

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 all these sources and introduce Retrieval-Augmented Reasoning Consistency (R^2C)—a novel UQ method for RAR. The core idea of R^2C 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 one another's inputs, enabling us to capture uncertainty arising from both components. Experiments on five popular RAR systems across diverse QA datasets show that R^2C improves AUROC by over 5% on average compared to the state-of-the-art UQ baselines. Extrinsic evaluations using R^2C as an external signal further confirm its effectiveness for two downstream tasks: in Abstention, it achieves ~5% gains in both F1Abstain and AccAbstain; in Model Selection, it improves the exact match by ~7% over single models and ~3% over selection methods.

ACM Reference Format:

Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. 2018. Uncertainty Quantification for Retrieval-Augmented Reasoning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 14 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Retrieval-augmented generation (RAG) is widely used for knowledge-intensive tasks, but remains limited in addressing complex multi-step reasoning [24, 44, 49, 50]. 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) [19, 24, 52, 54]. 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) a model has in its predictions [16, 18, 25, 28, 34, 65]. 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, 11, 60]. The limited work on UQ for RAG [38, 48] 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: *retriever*, which may provide irrelevant or partially relevant retrieved documents and potentially mislead the model's reasoning and response generation processes; and *generation*, 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 **Retrieval-Augmented Reasoning Consistency** (R^2C), a novel UQ method that, unlike previous approaches, accounts for multiple sources of uncertainty in RAR. The central idea of R^2C 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 perturbations enable models to arrive at diverse final responses for uncertain generations. The uncertainty 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^2C 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^2C on two

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

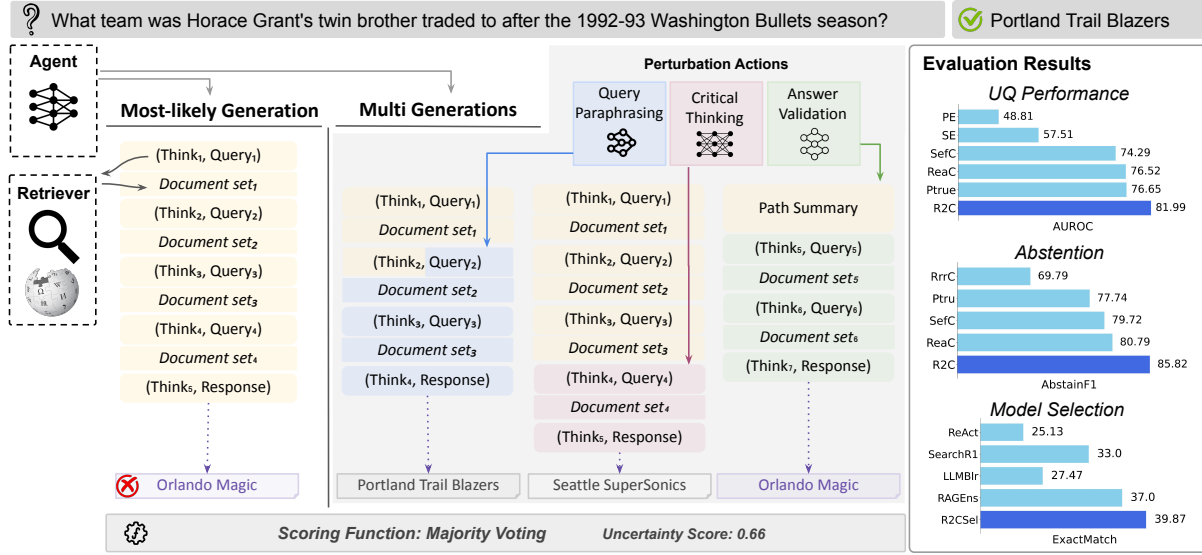


Figure 1: R²C overview. Given a user query, the agent (LLM) first generates the most-likely reasoning path leading to the most-likely response (left, yellow). To estimate uncertainty, R²C creates multiple perturbed generations by randomly altering states in the reasoning path (middle, gray). The uncertainty score is then derived via majority voting. R²C significantly outperforms established UQ methods and achieves significant improvements on two downstream tasks: abstention and model selection.

