Uncertainty Quantification for Retrieval-Augmented Reasoning

Heydar Soudani
Radboud University
Nijmegen, The Netherlands
heydar.soudani@ru.nl

Hamed Zamani
University of Massachusetts Amherst
Amherst, United States
zamani@cs.umass.edu

Faegheh Hasibi
Radboud University
Nijmegen, The Netherlands
faegheh.hasibi@ru.nl

Abstract

Retrieval-augmented reasoning (RAR) is a recent evolution of retrieval-augmented generation (RAG) that employs multiple reasoning steps for retrieval and generation. While effective for some complex queries, RAR remains vulnerable to errors and misleading outputs. Uncertainty quantification (UQ) offers methods to estimate the confidence of systems' outputs. These methods, however, often handle simple queries with no retrieval or single-step retrieval, without properly handling RAR setup. Accurate estimation of UQ for RAR requires accounting for all sources of uncertainty, including those arising from retrieval and generation. In this paper, we account for these sources and introduce Retrieval-Augmented Reasoning Consistency (R^{2}C), a novel UQ method for RAR. The core idea of R^{2}C is to perturb the multi-step reasoning process by applying various actions to reasoning steps. These perturbations alter the retriever's input, which shifts its output and consequently modifies the generator's input at the next step. Through this iterative feedback loop, the retriever and generator continuously reshape each other's inputs, enabling us to capture uncertainty arising from both components. Experiments on five popular RAR systems across diverse QA datasets show that R^{2}C improves AUROC by over 5% on average compared to the state-of-the-art UQ baselines. Extrinsic evaluations using R^{2}C as an external signal further confirm its effectiveness for two downstream tasks: in the Abstention task, it achieves ~5% gains in both F1Abstain and AcAbstain; in Model Selection, it improves exact match by ~7% over single models and ~3% over selection methods. Code is available on https://github.com/HeydarSoudani/R2C.

CCS Concepts

• **Computing methodologies → Natural language generation**; • **Information systems → Question answering**; **Language models**.

Keywords

Uncertainty Quantification, Retrieval Augmented Generation, Reasoning Consistency

ACM Reference Format:
Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. 2026. Uncertainty Quantification for Retrieval-Augmented Reasoning. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3805712.3809717

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '26, Melbourne, VIC, Australia
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809717

1 Introduction

Retrieval-augmented generation (RAG) is widely used for knowledge-intensive tasks, but remains limited in addressing complex multi-step reasoning [22, 24, 29, 53, 58]. Recent work has explored combining RAG with reasoning, where LLMs are prompted or trained to use search engines as tools during their reasoning process; a paradigm referred to as retrieval-augmented reasoning (RAR) [23, 29, 60, 62]. However, RAR models are still prone to producing incorrect responses, due to issues such as retrieving irrelevant documents in early steps, misinterpreting retrieved content, or misusing internal knowledge. Therefore, ensuring the trustworthiness of RAR outputs has become a critical challenge.

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)[1] a model has in its predictions [19, 21, 30, 34, 42, 75]. Recent methods of estimating the uncertainty of LLM outputs are designed for settings where the input consists solely of a query, meaning the LLM itself is the only source of uncertainty [3, 12, 70]. The limited work on UQ for RAG [46, 57] incorporates the document–response relationship into the uncertainty score, but these methods are only applicable for simple RAG settings, where documents are retrieved once and inserted into the input prompt for generation. As a result, existing UQ approaches are suboptimal for RAR.

A fundamental reason for the relatively poor performance of existing UQ methods for RAR models is that they primarily attribute uncertainty to the LLM's generative process; i.e., next-token prediction. In RAR systems, however, we have more sources of uncertainty: the *retriever*, which may provide irrelevant or partially relevant retrieved documents and potentially mislead the model's reasoning and response-generation processes; and the *generator*, where 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.

In this paper, we propose **R**etrieval-Augmented **R**easoning **C**onsistency (R^{2}C), a novel UQ method that, unlike previous approaches, accounts for multiple sources of uncertainty in RAR. The central idea of R^{2}C is to allow the model to explore diverse reasoning paths, queries, and documents and then measure the consistency of the resulting final answers. This is achieved by modeling RAR as a Markov Decision Process (MDP) and perturbing this process in a controlled way through a set of perturbation actions across various states. Three perturbation actions are designed to influence query generation, document retrieval, and LLMs' thinking process. These

[1]Some literature distinguishes between a model's uncertainty and its confidence in a response [37]. In this work, we follow the terminology used in most of the literature, where uncertainty is treated as a lack of confidence and the two terms are used interchangeably [39].

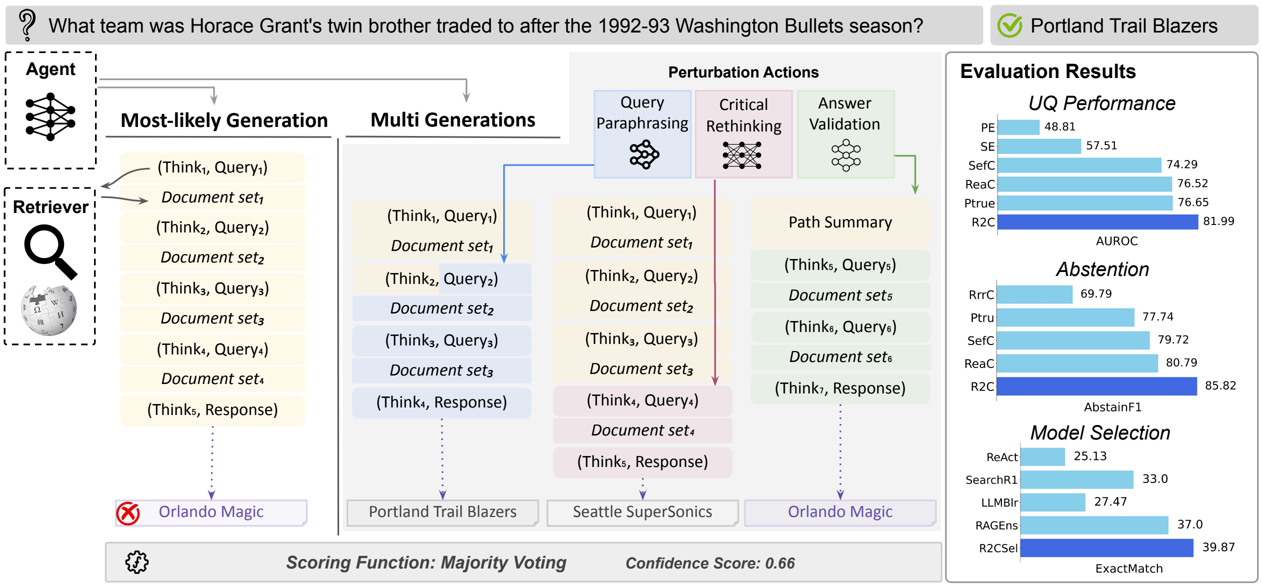

Figure 1: R^{2}C overview. Given a user query, the agent (LLM) first generates the most-likely reasoning path leading to the most-likely response (left, yellow). To estimate uncertainty, R^{2}C creates multiple perturbed generations by randomly altering states in the reasoning path (middle, gray). The confidence score is then derived via majority voting. R^{2}C significantly outperforms established UQ methods and achieves significant improvements on two downstream tasks: abstention and model selection.

perturbations enable models to arrive at diverse final responses for uncertain generations. The confidence score is then obtained by measuring consistency of the generated answers using majority voting; see Figure 1.

We conduct our experiments across multiple datasets and five RAR models. Our experiments show that R^{2}C significantly outperforms existing LLM-specific UQ methods, achieving on average more than a 5% improvement in AUROC compared to the state-of-the-art UQ methods. We further extrinsically evaluate R^{2}C on two downstream tasks: (i) Abstention:[2] the task of generating 'I don't know' when the model is uncertain about its output [16, 41], and (ii) Model Selection:[3] the task of selecting a final answer from a pool of candidates generated by multiple systems [8, 18]. Our experimental results indicate that R^{2}C delivers statistically significant gains over existing approaches: in Abstention, it achieves roughly 5% improvements on both F1Abstain and AccAbstain; in Model Selection, it increases exact match by about 7% relative to single RAR models and by about 3% compared to selection model baselines.

Given the strong performance of R^{2}C in both direct evaluation and extrinsic evaluation on downstream tasks, we investigate the factors that contribute to its effectiveness. We show that R^{2}C retrieves on average 25 unique documents for each score, compared to 16 documents retrieved by other UQ methods. It also achieves a query diversity of 0.35 compared to 0.30 of other methods, measured by the inverse of the average pairwise cosine similarity between queries [9, 78]. This diversity in queries and documents demonstrates that R^{2}C generates diverse, yet relevant reasoning paths through our controlled perturbation mechanism. As a result, this enables the method to achieve confidence scores comparable to baseline approaches while requiring only about 3 generations on average, 2.5 times fewer token generations than the 10 used by the baselines. This highlights that R^{2}C is not only the most effective method of its kind but also a relatively more efficient UQ approach. To summarize, the main contributions of this paper are:

(1) We propose R^{2}C, a novel theoretically grounded UQ method based on MDP; the first of its kind that captures different sources of uncertainty in RAR.
(2) We conduct extensive experiments on three datasets and five RAR methods, demonstrating the superiority of the proposed method on the UQ task with average AUROC of 82%.
(3) We show the effectiveness of our method on both model selection and abstention tasks, significantly outperforming baselines by at least 3%.
(4) We demonstrate that R^{2}C achieves an improvement in token efficiency by approximately 2.5 times.
(5) We show that diverse query and document generation strengthens UQ by capturing multiple uncertainty sources.

2 Related Work

Retrieval-Augmented (Reasoning) Models. RAG is a framework that combines the strengths of retrieval models and generative models [10, 56]. Broadly, RAG can be implemented in different ways. In the retrieve-then-generate paradigm, relevant documents are first retrieved based on the user's input and then incorporated into the model's prompt [43, 56]. In contrast, Active RAG allows retrieval to occur throughout the generation process, either in fixed intervals or dynamically, whenever additional information is needed [28, 59, 64]. Retrieval-Augmented Reasoning (RAR) is a recent extension of RAG

[2] Also referred to as selective prediction in the literature [13, 68].
[3] Also referred to as selection-based model ensemble in the literature [25].

that integrates retrieval with reasoning, aiming to improve the interaction between LLMs and retrievers [60, 62]. For example, Self-Ask [47] decomposes complex questions into follow-up queries and intermediate answers, while ReAct [74] defines a set of actions, such as search, lookup, and finish, to structure interactions with external resources. More recent and effective models such as ReSearch [6] and Search-R1 [29] are explicitly trained to seamlessly integrate external resources into reasoning. Despite their effectiveness on complex queries, RAR models remain prone to errors, such as retrieving irrelevant documents, misinterpreting content, or misusing internal knowledge.

