Uncertainty Quantification for Retrieval-Augmented Reasoning

Heydar Soudani[1], Hamed Zamani[2], Faegheh Hasibi[1]

Radboud University[1] University of Massachusetts Amherst[2]

1 Introduction

- **Retrieval-augmented Reasoning (RAR)**

  is a recent evolution of retrieval-augmented generation (RAG) that employs multiple reasoning steps for retrieval and generation. RAR explored combining RAG with reasoning, where LLMs are prompted or trained to use search engines as tools during their reasoning process.

- **Uncertainty Quantification (UQ)**

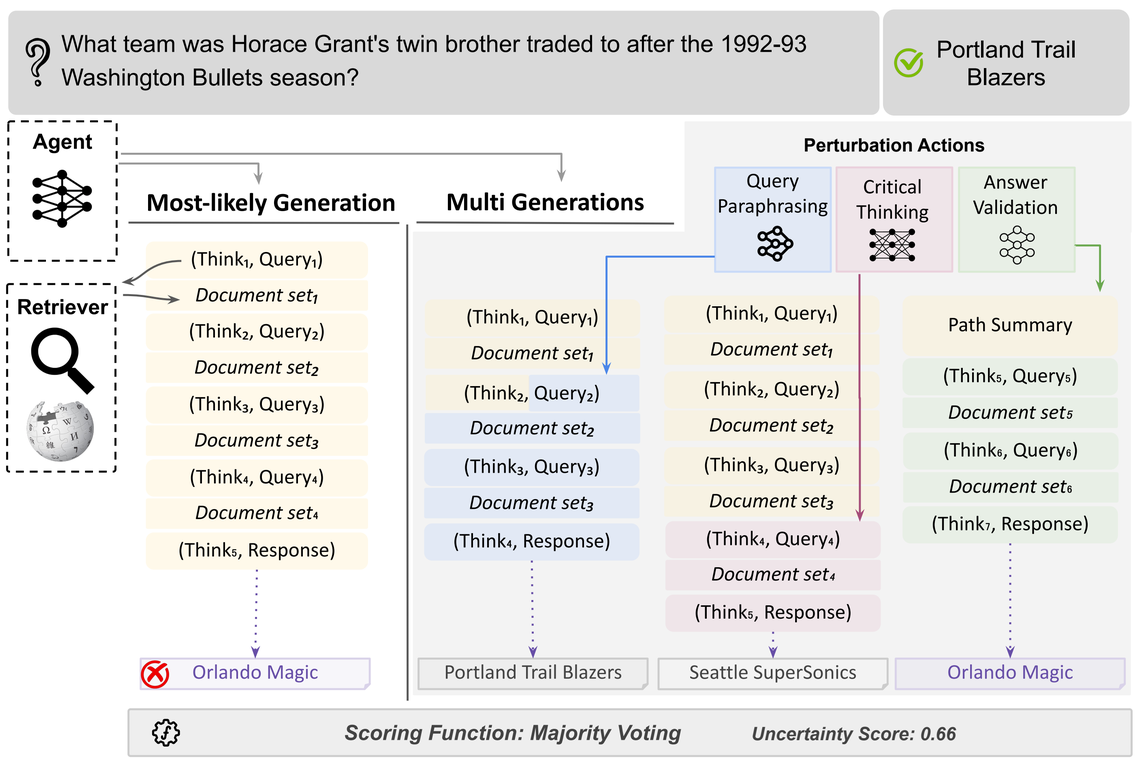
  is a widely studied task in machine learning, aimed at assessing the reliability of model outputs by measuring the degree of uncertainty (or lack of confidence) a model has in its predictions

- **Challenge:** In RAR systems we have more sources of uncertainty

  Retriever: which may provide irrelevant or partially relevant retrieved documents and potentially mislead the model's reasoning and response generation processes;
  Generation: the model's reasoning may deviate from the user query's intent and retrieved documents, leading it to formulate new search queries that fail to gather informative evidence.

2 Methodology

- **R2C: Retrieval-Augmented Reasoning Consistency**



- **Perturbation Actions**

A1: Query Paraphrasing (QP). as a query optimization mechanism that enables the system to explore alternative semantic formulations of the original query.

