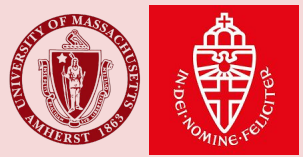
Uncertainty Quantification for Retrieval-Augmented Reasoning

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

[1]Radboud University

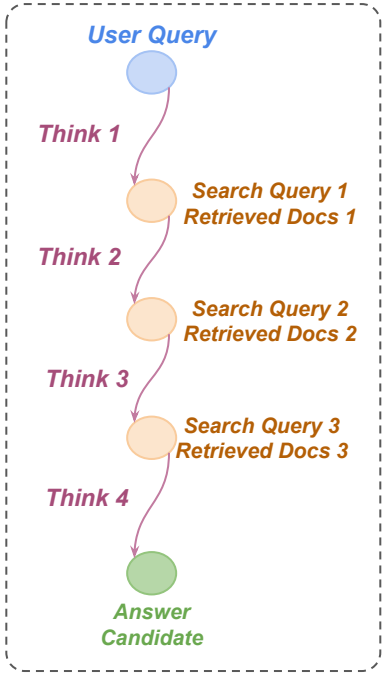
[2]University of Massachusetts Amherst

July 2026



Retrieval-Augmented Reasoning (RAR)

***Example:* Search-R1, GLM, GPT-oss, …**



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

Think 1: I need to find the type of profession that Chris Jericho and Gary Barlow have in common. I'll search for it
Search Query 1: Chris Jericho and Gary Barlow profession
Retrieved Docs 1:
Doc 1 (Title: "Chris Jericho") "Do You Wanna …"
Doc 2 (Title: "Fozzy") "Christopher Keith Irvine (born …"
…

Think 2: I found out that Chris Jericho is a professional wrestler, musician, author …
Search Query 2: Gary Barlow profession
Retrieved Docs 2:
Doc 1(Title: "Gary Barlow") hits, twelve number one …
Doc 2(Title: "Professional wrestling") Professional wrestling …"
…

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

Retrieval-Augmented Reasoning (RAR)

Example: Search-R1, GLM, GPT-oss, …

Prone to incorrect responses

- Retrieving irrelevant documents
- Misinterpreting retrieved content
- Misusing internal knowledge

Grounded and well-reasoned ≠ Trustworthy !!

Answer Candidate

wrestler, musician, author …
Answer: musician

Uncertainty Quantification (UQ)

Formal Definition

A task aimed at assessing the reliability of system outputs by measuring the degree of uncertainty, the system's lack of confidence in its own prediction

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR.

Uncertainty Quantification (UQ)

Formal Definition

A task aimed at assessing the reliability of system outputs by measuring the degree of uncertainty, the system's lack of confidence in its own prediction

Confidence vs. Uncertainty

- Some literature distinguishes uncertainty from confidence:
  - **U(x)**: dispersion over outputs given input *x*
  - **C(x, r)**: estimated correctness of *r* given *x*
- We follow the literature that treats uncertainty as a lack of confidence
- Use the two terms interchangeably

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR.

Type of Uncertainty

- **Epistemic**: Uncertainty due to the model not knowing something (reducible)
- **Aleatoric**: Uncertainty due to the data being fundamentally unclear (irreducible)

Total uncertainty: Encompasses epistemic and aleatoric uncertainty

Why Total?

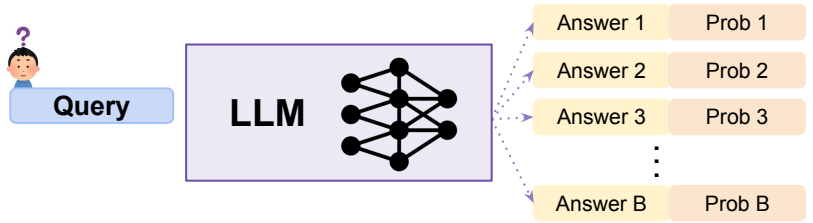
In language models these two parts are very hard to separate

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR.

UQ for LLMs

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

Multi Generations



1) Perturbing the input
2) Changing the decoding temperature

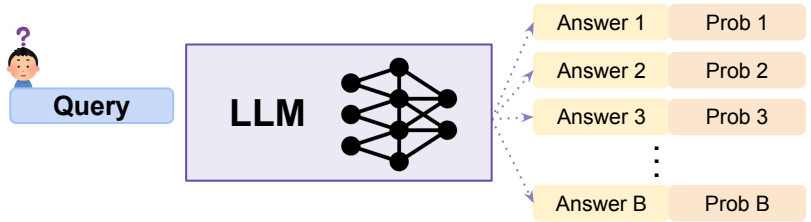
[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.
[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.
[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

UQ for LLMs

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

Score Function: Convert the generations to an uncertainty / confidence score

Multi Generations



1) Perturbing the input
2) Changing the decoding temperature

Score Function

White-box Methods	Black-box Methods	Trained Models

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.
[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.
[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

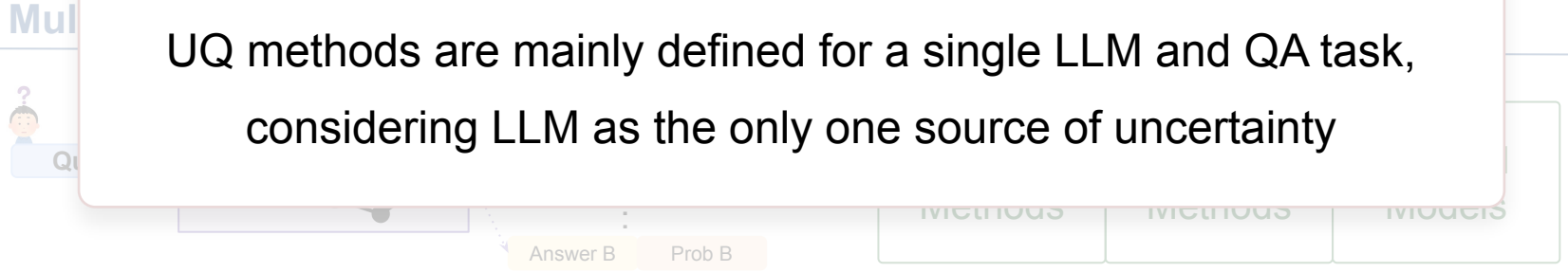
UQ for LLMs

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

Score Function: Convert the generations to an uncertainty / confidence score

!!! Limitation

UQ methods are mainly defined for a single LLM and QA task, considering LLM as the only one source of uncertainty

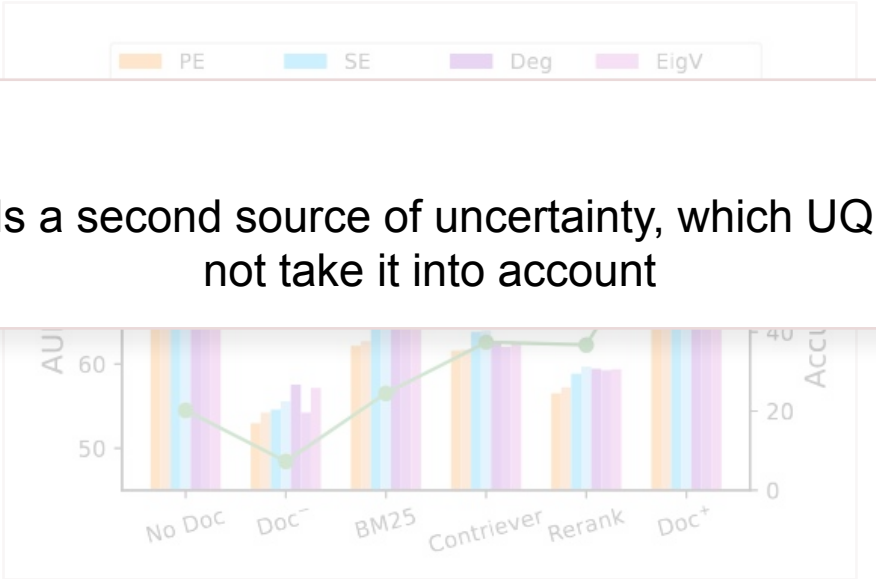


1) Perturbing the input
2) Changing the decoding temperature

[1] Kuhn et al. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in NLG. ICLR.
[2] Lin et al. (2024). Generating with Confidence: Uncertainty Quantification for Black-box LLMs. TMLR.
[3] Wang et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.