Uncertainty Quantification (UQ) and Confidence Estimation (CE). UQ and CE are closely related but conceptually distinct. In traditional machine learning, uncertainty is defined as a property of the model's predictive distribution given a specific input [37, 39]. Confidence, by contrast, reflects the model's belief in the correctness of a *particular prediction* for the given input. In other words, CE is defined as the task of quantifying how certain a model is about a specific generated response [39], while UQ is the task of capturing the degree of variability or unpredictability in the model's outputs (irrespective of a specific response) [17, 32]. Despite this distinction, existing work sometimes uses uncertainty to refer to estimated confidence [30, 39]. In this paper, we compute confidence scores for LLM-generated responses. Nevertheless, following common practice in the literature [5, 14, 30, 38, 42, 49, 67, 76, 77], we refer to this confidence score as an uncertainty measure and compare it against other UQ methods.

Broadly, existing UQ methods can be divided into two categories: white-box approaches, which leverage token-level probabilities and entropy [3, 12, 30, 34, 55, 70], and black-box approaches, which rely only on the final textual outputs [37, 61]. Most UQ methods focus on question answering and view the LLM as the only source of uncertainty, but in RAG, the retriever also contributes its own uncertainty, making this assumption incomplete. Limited research has explored UQ for RAG by modeling the document-response link, either via axioms [57] or utility models [46]. Some path-based approaches focus on assessing the consistency of reasoning paths for reasoning tasks [35, 48, 66]. However, these methods do not extend naturally to RAR that involves repeated retrieval during reasoning. Recent work has studied uncertainty propagation in multi-step decision-making by combining uncertainties from intermediate steps. SAUP [79] learns aggregation weights to merge per-step uncertainties, but it relies on ground-truth labels from the test domain. In contrast, we propose R^2C, a method that accumulates uncertainty over the entire reasoning path, while considering different sources of uncertainty including the retriever and the generator.

3 Preliminaries

RAR as Markov Decision Process. We formalize RAR as a stochastic Markov Decision Process (MDP), described by a quadruple (S, A, P, R), where S denotes a set of states, A represents a set of actions the agent can take in a state, $P(s_{t+1} \mid s_t, a_t)$ denotes the probability of transitioning from state s_t to state s_{t+1} given action a_t, and $R(s_t, a_t)$ is the reward received by the agent after taking action a_t in state s_t. To generate a factual response to a user query x, the agent π starts from an initial state s_0 corresponding to the query x. The agent then iteratively selects an action at each step until it chooses a halting action and generates the response. The agent assigns a probability $p_\pi(a_t \mid s_t)$ to each possible action based on the current state s_t.

In our unified formulation of MDP for RAR, the LLM π_θ acts as the agent. The environment can take various forms, such as a knowledge repository [74]. The set of possible actions in each state is $A = \{a_{\text{ret}}, a_{\text{ans}}\}$, where a_{ans} denotes the halting action, and a_{ret} represents the retrieval action. Each intermediate state s_t consists of a think τ_t followed by a search query q_t; i.e., $s_t = \langle \tau_t, q_t \rangle$. The final state s_N contains a think τ_N followed by a final response r. The transition probability $P(s_{t+1} \mid s_t, a_t)$ is determined by the LLM itself. An explicit reward function is not always required, as some agents operate without additional training.

Algorithm 1 $\mathbf{R^2C}$: Retrieval-Augmented Reasoning Consistency

Require: user query x, backbone LLM π_θ, number of generations B, set of main actions $A = \{a_{ret}, a_{ans}\}$, set of perturbation actions $A^* = \{a_{qp}, a_{cr}, a_{av}\}$
Ensure: Confidence score $C(x, r)$

for $t = 1$ **to** N **do** $s_t \leftarrow \pi_\theta(s_{t-1}, a_{t-1})$ ▷ Generate the most-likely path iteratively
$r \leftarrow s_N$ ▷ Capture the most-likely response
$R_G = \emptyset$ ▷ Initialize multi-generation response set
for $b = 1 \rightarrow B$ **do**
  $a^* \sim \mathcal{U}(\mathcal{A}^*)$ ▷ Sample an action from $\mathcal{A}^*$
  if $a^* = a_{av}$ **then** $s_t \leftarrow s_N$ ▷ Select the last state for a_{av}
  else $s_t \sim \mathcal{U}(s_1, s_{N-1})$ ▷ Sample a state
  end if
  $s_{t+1} = \pi_\theta(s_t, a^*)$ ▷ Apply action a^* at state s_t
  for $i = 1$ **to** N^b **do** $s_{t+i+1} \leftarrow \pi_\theta(s_{t+i}, a_{t+i})$
  $r^b \leftarrow s_{N^b}$ ▷ Capture the sampled response
  $R_G = R_G \cup \{r^b\}$ ▷ Update the sampled response set
end for
$c = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(r^b = r)$ ▷ Compute the confidence score
return c

Consistency-based Uncertainty Quantification. The core idea of consistency-based methods [21, 69] is to generate multiple responses for a given query by varying either the input prompt or the temperature parameter used in stochastic decoding [33]. The idea builds upon the hypothesis of self-consistency theory, where correct reasoning processes, even if they are diverse, tend to have greater agreement in their final answer [66]. The pairwise similarity among these responses is then computed and aggregated into a single confidence score [4, 37]. Formally, consider a model π parameterized with θ, generates the most-likely response r by setting the sampling temperature to less than one. Then, B additional responses $R_G = \{r^b\}_{b=1}^{B}$ are sampled using various sampling strategies such as increasing the temperature [66], changing the input [33], or altering the reasoning path [35, 48]. A transformation function ϕ is then applied to convert R_G into a confidence score. Our method builds on the reasoning path perturbation approach and employs *Majority Voting* [66] as the transformation function, where confidence is the degree of consistency, measured by the proportion of sampled responses that match the most-likely response:

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

4 Methodology

We propose **R**etrieval-Augmented **R**easoning **C**onsistency, **R^{2}C**, to address UQ in RAR models. R^{2}C is a consistency-based approach that performs UQ in two main stages, as illustrated in Figure 1: (i) generating the most-likely response, and (ii) sampling multiple generations. The core idea of R^{2}C is to perturb the reasoning paths of these multiple generations through a set of perturbation actions, denoted as A^*. In MDP terms, R^{2}C temporarily replaces the main action set A with A^* for a single state, allowing the RAR model to interleave its generation process and explore new reasoning trajectories, queries, and documents. We define three perturbation actions employed in R^{2}C: (i) Query Paraphrasing, a_{qp}, (ii) Critical Rethinking, a_{cr}, and (iii) Answer Validation, a_{av}. By employing a variety of actions to perturb the reasoning processes of LLMs, R^{2}C captures both epistemic (model) uncertainty and aleatoric (data) uncertainty. This perspective is supported by Liu et al. [39], who argue that reasoning uncertainty encompasses both epistemic and aleatoric components. In this work, we do not seek to disentangle these two sources of uncertainty; rather, we measure their combined effect as total uncertainty.

In the following sections, we first formally describe how R^{2}C perturbs the generation path and then describe perturbation actions.

4.1 R^{2}C: Retrieval-Augmented Reasoning Consistency

R^{2}C models uncertainty for RAR as an MDP, in which multiple response generations are produced by temporarily replacing the action set A with an alternative set A^* at a randomly selected state s_t. First, the most-likely generation is produced as an MDP iteratively:

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

where N denotes the length of the reasoning path, determined by the agent π_θ when it selects the halting action a_{ans}. The most-likely response r is obtained from the final state s_N. For example, in the left part of Figure 1, the most-likely response is "*Orlando Magic*," derived from the most-likely generation highlighted in yellow. The middle part of the figure shows the multi-generation process where an action is randomly selected and applied to a state.

To construct the sampled response set R_G, we first fix the number of generations to B. In each generation, we uniformly sample an action a^* from the perturbation action set A^* and a perturbation state s_t from the most-likely reasoning path $\{s_t\}_{t=1}^N$. An exception is the action a_{av}, for which the perturbation state is always set to s_N. The agent then transitions from state s_t to s_{t+1} given action a^*:

$$s_{t+1} \leftarrow \pi_\theta(s_t, a^*).$$

For the remainder of the path, the actions are sampled from the main action set A until reaching a new end state N^b, determined by the agent:

$$s_{t+i+1} \leftarrow \pi_\theta(s_{t+i}, a_{t+i})\,; \quad i = 1, \ldots, N^b.$$

The final sampled response r^b is obtained from the last state s_{N^b} and added to R_G. After B iterations, we obtain the response set $R_G = \{r^b\}_{b=1}^B$, on which the confidence score function (Eq. (1)) is applied to compute the final confidence score. Algorithm 1 provides a detailed description of the entire process in R^{2}C.

4.2 Perturbation Actions

A1: Query Paraphrasing (QP). Constructing effective search queries is critical for retrieving relevant documents; however, LLMs are not inherently optimized for this purpose [27, 35, 40]. The QP action (a_{qp}) is introduced as a query optimization mechanism that enables the system to explore alternative semantic formulations of the original query. Precisely, when action a_{qp} is applied, the think τ_t of the state s_t is preserved and only the query changes. Formally, the LLM π_θ takes action a_{qp} at state s_t, transitioning to state s_{t+1} with the same think τ_t and a new query q_{t+1}:

$$s_{t+1} = \langle \tau_t, q_{t+1} \rangle \leftarrow \pi_\theta(\langle \tau_t, q_t \rangle, a_{qp}).$$

Conceptually, QP perturbation tests whether the reasoning path is so fragile that paraphrasing the search query can alter its direction and lead to the retrieval of different documents. The QP action is implemented by prompting the LLM with a paraphrasing instruction.

A2: Critical Rethinking (CR). RAR models often suffer from the problem of *self-criticism*, where they fail to recognize that previously retrieved information is noisy or irrelevant [2, 27, 29, 35]. Consequently, they continue to build their reasoning on top of earlier steps, even when those steps are uninformative and lack relevant content. This issue becomes particularly severe when it occurs in the early stages of the reasoning path.

The CR action (a_{cr}) critically reassesses the reasoning states produced up to state s_t. When applied at state s_t, it introduces a new state s_{t+1} that the think τ_{t+1} explicitly evaluates the previously retrieved information as unhelpful, irrelevant, or misleading, and the accompanying search query q_{t+1} is formulated to support this critical assessment. Formally, in state s_t, the LLM π_θ is prompted with a critical rethinking instruction to generate a new state s_{t+1}:

$$s_{t+1} = \langle \tau_{t+1}, q_{t+1} \rangle \leftarrow \pi_\theta(\langle \tau_t, q_t \rangle, a_{cr}).$$

Conceptually, if the reasoning path so far has been incorrect, this action enables the system to adjust to a more reliable trajectory. If the path has been correct, CR strengthens its validity, thereby increasing confidence in the final outcome.

