Enhancing Knowledge Injection in Large Language Models for Efficient and Trustworthy Responses

Heydar Soudani, *Radboud University*

Doctoral Consortium

SIGIR 2025, Padova, Italy

Background

User Satisfaction in Information Access Systems

1. *Effectiveness:* Improve the correctness of system’s output
    - Knowledge Injection
2. *Trustworthiness:* Measure the reliability of the response as a score
    - Uncertainty Estimation

      To report to the user (and further improve effectiveness)

1. Effectiveness

Knowledge Injection

- *Fine-tuning (with synthetic data)*
- *Retrieval Augmented Generation (RAG)*
- Model Editing
- Retrieval Augmented Fine Tuning (RAFT)
- Parametric RAG

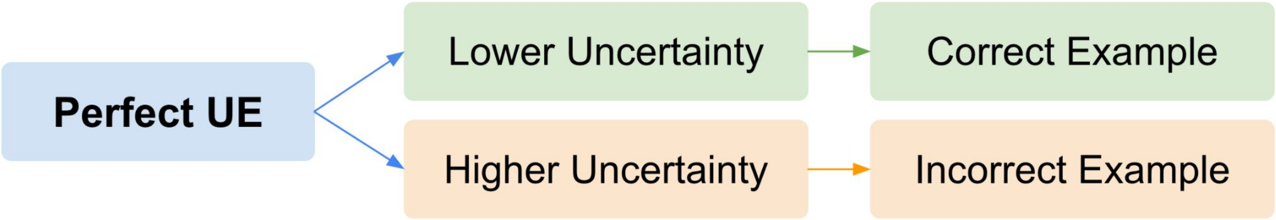
In scope for this presentation

[1] Meng et al., Mass-Editing Memory in a Transformer, ICLR 2023
[2] Zhang et al., RAFT: Adapting Language Model to Domain Specific RAG, Arxiv 2024
[3] Su et al., Parametric Retrieval Augmented Generation, SIGIR, 2025

2. Trustworthiness

Uncertainty Estimation (UE)

- Quantifies the correctness of the generated response without knowing the ground truth
- Assigns uncertainty to each (input, output) pair, representing its correctness (or truthfulness)



Perfect UE

Lower Uncertainty → Correct Example

Higher Uncertainty → Incorrect Example

Presentation Overview

- Knowledge Injection
  - Comparing RAG vs. Fine Tuning
- Trustworthiness
  - Performance of current UE methods in RAG setup
  - Proposing a new UE method for RAG-Reasoning systems
- Incorporating trustworthiness score to improve RAG system

RAG vs. Fine Tuning

Motivation: Customize LLMs on specific knowledge, access to textual data

Need: Assess the memorized knowledge in an LLM after injecting knowledge

Focus: Unpopular Factual Knowledge

(Kathy Saltzman, Occupation, Politician)

