# Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis

**Heydar Soudani[1], Evangelos Kanoulas[2], and Faegheh Hasibi[1]**

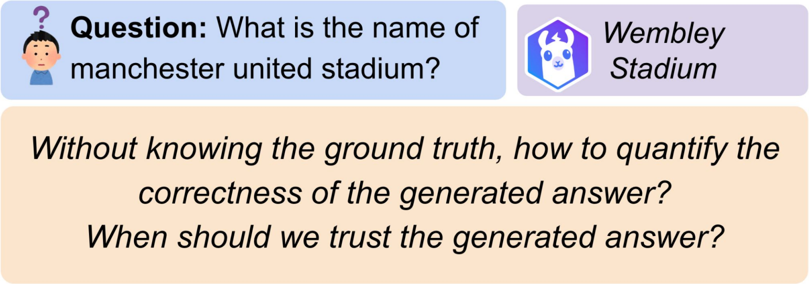
[1]Radboud University

[2]University of Amsterdam

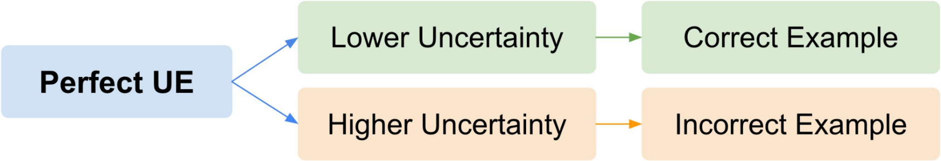
***ACL'25***

# Trustworthiness

- **Motivation:** LLMs have a tendency to produce inaccurate or misleading outputs
  - *Reasons:* Hallucination, Temporal Knowledge Shift, Noisy Information in RAG, …




- **Solution:** Uncertainty Estimation
  - Assigns uncertainty to each (input, output) pair, representing its correctness (or truthfulness)




heydar.soudani@ru.nl

# Background

## White-box

*Assumption:* Access to token probability

*Main Concept:* Entropy

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

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

Probability of each generation

## Black-box

*Assumption:* Rely solely on final outputs

*Main Concept:* Semantic Similarity

*Some Methods:*

- Sum of Eigenvalues
- Degree Matrix
- Eccentricity

[1] MARS: Meaning-Aware Response Scoring for Uncertainty Estimation in Generative LLMs, ACL, 2024
[2] Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models, TMLR, 2023

heydar.soudani@ru.nl

# 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***

heydar.soudani@ru.nl

# Research Questions

- **(RQ1)** How do UE methods perform when the input prompt includes non-parametric knowledge, such as in RAG?
- **(RQ2)** What properties can guarantee optimal performance of UE considering  LLMs' both parametric and non-parametric knowledge?
- **(RQ3)** Can the axiomatic framework guide us in deriving an optimal UE method?

heydar.soudani@ru.nl

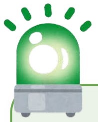
# Research Questions

- **(RQ1)** How do UE methods perform when the input prompt includes non-parametric knowledge, such as in RAG?
- **(RQ2)** What properties can guarantee optimal performance of UE considering  LLMs' both parametric and non-parametric knowledge?
- **(RQ3)** Can the axiomatic framework guide us in deriving an optimal UE method?

heydar.soudani@ru.nl

# Experimental Setup

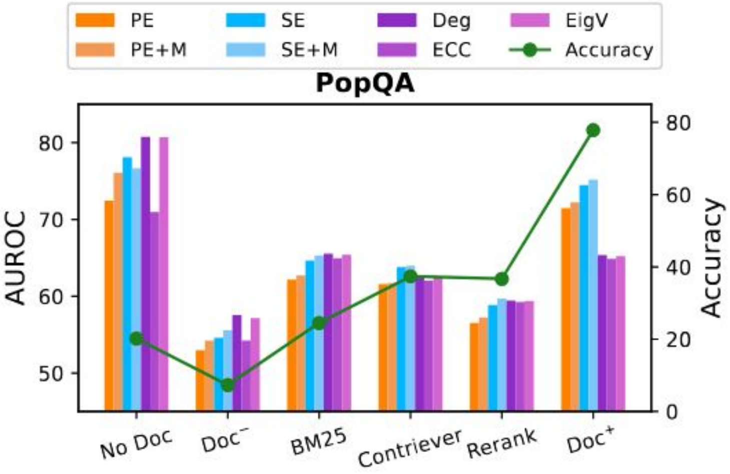
- **Retrievers**
  - **Doc**$^{-}$ : A weak synthetic retriever that returns irrelevant documents
  - **Doc**$^{+}$ : An idealized retriever that consistently ranks the gold document at the top
  - Several widely used retrievers: BM25, Contriever, Rerank
- **UE Methods**
  - White-box: PE, SE, MARS
  - Black-box: Dig, EigV, ECC
- **Evaluation Metrics**
  - Correctness: Exact Match
  - Corr./Unc. Correlation: AUROC