UQ Methods Do Not Transfer to RAG

Adding retrieved context degrades UQ performance



!!! Reason

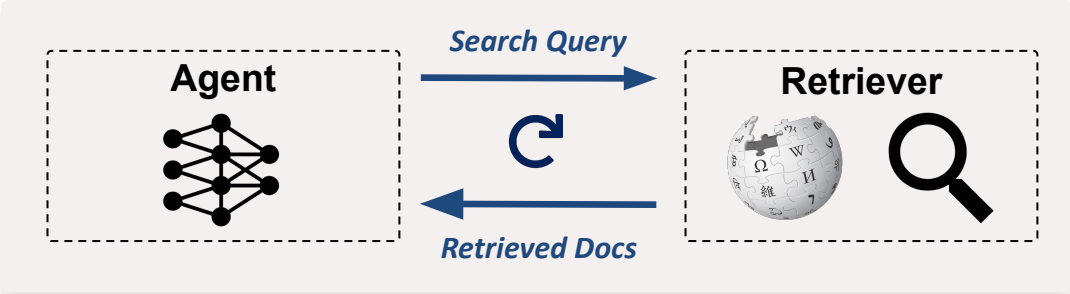
Retrieval adds a second source of uncertainty, which UQ methods do not take it into account

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

Research Gap: UQ for RAR

The problem gets even harder for RAR

- Two sources: Agent & Retriever
- Iterative interaction: each step's output is the next step's input
- Uncertainty propagates, accumulates, or discarded in the next steps



[1] Jin et al. (2025). Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. CoRR.
[2] Chen et al. (2025). ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. CoRR.
[3] Yao et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. ICLR.

The Core Challenge

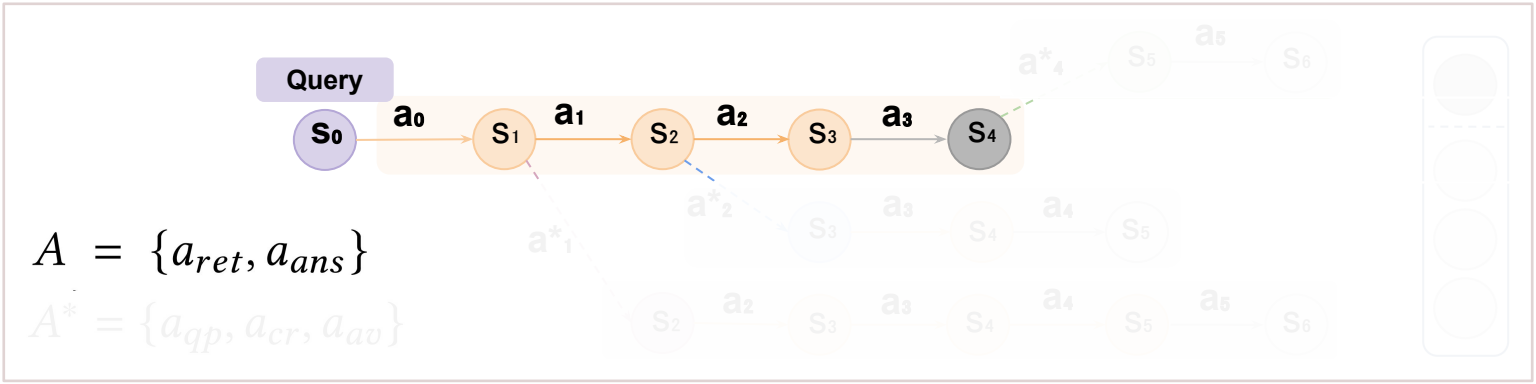
Accurate UQ for RAR must account for *both* sources of uncertainty; retriever and generator, across the whole reasoning path

Methodology

Reasoning Consistency for RAR: R2C

1) Model the sequential interaction in RAR as Markov Decision Process (MDP)

$$s_t \leftarrow \pi_\theta(s_{t-1}, a_{t-1}); \quad t = 1, \ldots, N,$$



$$A = \{a_{ret}, a_{ans}\}$$

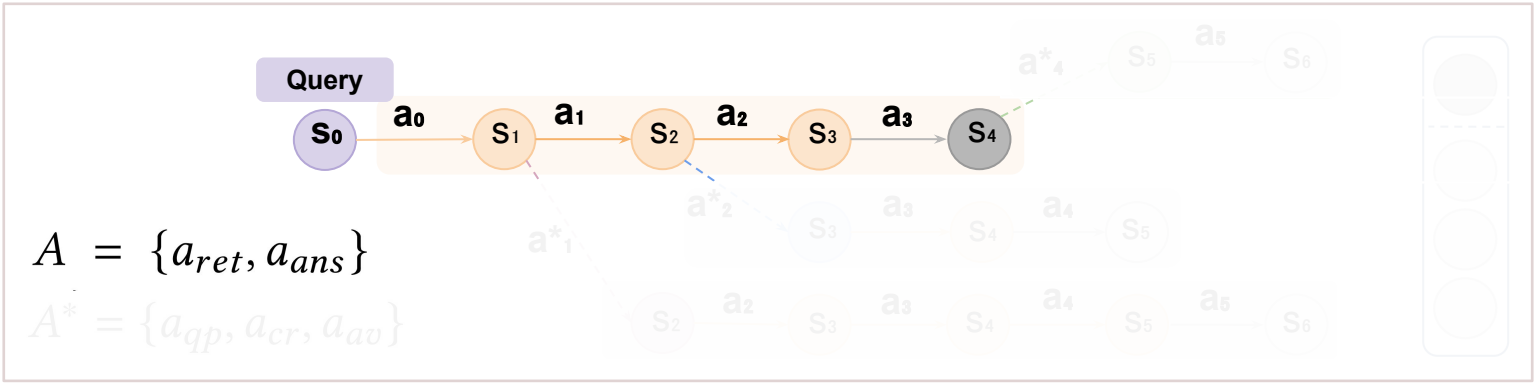
$$A^* = \{a_{qp}, a_{cr}, a_{av}\}$$

Reasoning Consistency for RAR: R2C

1) Model the sequential interaction in RAR as Markov Decision Process (MDP)

$$s_t \leftarrow \pi_\theta(s_{t-1}, a_{t-1}); \quad t = 1, \ldots, N,$$

Generate the most-likely trajectory → most-likely answer
This is the answer whose confidence we measure

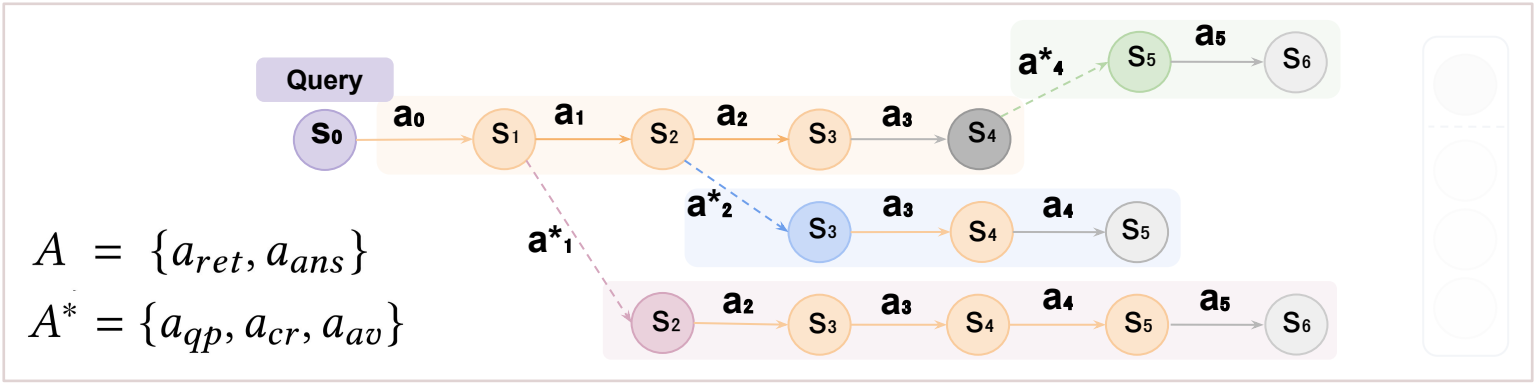


Reasoning Consistency for RAR: R2C

2) Generate multiple answers:

- At one random state, replace the ***main action set*** with the ***perturbation action set***
- Take one perturbed action, then continue with the main action set