downstream tasks: (i) Abstention:¹ the task of generating ‘I don’t know’ when the model is uncertain about its output [14, 33], and (ii) Model Selection:² the task of selecting a final answer from a pool of candidates generated by multiple systems [7, 15]. Our experimental results indicate that R²C 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 R²C in both direct evaluation and extrinsic evaluation on downstream tasks, we investigate the factors that contributes to its effectiveness. We show that R²C 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 other methods, measured by the inverse of the average pairwise cosine similarity between queries [8, 66]. This diversity in queries and documents demonstrates that R²C generates diverse, yet relevant reasoning paths through our controlled perturbation mechanism. As a result, this enables the method to achieve uncertainty 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 R²C is not only the most effective method of its kind but also a relatively more efficient uncertainty quantification approach. To summarize, the main contributions of this paper are:

- (1) We propose R²C, 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 dataset 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 R²C 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 [9]. 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 [35, 47]. In contrast, Active RAG allows retrieval to occur throughout the generation process, either in fixed intervals or dynamically, whenever additional information is needed [23, 51, 56]. Retrieval-Augmented Reasoning (RAR) is a recent extension of RAG that integrates retrieval with reasoning, aiming to improve the interaction between LLMs and retrievers [52, 54]. For example, Self-Ask [39] decomposes complex questions into follow-up queries and intermediate answers, while ReAct [64] defines a set of actions, such as search, look up, and finish, to structure interactions with external resources. More recent and effective models such as ReSearch [5] and Search-R1 [24] 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.

¹Also referred to as selective prediction in the literature [12, 59].

²Also referred to as selection-based model ensemble in the literature [20].

Algorithm 1 R²C: 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: Uncertainty score $U(x, r)$

```

1: for  $t = 1$  to  $N$  do  $s_t \leftarrow \pi_\theta(s_{t-1}, a_{t-1})$             $\triangleright$  Generate the
   most-likely path iteratively
2:  $r \leftarrow s_N$                                             $\triangleright$  Capture the most-likely response
3:  $R_G \leftarrow \emptyset$                                         $\triangleright$  Initialize multi-generation response set
4: for  $b = 1$  to  $B$  do
5:    $a^* \sim \mathcal{U}(A^*)$                                         $\triangleright$  Sample an action from  $A^*$ 
6:   if  $a^* = a_{av}$  then  $s_t \leftarrow s_N$                   $\triangleright$  Select the last state for  $a_{av}$ 
7:   else  $s_t \sim \mathcal{U}(s_1, s_{N-1})$                         $\triangleright$  Sample a state
8:   end if
9:    $s_{t+1} = \pi_\theta(s_t, a^*)$                                 $\triangleright$  Apply action  $a^*$  at state  $s_t$ 
10:  for  $i = 1$  to  $N^b$  do  $s_{t+i+1} \leftarrow \pi_\theta(s_{t+i}, a_{t+i})$ 
11:     $r^b \leftarrow s_{N^b}$                                     $\triangleright$  Capture the sampled response
12:     $R_G = R_G \cup \{r^b\}$                                 $\triangleright$  Update the sampled response set
13:  end for
14:  $u = 1 - \frac{1}{B} \sum_{b=1}^B \mathbb{I}(r^b = r)$             $\triangleright$  Compute the uncertainty score
15: return  $u$ 

```

Uncertainty Quantification. Broadly, existing UQ methods can be divided into two categories: white-box approaches, which leverage token-level probabilities and entropy [3, 11, 25, 28, 46, 60], and black-box approaches, which rely only on the final textual outputs [31, 53]. 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 [48] or utility models [38]. Some path-based approaches focus on assessing the consistency of reasoning paths in reasoning tasks [29, 40, 58]. 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 [67] learns aggregation weights to merge per-step uncertainties, but it relies on ground-truth labels from the test domain. In contrast, we propose R²C, 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, $P(s_{t+1} | s_t, a_t)$ 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 $p_\pi(a_t | s_t)$ 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 [64]. The set of possible actions in each state is $A = \{a_{ret}, a_{ans}\}$, 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 $P(s_{t+1} | s_t, a_t)$ is determined by the LLM itself. An explicit reward function is not always required, as some agents operate without additional training.