**Optimal Correlation with Correctness**

heydar.soudani@ru.nl

# UE for RAG




| LLM | Unc. | PopQA No Doc | Doc⁻ | BM25 | Cont. | ReRa. | Doc⁺ |
|---|---|---|---|---|---|---|---|
| Llama2-chat | PE | 1.29 | 1.11 * | 0.54 * | 0.46 * | 0.35 * | 0.34 * |
| | SE | 4.86 | 4.37 * | 3.45 * | 3.30 * | 3.13 * | 3.19 * |
| | PE+M | 1.59 | 1.34 * | 0.65 * | 0.55 * | 0.44 * | 0.45 * |
| | SE+M | 5.38 | 4.71 * | 3.62 * | 3.43 * | 3.23 * | 3.27 * |
| | Deg | 0.52 | 0.32 * | 0.12 * | 0.09 * | 0.06 * | 0.05 * |
| | ECC | 0.71 | 0.54 * | 0.22 * | 0.17 * | 0.12 * | 0.10 * |
| | EigV | 4.25 | 2.28 * | 1.42 * | 1.31 * | 1.18 * | 1.17 * |
| Mistral-v0.3 | PE | 1.51 | 0.94 * | 0.84 * | 0.69 * | 0.62 * | 0.51 * |
| | SE | 5.66 | 3.73 * | 3.68 * | 3.53 * | 3.41 * | 3.26 * |
| | PE+M | 2.35 | 1.42 * | 1.26 * | 1.05 * | 0.92 * | 0.80 * |
| | SE+M | 6.47 | 4.05 * | 3.98 * | 3.77 * | 3.60 * | 3.45 * |
| | Deg | 0.48 | 0.05 * | 0.07 * | 0.06 * | 0.05 * | 0.03 * |
| | ECC | 0.68 | 0.03 * | 0.08 * | 0.08 * | 0.05 * | 0.04 * |
| | EigV | 4.18 | 1.08 * | 1.16 * | 1.17 * | 1.11 * | 1.08 * |

**Improvements on the proposed UE methods in the literature do not add up when considering RAG setup**

heydar.soudani@ru.nl

# Research Questions

- **(RQ1)** How do UE methods perform when the input prompt includes non-parametric knowledge, such as in RAG?
- **(RQ2)** What properties can guarantee optimal performance of UE considering  LLMs' both parametric and non-parametric knowledge?
- **(RQ3)** Can the axiomatic framework guide us in deriving an optimal UE method?

heydar.soudani@ru.nl

# Axiomatic Thinking

## Definition

- A set of formal constraints is defined based on desired properties, which are then used as a guide to search for an optimal solution

## Applications

- Information Retrieval
- Interpretability
- Preference Modeling

[1] An exploration of axiomatic approaches to information retrieval, SIGIR, 2005
[2] Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models, SIGIR, 2024
[3] Axiomatic preference modeling for longform question answering, EMNLP, 2023