A3: Answer Validation (AV). RAR models face difficulties validating their final response, detecting whether the generated response meets certain criteria of the query [20, 26, 52, 65]. One challenge is that the response is built upon a reasoning path that integrates both documents and the intermediate reasoning trajectory, which often leads the LLM to exhibit excessive confidence in its output. Another challenge is that different tasks and response types involve specific validation criteria, but RAR models are generally unaware of these requirements.

Table 1: Performance of UQ methods measured by AUROC. In each column, the best and second-best methods are indicated by bold and <u>underline</u>, 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 [47]			ReAct [74]			Search-o1 [36]			ReSearch [6]			Search-R1 [29]			Avg.
Uncer. M.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	
PE [30]	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 [34]	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 [12]	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 [70]	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 [34]	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 [37]	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 [37]	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 [37]	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 [35]	71.14	71.17	**81.28**	48.02	68.30	<u>75.99</u>	65.95	73.87	<u>77.95</u>	68.30	71.63	74.25	68.92	71.08	70.77	70.57
SelfC [66]	74.33	69.06	68.40	80.73	75.34	72.96	<u>81.26</u>	72.34	76.87	82.01	72.14	77.89	71.63	69.04	70.36	74.29
ReaC [48]	<u>77.02</u>	70.90	<u>76.75</u>	<u>81.53</u>	<u>76.94</u>	74.74	81.09	<u>74.97</u>	77.01	82.50	75.22	77.86	73.22	72.79	<u>75.29</u>	76.52
P(true) [30]	76.51	<u>77.66</u>	78.65	75.07	73.45	71.51	78.57	73.19	74.83	<u>84.35</u>	<u>77.77</u>	<u>81.42</u>	<u>75.73</u>	<u>76.25</u>	74.80	<u>76.65</u>
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**

We introduce the AV action (a_{av}) to validate the final response by prompting the LLM to reconsider its generation once a response has been produced, based on predefined criteria. Specifically, the LLM first generates a ***query-aware reasoning path summary*** [1, 36, 54], and then evaluates the final response using two criteria: (i) *Groundedness*: is the response supported by the retrieved documents? and (ii) *Correctness*: does the response appropriately and sufficiently address the query, given the available evidence? Formally, let $D = \{D_1, \ldots, D_{N-1}\}$ be the set of documents retrieved at states $[s_1, s_2, \ldots s_{N-1}]$. A model $\mathcal{M}$ generates the summary $\hat{S}$ of these documents: $\hat{S} = \mathcal{M}(x, D)$. The state s_N is then updated with this summary, denoted as $\hat{s}_N$. With the updated state $\hat{s}_N$, the LLM π_θ is then instructed to generate a new state s_{N+1}:

$$s_{N+1} \leftarrow \pi_\theta(\hat{s}_N, a_{av}).$$

In principle, if the final response r is validated as correct, the system outputs it directly in state s_{N+1}. Otherwise, if the validation indicates that the answer is incorrect or incomplete, the system begins a new reasoning path starting from $\langle \tau_{N+1}, q_{N+1} \rangle$.

5 Experimental Setup

Our experiments consist of evaluation of uncertainty scores estimated by UQ methods as well as extrinsic evaluation on Abstention and Model Selection. In the following, we review our experimental setup for each of these tasks.

5.1 Direct Evaluation of UQ Estimations

Datasets: We evaluate R^2C on both single-hop and multi-hop QA tasks using the PopQA [43], HotpotQA [73], and Musique [63] datasets. Following prior work [4, 28, 44, 64, 74], we randomly sample 500 queries from each dataset as the test set. We will release our sampled queries to improve reproducibility of our work. For the retrieval corpus, we use the 2018 Wikipedia dump [31].[4] The number of retrieved documents is fixed to three across all models [29, 43, 47, 74].

Evaluation Metrics: To evaluate the quality of outputs, we follow Jin et al. [29] and report the exact match, where a prediction is counted as correct if and only if it exactly matches one of the ground-truth responses. For evaluating UQ methods, we follow prior work on UQ and use the threshold-free metric AUROC, which captures the correlation between uncertainty scores and response correctness [3, 30, 34]. As suggested by Perez-Beltrachini and Lapata [46], significant differences between two AUROC values are assessed using the paired De Long test [11].

Models: In line with prior work in RAR [6, 29], we employ *Qwen-2.5-7B-Instruct* [72] as the generator LLM and path summary generator for action a_{av}. For UQ, we sample 10 responses per query with a temperature of $T = 1.0$, while for correctness evaluation we generate the most-likely generation with $T = 0.7$ [3, 57]. Retrieval is performed using a two-stage re-ranking pipeline: BM25 [51] is used for initial retrieval, followed by re-ranking with the pre-trained cross-encoder model `ms-marco-MiniLM-L-6-v2` from the `sentence-transformers` library.[5] All experiments are conducted on four Nvidia A100 GPUs, each with 40 GB memory, requiring ~1500 GPU hours in total.

Baselines: We use two sets of baselines: (1) ***Path-based*** methods, which focus on generating multiple responses based on diverse reasoning paths and differ mainly in how they initiate new generations relative to the most-likely generation. Self-Consistency (SelfC) [66] ignores the most-likely generation and instead produces a diverse set of reasoning paths, starting from scratch. Reasoning Consistency

[4] https://huggingface.co/datasets/PeterJinGo/wiki-18-corpus
[5] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

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 [47]			ReAct [74]			Search-o1 [36]			ReSearch [6]			Search-R1 [29]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
Abstain Accuracy																
RrrC [35]	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) [30]	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 [66]	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 [48]	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 [35]	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) [30]	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 [66]	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 [48]	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**

(ReaC) [48] randomly truncates the most-likely reasoning path at different random steps and regenerates the response based on the subsequent reasoning steps. Retrieval-Retained Reasoning Consistency (RrrC) [35] applies truncation only after the last retrieved document. In all cases, the generated responses are aggregated into a final score using majority voting. 2) ***Estimation-based*** methods include both white-box and black-box approaches. The white-box methods are PE [30], SE [34], MARS [3], LARS [70], and SAR [12]. The black-box methods are NumSS [34], EigV, ECC, Deg [37], and P(true) [30]. All of these methods rely on generations obtained in the same way as SelfC. All UQ methods are implemented using TruthTorchLM [71].

5.2 Extrinsic UQ Evaluation via Abstention

Task Formulation. Declining to respond due to uncertainty is an important application of UQ [16]. An abstention function determines whether the model should withhold an answer. In our setup, this decision is guided by confidence scores. We introduce a threshold τ_{abs}: if the confidence is less than τ_{abs}, the model abstains; otherwise, it produces an answer. Formally, given a response r to a query x and a confidence function C, we define the abstention function f_{abs} as:

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

Baselines & Evaluation Metrics. As baselines, we implement the abstention task using confidence scores derived from different UQ methods, including P(true), RrrC, ReaC, SelfC, and R^2C. For evaluation, we follow Feng et al. [16] and report two metrics: *AbstainAccuracy* and *AbstainF1*. These metrics assess how well a model balances answer correctness with appropriate abstention. We define a confusion matrix with four outcomes: (A) answered correctly, (B) abstained correctly, (C) answered incorrectly, and (D) abstained incorrectly. Based on this matrix, *AbstainAccuracy* is defined as $\frac{A+D}{A+B+C+D}$ and measures whether abstention decisions are correct overall. Ideally, a model should abstain when it would otherwise answer incorrectly, and provide an answer when it is correct. *AbstainF1* captures the trade-off between reliability and answer coverage. It is computed as the harmonic mean of precision ($\frac{D}{B+D}$) and recall ($\frac{D}{C+D}$). This metric penalizes both unnecessary abstentions and incorrect answers, providing a balanced evaluation of abstention behavior. The threshold τ_{abs} for each UQ method is selected via a sweep over the validation sets.

5.3 Extrinsic UQ Evaluation via Model Selection

Task Formulation. Model selection (or selection-based model ensemble [8, 25]) aims to select a final response for a question based on multiple candidate responses generated by different systems [8]. Formally, given a user query x and a set of systems $\{S_1, S_2, \ldots, S_M\}$, each system S_i produces a response candidate r_i. A model selection method $\mathcal{M}(x, R)$ then selects the final response $\hat{r}$ from all candidate responses $R = \{r_1, r_2, \ldots, r_M\}$.

R^2C Select. The proposed R^2C Select utilizes the confidence scores derived from R^2C. Our method groups semantically similar responses into K clusters. Following [35, 71], we use *Qwen-2.5-7B-Instruct* to compute pairwise semantic similarities between candidate responses and cluster similar ones. We then assign a confidence score to each cluster g_i:

$$R_G = \{\langle r_1, c_{g_1} \rangle, \langle r_2, c_{g_1} \rangle, \ldots, \langle r_K, c_{g_K} \rangle\}.$$

The confidence score of each cluster $g = \{r_i\}_{i=1}^{m}$ is computed by aggregating the confidence scores of its members: $c_g = \sum_{r_i \in g} C(x, r_i)$, where the C function provides the confidence score for the response r_i. The final response $\hat{r}$ is the response with the highest confidence score: $\hat{r} = \arg\max_j c_{g_j}$. If no clustering is applied, each response constitutes a cluster and $M = K$; we refer to this variation as *R^2C Select w/o clustering*.

Baselines & Evaluation Metrics. We evaluate R^2C Select performance against both single RAG and RAR systems, as well as existing selection-based ensemble approaches. As baselines, we

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 [64]	34.6	27.2	5.4	22.4
FLARE [28]	31.6	25.2	8.4	21.7
DRAGIN [59]	28.6	23.2	4.6	18.8
Retrieval-Augmented Reasoning (RAR)				
SelfAsk [47]	35.6	33.0	10.4	26.3
ReAct [74]	36.8	27.8	10.8	25.1
Search-o1 [36]	33.2	29.0	10.0	24.1
ReSearch [6]	38.6	38.8	16.6	31.3
Search-R1 [29]	41.6	41.4	16.0	33.0
Model Selection RAR				
Random	31.4	31.6	10.6	24.5
LLMBlender [25]	34.4	36.0	12.0	27.5
- *w/o clustering*	35.6	31.8	9.6	25.7
- *w/o clus. & conf.*	32.0	26.0	8.6	22.2
RAGEnsemble [7]	<u>45.6</u>	<u>46.0</u>	<u>19.4</u>	<u>37.0</u>
- *w/o clustering*	44.4	40.8	18.0	34.4
- *w/o clus. & conf.*	43.2	37.2	13.0	31.1
R^{2}CSelect (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

consider LLMBlender [25] and RAGEnsemble [7]. LLMBlender is a trained reward model that given a user query and a set of candidate responses, ranks the responses accordingly. RAGEnsemble is an instruction-based approach that selects a single final answer from a set of candidate responses. For all selection methods, including the baselines and R^{2}C Select, the response candidates are obtained from SelfAsk [47], ReAct [74], Search-o1 [36], ReSearch [6], and Search-R1 [29]. For evaluation, we report the correctness of the final answer using exact match.