Consistency-based Uncertainty Quantification. The core idea of consistency-based methods is to generate multiple responses for a given query by varying either the input prompt or the temperature parameter used in stochastic decoding [27]. The pairwise similarity among these responses is then computed and aggregated into a single uncertainty score [4, 31]. 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 [58], changing the input [27], or altering the reasoning path [29, 40]. A transformation function ϕ is then applied to convert R_G into an uncertainty score. Our method builds on the reasoning path perturbation approach and employs *Majority Voting* [58] as the transformation function, where the degree of consistency is measured by the proportion of sampled responses that match the most-likely response:

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

Accordingly, the uncertainty score is defined as:

$$U(x, r) = 1 - C(x, r). \quad (2)$$

4 Methodology

We propose **Retrieval-Augmented Reasoning Consistency**, R²C, to address UQ in RAR models. R²C 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 R²C is to perturb the reasoning paths of these multiple generations through a set of perturbation actions, denoted as A^* . In MDP terms, R²C 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 R²C: (i) Query Paraphrasing, a_{qp} , (ii) Critical Rethinking, a_{cr} , and (iii) Answer Validation, a_{av} . In the following sections, we first formally describe how R²C perturbs the generation path and then describe perturbation actions.

4.1 R²C: Retrieval-Augmented Reasoning Consistency

R²C models UQ 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:

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

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:

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

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 majority voting function (Eq. (1)) and the uncertainty scoring function (Eq. (2)) are applied to compute the final uncertainty score. Algorithm 1 provides a detailed description of the entire process in R²C.

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 [22, 29, 32]. 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} :

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

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, as shown in Figure 7 in Appendix A.

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, 22, 24, 29]. 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} :

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

Conceptually, if the reasoning path so far has been incorrect, this action enables the system to adjust toward a more reliable trajectory. If the path has been correct, CR strengthens its validity, thereby increasing confidence in the final outcome. Figure 8 in Appendix A illustrates the instruction prompt used for this action.

A3: Answer Validation (AV). [17, 21, 43] RAR models face difficulties with validating their final response, detecting whether the generated response meets certain criteria of the query [17, 21, 43, 57]. 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.

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, 30, 45], 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 $D = \{D_1, \dots, D_{N-1}\}$ be the set of documents retrieved at states $[s_1, s_2, \dots, s_{N-1}]$. 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} :

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

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 $\langle \tau_{N+1}, q_{N+1} \rangle$. The prompt instructions for the path summary model and the action are presented in Figures 9 and 10 in Appendix A, respectively.

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 R²C on both single-hop and multi-hop QA tasks using the PopQA [35], HotpotQA [63], and Musique [55] datasets. Following prior work [4, 23, 36, 56, 64], we randomly sampled 500 queries from each dataset as the test set. We will release our sampled queries for improving reproducibility of our work. For the retrieval corpus, we use the 2018 Wikipedia dump [26].³

³<https://huggingface.co/datasets/PeterJinGo/wiki-18-corpus>

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

RAG	SelfAsk [39]			ReAct [64]			Search-o1 [30]			ReSearch [5]			Search-R1 [24]			Avg.
Uncer. M.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	Popqa	Hotp.	Musi.	
PE [25]	55.34	59.11	48.61	36.75	39.95	41.14	39.86	53.44	51.59	40.93	61.03	56.69	48.49	48.27	50.90	48.81
SE [28]	64.26	68.01	54.68	49.73	40.37	41.69	67.56	65.84	54.61	53.88	64.72	63.43	59.11	53.45	61.38	57.51
MARS [3]	54.70	59.48	51.53	41.29	40.42	41.97	43.28	57.69	56.80	40.03	61.76	57.58	48.33	49.28	49.94	50.27
SAR [11]	51.65	63.05	52.31	29.56	32.90	31.54	40.92	52.37	44.66	41.67	62.75	51.47	45.87	46.40	45.22	46.16
LARS [60]	73.97	70.03	66.45	79.95	68.28	71.87	83.79	71.42	66.19	76.95	66.12	71.64	71.54	67.08	71.62	71.79
NumSS [28]	71.31	65.19	62.75	74.90	63.73	62.59	78.88	63.74	62.12	80.42	64.46	69.34	69.76	64.10	65.72	67.93
EigV [31]	70.53	66.80	62.13	76.48	66.44	57.44	79.72	68.58	64.75	80.63	66.43	66.31	69.33	66.54	64.27	68.43
ECC [31]	72.89	69.95	64.11	80.27	69.51	61.92	81.98	69.27	69.65	81.85	68.49	72.55	70.87	67.76	67.65	71.25
Deg [31]	70.53	66.79	61.77	76.70	67.70	58.12	80.69	68.87	66.29	81.67	67.15	67.85	69.51	66.75	64.53	68.99
RrrC [29]	71.14	71.17	81.28	48.02	68.30	<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 [58]	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 [40]	<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) [25]	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>
R²C(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

The number of retrieved documents is fixed to three across all models [24, 35, 39, 64].