Subject **Relationship** **Object**

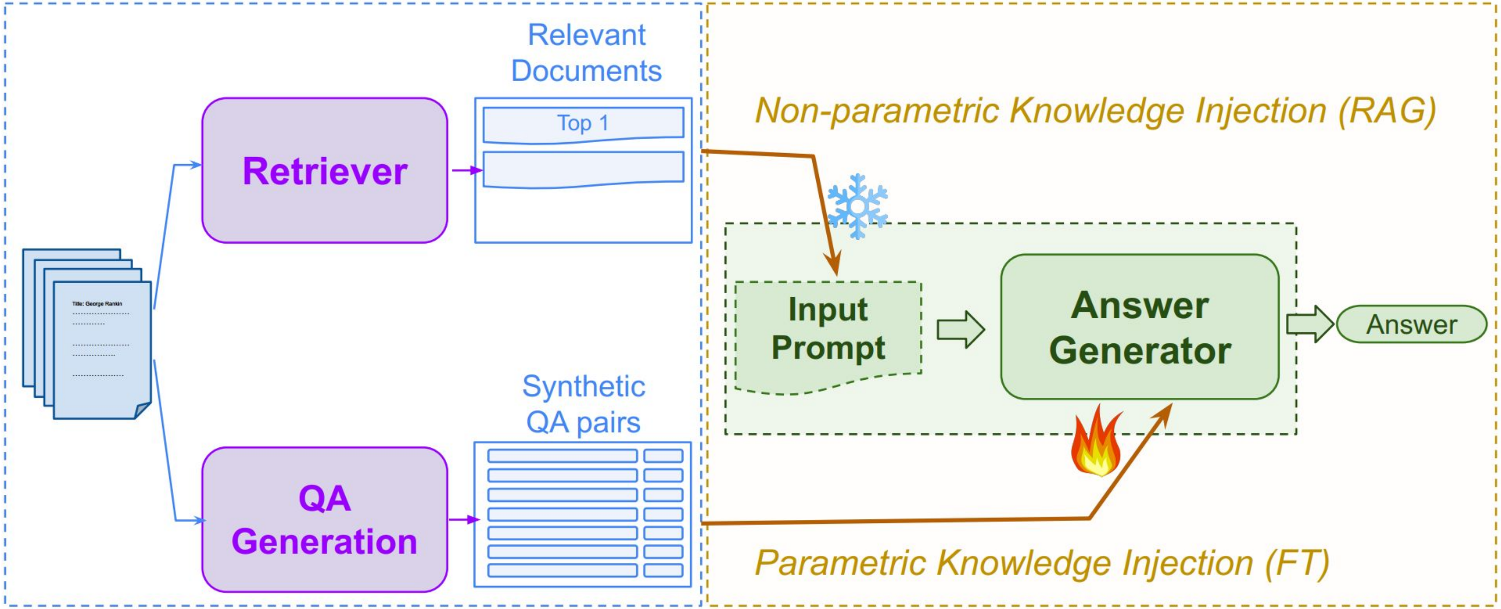
Q: What is the occupation of Kathy Saltzman?

A: Politician

Popularity: Wikipedia pageviews

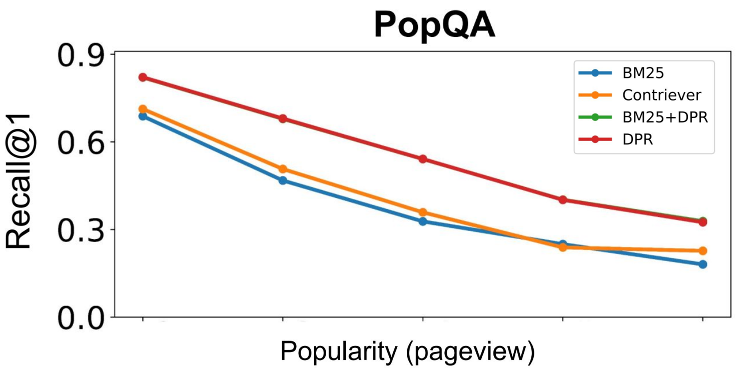
[4] Soudani, et, al. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge, SIGIR-AP, 2024