6 Results

We present a set of experiments that address the following research questions: **RQ1**: How does R^{2}C perform in quantifying uncertainty for different RAR models? (Sec. 6.1), **RQ2**: How does R^{2}C perform as an external signal on downstream tasks, such as Abstention and Model Selection? (Sec. 6.2), **RQ3**: What factors contribute to the effectiveness of R^{2}C? (Sec. 6.3), **RQ4**: How does R^{2}C balance effectiveness and efficiency? (Sec. 6.4), **RQ5**: What is the effect of different actions in the performance of R^{2}C? (Sec. 6.5). **RQ6**: Does R^{2}C generalize across different backbone LLMs? (Sec. 6.6)

6.1 Uncertainty Quantification Performance

RQ1 evaluates the performance of R^{2}C compared to other UQ methods. Table 1 presents results on five RAR systems across three datasets. The findings indicate that most white-box methods, i.e.,

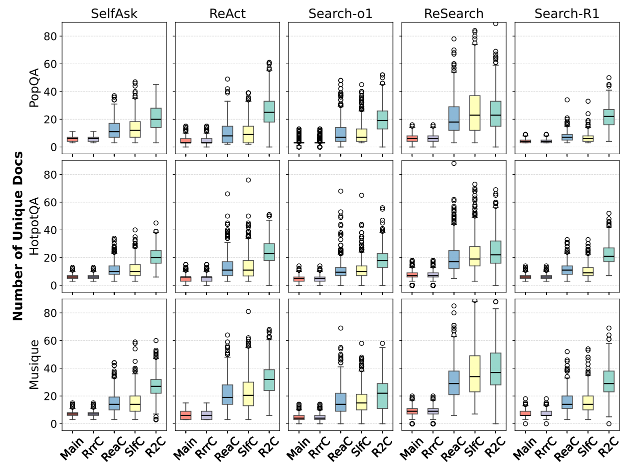

Figure 2: Distribution of the number of unique retrieved documents for the most-likely path (main) and multi-generations.

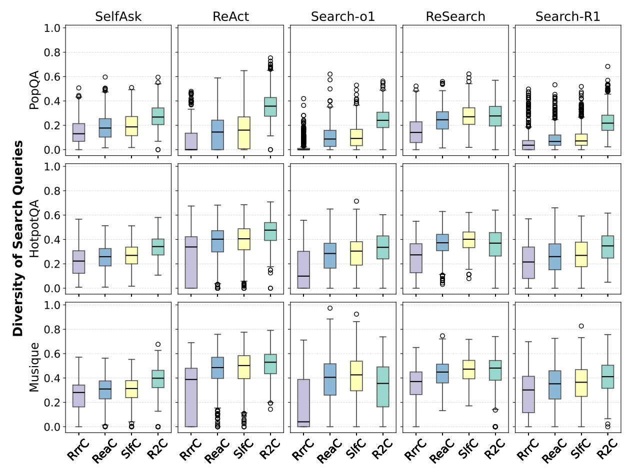

Figure 3: Distribution of diversity scores for search queries generated for each user query across reasoning paths.

PE [30], SE [34], MARS [3], and SAR [12], perform relatively poorly, with AUROC values ranging from about 30 to 60. This weakness stems from their overreliance on token probabilities. In contrast, black-box methods, such as NumSS [34], EigV, ECC, and Deg [37], generally outperform white-box methods, reaching average AUROC values between roughly 60 and 80. Interestingly, P(true) ranks as the second-best method in terms of average AUROC across all approaches, highlighting that black-box methods become much more effective when using an LLM as the scoring function. This advantage largely comes from their stronger reliance on textual diversity. The supervised method LARS, optimized for QA with no reasoning, achieves AUROC scores between 65 and 85 across all cases, surpassing both white-box and black-box approaches on average. This finding highlights the potential of supervised UQ as a promising direction for future research.

Path-based methods are another group of approaches that we evaluate, including the proposed R^{2}C. RrrC achieves an average AUROC of 70.57, which is the lowest among these methods. This indicates that simply keeping the documents in the reasoning path and regenerating the last state is not effective. SelfC and ReaC

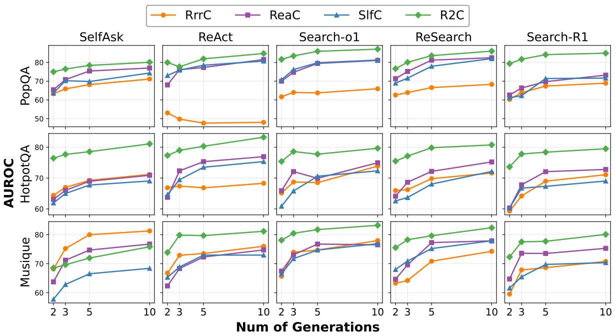

Figure 4: Performance of UQ methods with varying numbers of generations.

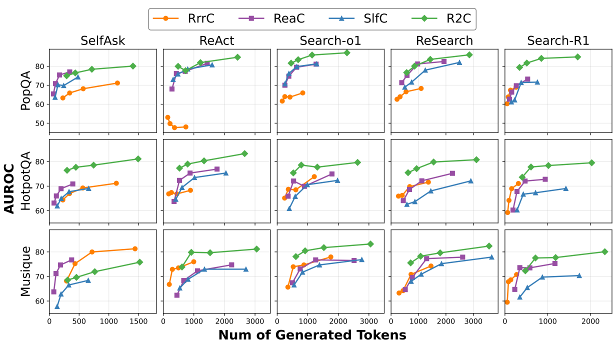

Figure 5: Performance of UQ methods with varying number of generated tokens, illustrating the trade-off between effectiveness and efficiency.

perform roughly on par, and ReaC outperforms SelfC on average, showing that regenerating from the top of the reasoning path does not necessarily guarantee a better uncertainty score. Finally, **R^{2}C** outperforms all methods by a large margin, with an absolute improvement of 5% on average. These findings suggest that applying actions to the reasoning path enables exploration of a wider variety of possible reasoning states, leading to more reliable scores.

6.2 Extrinsic Evaluation Results

RQ2 evaluates the performance of R^{2}C in Abstention and Model Selection. For Abstention, Table 2 shows the performance on RAR models and datasets. Abstain Accuracy measures whether both abstentions and non-abstentions are detected correctly, while AbstainF1 captures the effectiveness of identifying cases where the model should abstain. Considering Abstain Accuracy, R^{2}C outperforms the baselines in all cases except on the Musique dataset, where it performs on par with the second-best baseline in ReAct, Search-o1, and ReSearch. Considering Abstain F1, R^{2}C significantly outperforms the best baseline in all setups, expect for ReAct on Musique. This indicates that R^{2}C scores are more reliable at identifying cases where abstention is appropriate. Overall, compared to other UQ methods, R^{2}C achieves significantly better performance in most cases—on average, about 5% higher than the second-best model. These results demonstrate that the uncertainty scores generated by R^{2}C are reliable enough for the system to decide when to refrain from answering.

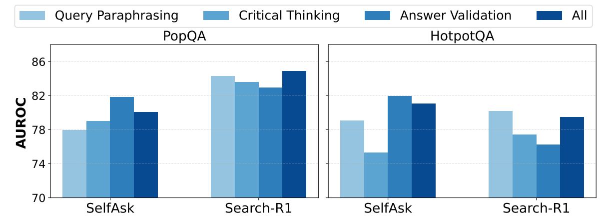

Figure 6: Performance of R^{2}C with different action sets.

For model selection, Table 3 reports the results of individual RAG and RAR systems, as well as selection-based ensemble models. Among individual systems, RAR models outperform RAG models for both simpler single-hop questions (PopQA) and more complex multi-hop queries (HotpotQA and Musique). Within selection-based ensemble models, LLMBLender performs poorly, even worse than fine-tuned single RAR models such as ReSearch and Search-R1. In contrast, RAGEnsemble surpasses both LLMBLender and all single RAR models, highlighting the advantage of an instruction-tuned model for selecting the final response. Finally, **R^{2}C Select** achieves the best performance, significantly outperforming all RAR and RAG systems, with 3.7% average improvement on HotpotQA and Musique datasets. **These results confirm that the R^{2}C score is a reliable criterion, not only within a single system but also across different systems.** Interestingly, using R^{2}C confidence scores alone, without clustering, improves selection-based ensemble models in comparable settings, highlighting the informativeness of R^{2}C scores even for models that are not trained on confidence.

6.3 Explaining the Strength of R^{2}C

RQ3 examines why R^{2}C outperforms other UQ methods. To this end, we analyze the diversity of reasoning paths as an indicator of how effectively R^{2}C captures the uncertainty arising from both the generator and the retriever, i.e., the two key sources of uncertainty in our framework. We report the number of unique documents retrieved during the multi-generation step, as shown in Figure 2. On average, R^{2}C retrieves 24.71 unique documents for a single confidence score, whereas RrrC, SelfC, and ReaC retrieve 5.81, 15.35, and 16.4 documents, respectively.

We further report on the diversity of generated search queries, as a proxy for the LLM's thinking process. Query diversity is measured based on the average pair-wise similarity of generated queries [9, 78]. Formally, given n search queries $\{q_1, \ldots, q_n\}$, query diversity is computed as:

$$\text{Query Diversity} = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i<j}^{n} cos(e_i, e_j),$$

where $cos(.)$ represents cosine similarity function, and e_i denotes the normalized embedding of query q_i (i.e., $|e_i| = 1$) obtained from `sentence-transformers/all-MiniLM-L6-v2`. Figure 3 presents the results. On average, R^{2}C achieves a diversity score of 0.35, while RrrC, SelfC, and ReaC achieve 0.20, 0.28, and 0.30, respectively. These results indicate that R^{2}C generates more diverse search queries when estimating confidence. These findings indicate that R^{2}C effectively captures the uncertainty of both the retriever and the generator by sequentially diversifying their inputs.

Table 4: Performance of UQ methods measured by AUROC across different LLMs. For each LLM and for 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 SelfC and ReaC, respectively, which are the two best-performing methods on average.