Evaluation Metrics: To evaluate the quality of outputs, we follow Jin et al. [24] 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, 25, 28]. As suggested by Perez-Beltrachini and Lapata [38], significant differences between two AUROC values are assessed using the paired De Long test [10].

Models: In line with prior work in RAR [5, 24], we employ *Qwen-2.5-7B-Instruct* [62] 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, 48]. Retrieval is performed using a two-stage re-ranking pipeline: BM25 [42] 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.⁴ All experiments are conducted on four Nvidia A100 GPU, 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) [58] ignores the most-likely generation and instead produces diverse set of reasoning paths, starting from scratch. Reasoning Consistency (ReaC) [40] 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) [29] 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 [25], SE [28], MARS [3], LARS [60], and SAR [11]. The black-box methods are NumSS [28], EigV, ECC, Deg [31], and P(true) [25]. All of these methods rely on generations obtained in the same way as SelfC. All UQ methods are implemented using TruthTorchLM [61].

5.2 Extrinsic UQ Evaluation via Abstention

Task Formulation. Declining to respond due to uncertainty is an important application of UQ [14]. An abstention function determines whether the model should withhold an answer. In our setup, this decision is guided by uncertainty scores. We introduce a threshold τ_{abs} : if the uncertainty exceeds τ_{abs} , the model abstains; otherwise, it produces an answer. Formally, given a response r to a query x and an uncertainty function U , we define the abstention function f_{abs} as:

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

Baselines & Evaluation Metrics. As baselines, we implement the abstention task using uncertainty scores derived from different UQ methods, including P(true), RrrC, ReaC, SelfC, and R²C. For evaluation, we follow Feng et al. [14] and report results on AbstainAccuracy and AbstainF1. *AbstainAccuracy* evaluates whether the abstain decisions are correct, and *AbstainF1* captures the balance between reliability and answer coverage. The evaluation metrics are explained in detail in Appendix B.1. The threshold τ_{abs} for each UQ method is selected by performing sweep using validation sets; see Appendix B.2 for more details.

5.3 Extrinsic UQ Evaluation via Model Selection

Task Formulation. Model selection (or selection-based model ensemble [7, 20]) aims to select a final response for a question based on multiple candidate responses generated by different systems [7].

⁴<https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-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 [39]			ReAct [64]			Search-o1 [30]			ReSearch [5]			Search-R1 [24]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
<i>Abstain Accuracy</i>																
RrrC [29]	61.4	65.6	73.0	60.8	72.2	82.4	56.6	66.2	68.2	59.2	64.4	64.0	60.6	64.0	63.6	65.48
P(true) [25]	<u>70.6</u>	<u>70.8</u>	<u>80.4</u>	70.4	67.2	75.0	72.6	65.8	75.0	<u>78.8</u>	<u>71.4</u>	78.2	<u>70.8</u>	<u>70.0</u>	74.4	72.76
SelfC [58]	68.6	64.2	76.2	74.0	<u>75.8</u>	87.6	75.0	69.0	90.2	75.2	68.6	83.6	65.4	62.8	77.8	74.27
ReaC [40]	69.2	64.6	<u>80.4</u>	<u>74.4</u>	75.4	<u>89.4</u>	<u>75.6</u>	<u>71.4</u>	87.4	77.2	68.6	80.8	67.8	68.8	<u>80.6</u>	<u>75.44</u>
R²C (our)	77.2[†]	74.4[†]	88.6[†]	77.0	76.8	90.4	80.4[†]	74.4	<u>89.4</u>	81.6[†]	74.8[†]	84.0	77.0[†]	73.4	84.4[†]	80.25
<i>Abstain F1</i>																
RrrC [29]	62.52	70.54	82.75	71.59	80.82	89.69	56.33	71.50	78.99	54.86	63.37	73.68	55.12	61.20	73.92	69.79
P(true) [25]	<u>73.60</u>	<u>75.33</u>	88.27	72.07	73.46	84.58	76.50	71.73	84.51	81.20	73.46	85.82	<u>71.03</u>	<u>71.15</u>	83.37	77.74
SelfC [58]	72.60	69.51	85.71	<u>79.10</u>	<u>83.80</u>	93.18	80.50	78.07	94.64	<u>81.71</u>	<u>75.19</u>	90.57	62.31	62.34	86.51	79.72
ReaC [40]	73.54	70.45	<u>88.41</u>	79.01	83.17	<u>94.19</u>	81.24	<u>79.48</u>	93.03	81.25	72.60	88.43	67.60	71.11	<u>88.27</u>	80.79
R²C (our)	82.35[†]	81.76[†]	93.77[†]	83.69[†]	85.08[†]	94.81	86.42[†]	82.22[†]	<u>94.27[†]</u>	84.40[†]	78.71[†]	<u>90.49[†]</u>	80.80[†]	77.57[†]	90.97[†]	85.82

