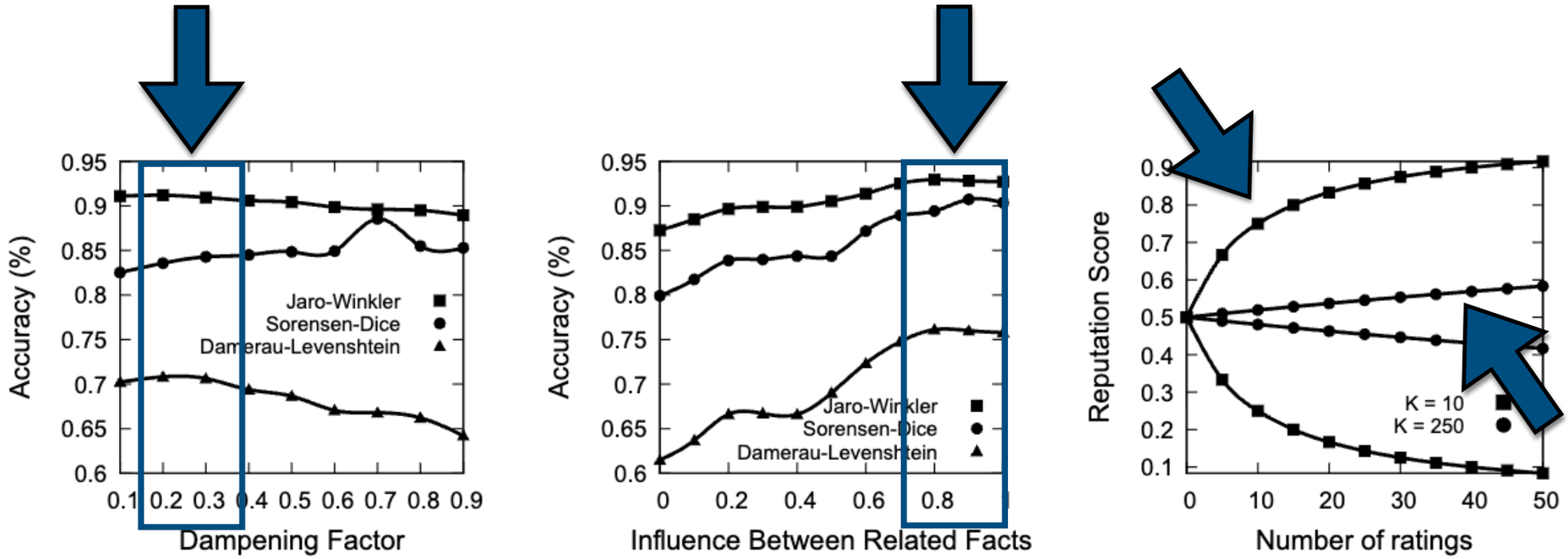
(a) Evaluation of γ. (b) Evaluation of ρ. (c) Evaluation of k.

Fig. 3: Evaluation for different γ, ρ, and k values.

The optimal range for the influence between related facts resides in the [0.8,0.9] interval.

Skald

Assessing adjustable parameters



(a) Evaluation of γ. (b) Evaluation of ρ. (c) Evaluation of k.

Fig. 3: Evaluation for different γ, ρ, and k values.

Low *k* increased reputation evolution speed. High *k* has the opposite effect.

Reputation-Aware Truth Discovery for Robust Evidence Fusion in RAG Pipelines

◉ **Conclusions:**

- ❖ Skald provides an evidence-fusion layer for RAG.
- ❖ Consolidates evidence before generation.
- ❖ Outperforms existing Truth Discovery algorithms.

◉ **Future Work**:

- ❖ Perform an end-to-end evaluation with a complete system.
- ❖ Support unstructured evidence.

Reputation-Aware Truth Discovery for Robust Evidence Fusion in RAG Pipelines

◉ **Conclusions:**

- ❖ Skald provides an evidence-fusion layer for RAG.
- ❖ Consolidates evidence before generation.
- ❖ Outperforms existing Truth Discovery algorithms.

◉ **Future Work**:

- ❖ Perform an end-to-end evaluation with a complete system.
- ❖ Support unstructured evidence.

◉ Skald is publicly available at:

- ❖ **GitHub**: https://github.com/dissect-project/skald

◉ Contact: diogo.p.ribeiro@inesctec.pt