Perturbations stress-test how sensitive the trajectory is

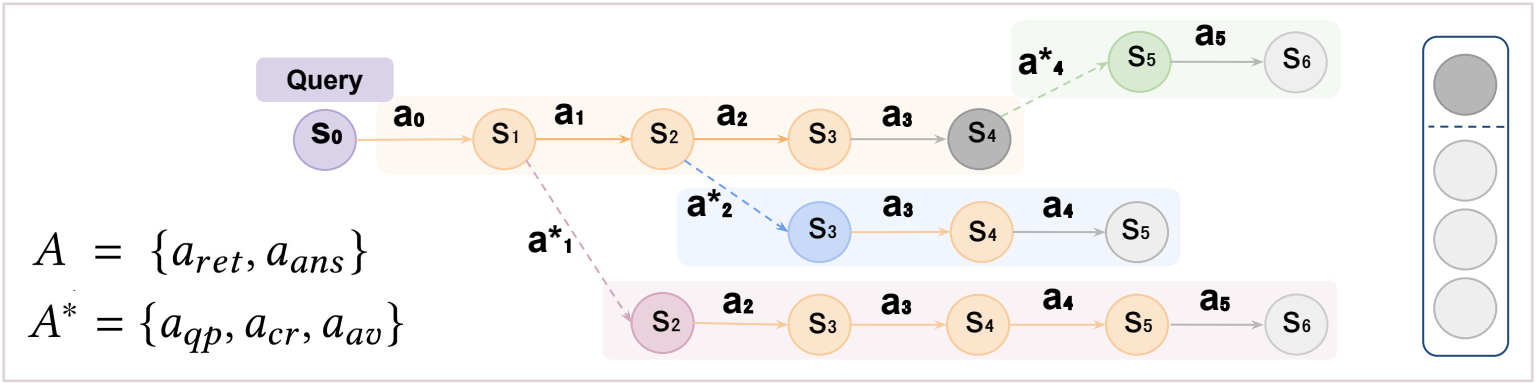


Reasoning Consistency for RAR: R2C

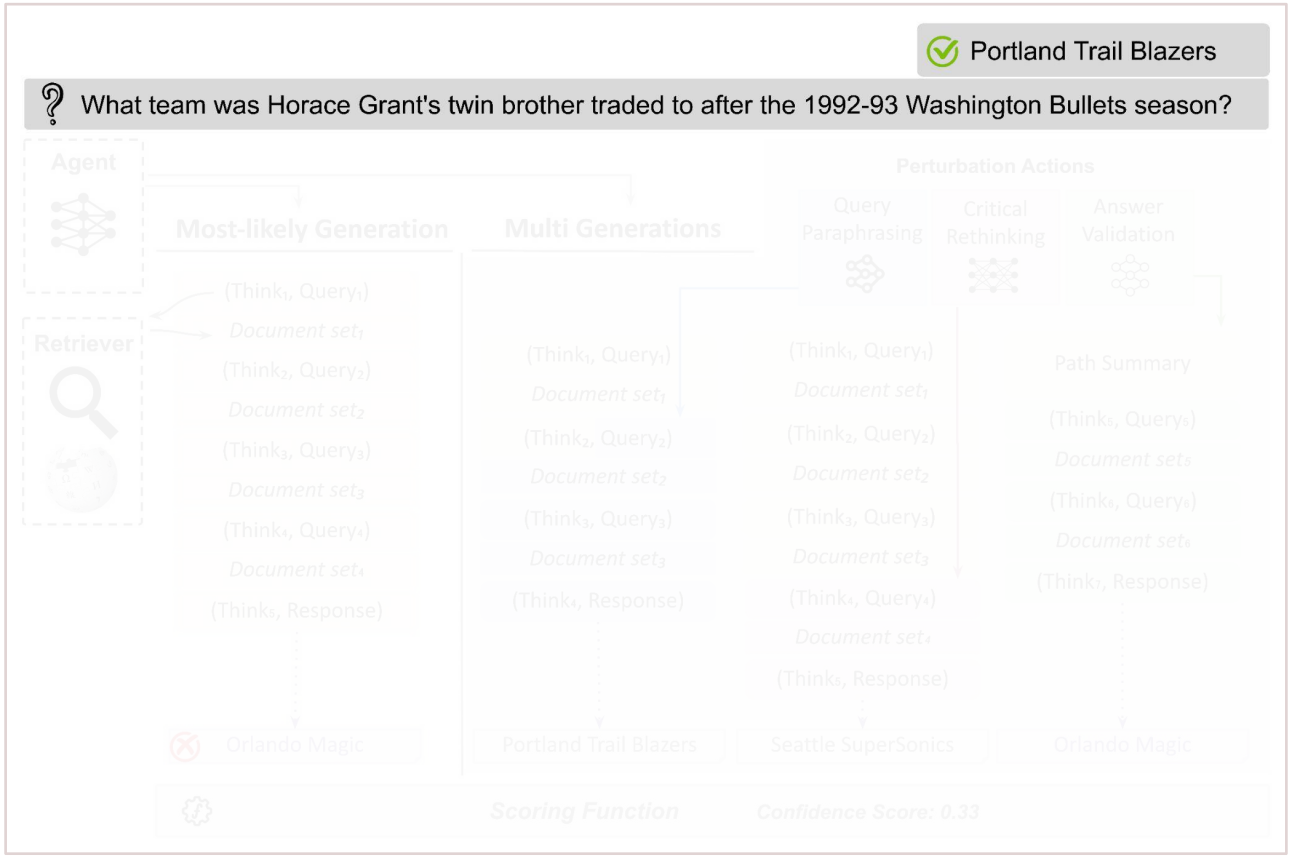
3) Compute the consistency score:

In our case: proportion of sampled responses that match the most-likely response

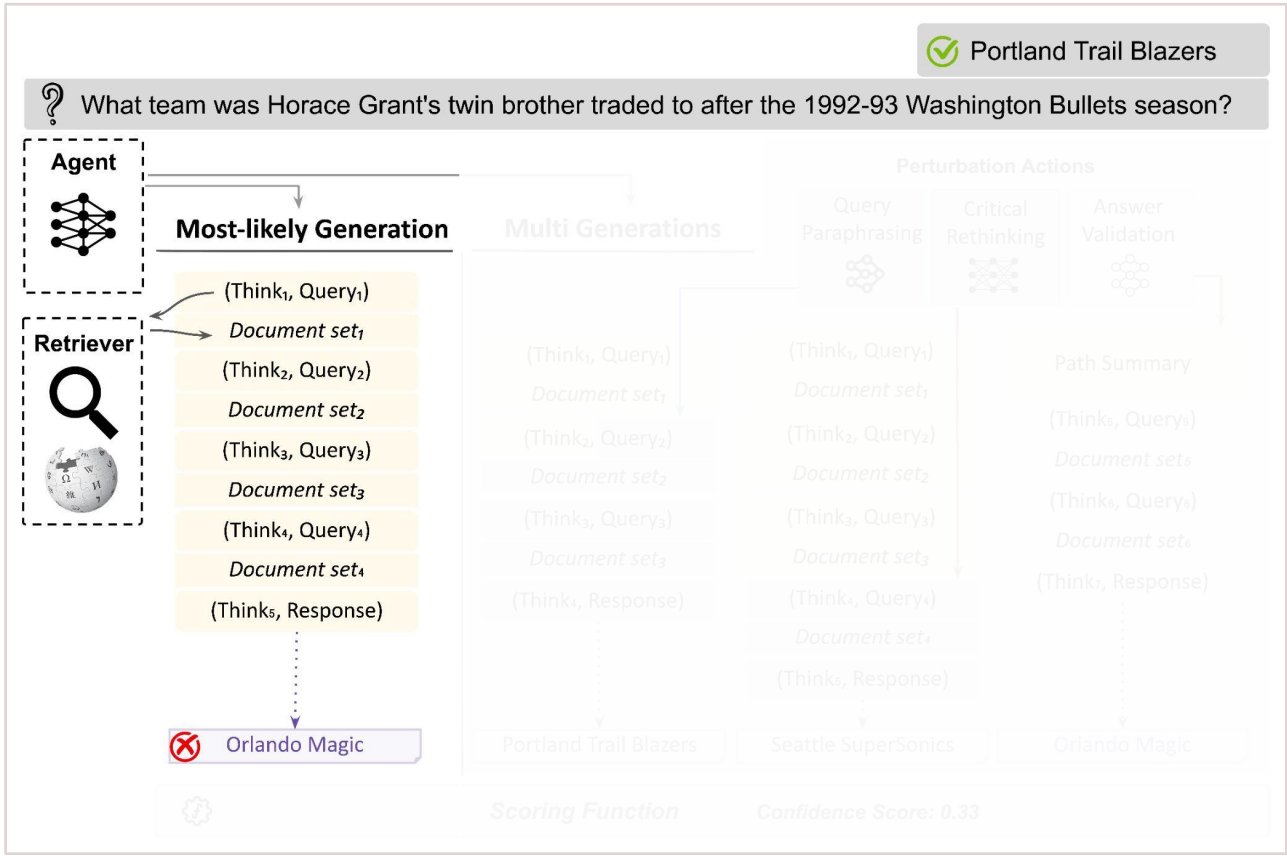
$$C(x, r) = \phi(R_G, r) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(r^b \equiv r)$$



