Axiomatic Analysis of Uncertainty Estimation For Retrieval Augmented Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are valued for their strong performance across various tasks, but they also produce inaccurate or misleading outputs. Uncertainty Estimation (UE) quantifies the model's confidence and helps users assess response reliability. However, existing UE methods have not been thoroughly examined in scenarios like Retrieval-Augmented Generation (RAG), where the input prompt includes non-parametric knowledge. This paper shows that current UE methods cannot reliably assess correctness in the RAG setting. We further propose an axiomatic framework to identify deficiencies in existing methods and guide the development of improved approaches. Our framework introduces five constraints that an effective UE method should meet after incorporating retrieved documents into the LLM's prompt. Experimental results reveal that no existing UE method fully satisfies all the axioms, explaining their suboptimal performance in RAG. We further introduce a simple yet effective calibration function based on our framework, which not only satisfies more axioms than baseline methods but also improves the correlation between uncertainty estimates and correctness.

1 Introduction

Large language models (LLMs) specializing in natural language generation (NLG) have recently demonstrated promising capabilities across various tasks (Mallen et al., 2023). However, LLMs are prone to generating incorrect information for multiple reasons, such as lack of parametric knowledge, temporal knowledge shifts (Zhao et al., 2024a; Kordjamshidi et al., 2024), or noisy information introduced through retrieved documents in Retrieval-Augmented Generation (RAG) (Soudani et al., 2024; Min et al., 2023). As a result, the trustworthiness of LLM-generated responses has become a critical concern, directly impacting user satisfaction (Hou et al., 2024; Mahaut et al., 2024).

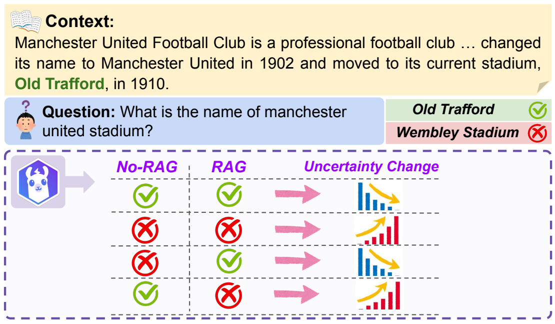

Figure 1: For a given query, uncertainty can be compared with and without RAG. For instance, when the LLM answers correctly on its own, and RAG provides supporting evidence, uncertainty should decrease. These intuitive principles form a framework for evaluating and interpreting uncertainty behavior in RAG.

Uncertainty Estimation (UE) is a widely studied approach for assessing the reliability of LLM outputs. A UE method assigns an uncertainty score to each (input, output) pair, reflecting its truthfulness. Ideally, a perfect UE method would assign lower uncertainty to correct samples and higher uncertainty to incorrect ones (Duan et al., 2024). While existing UE methods mainly focus on scenarios where the input is just a query, real-world applications like RAG involve non-parametric knowledge in more complex prompts (Huang et al., 2024). Research shows that non-parametric knowledge significantly influences LLM responses, often aligning them with the provided context (Cuconasu et al., 2024; Mallen et al., 2023). Despite this, it is unclear how current UE methods account for non-parametric knowledge.

In this paper, we investigate a critical question: (***RQ1***) *How do UE methods perform when the input prompt includes non-parametric knowledge, such as in RAG?* We study UE in the context of RAG with retrievers of varying effectiveness: (i) a deliberately weak synthetic retriever that returns irrelevant documents, (ii) an idealized retriever that consistently ranks the gold document at the top, and (iii) several widely used retrievers with varying performance levels. Our findings unveil that the performance of existing UE methods is inconsistent and mainly deteriorates when non-parametric knowledge is included in the input prompt (Figure 2). Most notably, improvements on the proposed UE methods in the literature do not add up when considering RAG setup (Table 1).

Against this background, it is clear that UE requires a methodological departure; existing methods are developed without paying attention to the specific properties that UE must satisfy in RAG setups. The question that arises here is ***(RQ2)*** *What properties can guarantee optimal performance of UE considering LLMs' both parametric and non-parametric knowledge?* We approach this question theoretically using axiomatic thinking, proven effective in various fields and tasks, including information retrieval (Fang and Zhai, 2005; Bondarenko et al., 2022), interpretability (Chen et al., 2024), and preference modeling (Rosset et al., 2023). In axiomatic thinking, a set of formal constraints is defined based on desired properties, which are then used as a guide to search for an optimal solution. In this work, we define an axiomatic framework for UE and establish five axioms considering the desired behavior of a UE method with and without external knowledge. Our axiomatic analysis reveals that current UE methods can satisfy only two of the axioms, violating the remaining three axioms in majority of cases.

The axiomatic framework helps explaining deficiencies of existing UE methods for RAG setup. The next question is ***(RQ3)*** *Can the axiomatic framework guide us in deriving an optimal UE method?* We use the constraints of the axiomatic framework to define a calibration function based on three components. We implement three instantiations of this function and apply it to different UE methods on a number of representative datasets. The results show that the derived functions are not only more stable than the existing UE methods but also improve overall performance with respect to AUROC. This highlights two key insights: first, satisfying the axioms leads to performance improvements, and second, existing UE methods can still be used for RAG by incorporating an axiomatically informed coefficient.

The main **contributions** of this paper include:
(1) Analyzing existing UE methods and showing their deficiencies in RAG setup.
(2) Proposing an axiomatic framework for UE with five formalized constraints and demonstrating deficiencies of existing methods in satisfying them.
(3) Introducing a calibration function guided by axioms and showing consistent improvements of the UE methods as a result of alignment with axioms.

2 Background

UE methods are typically divided into white-box approaches, which utilize token probabilities and entropy (Kadavath et al., 2022; Kuhn et al., 2023), and black-box approaches, which rely solely on final outputs (Lin et al., 2024; Band et al., 2024). This section reviews methods of both categories that are explored in this paper. For further details on related work, see Appendix A.

2.1 White-box Methods

Predictive Entropy (PE) for generative models quantifies uncertainty as the entropy of responses for an LLM input. The entropy is maximized when all outcomes are equally likely, indicating low informativeness (Kadavath et al., 2022; Kuhn et al., 2023). Given an LLM parametrized by θ and an input x, the LLM uncertainty is estimated by computing entropy using Monte-Carlo approximation:

$$PE(x, \theta) = -\frac{1}{B}\sum_{b=1}^{B} \ln P\left(r_b \mid x, \theta\right), \quad (1)$$

where r_b is a beam-sampled response and B is the number of samples. The probability of generating a response $r = \{r^1, r^2, ..., r^N\}$, comprising N tokens, given the input x is computed as the product of the conditional probabilities of each token, given its preceding tokens and the input x. For a model with parameters θ, the sequence probability is defined as:

$$P(r \mid x, \theta) = \prod_{n=1}^{N} P\left(r^n \mid r^{<n}, x; \theta\right), \quad (2)$$

where $r^{<n}$ denotes the tokens generated before r^n.