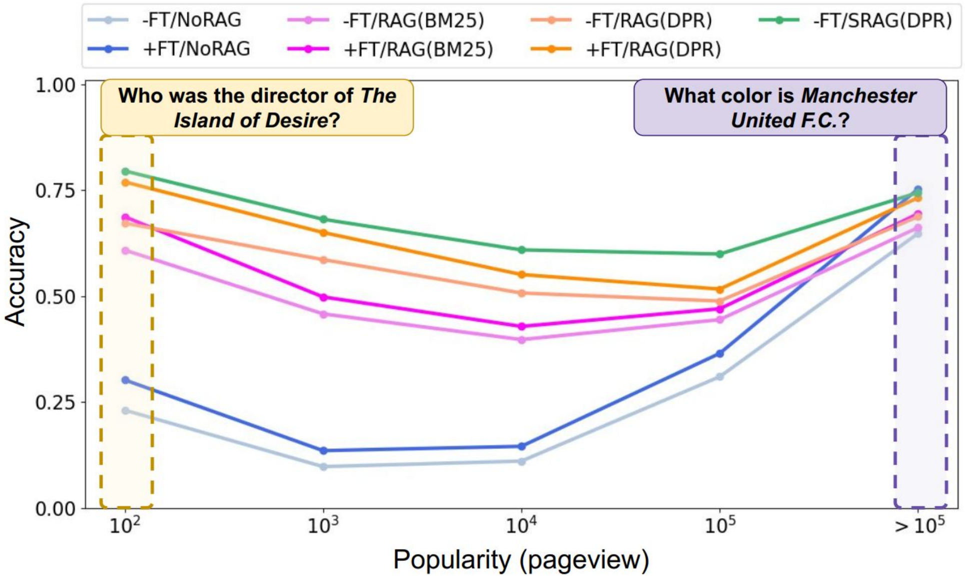
Methodology: Evaluation Framework



[4] Soudani, et, al. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge, SIGIR-AP, 2024

Results





Who was the director of *The Island of Desire*?

What color is *Manchester United F.C.*?

- RAG is particularly beneficial for less popular entities
- Fine-tuned LMs with RAG either outperform or match vanilla LMs with RAG

[4] Soudani, et, al. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge, SIGIR-AP, 2024

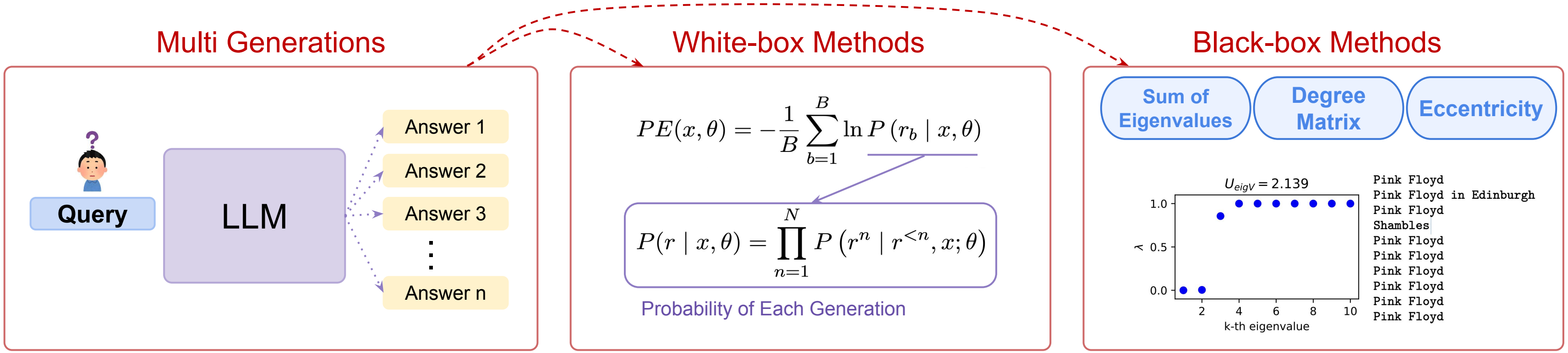
Presentation Overview

- Knowledge Injection
    - Comparing RAG vs. Fine Tuning
- Trustworthiness Measurement
    - Performance of current UE methods in RAG setup
    - Proposing a new UE method for RAG-Reasoning systems
- Incorporating trustworthiness score to improve RAG system

Background: Uncertainty Estimation

Main Idea: The more diverse the outputs, the more uncertain the model is

Evaluation: There is no ground-truth, assess by correlation with correctness, within-dataset