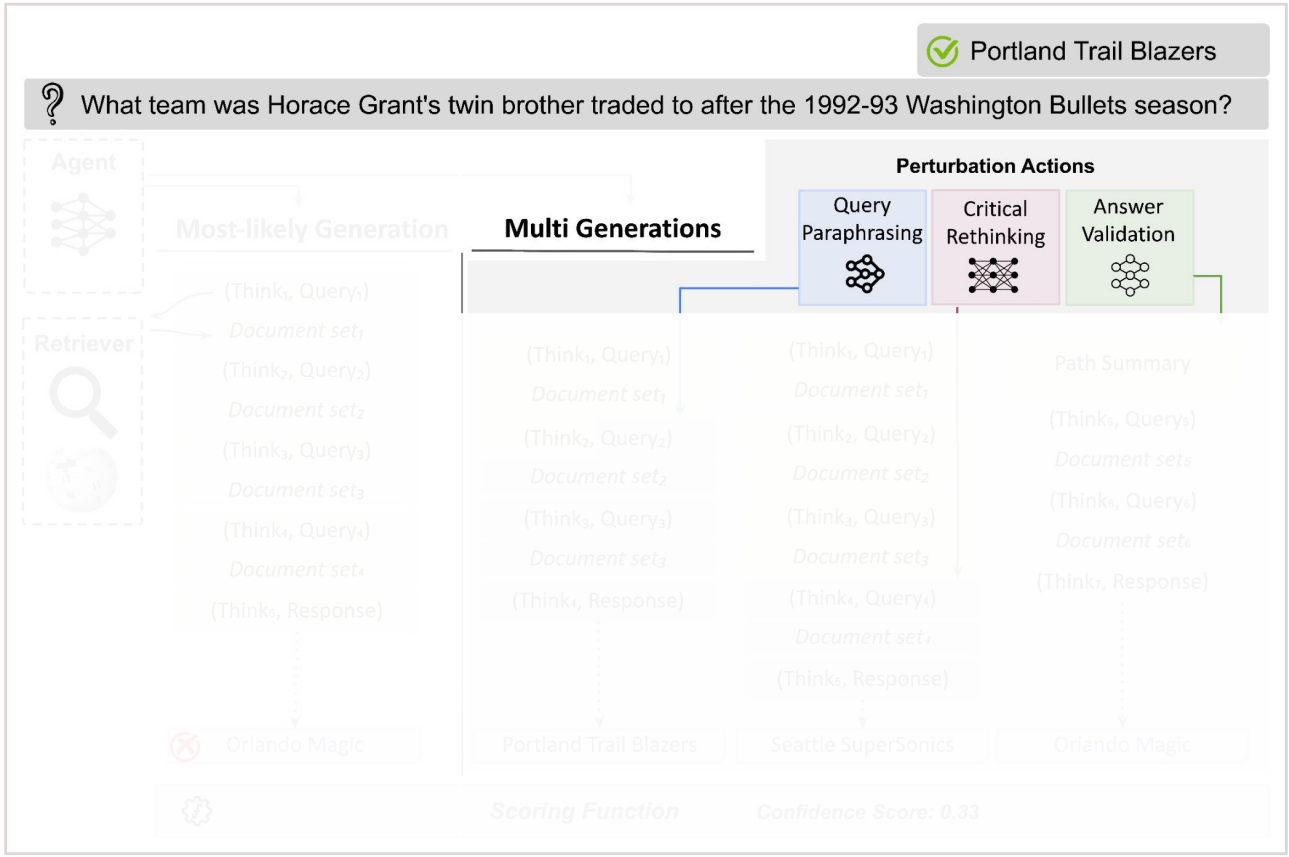
R2C: Example


Portland Trail Blazers
What team was Horace Grant's twin brother traded to after the 1992-93 Washington Bullets season?
Agent
Retriever
Most-likely Generation
Multi Generations
Perturbation Actions
Query Paraphrasing
Critical Rethinking
Answer Validation
Path Summary
Orlando Magic
Portland Trail Blazers
Seattle SuperSonics
Orlando Magic
Scoring Function
Confidence Score: 0.33