	Gemma2-9b-it										GPT-4o-mini									
RAG	SelfAsk [47]			ReAct [74]			Search-o1 [36]			Avg.	SelfAsk [47]			ReAct [74]			Search-o1 [36]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.		Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
RrrC [35]	76.64	71.01	76.16	57.74	72.93	75.32	55.72	63.96	64.33	68.20	63.07	70.96	83.69	53.21	58.16	54.78	70.49	77.27	84.04	68.41
P(true) [30]	71.33	74.90	73.44	76.56	73.08	75.26	76.32	72.45	74.62	74.22	–	–	–	–	–	–	–	–	–	–
SelfC [66]	76.82	72.70	73.59	81.79	74.34	76.84	79.17	70.59	76.07	75.77	76.59	58.28	54.04	81.08	73.32	75.83	78.52	74.39	81.83	72.65
ReaC [48]	76.24	79.77	78.73	80.16	71.33	**80.80**	78.45	73.61	75.77	77.21	75.53	60.58	54.66	81.53	74.34	81.10	79.41	75.09	85.09	74.15
R^2C (our)	**80.29**‡	**81.86**	**80.31**	**82.87**	**80.62**†‡	80.37	**87.65**†‡	**76.95**†‡	**77.39**	**80.92**	**86.56**†‡	**87.82**†‡	**84.28**†‡	**81.56**	**75.84**	**81.94**†	**82.48**†	**84.59**†‡	**86.29**†	**83.48**

6.4 Effectiveness vs. Efficiency

RQ4 explores the trade-off between the effectiveness and efficiency of R^2C. Figure 4 presents the relationship between AUROC and the number of response generations across various datasets and RAR models. The results demonstrate that R^2C consistently outperforms other methods, even with a smaller number of generations. On average, across all datasets and RAR models, R^2C achieves an AUROC performance of about 77% with only three generations, which is comparable to the performance of SelfC and ReaC, requiring 10 generations to reach a similar level.

For a deeper exploration of the trade-offs between effectiveness and efficiency, we measure efficiency using the total number of generated tokens, following prior work [46]. Figure 5 reports AUROC performance under equal token-generation budgets. It shows R^2C achieves higher AUROC scores than other methods given the same number of generated tokens for multi-hop datasets. For the single-hop dataset, PopQA, R^2C performs comparably to SelfAsk, ReAct, and ReSearch, but surpasses them on Search-o1 and Search-R1. Moreover, R^2C produces on average around 700 tokens with three generations, reaching the same AUROC score as the baselines with approximately 1,700 tokens with 10 generations. These findings demonstrate that R^2C improves efficiency by roughly 2.5 times. Overall, these results indicate that R^2C outperforms other UQ methods in terms of both effectiveness and token generation efficiency. Beyond improving token-generation efficiency, R^2C is well suited for real-world deployment because it enables parallel generation of fully independent responses, thereby addressing the typical latency concerns associated with sampling-based methods. In practice, the confidence score can be computed at the cost of at most one additional generation beyond the most-likely response.

6.5 Action Selection

RQ5 investigates the role of the action set in R^2C. Figure 6 illustrates the performance of R^2C with different action sets. In the first three bars, only a single action is used, meaning that in each generation, the cut point in the reasoning path is chosen randomly, while the action itself remains fixed. The fourth bar represents our main setup, where both the action and the cut point are selected randomly. For Search-R1 and PopQA, the main setup outperforms all other configurations, whereas in HotpotQA, QP performs slightly better. For SelfAsk, the AV action achieves the best performance, with the main setup ranking second. These results indicate that while our main setup is generally robust, there are still potentials to design action configurations better suited to specific RAR systems. To further improve R^2C, future work could focus on learning or adaptively selecting perturbations (e.g., via reinforcement learning), which we expect would enhance both effectiveness and efficiency.

6.6 Generalization Across Backbone LLMs

RQ6 investigates the robustness of R^2C to changes in the generation model. Table 4 summarizes the performance of UQ methods under alternative LLM backbones. To ensure diversity, we evaluate one open-source LLM, *Gemma2-9b-it*,[6] and one closed-source LLM, *GPT-4o-mini*. We do not evaluate ReSearch and Search-R1 in this setting, as their fine-tuned models are only available for *Qwen* family. With *Gemma2-9b-it* as the backbone, R^2C outperforms the second-best method, ReaC, by more than 3%, indicating that the effectiveness of R^2C is largely independent of the choice of LLM.

For *GPT-4o-mini*, R^2C outperforms ReaC by more than 9%. The P(true) metric cannot be reported in this setting because token-level probabilities are not accessible. In the case of SelfAsk, we observe a different performance ordering, in which RrrC emerges as the second-best method after R^2C, most notably on the Musique dataset. This shift is primarily due to the reduced effectiveness of SelfAsk when *GPT-4o-mini* is used as the backbone compared to *Qwen-2.5-7B-Instruct*, despite additional prompt engineering. In particular, SelfAsk achieves exact match accuracies of 18.6%, 12.2%, and 2.0% on PopQA, HotpotQA, and Musique, respectively, which are lower than the corresponding results reported in Table 3. Overall, R^2C consistently achieves the highest average performance by a substantial margin, demonstrating strong generalization across different LLM backbones.

7 Conclusions and Future Work

This paper introduces a novel and theoretically grounded UQ method for retrieval-augmented reasoning (RAR) systems, called Retrieval-Augmented Reasoning Consistency (R^2C). We argue that an effective UQ method should account for different sources of uncertainty and accurately reflect them in its final score. R^2C models uncertainty stemming from both the retriever and the generator by

[6] https://huggingface.co/google/gemma-2-9b-it

perturbing the reasoning process through a series of actions, including query paraphrasing, critical rethinking, and answer validation. Comprehensive experiments conducted on three datasets and five RAR models demonstrate that R^2C improves AUROC by more than 5% on average compared to state-of-the-art UQ baselines. Moreover, when used as an external signal in two downstream tasks, R^2C consistently proves effective: in Abstention, it yields around 5% gains in both F1Abstain and AccAbstain; in Model Selection, it increases exact match by approximately 7% over individual models and about 3% over selection methods. While this paper focuses on UQ for RAR models, the underlying concept of modeling and stimulating multiple sources of uncertainty is broadly applicable. Future work can extend this approach to other domains involving multiple sources of uncertainty, such as vision-language models. Moreover, in this work, we focus on short-form QA, where the final answer is an entity. Future research can explore UQ for long-form generation, which represents a more realistic scenario.

Acknowledgments

This work was supported in part by the project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO), in part by the Center for Intelligent Information Retrieval, and in part by the Office of Naval Research contract #N000142412612. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] Pierre Achkar, Tim Gollub, and Martin Potthast. 2025. Ask, Retrieve, Summarize: A Modular Pipeline for Scientific Literature Summarization. *CoRR* (2025).
[2] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations, ICLR*. OpenReview.net.
[3] Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. MARS: Meaning-Aware Response Scoring for Uncertainty Estimation in Generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics ACL*. 7752–7767.
[4] Yavuz Faruk Bakman, Duygu Nur Yaldiz, Sungmin Kang, Tuo Zhang, Baturalp Buyukates, Salman Avestimehr, and Sai Praneeth Karimireddy. 2025. Reconsidering LLM Uncertainty Estimation Methods in the Wild. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 29531–29556.
[5] Evan Becker and Stefano Soatto. 2024. Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441* (2024).
[6] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. *CoRR* (2025).
[7] Yifei Chen, Guanting Dong, Yutao Zhu, and Zhicheng Dou. 2025. Revisiting RAG Ensemble: A Theoretical and Mechanistic Analysis of Multi-RAG System Collaboration. *arXiv preprint arXiv:2508.13828* (2025).
[8] Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S. Yu. 2025. Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. *CoRR* abs/2502.18036 (2025). https://doi.org/10.48550/ARXIV.2502.18036 arXiv:2502.18036
[9] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–35.
[10] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. 719–729.
[11] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* (1988), 837–845.
[12] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5050–5063.
[13] Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. 2025. UProp: Investigating the Uncertainty Propagation of LLMs in Multi-Step Agentic Decision-Making. *CoRR* abs/2506.17419 (2025).
[14] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *Findings of the Association for Computational Linguistics: ACL*. 9367–9385.
[15] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.* 630, 8017 (2024), 625–630.
[16] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 14664–14690.
[17] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 56, Suppl 1 (2023), 1513–1589.
[18] Neel Guha, Mayee F. Chen, Trevor Chow, Ishan S. Khare, and Christopher Ré. 2024. Smoothie: Label Free Language Model Routing. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS*.
[19] James Harrison, John Willes, and Jasper Snoek. 2024. Variational Bayesian Last Layers. In *The Twelfth International Conference on Learning Representations, ICLR*. OpenReview.net.
[20] Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics, 10371–10393.
[21] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. https://openreview.net/forum?id=byxXa99PtF
[22] Mohanna Hoveyda, Harrie Oosterhuis, Arjen P. de Vries, Maarten de Rijke, and Faegheh Hasibi. 2025. Adaptive Orchestration of Modular Generative Information Access Systems. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. 3899–3910.
[23] Shayekh Bin Islam, Md. Asib Rahman, K. S. M. Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md. Rizwan Parvez. 2024. Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*. Association for Computational Linguistics, 14231–14244.
[24] Jelle, de Vries Arjen P., de Rijke Maarten, Hasibi Faegheh Hoveyda Mohanna, and Piepenbrock. 2026. OrLog: Resolving Complex Queries with LLMs and Probabilistic Reasoning. In *Advances in Information Retrieval (ECIR 2026)*. 98–114.
[25] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 14165–14178.
[26] Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Xin Zhao, Yang Song, and Tao Zhang. 2025. RAG-Star: Enhancing Deliberative Reasoning with Retrieval Augmented Verification and Refinement. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*. Association for Computational Linguistics, 7064–7074.
[27] Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. s3: You Don't Need That Much Data to Train a Search Agent via RL. *arXiv preprint arXiv:2505.14146* (2025).
[28] Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7969–7992.

[29] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *CoRR* (2025).

[30] Saurav Kadavath et al. 2022. Language Models (Mostly) Know What They Know. abs/2207.05221 (2022).

[31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 6769–6781.

[32] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017).

[33] Anton Korikov, Pan Du, Scott Sanner, and Navid Rekabsaz. 2025. Batched Self-Consistency Improves LLM Relevance Assessment and Ranking. *CoRR* abs/2505.12570 (2025).

[34] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations ICLR*.

[35] Baixuan Li, Yunlong Fan, Tianyi Ma, Miao Gao, Chuanqi Shi, and Zhiqiang Gao. 2025. RASPberry: Retrieval-Augmented Monte Carlo Tree Self-Play with Reasoning Consistency for Multi-Hop Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*. 11258–11276.

[36] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. *CoRR* (2025).

[37] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Trans. Mach. Learn. Res.* 2024 (2024).

[38] Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993* (2024).

[39] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 6107–6117.

[40] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query Rewriting in Retrieval-Augmented Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5303–5315.

[41] Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING*. Association for Computational Linguistics, 9329–9345.

[42] Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *9th International Conference on Learning Representations, ICLR*.

[43] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9802–9822.

[44] Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive Retrieval Without Self-Knowledge? Bringing Uncertainty Back Home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 6355–6384.