Semantic Entropy (SE) (Kuhn et al., 2023) extends PE by incorporating the semantic meaning of sampled responses. In this approach, generated samples are clustered into semantic clusters $c_i \in C$, and SE is defined as:

$$SE(x, \theta) = -\frac{1}{|C|}\sum_{i=1}^{|C|} \log \tilde{P}\left(c_i \mid x, \theta\right), \quad (3)$$

where c_i represents a semantic cluster, containing semantically similar responses. The cluster score $\tilde{P}(c_i|.)$ is computed as:

$$\tilde{P}(c_i \mid x, \theta) = \sum_{r \in c_i} P(r \mid x, \theta).$$

Length Normalization and Semantic Awareness are two important components in UE. It has been observed that the sequence probability in Equation (2) is biased against longer generations (Malinin and Gales, 2021). To address this, a length-normalized probability is introduced to generate equal weighting of tokens and reduce bias toward shorter sequences:

$$P_{\text{ln}}(r \mid x, \theta) = \prod_{n=1}^{N} P\left(r^n \mid r^{<n}, x; \theta\right)^{\frac{1}{N}}.$$

MARS (Bakman et al., 2024) and TokenSAR (Duan et al., 2024) further refined this approach by incorporating semantic importance. These approaches assign weights based on each token's contribution, resulting in the meaning-aware probability:

$$P_{\text{me}}(r \mid x, \theta) = \prod_{n=1}^{N} P\left(r^n \mid r^{<n}, x; \theta\right)^{w(r,x,N,n)}$$

where $w(r, x, N, n)$ is the importance weight for the n-th token. Both the length-normalized and meaning-aware probabilities can be used in the PE (1) and SE (3) equations.

2.2 Black-box Methods

We examine state-of-the-art semantic similarity-based methods (Lin et al., 2024), following these steps: (i) generate B sampled responses $\{r_1, \ldots, r_B\}$ for a given input x; (ii) compute pairwise similarity scores $a_{i,j} = a(r_i, r_j)$ between the responses; and (iii) derive uncertainty from these scores. Three approaches are proposed for computing uncertainty scores, described below.

Sum of Eigenvalues (EigV) (Lin et al., 2024). SE groups responses into semantic equivalence subsets and uses their count (*NumSet*) as an uncertainty metric; greater diversity implies higher uncertainty. To compute a more nuanced and continuous value for uncertainty than *NumSet*, Lin et al. (2024) define uncertainty as:

$$U_{\text{EigV}}(x) = \sum_{k=1}^{B} \max\left(0, 1 - \lambda_k\right), \quad (4)$$

where $\lambda_1, \ldots, \lambda_B$ are the eigenvalues of symmetric normalized Graph Laplacian (von Luxburg, 2007), defined as:

$$L := I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

Here, W represents a symmetric weighted adjacency matrix for a graph, where each node represents a response r_i for input x and weights are $w_{i,j} = (a_{i,j} + a_{j,i})/2$. The degree matrix D is defined as

$$D_{i,j} = \begin{cases} \sum_{j' \in [B]} w_{i,j'} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \quad (5)$$

Degree Matrix (Deg) relies on the degree matrix in Eq. (5) to computer uncertainty. Here, the intuition is that D reflects node connectivity, and nodes with higher degrees indicate confident regions in the LLM (Lin et al., 2024). Building on this, the uncertainty score is computed by:

$$U_{\text{Deg}}(x) = \text{trace}(BI - D)/B^2.$$

Eccentricity (ECC) is defined as the average distance of response embeddings from their centroid, which can serve as an uncertainty measure. Since access to the embeddings is not possible in black-box LLMs, the embeddings are driven from graph Laplacian. Let $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^B$ be the k smallest eigenvectors of L. For each response r_j, define the embedding as $\mathbf{v}_j = [u_{1,j}, \ldots, u_{k,j}]$ (Ng et al., 2001), and its centroid as $\mathbf{v}'_j = \mathbf{v}_j - \frac{1}{B} \sum_{j'=1}^{B} \mathbf{v}_{j'}$. Uncertainty is computed as:

$$U_{\text{ECC}}(x) = \left\| \left[\mathbf{v}_1'^{\top}, \ldots, \mathbf{v}_B'^{\top} \right] \right\|_2.$$

3 Axiomatic Framework

The assumption of an axiomatic framework for UE is that by satisfying a set of formal constraints, a UE method would likely have an optimal correlation with correctness for both RAG and non-RAG setups. To define the framework, we introduce five *axioms* based on a set of *functions* that form our search space for an optimal UE. These axioms, while necessary, do not represent an exhaustive set, as increasing the number of axioms can, in reality, introduce stringent, contradictory, or biased constraints. In the following, we introduce the functions and constraints of our axiomatic framework.

3.1 Functions

We define UE as the task of learning a function $\mathcal{U}$ that predicts a score s, quantifying the LLM's uncertainty for its output (Liu et al., 2024). Formally, let x be the input given to a generative LLM $\mathcal{M}_\theta$, parameterized by θ. The uncertainty estimator function is formulated as follows:

$$\mathcal{U} : \mathcal{M}_\theta(x), r \mapsto s$$

where the input consists of an LLM with the given input x and a generated response r. In a non-RAG setting, the input x is only the query q, while for the RAG setup, the input x consists of a query q and a context c, denoted as $\mathcal{M}_\theta(q, c) = r$. We define context c broadly, including an individual document or a set of documents.

Before defining the axioms, we introduce functions that formalize the relation between a context, a query, and an LLM-generated response. These functions, defined based on Natural Language Inference (NLI) (Pavlick and Callison-Burch, 2016; Williams et al., 2018), are as follows:

Entailment ($c \vDash (q, r)$): Given the context c, a human can infer that r is the correct response to q; i.e., the premise c entails the hypothesis (q, r) (asymmetric relation).

Contradiction ($c \perp (q, r)$): Given the context c, a human can infer that r is an incorrect response to q; i.e., the premise c contradicts the hypothesis (q, r) and vice versa (symmetric relation).

Independence ($c \# (q, r)$): Given the context c, a human cannot infer any information about the correctness of response r to query q; i.e., the premise c does not guarantee the truth or falsity of hypothesis (q, r) and vice versa (symmetric relation).

Equivalence ($r_1 \equiv r_2$): Two LLM responses, r_1 and r_2, convey the same meaning; i.e., the premise r_1 entails the hypothesis r_2 and vice versa (symmetric relation).