Formally, given a user query x and a set of systems $\{S_1, S_2, \dots, S_M\}$, 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 $R = \{r_1, r_2, \dots, r_M\}$.

R²C Select. The proposed R²C Select utilizes the uncertainty scores derived from R²C. Our method groups semantically similar responses into K clusters. Following [29, 61], we use *Qwen-2.5-7B-Instruct* to compute pairwise semantic similarities between candidate responses and cluster similar ones together. We then assign an uncertainty score to each cluster c_i :

$$R_C = \{\langle r_1, u_{c_1} \rangle, \langle r_2, u_{c_1} \rangle, \dots, \langle r_K, u_{c_K} \rangle\}.$$

The uncertainty score of each cluster $c = \{r_i\}_{i=1}^m$ is computed by aggregating the uncertainty scores of its members: $u_c = \sum_{r_i \in c} U(x, r_i)$, where the U function provides the uncertainty score for the response r_i . The final response $\hat{r}$ is the response with the lowest uncertainty score: $\hat{r} = \arg \min_j u_{c_j}$. If no clustering is applied, each response constitutes a cluster and $M = K$; we refer to this variation as *R²C Select w/o clustering*.

Baselines & Evaluation Metrics. We evaluate R²C Select performance against both single RAG and RAR systems, as well as existing selection-based ensemble approaches. As baselines, we consider LLMBLender [20] and RAGEnsemble [6]. 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 R²C Select, the response candidates are obtained from SelfAsk [39], ReAct [64], Search-o1 [30], ReSearch [5], and Search-R1 [24]. For evaluation, we report the correctness of the final answer using exact match.

6 Results

We present a set experiments that address the following research questions: **RQ1**: How does R²C perform in quantifying uncertainty for different RAR models? (Sec. 6.1), **RQ2**: How does R²C 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 R²C? (Sec. 6.3), **RQ4**: How does R²C balance effectiveness and efficiency? (Sec. 6.4), **RQ5**: What is the effect of different actions in the performance of R²C? (Sec. 6.5).

6.1 Uncertainty Quantification Performance

RQ1 evaluates the performance of R²C 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., PE [25], SE [28], MARS [3], and SAR [11], 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 [28], EigV, ECC, and Deg [31], 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 as a promising direction for future research.

Path-based methods are another group of approaches that we evaluate, including the proposed R²C. 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 perform relatively 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, **R²C** 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.

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
<i>Vanilla LLM & RAG</i>				
Direct	18.8	20.8	2.6	14.1
CoT	17.6	22.2	5.8	15.2
Vanilla RAG	30.2	18.6	4.4	17.7
IRCoT [56]	34.6	27.2	5.4	22.4
FLARE [23]	31.6	25.2	8.4	21.7
DRAGIN [51]	28.6	23.2	4.6	18.8
<i>Retrieval-Augmented Reasoning (RAR)</i>				
SelfAsk [39]	35.6	33.0	10.4	26.3
ReAct [64]	36.8	27.8	10.8	25.1
Search-o1 [30]	33.2	29.0	10.0	24.1
ReSearch [5]	38.6	38.8	16.6	31.3
Search-R1 [24]	41.6	41.4	16.0	33.0
<i>Model Selection RAR</i>				
Random	31.4	31.6	10.6	24.5
LLMBLender [20]	34.4	36.0	12.0	27.5
- w/o clustering	35.6	31.8	9.6	25.7
- w/o clus. & unc.	32.0	26.0	8.6	22.2
RAGEnsemble [6]	<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. & unc.	43.2	37.2	13.0	31.1
R²CSelect (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