A2: Critical Rethinking (CR). *critically reassesses the reasoning states*

A3: Answer Validation (AV). *validate the final response by prompting the LLM to reconsider its generation once a response has been produced, based on predefined criteria.*

3 Experimental Setup

- ***Backbone LLM***
  - *Qwen2.5-instruct 7B*
- ***Retriever***
  - *Re-Rank: BM25 + Cross-encoder*
- ***RAR Models***
  - Prompt-based: *SelfAsk, ReAct, Search-o1*
  - Finetuned: *ReSearch, Search-R1*
- ***Datasets:***
  - *PopQA, HotpotQA, Musique*

4 Results

- **Direct Evaluation of UQ Estimations**

Table 1: Performance of UQ methods measured by AUROC. In each column, the best and second-best methods are indicated by bold and underline, respectively. Superscripts † and ‡ denote statistically significant differences according to the DeLong test ($p < 0.05$), compared to ReaC and P(true), respectively, which are the two best-performing methods on average.

RAG	SelfAsk [39]			ReAct [64]			Search-o1 [30]			ReSearch [5]			Search-R1 [24]			Avg.
Uncer. M.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	
PE [25]	55.34	59.11	48.61	36.75	39.95	41.14	39.86	53.44	51.59	40.93	61.03	56.69	48.49	48.27	50.90	48.81
SE [28]	64.26	68.01	54.68	49.73	40.37	41.69	67.56	65.84	54.61	53.88	64.72	63.43	59.11	53.45	61.38	57.51
MARS [3]	54.70	59.48	51.53	41.29	40.42	41.97	43.28	57.69	56.80	40.03	61.76	57.58	48.33	49.28	49.94	50.27
SAR [11]	51.65	63.05	52.31	29.56	32.90	31.54	40.92	52.37	44.66	41.67	62.75	51.47	45.87	46.40	45.22	46.16
LARS [60]	73.97	70.03	66.45	79.95	68.28	71.87	83.79	71.42	66.19	76.95	66.12	71.64	71.54	67.08	71.62	71.79
NumSS [28]	71.31	65.19	62.75	74.90	63.73	62.59	78.88	63.74	62.12	80.42	64.46	69.34	69.76	64.10	65.72	67.93
EigV [31]	70.53	66.80	62.13	76.48	66.44	57.44	79.72	68.58	64.75	80.63	66.43	66.31	69.33	66.54	64.27	68.43
ECC [31]	72.89	69.95	64.11	80.27	69.51	61.92	81.98	69.27	69.65	81.85	68.49	72.55	70.87	67.76	67.65	71.25
Deg [31]	70.53	66.79	61.77	76.70	67.70	58.12	80.69	68.87	66.29	81.67	67.15	67.85	69.51	66.75	64.53	68.99
RrrC [29]	71.14	71.17	**81.28**	48.02	68.30	75.99	65.95	73.87	77.95	68.30	71.63	74.25	68.92	71.08	70.77	70.57
SelfC [58]	74.33	69.06	68.40	80.73	75.34	72.96	81.26	72.34	76.87	82.01	72.14	77.89	71.63	69.04	70.36	74.29
ReaC [40]	77.02	70.90	76.75	81.53	76.94	74.74	81.09	74.97	77.01	82.50	75.22	77.86	73.22	72.79	75.29	76.52
P(true) [25]	76.51	77.66	78.65	75.07	73.45	71.51	78.57	73.19	74.83	84.35	77.77	81.42	75.73	76.25	74.80	76.65
R^2C(our)	**80.08**	**81.09**†	75.82	**84.75**‡	**83.25**†‡	**81.16**†‡	**87.09**†‡	**79.66**†‡	**83.22**†‡	**86.02**†	**80.76**†	**82.39**	**84.92**†‡	**79.51**†	**80.08**†‡	**81.99**

- **Extrinsic UQ Evaluation via Abstention**

Table 2: Abstention performance measured by *AbstainAccuracy* and *AbstainF1* at a 0.9 confidence threshold. For each column, the best and second-best methods are indicated in bold and underlined, respectively. A superscript † denotes a statistically significant difference compared to ReaC based on the McNemar test for Accuracy and the Bootstrap test for F1 ($p < 0.05$).