3.2 Axioms

The axioms are defined based on two key assumptions to ensure the validity of axioms and the three aforementioned functions:

Assumption 1. *The context c is trustworthy and contains factually correct information.*

Assumption 2. *The context c given to the LLM for the query q does not contain contradictory information about the query q.*

We now define five constraints that any reasonable UE method should satisfy, considering LLM's both parametric and non-parametric knowledge. Our working hypothesis is that UE is a proxy for the correctness of the model (Bakman et al., 2024). Two of these constraints are proven based on this hypothesis, and three of them are intuitively driven.

Theorem 1 (Positively Consistent)**.** *$\forall q, c$ if $\mathcal{M}_\theta(q) = r_1$, $\mathcal{M}_\theta(q,c) = r_2$, $r_1 \equiv r_2$, $c \vDash (q, r_2)$, then $\mathcal{U}(\mathcal{M}_\theta(q), r_1) > \mathcal{U}(\mathcal{M}_\theta(q, c), r_2)$.*

This constraint states that if applying RAG does not alter the LLM's response and the RAG context supports LLM's generated response r_2, the LLM's internal belief aligns with the context. In such a scenario, the uncertainty after applying RAG should be lower than before, as the retrieved context reinforces the LLM's prior knowledge. For instance, consider the example in Figure 1. Given the query, "*What is the name of Manchester United's stadium?*" if the LLM initially generates the correct response, "*Old Trafford*," and the input context mentions "*Old Trafford*" as the name of the stadium, then the uncertainty value after applying RAG should be lower than before.

Theorem 2 (Negatively Consistent)**.** *$\forall q, c$ if $\mathcal{M}_\theta(q) = r_1$, $\mathcal{M}_\theta(q, c) = r_2$, $r_1 \equiv r_2$, $c \perp (q, r_2)$, then $\mathcal{U}(\mathcal{M}_\theta(q), r_1) < \mathcal{U}(\mathcal{M}_\theta(q, c), r_2)$.*

This constraint states that if the LLM's response remains unchanged after applying RAG, but the retrieved context c contradicts the generated response r_2, then the LLM's internal belief does not align with the context. In such a case, the uncertainty after applying RAG should be higher than before, as the retrieved information challenges the LLM's internal belief. For example, in Figure 1, if LLM's response before and after RAG is "*Wembley Stadium*," and RAG context contradicts the LLM's response, then the uncertainty of the RAG response should increase. This means that although the LLM persists with its incorrect response, it does so with a lower confidence.

Theorem 3 (Positively Changed)**.** *$\forall q, c$ if $\mathcal{M}_\theta(q) = r_1$, $\mathcal{M}_\theta(q, c) = r_2$, $\neg(r_1 \equiv r_2)$, $c \perp (q, r_1)$, $c \vDash (q, r_2)$, then*

$$\mathcal{U}(\mathcal{M}_\theta(q), r_1) > \mathcal{U}(\mathcal{M}_\theta(q, c), r_2).$$

Theorem 3 directly follows from the statement in the following lemma:

Lemma 1. *If $\mathcal{M}_\theta(x_1) = r_1$, $\mathcal{M}_\theta(x_2) = r_2$, r_1 is $False$, r_2 is $True$, then*

$$\mathcal{U}(\mathcal{M}_\theta(x_1), r_1) > \mathcal{U}(\mathcal{M}_\theta(x_2), r_2).$$

Proof. Given Assumptions 1 and 2 and $c \perp (q, r_1)$, then response r_1 is $False$. Similarly, given that $c \vDash (q, r_2)$, then response r_2 is $True$. Given these events and Lemma 1, then $\mathcal{U}(\mathcal{M}_\theta(q), r_1) > \mathcal{U}(\mathcal{M}_\theta(q, c), r_2)$. □

This constraint states that if the LLM's response changes from r_1 to r_2 after applying RAG, and the RAG context c supports r_2 while contradicting r_1, then the estimated uncertainty for r_2 should be lower than one for r_1. For example, consider the case illustrated in Figure 1. If the LLM initially generates "*Wembley Stadium*" but then, after seeing a context containing the correct response, changes its output to "*Old Trafford*," the uncertainty of "*Old*

Trafford" with RAG should be lower than the uncertainty of "*Wembley Stadium*" without RAG.

Theorem 4 (Negatively Changed)**.** $\forall q, c$ *if* $\mathcal{M}_\theta(q) = r_1$, $\mathcal{M}_\theta(q, c) = r_2$, $\neg(r_1 \equiv r_2)$, $c \vDash (q, r_1)$, $c \perp (q, r_2)$, *then*

$$\mathcal{U}(\mathcal{M}_\theta(q), r_1) < \mathcal{U}(\mathcal{M}_\theta(q, c), r_2).$$

This theorem follows from the statement in the Lemma 1 with the following proof.

Proof. The proof is similar to that of Theorem 3. Given Assumptions 1 and 2 and $c \vDash (q, r_1)$, then response r_1 is correct. Similarly, response r_1 is incorrect because $c \perp (q, r_1)$. Based on Lemma 1 and these events, then $\mathcal{U}(\mathcal{M}_\theta(q), r_1) < \mathcal{U}(\mathcal{M}_\theta(q, c), r_2)$. □

This constraint states that if the LLM's response changes from r_1 to r_2 after applying RAG, where r_1 is correct, and r_2 is incorrect, then the estimated uncertainty of r_2 should be higher than the one for r_1. In the example of Figure 1, the LLM generates the correct response "*Old Trafford*" and changes its response to "*Wembley Stadium*" in the RAG setup, which is incorrect. In this scenario, the uncertainty of the RAG response should be higher than that of the original response without RAG.

Theorem 5 (Neutrally Consistent)**.** $\forall q, c$ *if* $\mathcal{M}_\theta(q) = r_1$, $\mathcal{M}_\theta(q, c) = r_2$, $r_1 \equiv r_2$, $c\#(q, r_1)$, *then* $\mathcal{U}(\mathcal{M}_\theta(q), r_1) \approx \mathcal{U}(\mathcal{M}_\theta(q, c), r_2)$.

This constraint states that if the LLM's response remains unchanged after applying RAG, and the retrieved context c is unrelated to the query and responses r_1 and r_2, then the context neither supports nor contradicts the LLM's belief. In this case, the estimated salary should remain similar. For example, consider the query "*Who wrote the book The Origin of Species?*". If, in the RAG setup, the LLM is provided with the context shown in Figure 1, which is unrelated to the query, then as long as the response remains unchanged, the uncertainty value should remain unaffected.

3.3 Instantiation

To empirically examine UE methods against these axioms, we need to define a specific instantiation of functions in our framework (cf. Sec. 3.1). We introduce two instantiations of these functions: *reference-based* and *reference-free*. The reference-based instantiation assumes the existence of a benchmark containing ground truth responses to queries. Such a benchmark is not available for reference-free instantiation.