6.2 Extrinsic Evaluation Results

RQ2 evaluates the performance of R²C 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 captures the effectiveness of detecting abstain samples. Considering Abstain Accuracy, R²C outperforms the baselines in all cases except on the Musique dataset and performs on par with the second-best baseline in ReAct, Search-o1, and ReSearch. Considering Abstain F1, R²C significantly outperforms the best baseline in all setup, except for ReAct on Musique. This indicates that R²C scores are more reliable in detecting than correct abstentions than correct non-abstentions. Overall, compared to other UQ methods, R²C 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 R²C are reliable enough for the system to decide when to refrain from answering.

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, **R²C Select** achieves

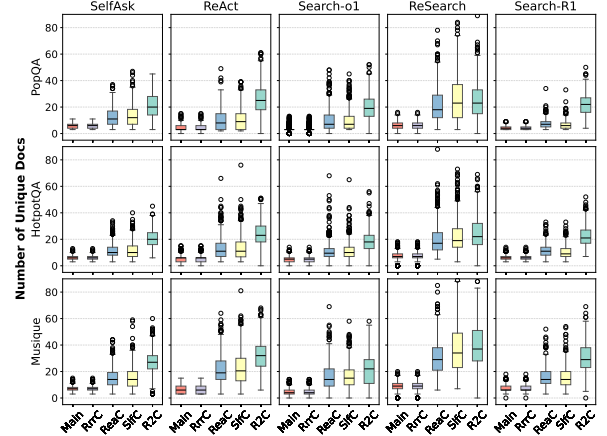


Figure 2: Distribution of the number of unique retrieved documents for the most-likely path (main) and multi-generations.

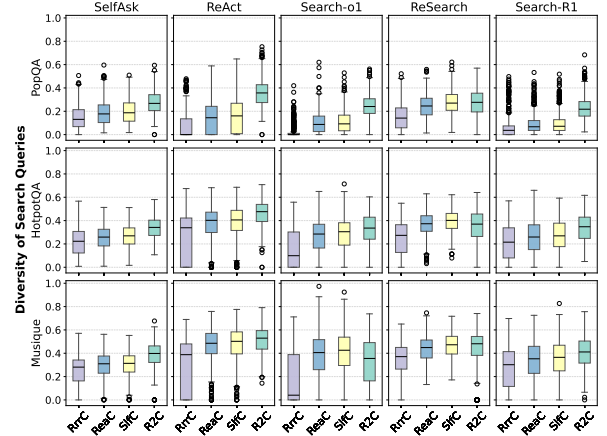


Figure 3: Distribution of diversity scores for search queries generated for each user query across reasoning paths.

the best performance, significantly outperforming all RAR and RAG systems, with 3.7% average improvement on HotpotQA and Musique datasets. **These results confirm that R²C score is a reliable criterion, not only within a single system but also across different systems.** Interestingly, using R²C uncertainty scores alone, without clustering, improves selection-based ensemble models in comparable settings, highlighting informativeness of R²C scores even for models that are not trained on uncertainty.

6.3 Explaining the Strength of R²C

RQ3 examines why R²C outperforms other UQ methods. To this end, we analyze the diversity of reasoning paths as an indicator of how effectively R²C 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, R²C retrieves 24.71 unique documents for a single uncertainty 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 LLM’s thinking process. Query diversity is measured

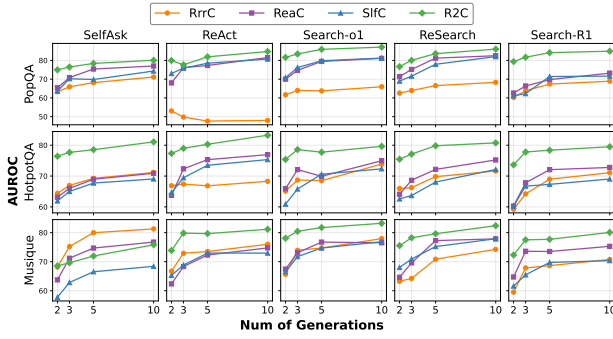


Figure 4: Performance of UQ methods with varying numbers of generations.