[45] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2010. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology, MLSB (JMLR Proceedings, Vol. 8)*. JMLR.org, 65–81.

[46] Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty Quantification in Retrieval Augmented Question Answering. *arXiv preprint arXiv:2502.18108* (2025).

[47] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*.

[48] Zhenting Qi, Mingyuan MA, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2025. Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solver. In *The Thirteenth International Conference on Learning Representations.*

[49] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558* (2022).

[50] Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-Evaluation Improves Selective Generation in Large Language Models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops (Proceedings of Machine Learning Research, Vol. 239)*. PMLR, 49–64.

[51] Stephen E. Robertson and Steve Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. 232–241.

[52] Alireza Salemi, Mukta Maddipatla, and Hamed Zamani. 2025. CIIR@LiveRAG 2025: Optimizing Multi-Agent Retrieval Augmented Generation through Self-Training. *CoRR* abs/2506.10844 (2025). https://doi.org/10.48550/ARXIV.2506.10844 arXiv:2506.10844

[53] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. ACM, 2395–2400.

[54] Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. 2025. Search and Refine During Think: Autonomous Retrieval-Augmented Reasoning of LLMs. *arXiv preprint arXiv:2505.11277* (2025).

[55] Heydar Soudani. 2025. Enhancing Knowledge Injection in Large Language Models for Efficient and Trustworthy Responses. In *Proceedings of the 48th International ACM Conference on Research and Development in Information Retrieval, SIGIR*.

[56] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP 2024*. 12–22.

[57] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why Uncertainty Estimation Methods Fall Short in RAG: An Axiomatic Analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*. 16596–16616.

[58] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2026. A Survey on Recent Advances in Conversational Data Generation. *ACM Comput. Surv.* 58 (4 2026). Issue 10. https://doi.org/10.1145/3795686

[59] Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12991–13013.

[60] Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Yang Song, and Han Li. 2025. ReARTeR: Retrieval-Augmented Reasoning with Trustworthy Process Rewarding. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*. ACM, 1251–1261.

[61] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 5433–5442.

[62] Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025. RARE: Retrieval-Augmented Reasoning Enhancement for Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 18305–18330.

[63] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Trans. Assoc. Comput. Linguistics* 10 (2022), 539–554.

[64] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10014–10037.

[65] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*. Association for Computational Linguistics, 9440–9450.

[66] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR*.

[67] Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In *Findings of the Association for Computational Linguistics: EMNLP*. 6886–6898.

[68] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*. Association for Computational Linguistics, 1040–1051.

[69] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063* (2023).

[70] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. 2025. Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

[71] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Sungmin Kang, Alperen Öziş, Hayrettin Eren Yildiz, Mitash Ashish Shah, Zhiqi Huang, Anoop Kumar, Alfy Samuel, Daben Liu, et al. 2025. TruthTorchLM: A Comprehensive Library for Predicting Truthfulness in LLM Outputs. *arXiv preprint arXiv:2507.08203* (2025).

[72] An Yang et al. 2024. Qwen2.5 Technical Report. *CoRR* abs/2412.15115 (2024).

[73] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*. Association for Computational Linguistics, 2369–2380.

[74] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR*.

[75] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via Uncertainty Quantification. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS*.

[76] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. 2024. Reasoning in Flux: Enhancing Large Language Models Reasoning through Uncertainty-aware Adaptive Guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2401–2416.

[77] Boxuan Zhang and Ruqi Zhang. 2025. CoT-UQ: Improving Response-wise Uncertainty Quantification in LLMs with Chain-of-Thought. In *Findings of the Association for Computational Linguistics: ACL 2025*. 26114–26133.

[78] Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2025. Evaluating the Evaluation of Diversity in Commonsense Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 24258–24275.

[79] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. 2025. Uncertainty Propagation on LLM Agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*. 6064–6073.

A Prompt Details

An important component of the R^2C method is the Perturbation Actions discussed in Section 4.2. These actions are implemented by prompting the LLM with specific instructions. Figures 7, 8, and 9 show the prompts used for Query Paraphrasing, Critical Rethinking, and Answer Validation, respectively. Moreover, the Answer Validation action includes a Reasoning Path Summarization component, which is also implemented through prompting an LLM. The prompt for this component is shown in Figure 10. Finally, Figure 11 presents the semantic equivalence prompt, which is used to assess the equality of two responses for majority voting and clustering in the Model Selection process described in Section 5.3, following pervious work [3, 35, 70].

Query Paraphraser

```
You are an expert in information retrieval.
Given an original search query, generate {n} semantically diverse and effective paraphrased search queries that capture the same intent but use different wording or structure.
These paraphrased queries should be suitable for improving search engine results by covering various phrasings a user might employ.
Do not add extra information in the new queries.

Here are some examples:
<original_query>
popular industry in the neighborhood of Willow Vale, New South Wales </original_query>
<paraphrased_query>
What industries are most common in Willow Vale, NSW?
</paraphrased_query>
<paraphrased_query>
Main economic activities in the Willow Vale area of New South Wales </paraphrased_query>

<original_query> {Original Search Query} </original_query>
```

Figure 7: Prompt for query paraphrasing action.

Critical Re-Thinker

```
You are a highly capable critical rethinker agent.
Given an original search query, you are tasked to critically assess the search query, and then generate a new and creative search query to support your critical thought.
you are also tasked to return one reasoning thought for the new search query, explaining why the new query works better.
The search query should be precise and focused.
Your output must include:
- One complete reasoning step that strongly rejects the entire retrieved information as unhelpful, irrelevant, or misleading, wrapped in a single pair of <critical_rethinking> and </critical_rethinking> tags.
- One creative and fundamentally new search query, wrapped in <search> and </search> tags.

Only use the following format, in this exact order:
<critical_rethinking> one complete reasoning step that strongly rejects the entire retrieved information as unhelpful, irrelevant, or misleading </critical_rethinking>
<search> a creative, focused, and fundamentally new search query </search>

Here are some examples:
<original_query> popular industry in the neighborhood of ... </original_query>
<critical_rethinking> The query "popular industry in the ...</critical_rethinking>
<search> economic activities and land use patterns in Southern ... </search>

<original_query> {Original Search Query} </original_query>
```

Figure 8: Prompt for critical rethinking action.

Answer Validator

```
You are a highly capable response validation agent.
Given a user query and a summary of the retrieved documents used during the reasoning process, your task is to verify whether the generated response satisfies two evaluation criteria. The two criteria are: 1) Is the response grounded in the provided information? 2) Does the response correctly and fully answer the user query?

You must provide a single, coherent reasoning step that examines both criteria and suggests how the response could be improved.
After your reasoning, you must return a precise search query that can help retrieve better information to improve the answer.
The document summary will be enclosed in <information> and </information> tags. This content is read-only: NEVER generate, modify, or repeat the <information> tags.
The predicted answer will be enclosed in <prediction> and </prediction> tags. This content is also read-only: NEVER generate, modify, or repeat the <prediction> tags.
The search query must be focused, informative, and aimed at enhancing the predicted answer.

Your output must include:
- One complete reasoning step that: (a) references the predicted answer, (b) evaluates it using the two criteria, and (c) proposes specific improvements. Wrap this reasoning inside a single pair of <think> and </think> tags.
- One search query that would help improve the answer. Wrap it inside <search> and </search> tags.
Only use the following format, in this exact order:
<think> one complete reasoning step that assesses the answer </think>
<search> a creative and focused search query </search>

Question: {User Query}
<information> {Documents Summary} </information>
<prediction> {Prediction} </prediction>
```

Figure 9: Prompt for answer validation action.

Reasoning Path Summarization

```
You are a highly capable summarization agent.
Your task is to generate a detailed summary based on the provided information and the user's query.

Your output must include:
- One complete and detailed summary, wrapped in a single pair of <summary> and </summary> tags.
Your output must follow this exact format and order:
<summary> one complete summary of the information considering the user query </summary>

<information> {Documents} </information>
<user_query> {User Query} </user_query>
```

Figure 10: Prompt for reasoning path summarization used in the answer validation action.

B Abstention Task

B.1 Evaluation Metrics

We evaluate R^2C on the abstention task in Section 5.2. Following Feng et al. [16], we adopt their definitions for the evaluation metrics. Assume we have a confusion matrix with four elements, each denoted by a character: (A) Answered Correct, (B) Abstained Correct, (C) Answered Incorrect, and (D) Abstained Incorrect. Based on these, four metrics are defined for the abstention task:

(1) **Reliable Accuracy:** $\frac{A}{A+C}$, measures how trustworthy the LLM's generated (non-abstained) answers are; that is, among all answered questions, how many are correct?
(2) **Effective Reliability:** $\frac{A-C}{A+B+C+D}$, balances reliability and coverage; that is, across all questions, what proportion are answered correctly minus those answered incorrectly?

Table 5: Abstention performance measured by the threshold-free metric AUARC. For each column, the best and second-best methods are indicated in bold and underlined, respectively. The superscript † denotes a statistically significant difference compared to ReaC based on the bootstrap test ($p < 0.05$).

RAG	SelfAsk [47]			ReAct [74]			Search-o1 [36]			ReSearch [6]			Search-R1 [29]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
RrrC [35]	52.68	45.25	19.42	41.77	40.25	18.71	44.96	41.58	16.43	45.26	52.13	24.44	56.50	54.70	22.19	38.42
P(true) [30]	56.26	49.53	18.35	56.53	38.82	15.63	54.47	39.75	15.37	60.42	55.03	29.54	61.11	58.72	24.55	42.27
SelfC [66]	55.28	44.63	15.35	59.88	45.13	20.88	56.09	41.83	21.10	62.36	54.38	29.55	58.31	51.53	24.76	42.67
ReaC [48]	56.47	48.32	18.72	60.81	45.48	20.52	56.98	42.55	21.61	**63.13**	55.21	29.63	54.02	55.73	27.25	43.83
R^2C (our)	**58.87**†	**53.81**†	**19.60**	**62.74**†	**47.62**†	**24.45**†	**60.72**†	**46.65**†	**21.84**	61.19	**59.25**†	**31.91**	**66.06**†	**60.90**†	**31.58**†	**47.15**

Semantic Equivalence of Responses

We are evaluating answers to the question: In what school district is Governor John R. Rogers High School, named after John Rankin Rogers, located?
Here are two possible answers:
Possible Answer 1: Puyallup School District of Washington
Possible Answer 2: Puyallup School District
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response: Yes

We are evaluating answers to the question: Which team featured in both the 2012 and 2011 Cops del Rey Finals?
Here are two possible answers:
Possible Answer 1: Barcelona
Possible Answer 2: Spain
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response: No

We are evaluating answers to the question: Which genus of flowering plant is found in an environment further south, Crocosmia or Cimicifuga?
Here are two possible answers:
Possible Answer 1: Crocosmia
Possible Answer 2: Iridaceae
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response: No

We are evaluating answers to the question: Is It Just Me? was a single by the English rock band from what Suffolk city?
Here are two possible answers:
Possible Answer 1: Lowestoft
Possible Answer 2: Lowestoft, Suffolk
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response: Yes

We are evaluating answers to the question: In what year did the man who shot the Chris Stockley, of The Dingoes, die?
Here are two possible answers:
Possible Answer 1: 1987
Possible Answer 2: The year of 1987
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response: Yes

We are evaluating answers to the question: {User Query}
Here are two possible answers:
Possible Answer 1: {Answer A}
Possible Answer 2: {Answer B}
For this question, is Possible Answer 1 semantically equivalent to Possible Answer 2? Respond with Yes or No.
Response:

Figure 11: Prompt designed to evaluate the semantic equivalence between two responses to a user query.