Multi Generations

Query

LLM

Answer 1

Answer 2

Answer 3

⋮

Answer n

White-box Methods

$$PE(x,\theta) = -\frac{1}{B}\sum_{b=1}^{B} \ln \underline{P\left(r_b \mid x,\theta\right)}$$

$$P(r \mid x,\theta) = \prod_{n=1}^{N} P\left(r^n \mid r^{<n}, x;\theta\right)$$

Probability of Each Generation

Black-box Methods

Sum of Eigenvalues

Degree Matrix

Eccentricity

```
Pink Floyd
Pink Floyd in Edinburgh
Pink Floyd
Shambles
Pink Floyd
Pink Floyd
Pink Floyd
Pink Floyd
Pink Floyd
Pink Floyd
```

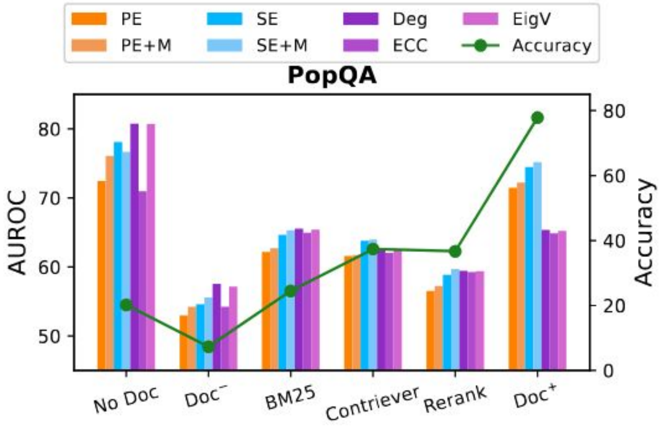
[5] Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction, ICLR, 2021

[6] Lin et al., Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models, TMLR, 2023

Uncertainty Estimation for RAG

Challenge

While existing UE methods mainly focus on scenarios where the input is just a query, ***it is unclear how current UE methods account for non-parametric knowledge***



The performance of existing UE methods is inconsistent and mainly deteriorates in RAG setup

[7] Soudani, et, al. Why Uncertainty Estimation Methods Fall Short in RAG:An Axiomatic Analysis, ACL (findings), 2025

Takeaways

- None of the existing UE methods *are optimal for RAG*
- Proposing an **Axiomatic Evaluation Framework**:

  1) Logical relationship between document-response

  2) Equivalence of LLM-generated and RAG-generated responses
- Introducing a calibration function guided by axioms
  - satisfying the axioms leads to performance improvements

[7] Soudani, et, al. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis, ACL (findings), 2025

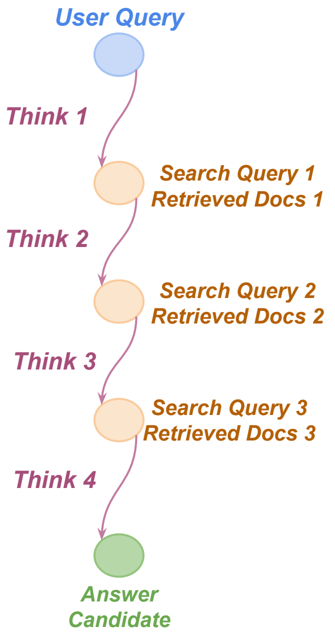
Follow-up Questions

- Is not it more effective to develop a standalone UE method for RAG that adheres to the prescribed axioms rather an a calibration function?
- Our calibration function works for a simple RAG setup (context, question, answer). What about RAG setup that involves multiple reasoning and retrieval steps (RAG-Reasoning systems)?

Presentation Overview

- Knowledge Injection
  - Comparing RAG vs. Fine Tuning
- Trustworthiness
  - Performance of current UE methods in RAG setup
  - Proposing a new UE method for RAG-Reasoning systems
- Incorporating trustworthiness score to improve RAG system

RAG-Reasoning Systems



Search-R1

User Query: What type of profession does Chris Jericho and Gary Barlow have in common?