Reference-based. In this setup, we rely on ground truth labels to check the condition of each axiom. We assume that for every q, the correct response $\hat{r}$ is available in our ground truth. The implementation of *Entailment* and *Contraction* functions then boils down to comparing the generated response r against the ground truth response $\hat{r}$. The comparison is performed using a matching function $\mathcal{M}(r_1, r_2)$, which assesses whether the two responses are equivalent. This function is also used to implement the *Equivalence* function. For datasets containing factual queries with short responses, $\mathcal{M}(.)$ is an Exact Match (EM) function, which returns *True* if and only if the two responses are identical on a token-by-token basis (Mallen et al., 2023). Using this setup, the following conditions can be inferred for our axioms:

Axiom 1. $\mathcal{M}(r_1, r_2) = True$, $\mathcal{M}(r_2, \hat{r}) = True$.
Axiom 2. $\mathcal{M}(r_1, r_2) = True$, $\mathcal{M}(r_2, \hat{r}) = False$.
Axiom 3. $\mathcal{M}(r_1, r_2) = False$, $\mathcal{M}(r_1, \hat{r}) = False$, $\mathcal{M}(r_2, \hat{r}) = True$.
Axiom 4. $\mathcal{M}(r_1, r_2) = False$, $\mathcal{M}(r_1, \hat{r}) = True$, $\mathcal{M}(r_2, \hat{r}) = False$.
Axiom 5. $\mathcal{M}(r_1, r_2) = True$, c is not relevant to q.

Reference-free. Since access to the correctness labels of LLM's responses limits the applicability of axioms to unseen queries, we propose a reference-free implementation of axioms. Specifically, we leverage an NLI classifier to assess the relationship between the generated response and the context. Following (Kuhn et al., 2023; Lin et al., 2024), we implement *Entailment* by merging entailment and neutral classes into a single class. The contradiction class of the NLI classifier is considered as *Contradiction*. Similar to the reference-based instantiation, function $\mathcal{M}(.)$ is used for *Equivalence*. Using these definitions, all axioms are implemented as formalized, except for Axiom 5, which is implemented similarly to the reference-based setup due to the limitations of existing NLI methods in predicting the neutral relation.

4 Derivation of a Calibration Function

In this section, we search over the function space of our axiomatic framework to drive calibration function that improves existing UE methods. To recap, our formal constraints are built around four functions that are examined for LLM responses without and with RAG; i.e., r_1 and r_2. These functions are of two types: (i) Equivalence that examines the relation between two LLM generated responses,

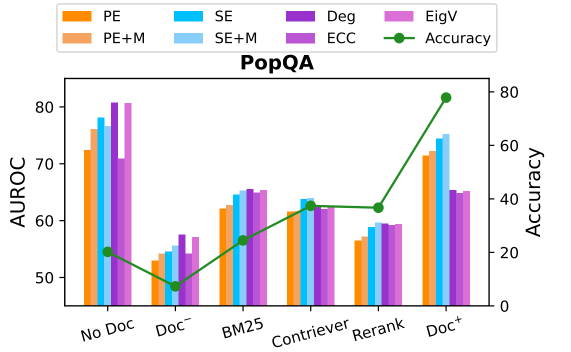

Figure 2: Comparison of AUROC between no-RAG and RAG settings for Llama2-chat.

represented as $\mathcal{E}(r_1, r_2)$, and (ii) other functions that examine entailment, contradiction, and independence relations between context, query, and an LLM generated response, represented as $\mathcal{R}(c, q, r)$. We define a calibration coefficient by searching the space of our axiomatic constraints using these two types of functions:

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

where k_1, k_2, k_3 are hyper parameters, and r_1, r_2 represent LLM generated responses without and with RAG, respectively. The calibrated UE function for RAG is then defined as:

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

The hyper parameters k_1–k_4 are set to satisfy the axioms using a validation set. This calibration enables increasing the uncertainty score of RAG for samples associated with axioms 2 and 4 while decreasing it for samples related to axioms 1 and 3.

4.1 Instantiation

We propose three instantiations of the calibration function, where three different models are used to implement $\mathcal{R}$.

CTI. The first model is based on the Context-sensitive Token Identification (CTI) task, which has been applied in self-citation and groundedness evaluation (Sarti et al., 2024; Qi et al., 2024). In this approach, each token in $r = \{r^1, r^2, \ldots, r^N\}$ is evaluated using a contrastive metric m (e.g., KL divergence, comparing the LLM's response distributions with and without the context. The resulting scores are $\{m_1, m_2, \ldots, m_N\}$, where $m_n = \mathrm{KL}(P(r^n \mid r^{<n}, (q, c); \theta) \parallel P(r^n \mid r^{<n}, q; \theta))$. These scores are converted into binary values via the selector function S_{CTI}. The overall relation score is then computed as:

$$\mathcal{R}(c, q, r) = \frac{1}{N} \sum_{n=1}^{N} S_{\text{CTI}}(m_n).$$

LLM	Unc.	No Doc	Doc⁻	BM25	Cont.	ReRa.	Doc⁺
Llama2-chat	PE	1.29	1.11 *	0.54 *	0.46 *	0.35 *	0.34 *
	SE	4.86	4.37 *	3.45 *	3.30 *	3.13 *	3.19 *
	PE+M	1.59	1.34 *	0.65 *	0.55 *	0.44 *	0.45 *
	SE+M	5.38	4.71 *	3.62 *	3.43 *	3.23 *	3.27 *
	Deg	0.52	0.32 *	0.12 *	0.09 *	0.06 *	0.05 *
	ECC	0.71	0.54 *	0.22 *	0.17 *	0.12 *	0.10 *
	EigV	4.25	2.28 *	1.42 *	1.31 *	1.18 *	1.17 *
Mistral-v0.3	PE	1.51	0.94 *	0.84 *	0.69 *	0.62 *	0.51 *
	SE	5.66	3.73 *	3.68 *	3.53 *	3.41 *	3.26 *
	PE+M	2.35	1.42 *	1.26 *	1.05 *	0.92 *	0.80 *
	SE+M	6.47	4.05 *	3.98 *	3.77 *	3.60 *	3.45 *
	Deg	0.48	0.05 *	0.07 *	0.06 *	0.05 *	0.03 *
	ECC	0.68	0.03 *	0.08 *	0.08 *	0.05 *	0.04 *
	EigV	4.18	1.08 *	1.16 *	1.17 *	1.11 *	1.08 *

(PopQA)

Table 1: Average uncertainty values for various settings. Lighter colors indicate lower uncertainty. Statistically significant differences are compared to *No Doc* are marked with $*$.