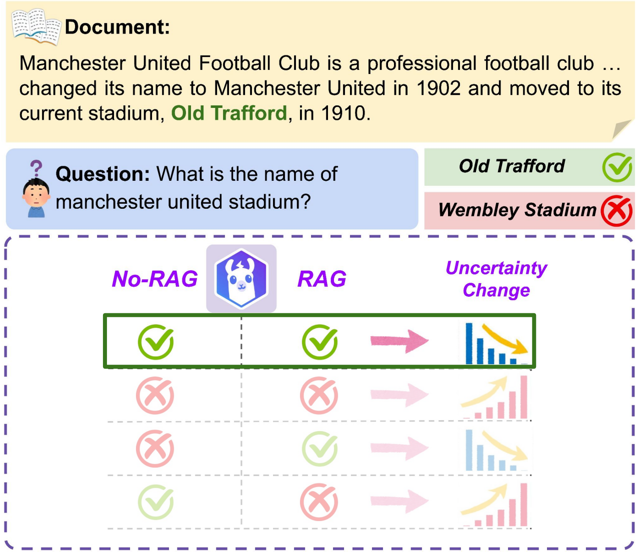
heydar.soudani@ru.nl

# Axiom 1: Positively Consistent

$$\forall q, c \ \text{ if } \ \mathcal{M}_\theta(q) = r_1, \ \mathcal{M}_\theta(q,c) = r_2, \ r_1 \equiv r_2, \ c \models (q, r_2),$$
$$\text{then } \ \mathcal{U}(\mathcal{M}_\theta(q), r_1) > \mathcal{U}(\mathcal{M}_\theta(q,c), r_2).$$

## Description

If an LLM generates the *same output* both before and after incorporating a context, and *the context logically supports that output*, ***the uncertainty in the RAG setup should decrease***.




heydar.soudani@ru.nl

# RQ2: Axiomatic Evaluation

| UE | PopQA | | |
|---|---|---|---|
| | BM25 | Contriever | Doc$^{+}$ |
| **Axiom 1:** Positively Consistent ↓ | | | |
| PE | 0.735 → 0.419 * | 0.735 → 0.408 * | 1.242 → 0.340 * |
| SE | 3.781 → 3.205 * | 3.791 → 3.158 * | 4.682 → 3.113 * |
| PE+M | 0.896 → 0.483 * | 0.881 → 0.458 * | 1.530 → 0.406 * |
| SE+M | 4.102 → 3.286 * | 4.091 → 3.248 * | 5.146 → 3.173 * |
| EigV | 1.951 → 1.166 * | 2.025 → 1.143 * | 4.074 → 1.078 * |
| ECC | 0.417 → 0.110 * | 0.426 → 0.094 * | 0.710 → 0.055 * |
| Deg | 0.220 → 0.048 * | 0.230 → 0.043 * | 0.496 → 0.022 * |
| **Axiom 2:** Negatively Consistent ↑ | | | |
| PE | 1.068 → 0.746 | 0.820 → 0.593 | 1.083 → 0.597 |
| SE | 4.163 → 3.548 * | 4.104 → 3.381 * | 4.388 → 4.107 |
| PE+M | 1.309 → 0.844 | 1.016 → 0.782 | 1.328 → 0.684 |
| SE+M | 4.599 → 3.700 * | 4.481 → 3.610 * | 4.764 → 4.221 |
| EigV | 2.453 → 1.338 * | 2.088 → 1.274 * | 2.758 → 1.910 |
| ECC | 0.541 → 0.197 * | 0.477 → 0.152 * | 0.503 → 0.443 |
| Deg | 0.286 → 0.101 * | 0.228 → 0.073 * | 0.343 → 0.254 |

| UE | BM25 | Contriever | Doc$^{+}$ |
|---|---|---|---|
| **Axiom 3:** Positively Changed ↓ | | | |
| PE | 1.375 → 0.347 * | 1.416 → 0.298 * | 1.342 → 0.268 * |
| SE | 4.889 → 3.015 * | 5.091 → 3.013 * | 4.884 → 3.051 * |
| PE+M | 1.708 → 0.398 * | 1.735 → 0.374 * | 1.604 → 0.340 * |
| SE+M | 5.514 → 3.072 * | 5.681 → 3.082 * | 5.379 → 3.099 * |
| EigV | 4.131 → 1.139 * | 4.733 → 1.114 * | 4.449 → 1.102 * |
| ECC | 0.790 → 0.085 * | 0.823 → 0.081 * | 0.780 → 0.072 * |
| Deg | 0.547 → 0.044 * | 0.588 → 0.035 * | 0.544 → 0.032 * |
| **Axiom 4:** Negatively Changed ↑ | | | |
| PE | 0.933 → 0.636 | 1.006 → 0.558 | 1.252 → 0.463 |
| SE | 4.152 → 3.552 * | 4.192 → 3.409 * | 4.830 → 3.690 * |
| PE+M | 1.164 → 0.714 * | 1.298 → 0.748 * | 1.689 → 0.747 |
| SE+M | 4.553 → 3.690 * | 4.653 → 3.608 * | 5.381 → 4.007 * |
| EigV | 2.593 → 1.449 * | 2.557 → 1.412 * | 3.567 → 1.449 * |
| ECC | 0.540 → 0.262 * | 0.548 → 0.220 * | 0.707 → 0.237 * |
| Deg | 0.320 → 0.128 * | 0.320 → 0.115 * | 0.463 → 0.140 * |