(3) **Abstain Accuracy:** $\frac{A+D}{A+B+C+D}$, evaluates whether abstention decisions are correct; ideally, LLMs should abstain when it would provide an incorrect answer and vice versa.

(4) **Abstain F1:** the harmonic mean of precision and recall, where precision = $\frac{D}{B+D}$ and recall = $\frac{D}{C+D}$, providing a balanced measure between reliability and answer coverage.

In this paper, we report *Abstain Accuracy* and *Abstain F1*, as our primary goal is to evaluate the abstention capability of the uncertainty scores.

B.2 Threshold Calibration

To determine the threshold τ_{abs}, we perform a parameter sweep using validation sets. To construct a validation set for each dataset, we subsample 100 examples from the training set of each dataset. The only exception is PopQA, which does not have a training set. For PopQA, we instead subsample from the original test set while ensuring that our validation and test sets do not overlap. (As a reminder, as described in Section 5.1, we sample 500 examples for the test set, and the PopQA dataset consists entirely of 14K test samples.) We then generate uncertainty scores for the validation set using R^2C as well as all baseline methods.

After obtaining the validation sets and corresponding uncertainty scores, we sweep the threshold values from 0.4 to 0.95 with an interval of 0.05, evaluating both AbstainAccuracy and AbstainF1. We conduct this procedure across all datasets and for all baseline methods, including RrrC, ReaC, SelfC, P(true), and our proposed approach. We first observe that both metrics exhibit similar behavioral patterns. Our results further show that all methods achieve their best performance at a threshold of 0.9; therefore, we set $\tau_{\text{abs}} = 0.9$.

B.3 Evaluation with AUARC

In Table 2, we present the performance of the abstention task using two threshold-based metrics: AbstainAccuracy and AbstainF1. The decision thresholds are determined through the detailed experiments described in Appendix B.2. However, some studies in the literature [13, 15, 50] adopt the threshold-free Area Under the Accuracy–Rejection Curve (AUARC) [45] as the evaluation metric. While AUARC has the advantage of being independent of a specific threshold, its final score is correlated with the model's overall accuracy. Considering these pros and cons, we also report the abstention task results using AUARC in Table 5. We observe that the model rankings remain consistent with those in Table 2, indicating that the thresholds for AbstainAccuracy and AbstainF1 were appropriately

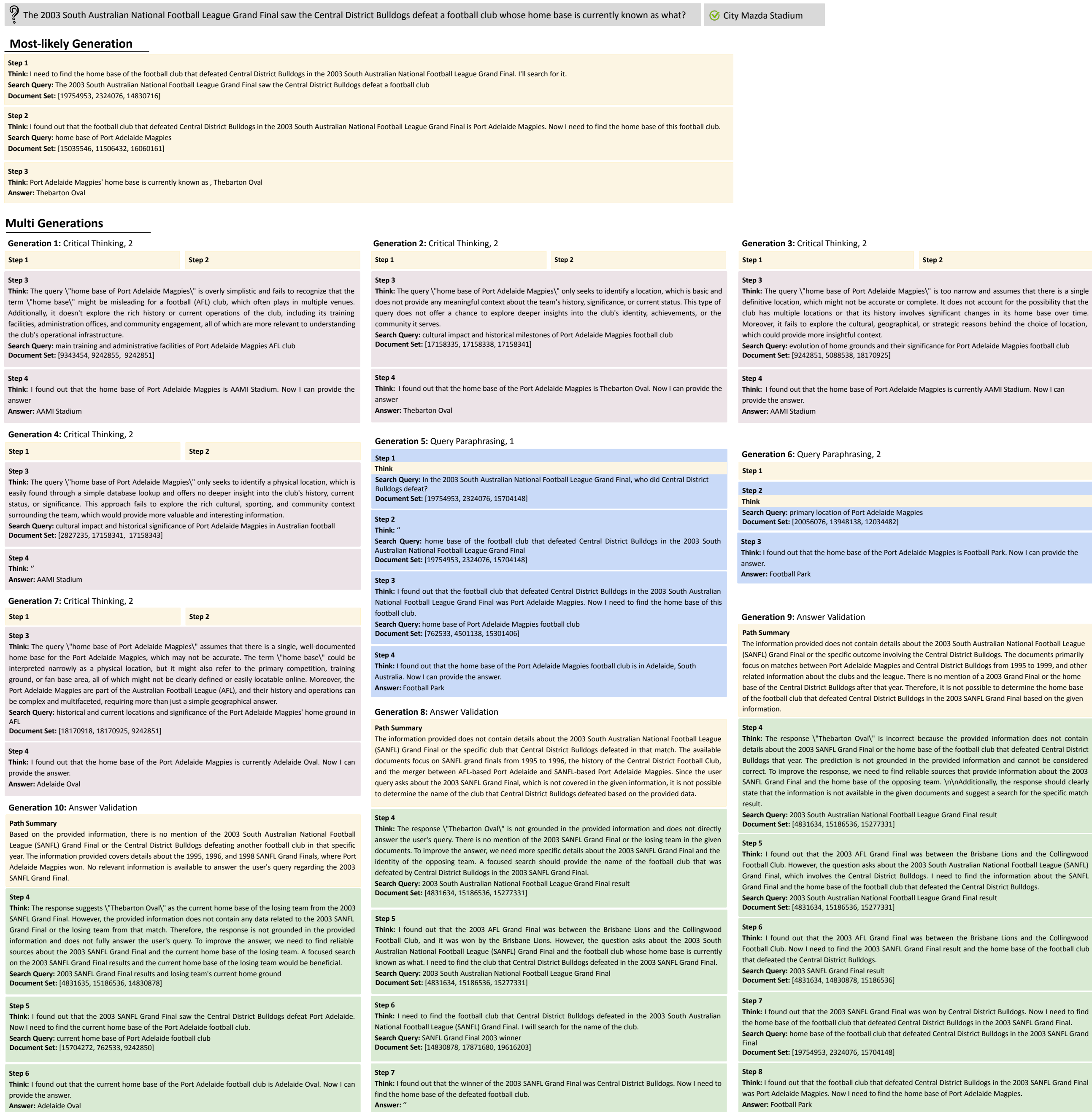

Figure 12: A case study of a successful sample from the HotpotQA dataset using the RAR model (Search-R1). The predicted answer, "Thebarton Oval," is incorrect, and the model assigns it a relatively high uncertainty score of 0.9, indicating a desirable correlation between uncertainty and correctness.

selected. Moreover, even with the threshold-free metric, the R^2C method continues to outperform the other baselines.

C Case Study

Figure 1 illustrates the workflow of the UQ for computing the uncertainty score, which consists of two main steps: Most-likely Generation and Multi-Generations. Based on this workflow, we present two case studies—one successful and one failed, in Figures 12 and 13, respectively. In the successful case, the most-likely response is "Thebarton Oval", which is incorrect. However, in the Multi-Generations step, only one out of ten generated responses is "Thebarton Oval". According to Equations (**??**) and (1), this results in an uncertainty score of 0.9, indicating high uncertainty. In contrast, in the failed case, the most-likely response is "Adrian Lyne", which is correct. Yet, in the Multi-Generations step, only four out

Who was the director of the 1987 American psychological thriller film in which the actor, who played Bob Charles in "Not Necessarily the News", appeared?

✓ Adrian Lyne

Most-likely Generation

Step 1
Think: I need to find the director of the 1987 American psychological thriller film in which the actor who played Bob Charles in \"Not Necessarily the News\" appeared. I'll search for it.
Search Query: actor who played Bob Charles in "Not Necessarily the News"
Document Set: [18789586, 9391749, 17877035]

Step 2
Think: I found out that the actor who played Bob Charles in \"Not Necessarily the News\" is Bryan Murray. Now I need to find the 1987 American psychological thriller film in which he appeared.
Search Query: 1987 American psychological thriller film in which Bryan Murray appeared
Document Set: [17906015, 18129571, 11885203]

Step 3
Think:
Search Query: director of the 1987 American psychological thriller film in which Bryan Murray appeared
Document Set: [17906015, 18129571, 11885203]

Step 4
Think: I found out that the director of the 1987 American psychological thriller film in which Bryan Murray appeared is Adrian Lyne.
Answer: Adrian Lyne

Multi Generations

Generation 1: query paraphrasing, 3

Step 1

Step 2

Step 3
Think
Search Query: Who directed the 1987 psychological thriller featuring Bryan Murray?
Document Set: [20557386, 18129571, 4092928]

Step 4
Think: I need to find the director of the 1987 American psychological thriller film in which Bryan Murray appeared. From the search, I found out that Bryan Murray appeared in the film \"The Bedroom Window". Now I need to find the director of this film.
Search Query: director of The Bedroom Window
Document Set: [20557386, 18129571, 20557385]

Step 5
Think: I need to find the director of the 1987 American psychological thriller film \"The Bedroom Window\" in which Bryan Murray appeared. From the search, I found out that the director of "The Bedroom Window" is Curtis Hanson.
Answer: Curtis Hanson

Generation 2: Query paraphrasing, 3

Step 1

Step 2

Step 3
Think
Search Query: 1987 American psychological thriller with Bryan Murray: name the director
Document Set: [20557386, 18129571, 11885203]

Step 4
Think: I found out that the 1987 American psychological thriller film in which Bryan Murray appeared is \"The Bedroom Window\". Now I need to find the director of this film.
Search Query: director of The Bedroom Window (1987 film)
Document Set: [20557386, 18129571, 20557385]

Step 5
Think: ''
Answer: Curtis Hanson

Generation 3: Answer Validation

Path Summary
Bryan Murray, the actor who played Bob Charles in the soap opera \"Fair City,\" appeared in the 1987 American psychological thriller film \"Fatal Attraction,\" directed by Adrian Lyne. \"Fatal Attraction\" starred Michael Douglas, Glenn Close, and Anne Archer, and it was based on a screenplay by James Dearden, adapted from his 1980 short film \"Diversion.\" This film was well-received critically and became a significant commercial success, grossing substantially at the box office despite initial controversies.