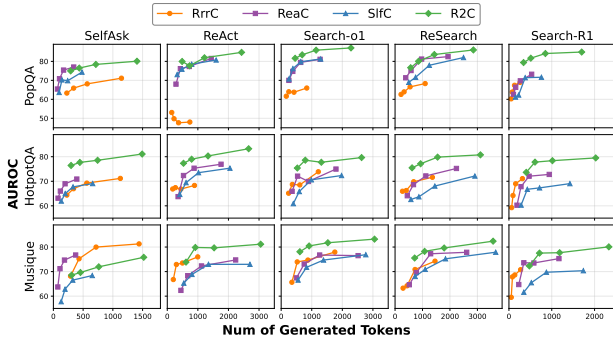


Figure 5: Performance of UQ methods with varying number of generated tokens, illustrating the trade-off between effectiveness and efficiency.

based on the average pair-wise similarity of generated queries [8, 66]. Formally, given n search queries $\{q_1, \dots, q_n\}$, 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 $\cos(\cdot)$ 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^2C achieves a diversity score of 0.35, while RrrC, SIFC, and ReaC achieve 0.20, 0.28, and 0.30, respectively. These results indicate that R^2C generates more diverse search queries when estimating uncertainty. These findings indicate R^2C effectively captures the uncertainty of both the retriever and the generator by sequentially diversifying their inputs.

6.4 Effectiveness vs. Efficiency

RQ4 explores the trade-off between 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 SIFC and ReaC, requiring 10 generations to reach a similar level.

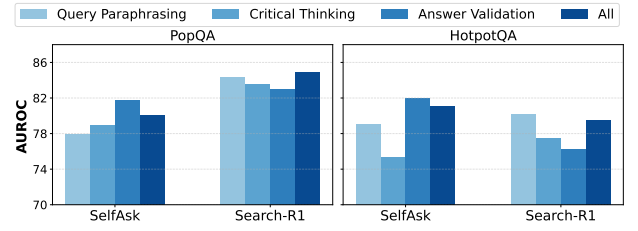


Figure 6: Performance of R^2C with different action sets.

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 [38]. Figure 5 reports AUROC performance for equal numbers of generated tokens. 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 of 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.

6.5 Action Selection

RQ5 investigates the roll of 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.

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 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 publication is part of 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).

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] 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).
- [6] Yifei Chen, Quanting 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).
- [7] Zhiyun 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
- [8] 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.
- [9] Florin Cucanasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, 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.
- [10] 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.
- [11] 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.
- [12] 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).
- [13] 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.
- [14] 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.
- [15] 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*.
- [16] James Harrison, John Willes, and Jasper Snoek. 2024. Variational Bayesian Last Layers. In *The Twelfth International Conference on Learning Representations, ICLR*. OpenReview.net.
- [17] 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.
- [18] 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>
- [19] 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.
- [20] 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.
- [21] 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.
- [22] 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).
- [23] 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.
- [24] 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).
- [25] Saurav Kadavath et al. 2022. Language Models (Mostly) Know What They Know. *abs/2207.05221* (2022).
- [26] 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.
- [27] Anton Korikov, Pan Du, Scott Sanner, and Navid Rekabsaz. 2025. Batched Self-Consistency Improves LLM Relevance Assessment and Ranking. *CoRR* abs/2505.12570 (2025).
- [28] 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*.
- [29] 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.
- [30] Xiaoxi Li, Quanting 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).
- [31] Zhen Lin, Shubendu Trivedi, and Jimeng Sun. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Trans. Mach. Learn. Res.* 2024 (2024).
- [32] 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.
- [33] 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.
- [34] Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *9th International Conference on Learning Representations, ICLR*.
- [35] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khoshabi, 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.
- [36] Viktor Moskvoretiskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Kononov, 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.
 - [37] 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.
 - [38] Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty Quantification in Retrieval Augmented Question Answering. *arXiv preprint arXiv:2502.18108* (2025).
 - [39] 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 2023*, 5687–5711.
 - [40] Zhenyong Qi, Mingyuan MA, Jiahang Xu, Li Lina Zhang, Fan Yang, and Mao Yang. 2025. Mutual Reasoning Makes Smaller LLMs Stronger Problem-Solver. In *The Thirteenth International Conference on Learning Representations*.
 - [41] Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-Evaluation Improves Selective Generation in Large Language Models. In *Proceedings of "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.
 - [42] 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.
 - [43] 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
 - [44] 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.
 - [45] 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).
 - [46] Heydar Soudani. 2025. Enhancing Knowledge Injection in Large Language Models for Efficient and Trustworthy Responses. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, 4211.
 - [47] 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.
 - [48] 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.
 - [49] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Data Augmentation for Conversational AI. In *Companion Proceedings of the ACM on Web Conference 2024, WWW*, 1234–1237.
 - [50] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. A Survey on Recent Advances in Conversational Data Generation. *CoRR abs/2405.13003* (2024).
 - [51] Weihang Su, Yichen Tang, Qingyao Ai, Zhijiang 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.
 - [52] 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.
 - [53] 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.
 - [54] 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.
 - [55] 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.
 - [56] 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.
 - [57] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Su. 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.
 - [58] 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*.
 - [59] 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.
 - [60] 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*.
 - [61] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Sungmin Kang, Alperen Öziş, Hayrettin Eren Yildiz, Mitash Ashish Shah, Zhiqi Huang, Anoop Kumar, Alf Samuel, Daben Liu, et al. 2025. TruthTorchLM: A Comprehensive Library for Predicting Truthfulness in LLM Outputs. *arXiv preprint arXiv:2507.08203* (2025).
 - [62] An Yang et al. 2024. Qwen2.5 Technical Report. *CoRR abs/2412.15115* (2024).
 - [63] 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.
 - [64] 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*.
 - [65] 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*.
 - [66] 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.
 - [67] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyao 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 R²C 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, 29, 60].