NLI. The second model employs an NLI-based approach that quantifies the relationship using entailment probability:

$$\mathcal{R}(c, q, r) = \mathcal{N}_{\models}(c, (q, r)).$$

MiniCheck. Finally, the third model employs MiniCheck (Tang et al., 2024), which performs sentence-level fact-checking using a fine-tuned model. It produces a score between 0 and 1 indicating how well the r is grounded in the c:

$$\mathcal{R}(c, q, r) = \text{MiniCheck}(c, (q, r)).$$

In all three instantiations, the equivalence function $\mathcal{E}(r_1, r_2)$ is an NLI classifier, wherein the entailment probability serves as a continuous measure of similarity between r_1 and r_2 (Kuhn et al., 2023); formally $\mathcal{E}(r_1, r_2) = \mathcal{N}_{\models}(r_1, r_2)$.

5 Experimental Setup

Datasets. We evaluate our approach on three open-book QA datasets, Natural Questions (NQ-open) (Lee et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). All datasets contain relevant passages for queries.

Experimental setup. Our experiments involve the reproduction of existing UE methods for the RAG setup. To ensure a fair comparison, we employ LLMs that are used in the original papers: Llama2-chat 7B and Mistral-7B. For uncertainty computation, 10 responses per query are generated with a temperature setting of $T = 1$; for correctness evaluation, the most likely response is considered.

Following (Kuhn et al., 2023), we use Deberta-large model fine-tuned on MNLI as NLI classifier. BM25, Contriever (Izacard et al., 2022), and BM25+Reranker are used as retrievers. Manually

UE	PopQA		
	BM25	Contriever	Doc$^+$
Axiom 1: Positively Consistent ↓			
PE	0.735 → 0.419 *	0.735 → 0.408 *	1.242 → 0.340 *
SE	3.781 → 3.205 *	3.791 → 3.158 *	4.682 → 3.113 *
PE+M	0.896 → 0.483 *	0.881 → 0.458 *	1.530 → 0.406 *
SE+M	4.102 → 3.286 *	4.091 → 3.248 *	5.146 → 3.173 *
EigV	1.951 → 1.166 *	2.025 → 1.143 *	4.074 → 1.078 *
ECC	0.417 → 0.110 *	0.426 → 0.094 *	0.710 → 0.055 *
Deg	0.220 → 0.048 *	0.230 → 0.043 *	0.496 → 0.022 *
Axiom 2: Negatively Consistent ↑			
PE	1.068 → 0.746	0.820 → 0.593	1.083 → 0.597
SE	4.163 → 3.548 *	4.104 → 3.381 *	4.388 → 4.107
PE+M	1.309 → 0.844	1.016 → 0.782	1.328 → 0.684
SE+M	4.599 → 3.700 *	4.481 → 3.610 *	4.764 → 4.221
EigV	2.453 → 1.338 *	2.088 → 1.274 *	2.758 → 1.910
ECC	0.541 → 0.197 *	0.477 → 0.152 *	0.503 → 0.443
Deg	0.286 → 0.101 *	0.228 → 0.073 *	0.343 → 0.254
Axiom 3: Positively Changed ↓			
PE	1.375 → 0.347 *	1.416 → 0.298 *	1.342 → 0.268 *
SE	4.889 → 3.015 *	5.091 → 3.013 *	4.884 → 3.051 *
PE+M	1.708 → 0.398 *	1.735 → 0.374 *	1.604 → 0.340 *
SE+M	5.514 → 3.072 *	5.681 → 3.082 *	5.379 → 3.099 *
EigV	4.131 → 1.139 *	4.733 → 1.114 *	4.449 → 1.102 *
ECC	0.790 → 0.085 *	0.823 → 0.081 *	0.780 → 0.072 *
Deg	0.547 → 0.044 *	0.588 → 0.035 *	0.544 → 0.032 *
Axiom 4: Negatively Changed ↑			
PE	0.933 → 0.636	1.006 → 0.558	1.252 → 0.463
SE	4.152 → 3.552 *	4.192 → 3.409 *	4.830 → 3.690 *
PE+M	1.164 → 0.714 *	1.298 → 0.748 *	1.689 → 0.747
SE+M	4.553 → 3.690 *	4.653 → 3.608 *	5.381 → 4.007 *
EigV	2.593 → 1.449 *	2.557 → 1.412 *	3.567 → 1.449 *
ECC	0.540 → 0.262 *	0.548 → 0.220 *	0.707 → 0.237 *
Deg	0.320 → 0.128 *	0.320 → 0.115 *	0.463 → 0.140 *

Table 2: Comparison of changes in average uncertainty values for Axioms 1–4 before (left) and after (right) applying RAG with Llama2-chat. Colors green and deep red indicate significant changes aligning or conflicting with axioms, respectively. Color shallow red represents non-significant changes conflicting with axioms. Significance is marked by $*$.

chosen relevant and irrelevant documents are denoted with Doc$^+$ and Doc$^-$, respectively.

Metrics. We report the Exact Match for correctness and AUROC (Bakman et al., 2024) to evaluate the correlation between uncertainty and correctness. Statistically significant differences are reported with p-value < 0.01 using Wilcoxon test. For further experimental details, refer to Appendix B.

6 Results

6.1 Uncertainty Changes with RAG

(RQ1) examines how the performance of UE methods and their associated uncertainty values vary with and without context in the input prompt. Figures 2 and 4 present accuracy and AUROC for different RAG settings. We observe inconsistent behavior of UE methods with and without RAG

Unc.	NQ-open	TriviaQA	PopQA
PE	2.072 → 2.248 *	0.872 → 1.155 *	0.897 → 0.909 *
SE	5.253 → 5.471 *	3.863 → 4.158 *	3.897 → 4.319 *
PE+M	4.791 → 4.805	1.415 → 1.699 *	1.031 → 1.130 *
SE+M	7.993 → 7.933	4.540 → 4.817 *	4.297 → 4.591
EigV	2.211 → 2.446 *	1.757 → 1.870 *	2.270 → 2.218
ECC	0.512 → 0.625 *	0.382 → 0.448 *	0.490 → 0.507
Deg	0.265 → 0.333 *	0.171 → 0.211 *	0.256 → 0.309

Table 3: Comparison of changes in average uncertainty values for Axiom 5 before (left) and after(right) applying RAG with Llama2-chat. Color coding and significance markers follow those in Table 2.

across different datasets, often displaying drop AUROC for RAG cases, except for *Doc*$^+$. While AUROC should be independent accuracy, the results suggest a correlation between the performance of the RAG method and AUROC; esp. when considering irrelevant and relevant documents. This indicates a bias of current UE methods towards RAG generations.

To assess this bias further, we report on average uncertainty values of these methods in Tables 1 and 5. The results reveal that incorporating any context results in lower uncertainty values. Even the inclusion of irrelevant contexts, which do not enhance accuracy, leads to a significant reduction in uncertainty scores. This suggests that current UE methods produce lower uncertainty values in the RAG setup without adequately accounting for the relevance of the context.