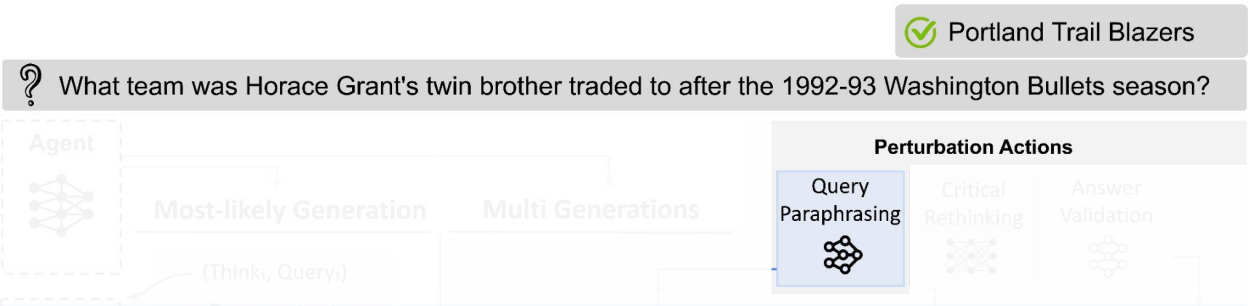
R2C: Most-likely Generation



R2C: Perturbation Action Set



R2C: Query Paraphrasing Action



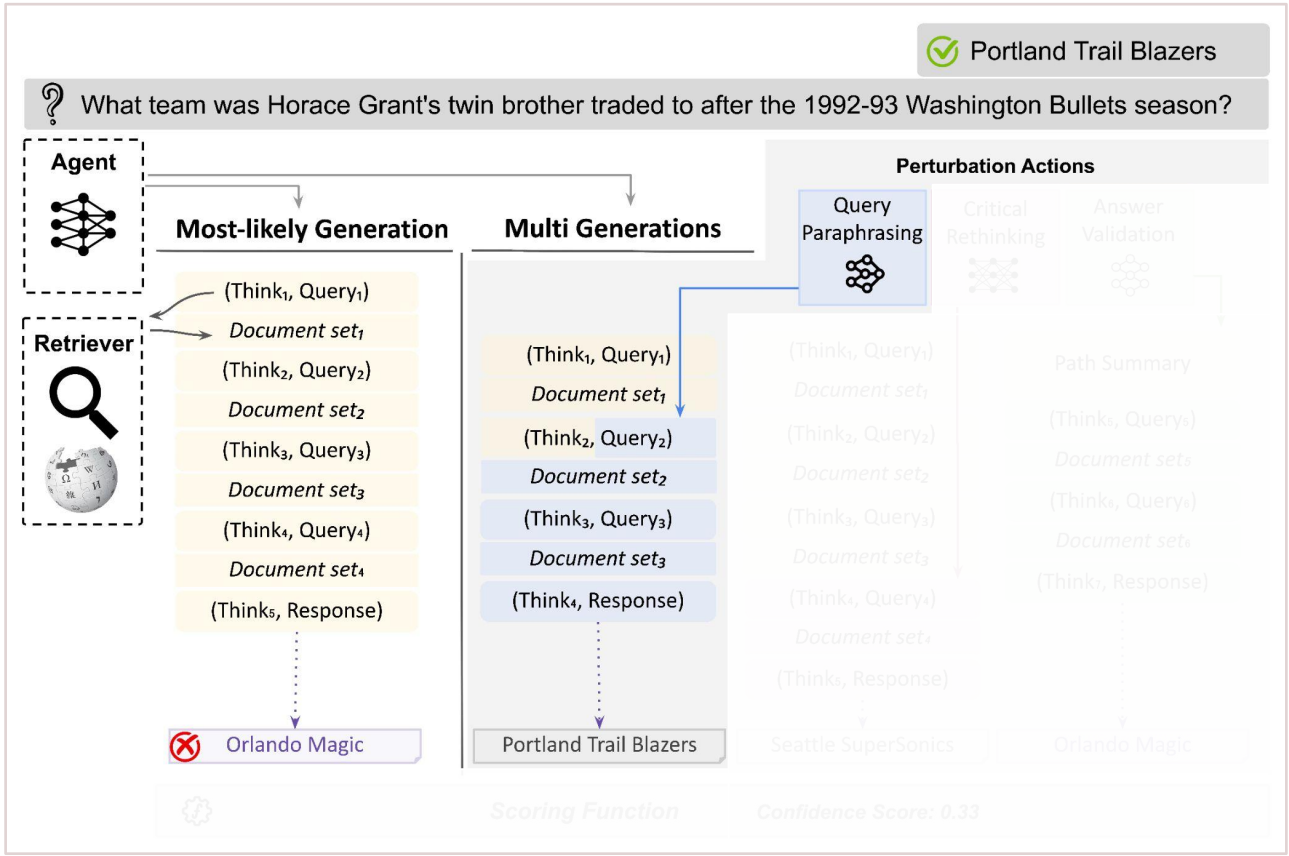
Query Paraphrasing

Uncertainty source: Retriever, directly alters its input

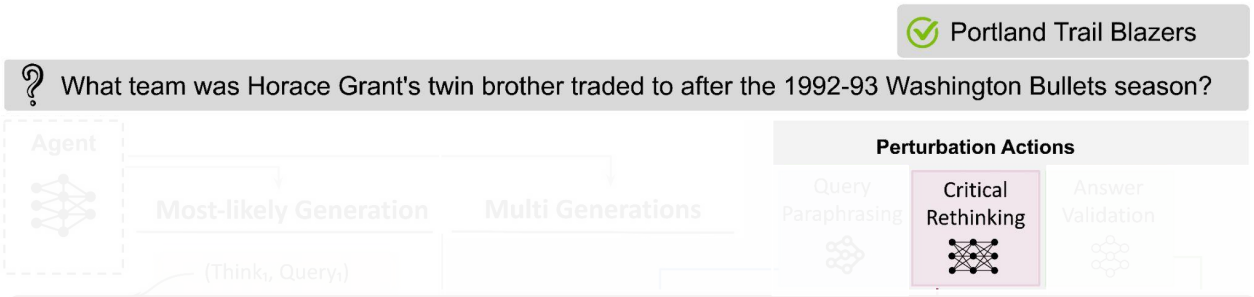
Why it redirects: LLMs aren't optimized to write good search queries; QP acts as query optimization

What it tests: Whether the reasoning path is fragile enough that rewording a query changes its direction

R2C: Query Paraphrasing Action



R2C: Critical Rethinking Action



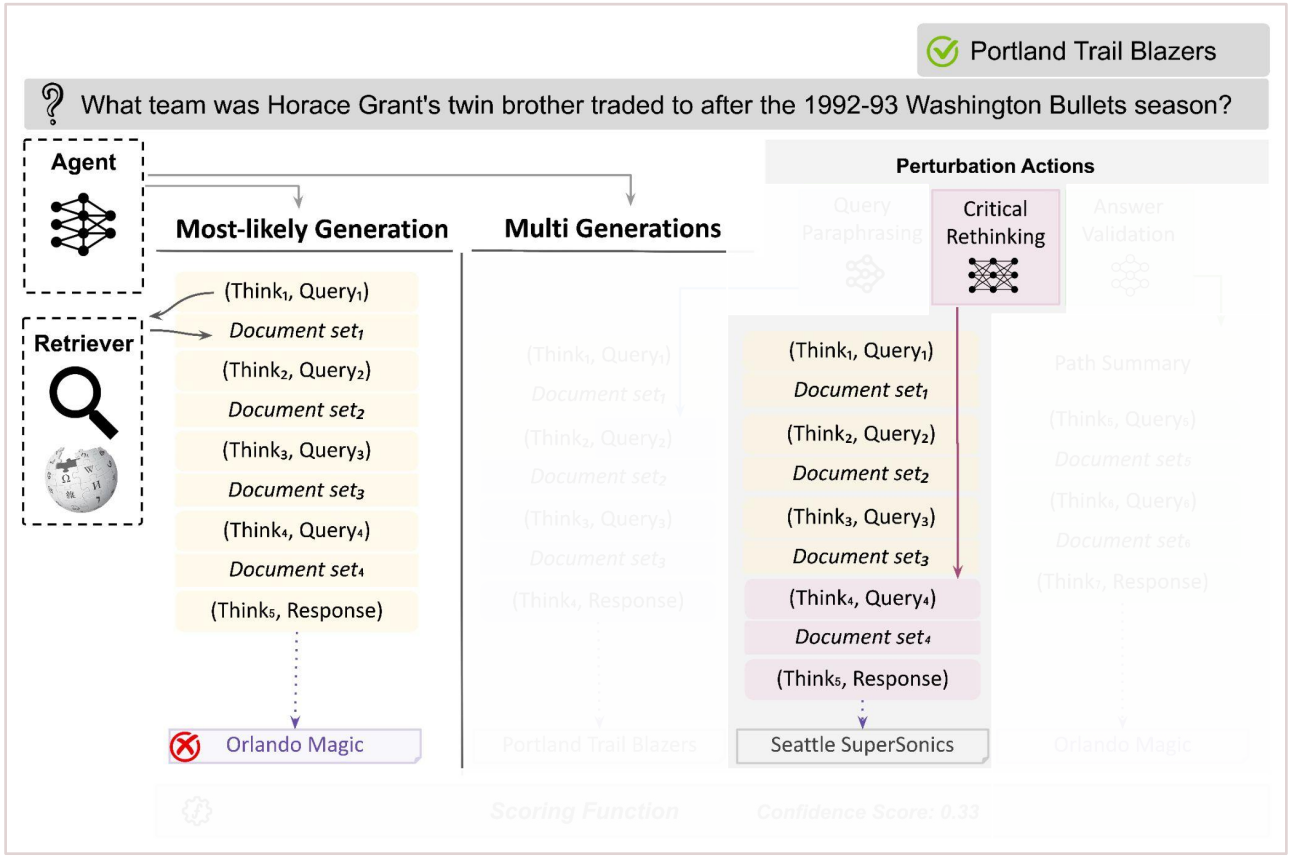
Critical Rethinking

Uncertainty source: Agent, injects a critical step into its reasoning

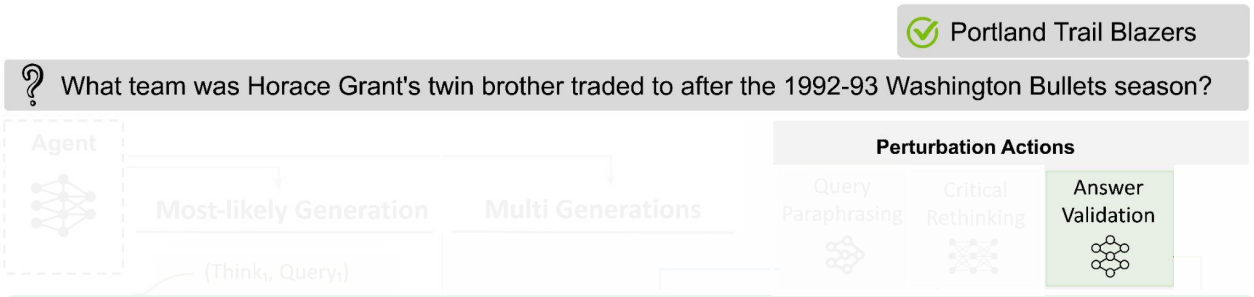
Why it redirects: RAR models under-criticize themselves and miss noisy or irrelevant retrieved evidence

What it tests: Whether the reasoning path can correct itself, An incorrect path may move toward a better one, while a correct path stays consistent

