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

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

Radboud University[1] University of Amsterdam[2]

1 Introduction

- **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.*

- **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?

Experimental Setup

Retrievers

- *Doc-: A weak synthetic*
- *Doc+: An idealized ret.*
- ***BM25***
- ***Contriever***
- ***Re-rank***

UE Methods

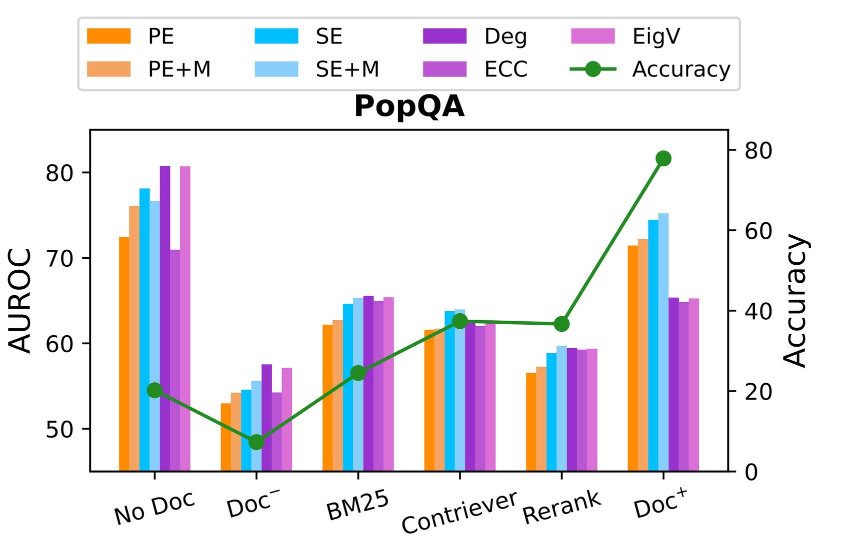
- ***White-box:***
  - *Predictive Entropy*
  - *Semantic Entropy*
  - *Meaning-Aware Ent.*
- ***Black-box:***
  - *Sum of Eignvalues*
  - *Degree Matrix*
  - *Eccentricity*

Evaluation Methods

- ***Correctness:***
  - *Exact Match*
- ***Corr./Unc. Correlation:***
  - *AUROC*

Optimal Correlation with Correctness

2 Uncertainty Estimation 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 *

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

3 Axiomatic Evaluation Framework

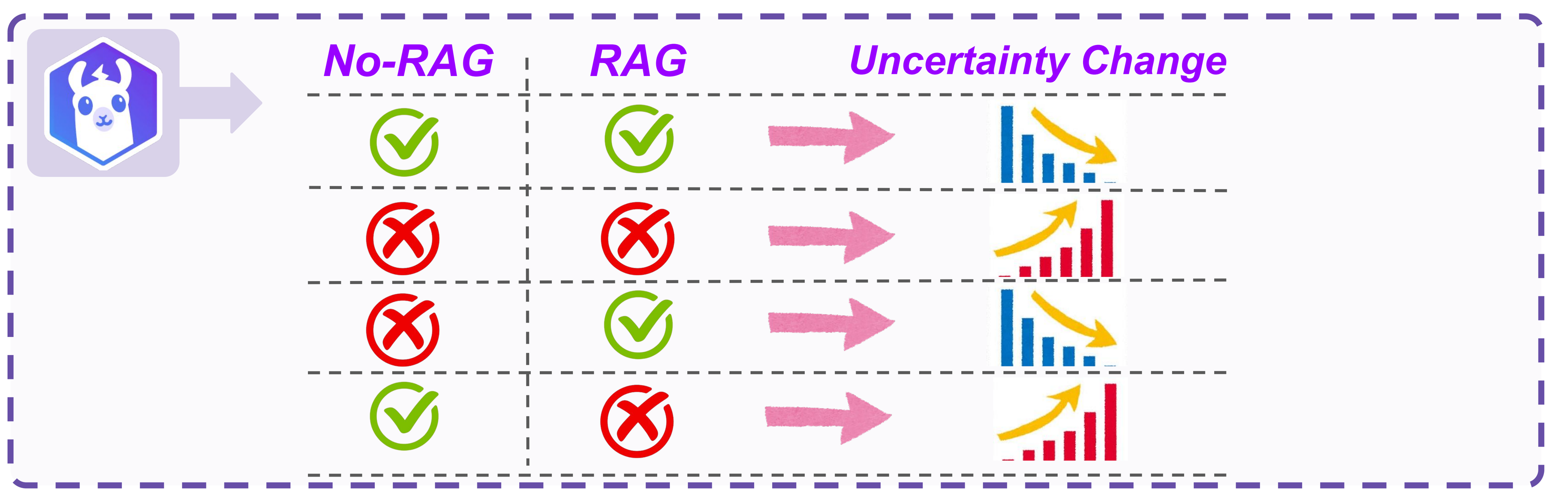
- Desired behavior of uncertainty estimation methods with and without RAG in an ***Example***

Context:
Manchester United Football Club is a professional football club … changed its name to Manchester United in 1902 and moved to its current stadium, **Old Trafford**, in 1910.