6.2 Axiomatic Evaluation

The second research question (**RQ2**) investigates properties (i.e., axioms) of UE methods that guarantee optimal performance, and assesses how these axioms are satisfied by current UE methods. Tables 2 and 6 present the change in the average uncertainty value Llama2-chat, from no-RAG to RAG, for Axioms 1–4 using the *Reference-based* implementation. The results indicate that Axioms 2 and 4 are largely unmet. Furthermore, MARS, although being a state-of-the-art white-box method, does not demonstrate improved axiom compliance. Similar trends are observed with Mistral and other datasets (see Table 7), underscoring the generality of these findings. Additionally, the *Reference-free* implementation of axioms (Table 9) strongly correlate with the *Reference-based* findings, confirming that UE methods completely fail to satisfy Axioms 2 and 4. This further shows the reliability of reference-free implementation for axiomatic evaluation of UE methods.

To evaluate Axiom 5, we add irrelevant context (*Doc*$^-$) for each query. Table 3 shows that only

UE	NQ-open					TriviaQA					PopQA				
	A1 (%)	A2 (%)	A3 (%)	A4 (%)	AUROC	A1 (%)	A2 (%)	A3 (%)	A4 (%)	AUROC	A1 (%)	A2 (%)	A3 (%)	A4 (%)	AUROC
PE	60.19	47.85	77.35	51.16	64.87	45.53	43.78	70.26	66.88	68.18	66.19	42.46	87.57	38.17	61.59
+CTI	61.49	44.17	76.43	53.88	65.38	46.00	43.78	69.23	68.47	69.29	69.63	39.73	87.95	38.17	63.04
+NLI	66.02	47.24	77.57	55.43	67.21	48.45	45.77	71.28	68.47	69.40	68.77	41.10	88.15	41.22	63.09
+MCH	76.05	37.42	83.75	51.93	69.85	51.36	49.25	74.10	69.75	71.92	69.34	39.73	89.48	39.70	64.31
SE	77.35	33.75	91.53	36.05	67.49	50.14	35.82	84.62	54.78	73.44	71.92	31.51	94.07	29.01	63.79
+CTI	77.02	25.76	89.47	40.31	67.09	56.54	39.30	79.74	56.69	72.65	78.51	26.03	91.21	26.72	62.58
+NLI	79.61	40.49	86.72	50.00	69.77	68.96	46.77	80.77	62.74	74.72	71.63	38.36	92.73	41.22	67.86
+MCH	88.02	32.52	91.53	46.90	**75.88**	73.28	49.75	82.82	67.20	**79.79**	77.94	31.51	94.07	41.22	**72.49**
EigV	65.37	12.88	88.56	24.42	63.94	37.16	24.38	86.15	39.17	70.00	55.30	6.85	92.93	20.61	62.42
+CTI	77.35	20.25	90.16	34.50	66.82	66.89	30.85	86.15	48.41	72.54	80.80	19.18	93.50	29.77	61.51
+NLI	80.91	27.61	91.76	35.27	69.44	60.21	41.79	87.69	51.59	73.58	73.35	35.62	95.60	38.17	67.60
+MCH	88.67	23.93	93.82	34.88	73.60	74.88	40.30	90.00	55.41	78.34	83.09	24.66	96.75	32.82	72.18
ECC	61.49	9.82	83.06	18.99	63.57	34.24	14.43	73.59	30.89	68.23	52.44	6.84	87.38	18.32	62.06
+CTI	75.73	23.31	87.18	37.98	67.37	65.47	31.84	77.69	53.19	69.92	78.80	23.29	90.82	34.35	61.75
+NLI	78.64	32.52	87.18	42.64	68.96	58.04	42.79	77.44	59.87	71.31	71.35	32.88	92.16	42.75	66.44
+MCH	86.08	26.99	89.93	39.54	71.81	72.44	41.29	82.31	58.92	74.94	79.37	21.92	94.84	35.87	71.39

Table 4: Percentage of samples passing the axioms before and after calibration for Contriver with Llama2-chat. The results show that as the number of samples passing the axioms increases, the AUROC also improves.

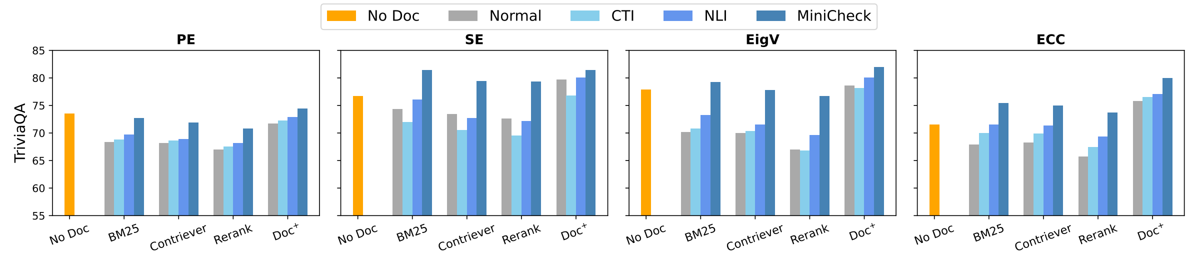

Figure 3: Comparison of AUROC between the no-RAG and calibrated RAG settings for Llama2-chat for TriviaQA. AUROC improves significantly, either surpassing the no-RAG setting or reducing the gap between them.

PE+M and SE+M partially satisfy Axiom 5 for Llama2. For Mistral (Table 8), all methods pass Axiom 5 for PopQA but not for the other datasets. These findings suggest that none of the existing UE methods fully satisfy Axiom 2, 4, and 5.

6.3 Axiomatic Calibration

Our third research question (**RQ3**) examines how our axiomatic framework can lead to designing an optimal UE method. Tables 4 and 10 present AUROC and percentage of samples passing the axioms 1–4 before and after applying our calibration method. Axiom 5 is not assessed, as retrievers tend to retrieve relevant documents. We perform the experiments on four representative (and not cherry-picked) UE methods, as the results generalize to other methods as well. The calibration function is implemented using the three models described in Section 4.1, and Contriever is employed for RAG.

The results show that calibration MiniCheck outperforms all implementations, improving percentages of all axioms for EigV and ECC and for most axioms in open-box methods. Most importantly, the results show as the percentage of samples satisfying the axioms increases, the AUROC improves, showing the empirical validity of our axioms in improving UE methods. Moreover, Figures 3 and 5 show that after calibration, the RAG AUROC becomes comparable to or even better than the *No Doc* baseline, suggesting that our calibration method successfully compensates for the inefficiencies of existing UE methods in RAG.