Step 5
Think: The response states that Adrian Lyne was the director of the 1987 film "Fatal Attraction," which aligns with the information provided. However, the answer should be more directly linked to the given information by specifying that Bryan Murray, the actor who played Bob Charles in "Not Necessarily the News," appeared in this film. The response meets the grounding criterion but falls short on completeness. To improve, the answer should explicitly connect the actor's role in "Not Necessarily the News" to their appearance in "Fatal Attraction". Specifically, the response should mention Bryan Murray by name and reference his role in "Not Necessarily the News."
Search Query: information on Bryan Murray's filmography including "Not Necessarily the News" and "Fatal Attraction
Document Set: [19266549, 1897855, 2789204]

Step 6
Think: ''
Search Query: director of the 1987 American psychological thriller film in which the actor, who played Bob Charles in \"Not Necessarily the News\", appeared
Document Set: [2789203, 2789207, 19505329]

Step 7
Think: I need to find the director of the 1987 American psychological thriller film in which the actor, who played Bob Charles in \"Not Necessarily the News\", appeared. I found out that the actor who played Bob Charles in \"Not Necessarily the News\" is Stuart Pankin. Now I need to find the film in which he appeared and was a psychological thriller.
Search Query: who was the director of the 1987 American psychological thriller film in which the actor, who played Bob Charles in \"Not Necessarily the News\", appeared
Document Set: [19505329, 2789203, 4999074]

Step 8
Think: I need to find the 1987 American psychological thriller film in which the actor, who played Bob Charles in \"Not Necessarily the News\", appeared. I found out that the actor who played Bob Charles in \"Not Necessarily the News\" is Stuart Pankin. Now I need to find the film in which he appeared and was a psychological thriller.
Search Query: 1987 American psychological thriller film in which Stuart Pankin appeared
Document Set: [2789205, 2789206, 17906015]

Step 9
Think: I need to find the director of the 1987 American psychological thriller film \"Fatal Attraction\" in which Stuart Pankin appeared.
Answer: Adrian Lyne

Generation 4: Answer Validation

Path Summary
The 1987 American psychological thriller film in which the actor who played Bob Charles in \"Not Necessarily the News\" appeared was \"Fatal Attraction\". The director of \"Fatal Attraction\" was Adrian Lyne. According to the information provided, Bryan Murray played Bob Charles in the soap opera \"Fair City\", not \"Not Necessarily the News\". However, the actor who played Bob Charles in \"Not Necessarily the News\" is not mentioned in the given documents. Since Bryan Murray appeared in \"Fatal Attraction\", the director of that film was Adrian Lyne.

Step 5
Think: The response identifies the director as Adrian Lyne, but it does not directly ground this conclusion in the provided information. It mentions \"Fatal Attraction\" as the 1987 American psychological thriller film featuring the actor who played Bob Charles in \"Not Necessarily the News\", and confirms that Adrian Lyne directed \"Fatal Attraction\". However, the response could be more clearly linked to the specific information provided about the actor's role. Additionally, the response could benefit from clarifying that the correct actor for Bob Charles in \"Not Necessarily the News\" is needed to ensure accuracy. \n\nTo improve, we should directly reference the connection between the actor who played Bob Charles in \"Not Necessarily the News\" and the film \"Fatal Attraction\".
Search Query: which actor played bob charles in not necessarily the news
Document Set: [2789203, 2789207, 2789204]

Step 6
Think: I found out that the actor who played Bob Charles in \"Not Necessarily the News\" appeared in the film \"Fatal Attraction\". Now I need to find the director of the 1987 American psychological thriller film \"Fatal Attraction\".
Search Query: director of Fatal Attraction
Document Set: [11885203, 11885231, 17819054]

Step 7
Think: I found out that the director of the 1987 American psychological thriller film \"Fatal Attraction\" is Adrian Lyne.
Answer: Adrian Lyne

Generation 5: Critical rethinking, 2

Step 1

Step 2

Step 3
Think: The query \"1987 American psychological thriller film in which Bryan Murray appeared\" is overly specific and relies on a known actor\u2019s appearance in a film, which can lead to retrieving only the film's title and basic plot details. This query does not explore the film's cultural impact, the director\u2019s style, or the broader genre characteristics of psychological thrillers from that era, thus limiting the depth and breadth of useful information.
Search Query: influential psychological thriller films of the 1980s and their impact on the genre
Document Set: [988461, 17449124, 158483]

Step 4
Think: I need to find the director of the 1987 American psychological thriller film in which the actor Bryan Murray appeared. I found out that Bryan Murray played Bob Charles in \"Not Necessarily the News\". Now I need to find the psychological thriller film he appeared in. From the search, I found out that Bryan Murray appeared in the film \"The Bedroom Window\" (1987). Now I need to find the director of this film.
Search Query: director of The Bedroom Window (1987 film)
Document Set: [20557386, 18129571, 20557385]

Step 5
Think: I need to find the director of the 1987 American psychological thriller film \"The Bedroom Window\" in which Bryan Murray appeared. From the search, I found out that the director of \"The Bedroom Window\" is Curtis Hanson.
Answer: Curtis Hanson

Generation 6: Critical rethinking, 1

Step 1

Step 2
Think: The query 'actor who played Bob Charles in \"Not Necessarily the News\"' is overly specific and relies on outdated media content that may not be well-documented or easily searchable. It assumes that Bob Charles was a character with significant screen time, which might not be accurate, and it focuses solely on one actor, potentially missing other related information or interpretations of the show.
Search Query: notable characters and actors in the comedy series \"Not Necessarily the News\"
Document Set: [1897854, 5064235, 5064238]

Step 3
Think: I need to find the director of the 1987 American psychological thriller film in which the actor who played Bob Charles in \"Not Necessarily the News\" appeared. I'll search for it.
Search Query: actor who played Bob Charles in \"Not Necessarily the News\"
Document Set: [2789203, 2789207, 18789586]

Step 4
Think: ''
Search Query: 1987 American psychological thriller film
Document Set: [18129571, 20557386, 11885203]

Step 5
Think: I found out that the 1987 American psychological thriller film in which Bryan Murray appeared is \"The Bedroom Window\". Now I need to find the director of this film.
Search Query: director of The Bedroom Window (1987 film)
Document Set: [20557386, 18129571, 20557385]

Step 6
Think: ''
Answer: Curtis Hanson

Generation 7: Answer Validation

Path Summary
The actor who played Bob Charles in \"Not Necessarily the News\" is Bryan Murray. Bryan Murray appeared in the 1987 American psychological thriller film \"Fatal Attraction,\" directed by Adrian Lyne. This film starred Michael Douglas, Glenn Close, and Anne Archer and was based on a short film called \"Diversion.\" It received generally positive critical response and became a significant commercial success at the box office.

Step 5
Think: The response states that Adrian Lyne was the director of the 1987 American psychological thriller film featuring Bryan Murray, who played Bob Charles in \"Not Necessarily the News\". Evaluating this against the criteria: \n1. The response is grounded in the provided information, as the document mentions that Bryan Murray appeared in \"Fatal Attraction,\" directed by Adrian Lyne.\n2. However, the response only partially answers the query, as it does not specify the film title. To fully satisfy the query, the answer should mention that Bryan Murray appeared in \"Fatal Attraction.\"\n\nTo improve the response, we should include the name of the film in the answer.
Search Query: title of the 1987 American psychological thriller film starring Bryan Murray
Document Set: [18129571, 20557386, 11885203]

Step 6
Think: I found out that the 1987 American psychological thriller film in which the actor who played Bob Charles in \"Not Necessarily the News\" appeared was \"Fatal Attraction\". Now I need to find the director of this film.
Search Query: director of Fatal Attraction
Document Set: [11885203, 11885231, 17819054]

Step 7
Think: I found out that the director of the 1987 American psychological thriller film \"Fatal Attraction\" is Adrian Lyne.
Answer: Adrian Lyne

Generation 8: Critical rethinking, 1

Step 1

Step 2
Think: The query \u2018actor who played Bob Charles in \"Not Necessarily the News\"' is overly specific and relies on a single piece of information that is likely to be obscure and difficult to verify. It fails to explore the broader context of the character, the show, or the actor\u2019s career, thereby limiting the scope of useful information that could be derived from the search results.
Search Query: background and significance of the character Bob Charles in \"Not Necessarily the News\"
Document Set: [2789203, 2789207, 1897855]

Step 4
Think: I found out that the actor who played Bob Charles in \"Not Necessarily the News\" is Stuart Pankin. Now I need to find the 1987 American psychological thriller film in which he appeared.
Search Query: 1987 American psychological thriller film in which Stuart Pankin appeared
Document Set: [2789205, 2789206, 17906015]

Step 5
Think: I found out that Stuart Pankin appeared in the film \"Fatal Attraction\" directed by Adrian Lyne. Now I can provide the answer.
Answer: Adrian Lyne

Generation 9: query paraphrasing, 2

Step 1

Step 2
Think
Search Query: 1987 psychological thriller movie featuring Bryan Murray
Document Set: [20557386, 18129571, 4092928]

Step 3
Think: I found out that the 1987 American psychological thriller film featuring Bryan Murray is \"The Bedroom Window\". Now I need to find the director of this film.
Search Query: director of The Bedroom Window (1987 film)
Document Set: [20557386, 18129571, 20557385]

Step 4
Think: I found out that the director of the 1987 American psychological thriller film \"The Bedroom Window\" is Curtis Hanson.
Answer: Curtis Hanson

Generation 10: query paraphrasing, 2

Step 1

Step 2
Think
Search Query: Bryan Murray's role in an American psychological thriller from 1987
Document Set: [17045675, 17045676, 18789586]

Step 3
Think: ''
Search Query: 1987 American psychological thriller film
Document Set: [18129571, 20557386, 11885203]

Step 4
Think: I found out that the 1987 American psychological thriller film in which Bryan Murray appeared is \"The Bedroom Window\". Now I need to find the director of this film.
Search Query: director of The Bedroom Window (1987 film)
Document Set: [20557386, 18129571, 20557385]

Step 5
Think: I found out that the director of the 1987 American psychological thriller film \"The Bedroom Window\" is Curtis Hanson.
Answer: Curtis Hanson

Figure 13: A case study of a failed sample from the HotpotQA dataset using the RAR model (Search-R1). Although the predicted answer, "Adrian Lyne," is correct, the model assigned a relatively high uncertainty score of 0.6, revealing a mismatch between uncertainty and correctness.

of ten responses are "Adrian Lyne", again leading to a relatively high uncertainty score. In this sample, we observe that the answer validation action performs correctly and generates a response similar to the most-likely one. However, the query paraphrasing action produces different responses, and two of the three critical rethinking actions also yields a different result. This observation supports the discussion in Section 6.5, which highlights the potential impact of action selection configurations—whether applied per sample, per model, or per dataset.