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.

B Abstention Task

B.1 Evaluation Metrics

We evaluate R²C on the abstention task in Section 5.2. Following Feng et al. [14], we adopt their definitions for the evaluation metrics. Assume we have a confusion matrix with four elements, each

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.

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?
- (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.

Table 4: 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 [39]			ReAct [64]			Search-o1 [30]			ReSearch [5]			Search-R1 [24]			Avg.
Uncer. M.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	Popqa	Hotpot	Musiq.	
RrrC [29]	52.68	45.25	<u>19.42</u>	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) [25]	56.26	<u>49.53</u>	18.35	56.53	38.82	15.63	54.47	39.75	15.37	60.42	55.03	29.54	<u>61.11</u>	<u>58.72</u>	24.55	42.27
SelfC [58]	55.28	44.63	15.35	59.88	45.13	<u>20.88</u>	56.09	41.83	21.10	<u>62.36</u>	54.38	<u>29.55</u>	58.31	51.53	24.76	42.67
ReaC [40]	<u>56.47</u>	48.32	18.72	<u>60.81</u>	<u>45.48</u>	20.52	<u>56.98</u>	<u>42.55</u>	<u>21.61</u>	63.13	<u>55.21</u>	29.63	54.02	55.73	<u>27.25</u>	<u>43.83</u>
R²C (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.

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 R²C as well as all baseline methods.

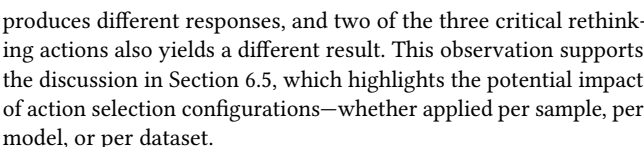
After obtaining the validation sets and corresponding uncertainty scores, we sweep the threshold values from 0.4 to 0.95 with an interval of 0.05, evaluating both AbstainAccuracy and AbstainF1. We conduct this procedure across all datasets and for all baseline methods, including RrrC, ReaC, SelfC, P(true), 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 [12, 13, 41] adopt the threshold-free Area Under the Accuracy-Rejection Curve (AUARC) [37] 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 4. We observe that the model rankings remain consistent with those in Table 2, indicating that the thresholds for AbstainAccuracy and AbstainF1 were appropriately selected. Moreover, even with the threshold-free metric, the R²C 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 (1) and (2), this results



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 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's 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's 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, 1749124, 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 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: 1. The response is grounded in the provided information, as the document mentions that Brian Murray appeared in "Fatal Attraction," directed by Adrian Lyne. 2. 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." 3. To 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 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: {1857854, 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 9: query paraphrasing, 2		
Step 1		
Step 2 Think: Search		