R2C: Critical Rethinking Action



R2C: Answer Validation Action



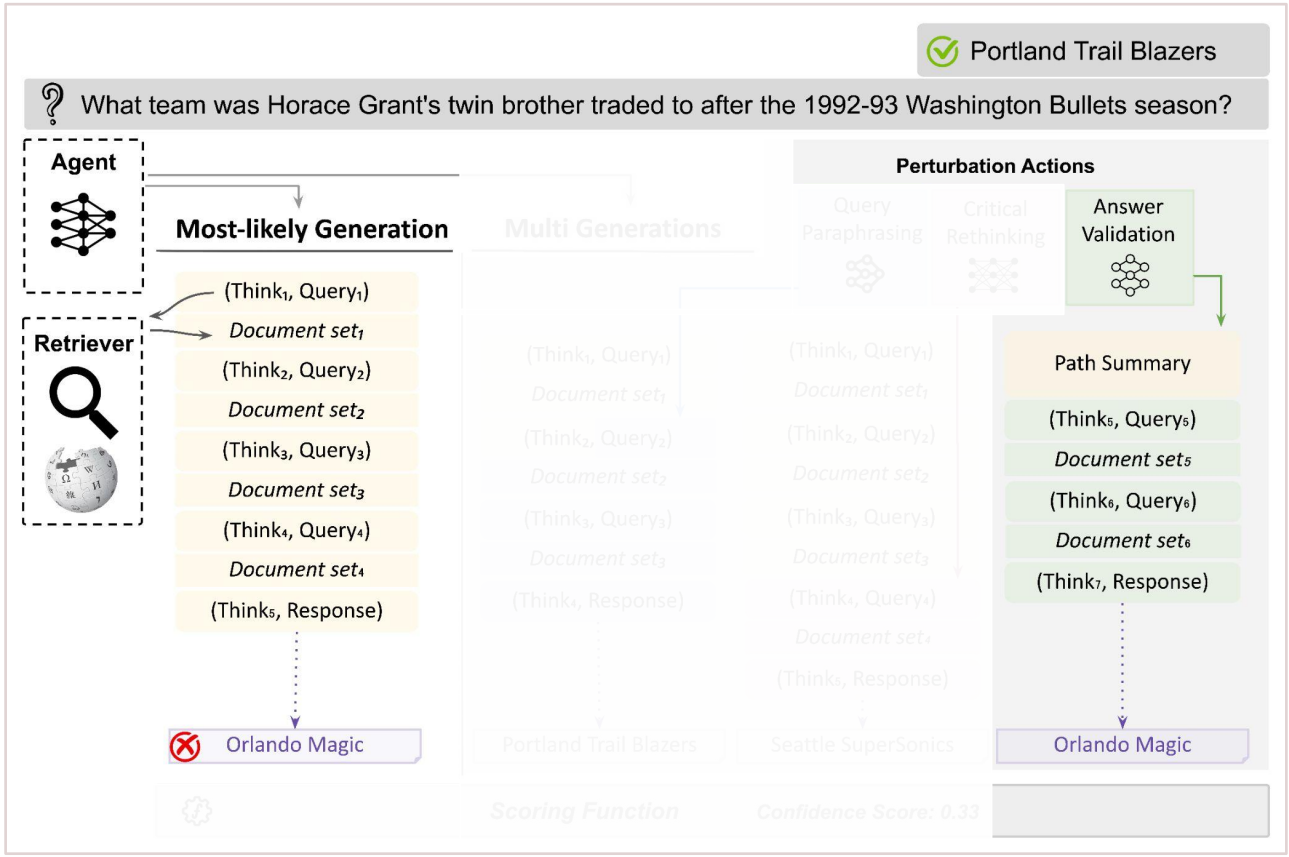
Answer Validation

Uncertainty source: Agent and retriever, judges the completed path before continuing

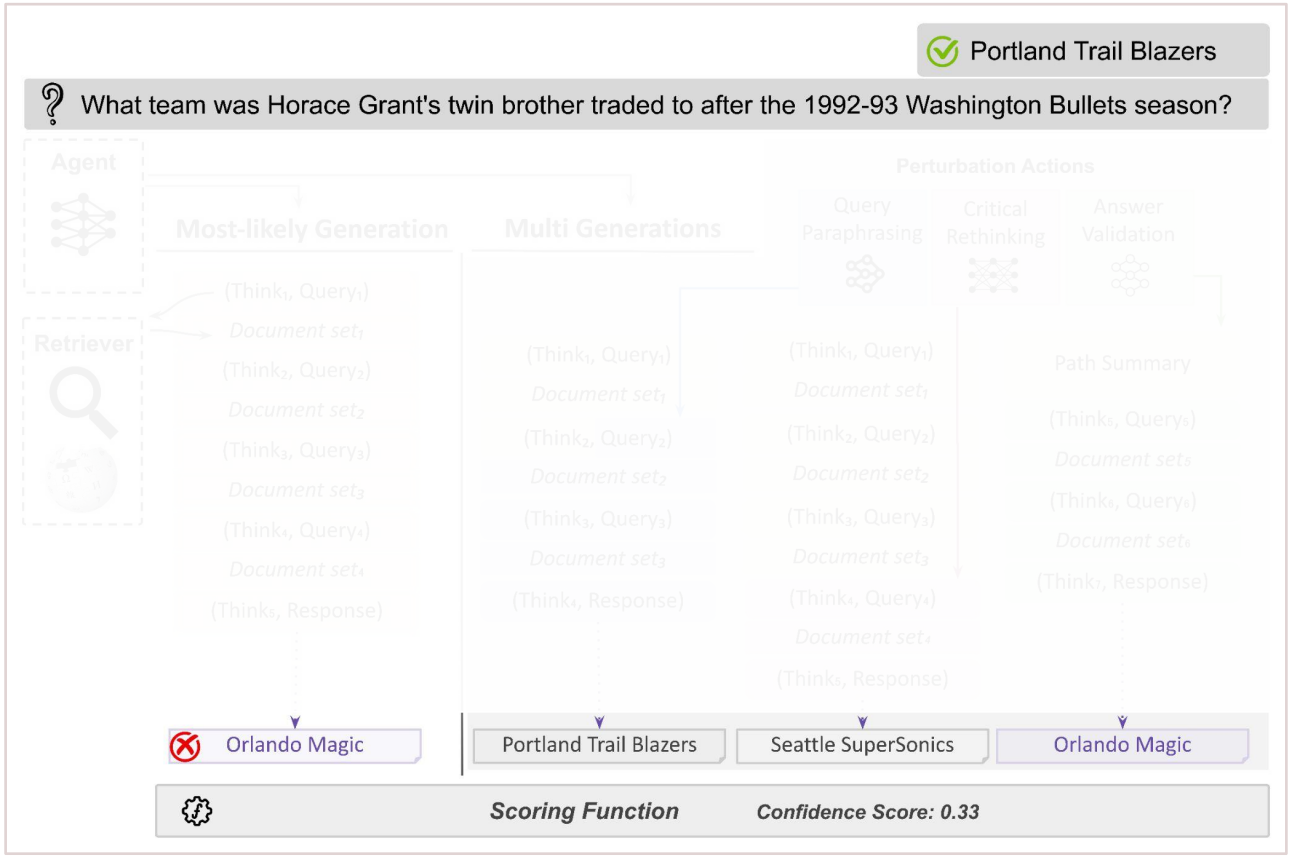
Why it redirects: RAR models struggle to verify that a response actually satisfies the query

What it tests: Whether the answer is valid. If it is, the model returns it; if not, the model begins a new reasoning process