**Axiom 5**

| Unc. | NQ-open | TriviaQA | PopQA |
|---|---|---|---|
| PE | 2.072 → 2.248 * | 0.872 → 1.155 * | 0.897 → 0.909 * |
| SE | 5.253 → 5.471 * | 3.863 → 4.158 * | 3.897 → 4.319 * |
| PE+M | 4.791 → 4.805 | 1.415 → 1.699 * | 1.031 → 1.130 * |
| SE+M | 7.993 → 7.933 | 4.540 → 4.817 * | 4.297 → 4.591 |
| EigV | 2.211 → 2.446 * | 1.757 → 1.870 * | 2.270 → 2.218 |
| ECC | 0.512 → 0.625 * | 0.382 → 0.448 * | 0.490 → 0.507 |
| Deg | 0.265 → 0.333 * | 0.171 → 0.211 * | 0.256 → 0.309 |

heydar.soudani@ru.nl

# Research Questions

- **(RQ1)** How do UE methods perform when the input prompt includes non-parametric knowledge, such as in RAG?
- **(RQ2)** What properties can guarantee optimal performance of UE considering LLMs' both parametric and non-parametric knowledge?
- **(RQ3)** Can the axiomatic framework guide us in deriving an optimal UE method?

heydar.soudani@ru.nl

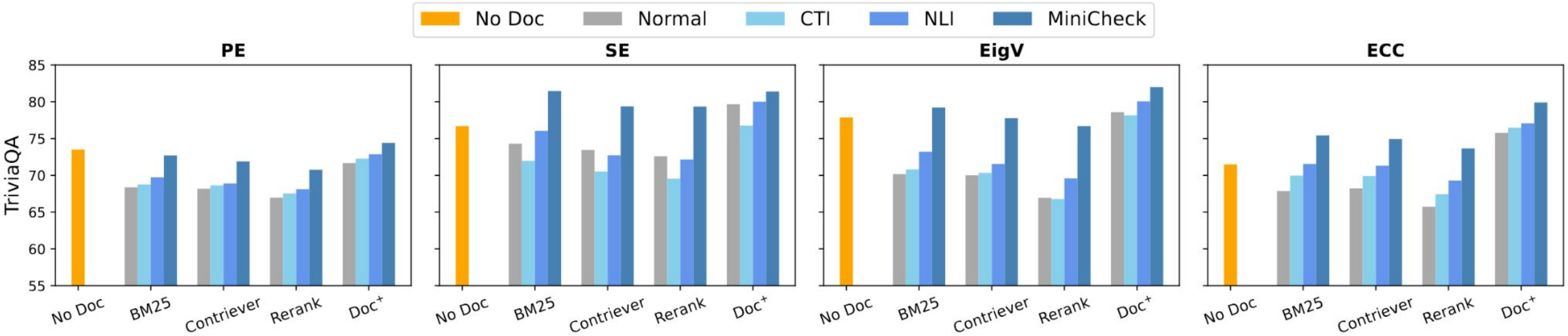
# Axiomatic Calibration

## Calibration Coefficient

$$\alpha_{\text{ax}} = k_1 \cdot \underbrace{\mathcal{E}(r_1, r_2)}_{\text{Equivalence of LLM \& RAG -generated responses}} + k_2 \cdot \underbrace{\mathcal{R}(c, q, r_1)}_{\text{Context / LLM-generated response Relationship}} + k_3 \cdot \underbrace{\mathcal{R}(c, q, r_2)}_{\text{Context / RAG-generated response Relationship}}$$

$$\mathcal{U}(\mathcal{M}_\theta(c, q), r_2)^{\text{cal}} = (k_4 - \alpha_{\text{ax}}) \cdot \mathcal{U}(\mathcal{M}_\theta(c, q), r_2).$$

# Axiomatic Calibration: Results




heydar.soudani@ru.nl

# Takeaways

- Existing UE methods *generated low uncertainty values* in the RAG setup without considering *the relevance of the given context to the query*
- None of the existing UE methods *pass all the proposed axioms*
- The result of the proposed ***calibration function*** shows
    - Satisfying the axioms leads to performance improvements

heydar.soudani@ru.nl