Question: What is the name of manchester united stadium?

Old Trafford

Wembley Stadium



Defined Axioms

1) ***Positively Consistent:*** 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.
2) ***Negatively Consistent:*** If an LLM generates the same output both before and after incorporating a context, but the context contradicts that output, the uncertainty in the RAG setup should increase.
3) ***Positively Changed:*** If an LLM generates a different output after incorporating a context, and the RAG output is supported by the context while the initial output contradicts it, the uncertainty of the RAG-generated output should decrease.
4) ***Negatively Changed:*** If an LLM generates a different output after incorporating a context, and the initial output is supported by the context while the RAG-generated output contradicts it, the uncertainty of the RAG-generated output should increase.
5) ***Neutrally Consistent:*** If an LLM generates the same output both before and after incorporating a context, and the context is irrelevant to the output, the uncertainty in the RAG setup should remain unchanged.

Results

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

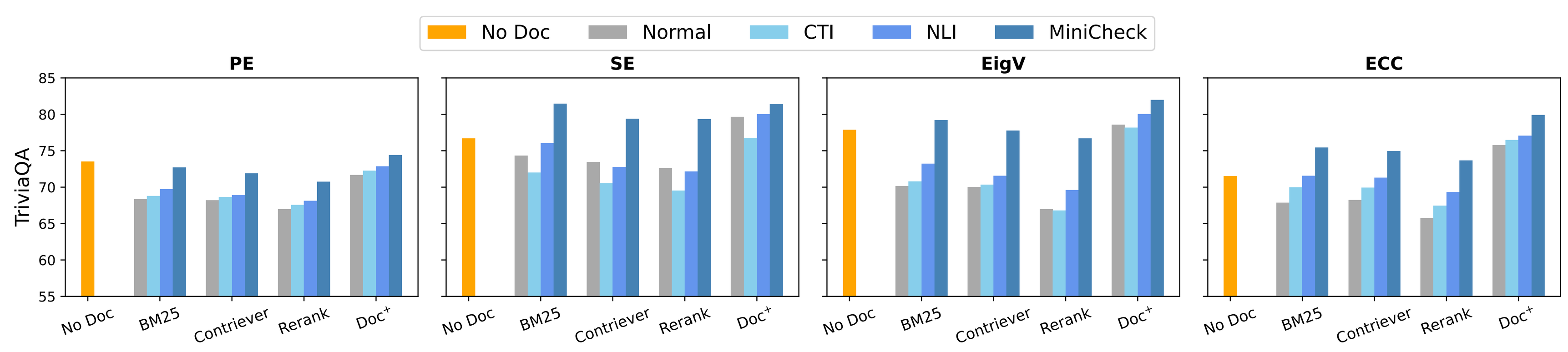
UE	PopQA		
	BM25	Contriever	Doc$^+$
Axiom 3: Positively Changed ↓			
PE	1.375 → 0.347 *	1.416 → 0.298 *	1.342 → 0.268 *
SE	4.889 → 3.015 *	5.091 → 3.013 *	4.884 → 3.051 *
PE+M	1.708 → 0.398 *	1.735 → 0.374 *	1.604 → 0.340 *
SE+M	5.514 → 3.072 *	5.681 → 3.082 *	5.379 → 3.099 *
EigV	4.131 → 1.139 *	4.733 → 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 *

4 Axiomatic Calibration

Calibration Coefficient

$$\alpha_{\text{ax}} = k_1 \cdot \mathcal{E}(r_1, r_2) + k_2 \cdot \mathcal{R}(c, q, r_1) + k_3 \cdot \mathcal{R}(c, q, r_2)$$

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



5 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 axioms*
- The result of the proposed *calibration function* shows satisfying the axioms leads to performance improvements

heydar.soudani@ru.nl

ACL 2025 VIENNA JULY 27 - AUGUST 1