7 Discussion and Conclusions

In this paper, we examined existing uncertainty estimation (UE) for the RAG setup and showed they systematically generated low uncertainty values in the RAG setup without considering the relevance of the given context to the query. We further proposed an axiomatic evaluation framework for UE in the RAG setup and defined five formal constraints that a UE method should satisfy when processing both parametric and non-parametric knowledge. These axioms were empirically validated across multiple representative datasets, UE methods, and LLMs. Our results showed that none of the existing UE methods pass all the axiom, pinpointing the problem in these methods. We further derived a calibration function for adjusting UE methods in the RAG setup and improvements in both axiomatic evaluation and correlation with correctness. Future work includes developing a UE method designed to naturally conform to the established axioms. Another direction is assessing these axioms in long-form responses and uncertainty-based applications, such as Active RAG.

Limitations

Axiomatic Uncertainty Estimator. In this study, we evaluate existing uncertainty estimation (UE) methods within the RAG setup and delineate the optimal behaviors that these methods should exhibit. Although we introduce a calibration function in Section 4, it may be more effective to develop an axiomatic UE model that inherently adheres to the prescribed axioms. Future research should leverage these principles in the construction of UE methods.

Comprehensiveness of the Axioms. As discussed in Section 3, while our current axioms address most cases, additional axioms may be needed to cover all sample types. For example, consider when an LLM produces a different output after incorporating a context, and both the initial and augmented responses contradict the context. In this scenario, our framework does not specify a change in uncertainty, though supplementary axioms might address this gap. Future research should develop axioms for such cases.

Scalability and Applications. We investigated the impact of incorporating context into the input prompt on uncertainty measures. However, we did not explore other input modalities, such as multi-modal RAG, or alternative response formats, such as long-form responses, each of which presents unique challenges. Furthermore, applications of uncertainty estimation, such as Adaptive RAG (Jiang et al., 2023; Cheng et al., 2024; Tao et al., 2024), hallucination detection (Geng et al., 2024), reasoning monitoring (Yin et al., 2024), and LLM-as-Judgment (Lee et al., 2024), fall outside the scope of this study. Future research should extend these findings to encompass diverse input types, response formats, and UE applications.

References

Enrique Amigó, Hui Fang, Stefano Mizzaro, and Chengxiang Zhai. 2020. Axiomatic thinking for information retrieval: introduction to special issue. *Inf. Retr. J.*, 23(3):187–190.

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*, pages 7752–7767.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. 2024. Linguistic calibration of long-form generations. In *Forty-first International Conference on Machine Learning, ICML*.

Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic retrieval experimentation with ir_axioms. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3131–3140.

Catherine Chen, Jack Merullo, and Carsten Eickhoff. 2024. Axiomatic causal interventions for reverse engineering relevance computation in neural retrieval models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pages 1401–1410.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 17153–17166.

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*, pages 719–729.

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), ACL*, pages 5050–5063.

Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56.

Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 480–487. ACM.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL*, pages 6577–6595.

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*.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 284–312.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050.

Zhengbao Jiang, Frank F. 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*, pages 7969–7992.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. abs/2207.05221.

Parisa Kordjamshidi, Qiang Ning, James Pustejovsky, and Marie-Francine Moens. 2024. Spatial and temporal language understanding: Representation, reasoning, and grounding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 39–46.

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*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2024. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pages 6086–6096.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *CoRR*.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. 2024. Factual confidence of llms: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 4554–4570.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR*.

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), ACL*, pages 9802–9822.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 12076–12100.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 [Neural InformationcProcessing Systems: Natural and Synthetic, NIPS]*, pages 849–856.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6037–6053.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Corby Rosset, Guoqing Zheng, Victor Dibia, Ahmed Awadallah, and Paul N. Bennett. 2023. Axiomatic preference modeling for longform question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 11445–11475.

Gabriele Sarti, Grzegorz Chrupala, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations, ICLR*.

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*, pages 12–22.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 8818–8847. Association for Computational Linguistics.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics, ACL*, pages 5984–5996.

Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards axiomatic explanations for neural ranking models. In *ICTIR '21: The 2021 ACM SIGIR*, pages 13–22.

Ulrike von Luxburg. 2007. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 1112–1122.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations ICLR*.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. abs/2406.11278.

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), ACL*, pages 2401–2416.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2024a. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics, ACL*, pages 15015–15040.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024b. Knowing what llms DO NOT know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL*, pages 7051–7063.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024c. Enhancing contextual understanding in large language models through contrastive decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL*, pages 4225–4237.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 3623–3643.

Appendix

A Related Work

Uncertainty Estimation (UE) seeks to quantify the confidence of LLMs in their predictions (Hou et al., 2024; Zhao et al., 2024b). UE methods are commonly divided into two groups: black-box and white-box approaches. Black-box methods rely solely on the LLM's outputs without accessing internal layers or generation logits. In addition to the semantic similarity-based methods discussed in Section 2, other black-box techniques exist. For example, verbalization methods prompt the model to explicitly report its confidence (e.g., "How confident are you that the answer is correct?"). Xiong et al. (2024) highlight that two key factors influence the quality of verbalized confidence: (i) the prompting strategy, which includes techniques such as vanilla, Chain-of-Thought (CoT), self-probing, multi-step, and Top-K prompting, and (ii) the sampling strategy, employing methods like self-random sampling, prompting-based elicitation, and misleading prompts to generate multiple responses. Additionally, Epi-M (Zhou et al., 2024) incorporates epistemic markers into the input prompt to facilitate well-calibrated confidence scores.

White-box approaches, by contrast, leverage access to next-token prediction probabilities for uncertainty calculation. Beyond the methods covered in Section 2, several techniques have been proposed. For instance, P(True) (Kadavath et al., 2022) measures the probability that a model assigns to the correctness of a given response by appending a sentence such as `Is the possible answer: (A) True (B) False. The possible answer is:` so that the probability of generating "True" or "False" serves as the measure. Similarly, P(IK) (Kadavath et al., 2022) estimates the likelihood that the model "knows" the correct answer, that is, the probability of generating the correct response when sampling at unit temperature. Furthermore, LARS (Yaldiz et al., 2024) introduces a learning-based approach by training a scoring model on token probabilities to enhance uncertainty prediction.

Axiomatic Evaluation. Axiomatic thinking refers to a problem-solving approach guided by a set of axioms closely aligned with conventional scientific methodologies (Amigó et al., 2020). More generally, this approach seeks solutions that satisfy all predefined axioms, that is, the desirable properties a solution should possess.