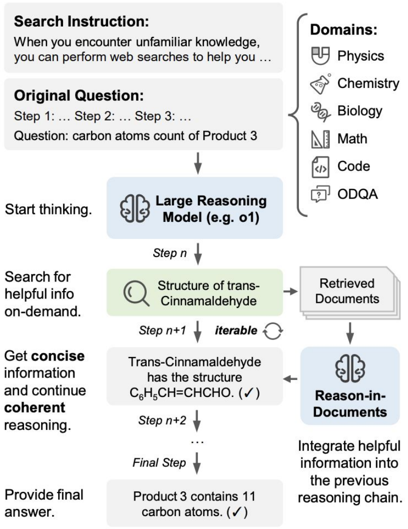
Think: I need to find the type of profession that Chris Jericho and Gary Barlow have in common. I'll search for it
Search Query: Chris Jericho and Gary Barlow profession
Information: Doc 2(Title: "Chris Jericho") ""Do You Wanna …

Think: I found out that Chris Jericho is a professional wrestler, musician, author …
Search Query: Gary Barlow profession
Information: Doc 1(Title: "Gary Barlow") hits, twelve number one …

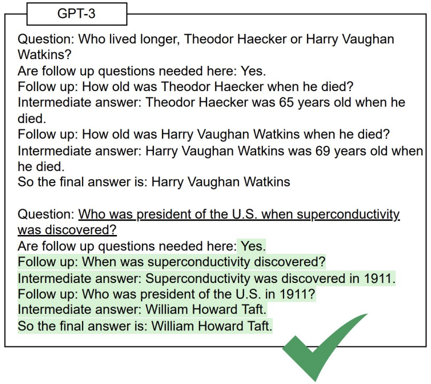
…

Think: I found out that Chris Jericho is a professional wrestler, musician, author …
Answer: musician

Search-O1


Search Instruction:
When you encounter unfamiliar knowledge, you can perform web searches to help you …
Original Question:
Step 1: … Step 2: … Step 3: …
Question: carbon atoms count of Product 3
Domains:
Physics
Chemistry
Biology
Math
Code
ODQA
Start thinking.
Large Reasoning Model (e.g. o1)
Step n
Search for helpful info on-demand.
Structure of trans-Cinnamaldehyde
Retrieved Documents
Step n+1
iterable
Get concise information and continue coherent reasoning.
Trans-Cinnamaldehyde has the structure C6H5CH=CHCHO. (✓)
Reason-in-Documents
Step n+2
…
Final Step
Integrate helpful information into the previous reasoning chain.
Provide final answer.
Product 3 contains 11 carbon atoms. (✓)

Self-Ask


GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?
Are follow up questions needed here: Yes.
Follow up: How old was Theodor Haecker when he died?
Intermediate answer: Theodor Haecker was 65 years old when he died.
Follow up: How old was Harry Vaughan Watkins when he died?
Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.
So the final answer is: Harry Vaughan Watkins