RAG	SelfAsk [39]			ReAct [64]			Search-o1 [30]			ReSearch [5]			Search-R1 [24]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
Abstain Accuracy																
RrrC [29]	61.4	65.6	73.0	60.8	72.2	82.4	56.6	66.2	68.2	59.2	64.4	64.0	60.6	64.0	63.6	65.48
P(true) [25]	70.6	70.8	80.4	70.4	67.2	75.0	72.6	65.8	75.0	78.8	71.4	78.2	70.8	70.0	74.4	72.76
SelfC [58]	68.6	64.2	76.2	74.0	75.8	87.6	75.0	69.0	**90.2**	75.2	68.6	83.6	65.4	62.8	77.8	74.27
ReaC [40]	69.2	64.6	80.4	74.4	75.4	89.4	75.6	71.4	87.4	77.2	68.6	80.8	67.8	68.8	80.6	75.44
R^2C (our)	**77.2**†	**74.4**†	**88.6**†	**77.0**	**76.8**	**90.4**	**80.4**†	**74.4**	89.4	**81.6**†	**74.8**†	**84.0**	**77.0**†	**73.4**	**84.4**†	**80.25**
Abstain F1																
RrrC [29]	62.52	70.54	82.75	71.59	80.82	89.69	56.33	71.50	78.99	54.86	63.37	73.68	55.12	61.20	73.92	69.79
P(true) [25]	73.60	75.33	88.27	72.07	73.46	84.58	76.50	71.73	84.51	81.20	73.46	85.82	71.03	71.15	83.37	77.74
SelfC [58]	72.60	69.51	85.71	79.10	83.80	93.18	80.50	78.07	**94.64**	81.71	75.19	**90.57**	62.31	62.34	86.51	79.72
ReaC [40]	73.54	70.45	88.41	79.01	83.17	94.19	81.24	79.48	93.03	81.25	72.60	88.43	67.60	71.11	88.27	80.79
R^2C (our)	**82.35**†	**81.76**†	**93.77**†	**83.69**†	**85.08**†	**94.81**	**86.42**†	**82.22**†	94.27†	**84.40**†	**78.71**†	90.49†	**80.80**†	**77.57**†	**90.97**†	**85.82**

- **Extrinsic UQ Evaluation via Model Selection**

Table 3: Model Selection performance measured by exact match. The superscript † denotes a statistically significant difference from the best-performing baseline (underlined), according to the Wilcoxon test ($p < 0.05$).

RAG System	PopQA	HotpotQA	Musique	Average
Vanilla LLM & RAG				
Direct	18.8	20.8	2.6	14.1
CoT	17.6	22.2	5.8	15.2
Vanilla RAG	30.2	18.6	4.4	17.7
IRCoT [56]	34.6	27.2	5.4	22.4
FLARE [23]	31.6	25.2	8.4	21.7
DRAGIN [51]	28.6	23.2	4.6	18.8
Retrieval-Augmented Reasoning (RAR)				
SelfAsk [39]	35.6	33.0	10.4	26.3
ReAct [64]	36.8	27.8	10.8	25.1
Search-o1 [30]	33.2	29.0	10.0	24.1
ReSearch [5]	38.6	38.8	16.6	31.3
Search-R1 [24]	41.6	41.4	16.0	33.0
Model Selection RAR				
Random	31.4	31.6	10.6	24.5
LLMBlender [20]	34.4	36.0	12.0	27.5
- *w/o clustering*	35.6	31.8	9.6	25.7
- *w/o clus. & unc.*	32.0	26.0	8.6	22.2
RAGEnsemble [6]	45.6	46.0	19.4	37.0
- *w/o clustering*	44.4	40.8	18.0	34.4
- *w/o clus. & unc.*	43.2	37.2	13.0	31.1
R^2CSelect (our)	**46.8**†	**50.2**†	**22.6**†	**39.9**
- *w/o clustering*	45.4	44.0	19.2	36.2
Ideal Model Selection	55.0	57.2	30.0	47.4

5 Takeaways

- R2C, a novel theocratically grounded UQ method based on MDP; the first of its kind that captures different sources of uncertainty in RAR
- We conduct extensive experiments on three dataset and five RAR methods, demonstrating the superiority of the proposed method on the UQ task
- We show the effectiveness of our method on both model selection and abstention tasks

heydar.soudani@ru.nl