R2C: Answer Validation Action

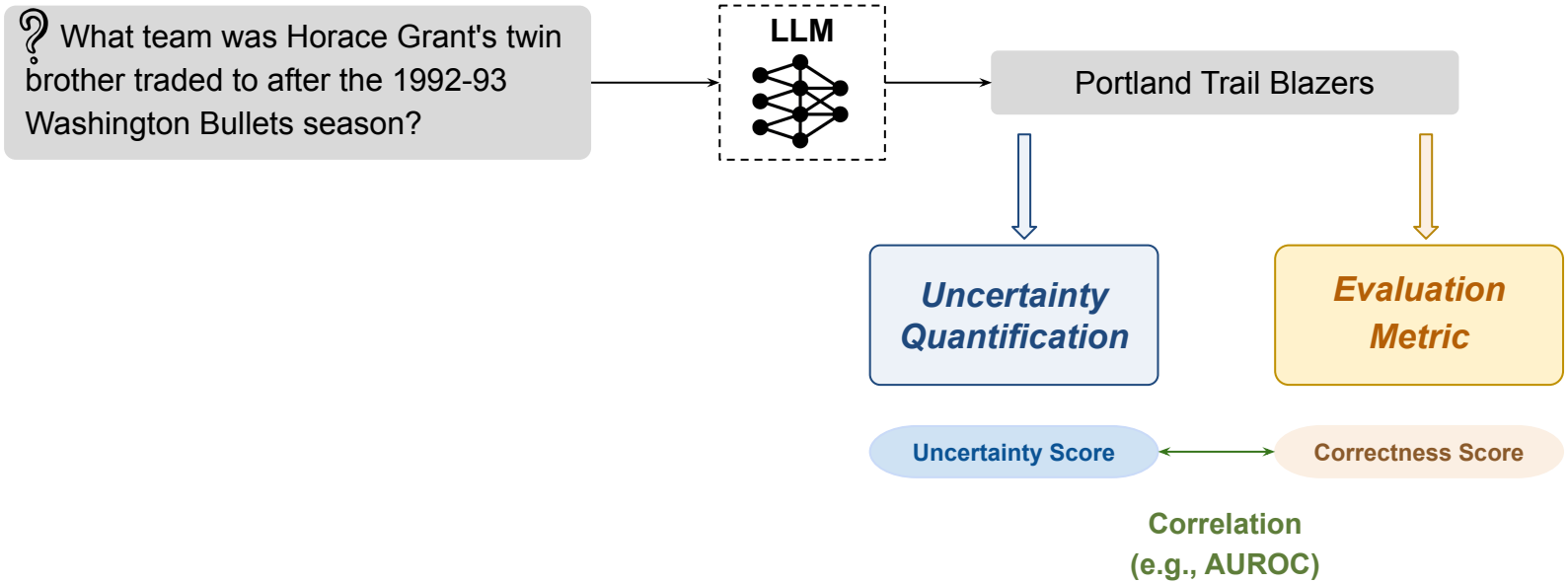


R2C: Compute the Score



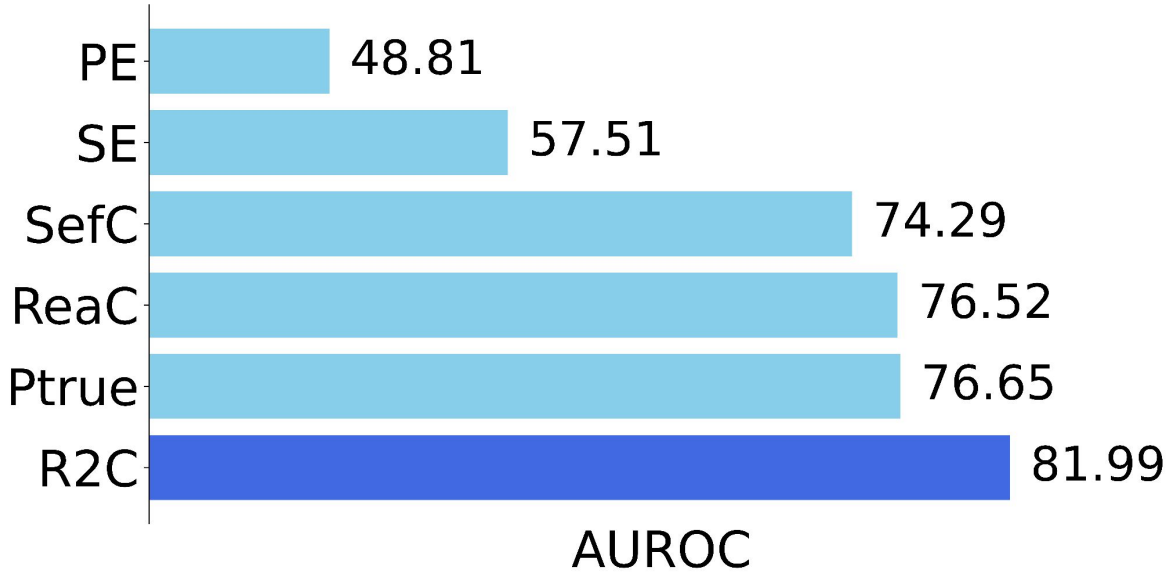
Evaluation

UQ Method Evaluation



Results

How does R2C perform in quantifying uncertainty for different RAR models?



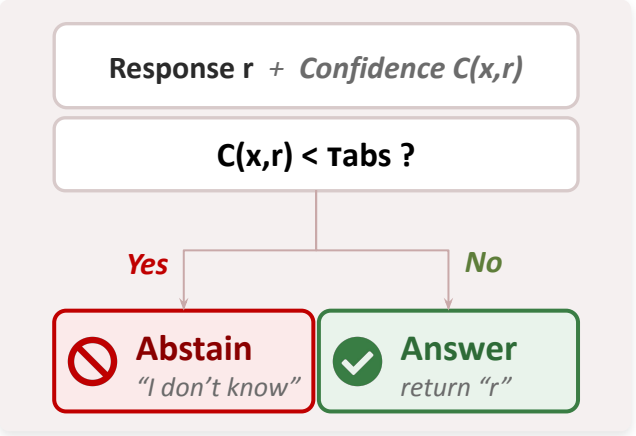
RAG	SelfAsk [47]			ReAct [74]			Search-o1 [36]			ReSearch [6]			Search-R1 [29]			Avg.
Uncer. M.	Popqa												p.	Musi.		
PE [30]	55.34												27	50.90		48.81
SE [34]	64.26												45	61.38		57.51
MARS [3]	54.70												28	49.94		50.27
SAR [12]	51.65												40	45.22		46.16
LARS [70]	73.97												08	71.62		71.79
NumSS [34]	71.31												10	65.72		67.93
EigV [37]	70.53												54	64.27		68.43
ECC [37]	72.89												76	67.65		71.25
Deg [37]	70.53												75	64.53		68.99
RrrC [35]	71.14												08	70.77		70.57
SelfC [66]	74.33												04	70.36		74.29
ReaC [48]	77.02												79	75.29		76.52
P(true) [30]	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	$81.09^{\dagger}$	75.82	$84.75^{\ddagger}$	$83.25^{\dagger\ddagger}$	$81.16^{\dagger\ddagger}$	$87.09^{\dagger\ddagger}$	$79.66^{\dagger\ddagger}$	$83.22^{\dagger\ddagger}$	$86.02^{\dagger}$	$80.76^{\dagger}$	82.39	$84.92^{\dagger\ddagger}$	$79.51^{\dagger}$	$80.08^{\dagger\ddagger}$	81.99

Task 1: Abstention

Declining to respond due to uncertainty

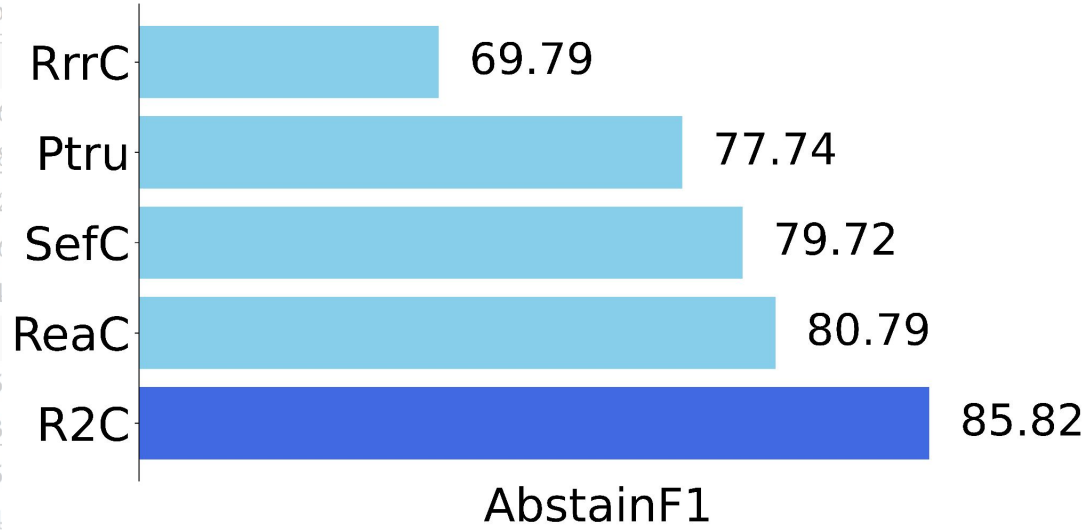
- ***C***: Confidence function
- **τ**: Threshold value
- ***f***: Abstention function

$$f_{\text{abs}}(r) = \begin{cases} \text{true}, & \text{if } C(x,r) < \tau_{\text{abs}} \\ \text{false}, & \text{otherwise.} \end{cases}$$