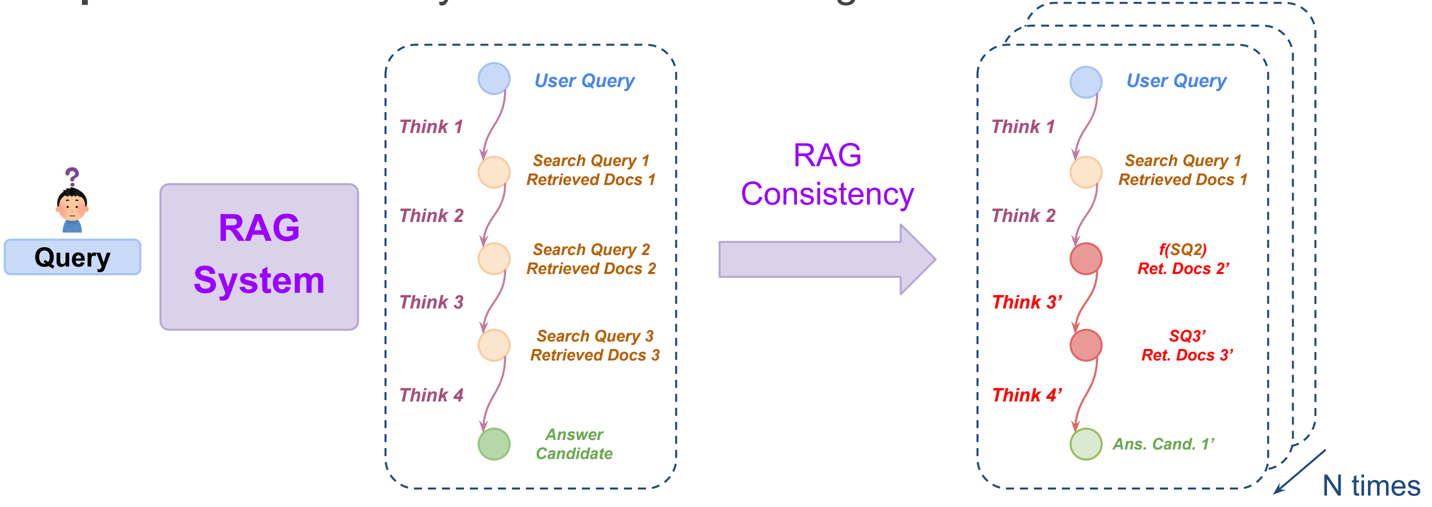
Question: Who was president of the U.S. when superconductivity was discovered?
Are follow up questions needed here: Yes.
Follow up: When was superconductivity discovered?
Intermediate answer: Superconductivity was discovered in 1911.
Follow up: Who was president of the U.S. in 1911?
Intermediate answer: William Howard Taft.
So the final answer is: William Howard Taft.

[8] Jin et al, Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning, Arxiv, 2025
[9] Li, Search-o1: Agentic Search-Enhanced Large Reasoning Models, Arxiv, 2025

UE for RAG-Reasoning Systems (Ongoing work)

RAG-Consistency

- **Idea:** Diversify the generations with applying changes in the reasoning-retrieval path
- **Scope:** Modern RAG systems with reasoning-retrieval interaction



Query

RAG System

User Query
Think 1
Search Query 1
Retrieved Docs 1
Think 2
Search Query 2
Retrieved Docs 2
Think 3
Search Query 3
Retrieved Docs 3
Think 4
Answer Candidate

RAG Consistency

User Query
Think 1
Search Query 1
Retrieved Docs 1
Think 2
f(SQ2)
Ret. Docs 2’
Think 3’
SQ3’
Ret. Docs 3’
Think 4’
Ans. Cand. 1’

N times

Initial Results (Ongoing work)

Search-R1	**Self-Consistency**	**Reasoning-Consistency**	**RAG-Consistency**
Bamboogle	58.01	67.04	74.76
PopQA	60.59	69.87	81.57
HotpotQA	55.97	61.98	69.63
2Wiki	63.52	62.19	70.40

Self-Ask	**Self-Consistency**	**Reasoning-Consistency**	**RAG-Consistency**
Bamboogle	59.03	71.62	80.39
PopQA	59.72	71.33	82.16
HotpotQA	59.58	75.55	82.30
2Wiki	55.65	66.16	76.99

Presentation Overview

- Knowledge Injection
    - Comparing RAG vs. Fine Tuning
- Trustworthiness Measurement
    - Performance of current UE methods in RAG setup
    - Proposing a new UE method for RAG-Reasoning systems
- Incorporating trustworthiness score to improve RAG system

Points of Discussion

Q1. Should we rely on UE methods to get a feedback for improving effectiveness or other approaches (e.g., LLM as judge). If yes, how?

- + RAG-Consistency shows promising results
- - It is computationally expensive
- - While it shows correlation with correctness, it does not necessarily lead to producing the correct response

Q2. What are other applications of UE/RAG-Consistency methods?

Q3. Any suggestions to make RAG-Consistency more efficient/accurate?

Thank You!