Abstention: Result

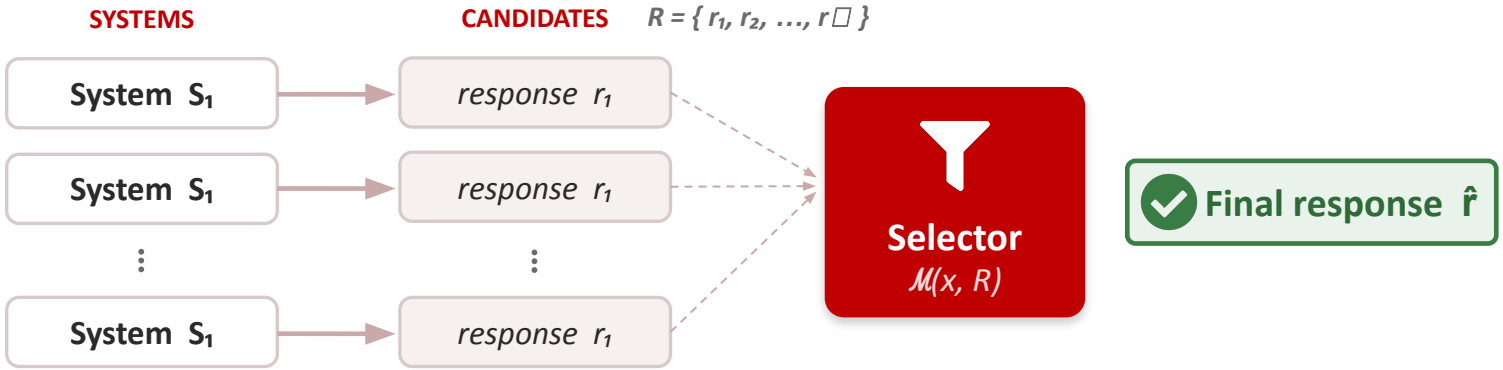
How does R2C perform as an external signal on downstream tasks (Abstention)?



RAG	SelfAsk [47]			ReAct [74]			Search-o1 [36]			ReSearch [6]			Search-R1 [29]			Avg.
Uncer. M.	Popqa	Hotp											qa	Hotpot	Musiq.	
Abstain Accuracy																
RrrC [35]	61.4	65.(											6	64.0	63.6	65.48
P(true) [30]	70.6	70.8											8	70.0	74.4	72.76
SelfC [66]	68.6	64.2											4	62.8	77.8	74.27
ReaC [48]	69.2	64.(											8	68.8	80.6	75.44
R^2C (our)	77.2†	74.4											)†	73.4	84.4†	80.25
Abstain F1																
RrrC [35]	62.52	70.5											2	61.20	73.92	69.79
P(true) [30]	73.60	75.3											)3	71.15	83.37	77.74
SelfC [66]	72.60	69.5											31	62.34	86.51	79.72
ReaC [48]	73.54	70.4											50	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

Task 2: Model Selection

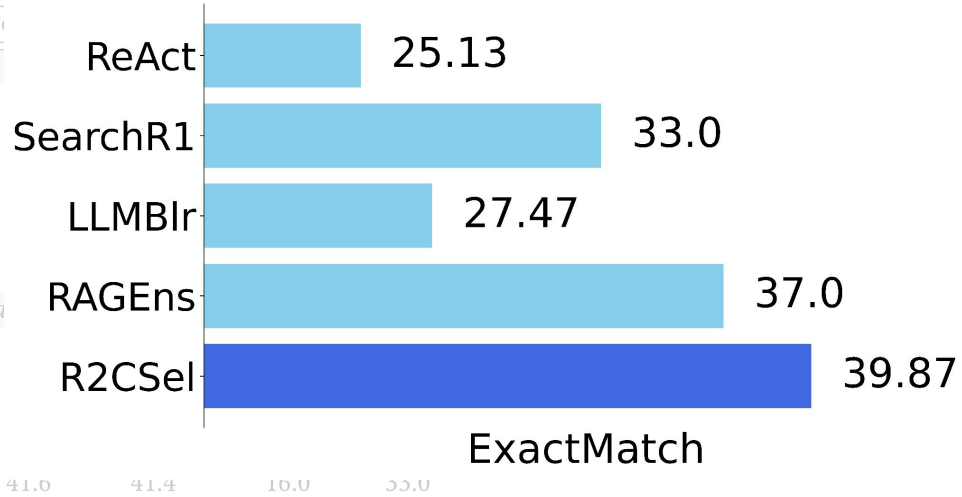
Select a final response based on multiple candidate responses generated by different systems



Model Selection: Result

How does R2C perform as an external signal on downstream tasks (Model Selection)?

RAG System	P
Vanilla LLM & RAG	
Direct	
CoT	
Vanilla RAG	
IRCoT [64]	
FLARE [28]	
DRAGIN [59]	
Retrieval-Augmented Rea	
SelfAsk [47]	
ReAct [74]	
Search-o1 [36]	
ReSearch [6]	
Search-R1 [29]	

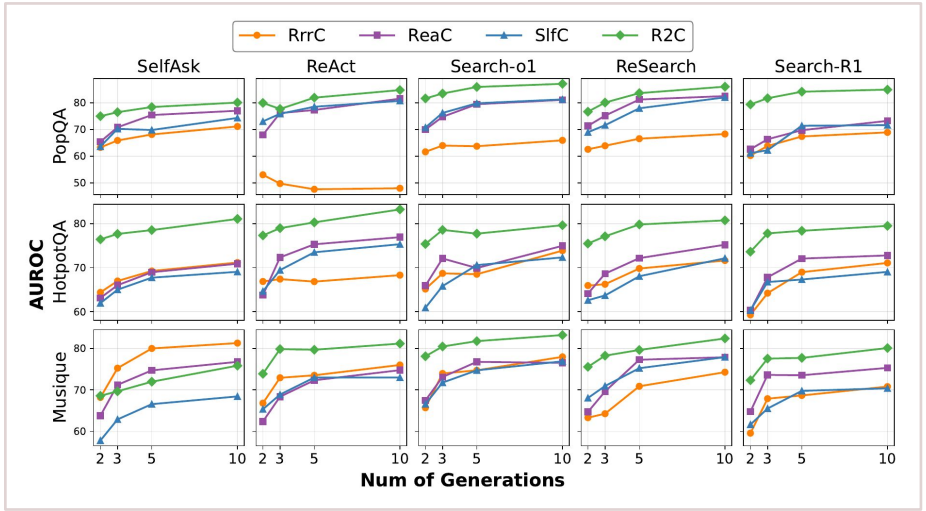


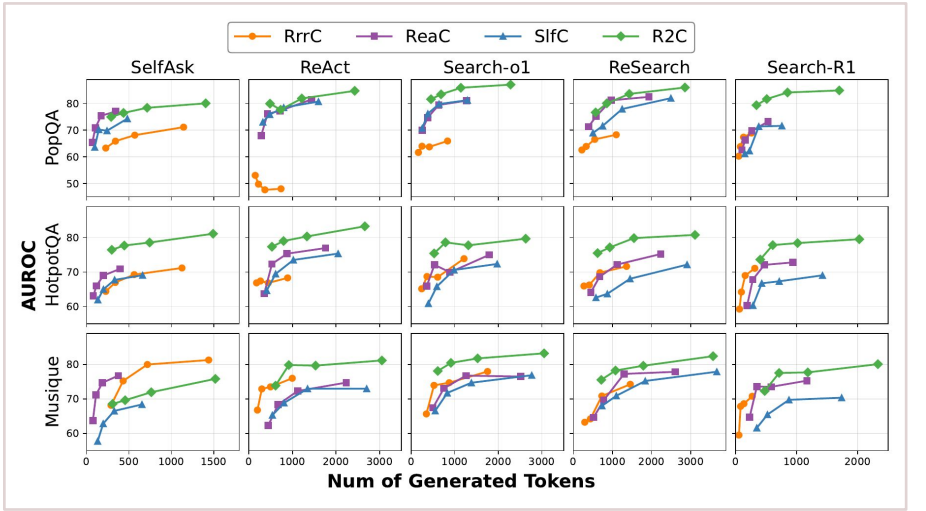
QA Musique	Average
10.6	24.5
12.0	27.5
9.6	25.7
8.6	22.2
19.4	37.0
18.0	34.4
13.0	31.1
22.6†	**39.9**
19.2	36.2
30.0	47.4

R2C score is a reliable criterion, not only within a single system but also across different systems

Effectiveness vs. Efficiency

How does R2C balance effectiveness and efficiency?





3 vs 10, generations to reach ~77% AUROC

~2.5× more token-efficient

Takeaways

- Existing UQ methods fall short for RAG and RAR systems
- A reliable UQ method must account for **all sources of uncertainty and their interaction**
- For RAR, **perturbation actions** effectively trigger these sources
- **R2C** improves UQ, and also downstream tasks: abstention and model selection
- Multi-generation methods are costly; R2C is **~2.5× more token-efficient**