Axiomatic thinking has been successfully applied to the study of Information Retrieval (IR), thereby contributing both to the theoretical understanding and the practical enhancement of existing retrieval models. The objective of Axiomatic IR is to establish formal constraints, or axioms, that delineate the essential properties an effective ranking model must satisfy (Völske et al., 2021). In this context, Fang et al. (2004) formally defined six fundamental constraints derived from empirical observations of common characteristics in traditional retrieval functions. These constraints correspond to intuitive retrieval heuristics, such as term frequency weighting, term discrimination weighting, and document length normalization. Building on this foundation, Fang and Zhai (2005) proposed an axiomatic framework for the development of retrieval models. Their framework comprises an inductive scheme for function definitions, which provides a common basis for the analytical comparison of different retrieval functions, as well as a set of formalized retrieval constraints adapted from (Fang et al., 2004). These axioms have been further examined in subsequent studies. For example, Chen et al. (2024) employed causal interventions to identify specific attention heads that encode a robust term frequency signal, thereby aligning with one of the original axioms.

Beyond IR, axiomatic approaches have been extended to other domains. For instance, Rosset et al. (2023) defined axioms representing the qualities that humans value in long-form answers, including usefulness, relevance, groundedness, truthfulness, and thoroughness. They generated training data corresponding to these principles and subsequently used it to train a preference model.

B Experimental Setup

Datasets. We conduct our experiments on three open-book Question Answering (QA) datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023). The NQ dataset comprises a large-scale collection of real-world queries derived from Google search data. Each entry includes a user query and the corresponding Wikipedia page that contains the answer. The NQ-open dataset (Lee et al., 2019), a subset of NQ, differs by removing the restriction of linking answers to specific Wikipedia passages, thereby emulating a more gen-

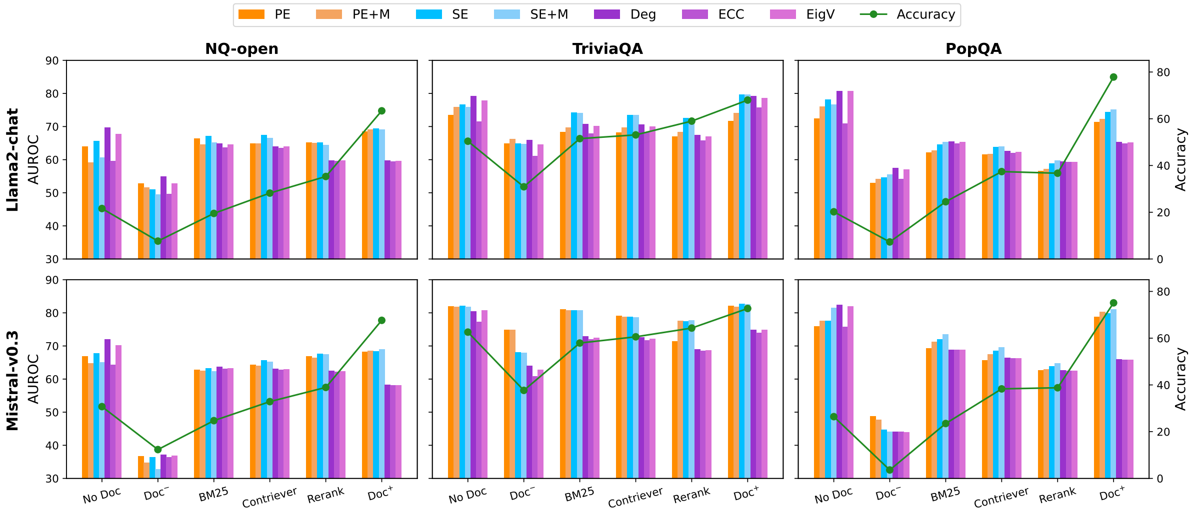

Figure 4: Comparison of AUROC between no-RAG and RAG settings.

LM	Unc.	NQ-open						TriviaQA						PopQA					
		No Doc	Doc⁻	BM25	Cont.	ReRa.	Doc⁺	No Doc	Doc⁻	BM25	Cont.	ReRa.	Doc⁺	No Doc	Doc⁻	BM25	Cont.	ReRa.	Doc⁺
Llama2-chat	PE	1.98	1.92	1.53 *	1.41 *	1.31 *	1.19 *	1.14	1.42 *	1.05 *	1.03	0.90 *	0.96 *	1.29	1.11 *	0.54 *	0.46 *	0.35 *	0.34 *
	SE	5.40	5.09 *	4.29 *	4.20 *	3.99 *	3.88 *	4.39	4.48 *	3.89 *	3.85 *	3.66 *	3.73 *	4.86	4.37 *	3.45 *	3.30 *	3.13 *	3.19 *
	PEM	3.90	3.89	3.33 *	3.26 *	3.12 *	2.97 *	1.74	2.06 *	1.64	1.64	1.46 *	1.51 *	1.59	1.34 *	0.65 *	0.55 *	0.44 *	0.45 *
	SEM	7.41	6.93 *	5.97 *	5.88 *	5.62 *	5.49 *	5.16	5.21	4.51 *	4.50 *	4.22 *	4.30 *	5.38	4.71 *	3.62 *	3.43 *	3.23 *	3.27 *
	Deg	0.52	0.36 *	0.16 *	0.13 *	0.09 *	0.07 *	0.31	0.29 *	0.17 *	0.15 *	0.11 *	0.16 *	0.52	0.32 *	0.12 *	0.09 *	0.06 *	0.05 *
	ECC	0.64	0.60 *	0.29 *	0.23 *	0.17 *	0.14 *	0.56	0.53 *	0.33 *	0.29 *	0.23 *	0.31 *	0.71	0.54 *	0.22 *	0.17 *	0.12 *	0.10 *
	EigV	3.06	2.48 *	1.57 *	1.42 *	1.28 *	1.21 *	2.52	2.21 *	1.65 *	1.57 *	1.41 *	1.68 *	4.25	2.28 *	1.42 *	1.31 *	1.18 *	1.17 *
Mistral-v0.3	PE	1.98	1.28 *	1.40 *	1.46 *	1.39 *	1.32 *	0.96	1.08 *	0.83 *	0.81 *	0.72 *	0.74 *	1.51	0.94 *	0.84 *	0.69 *	0.62 *	0.51 *
	SE	5.61	4.37 *	4.32 *	4.33 *	4.19 *	4.05 *	4.29	4.27 *	3.76 *	3.74 *	3.57 *	3.67 *	5.66	3.73 *	3.68 *	3.53 *	3.41 *	3.26 *
	PEM	4.25	2.51 *	3.29 *	3.61 *	3.48 *	3.36 *	1.73	1.88 *	1.51 *	1.54 *	1.36 *	1.41 *	2.35	1.42 *	1.26 *	1.05 *	0.92 *	0.80 *
	SEM	7.65	5.42 *	5.94 *	6.19 *	6.01 *	5.85 *	4.99	4.98 *	4.35 *	4.37 *	4.12 *	4.27 *	6.47	4.05 *	3.98 *	3.77 *	3.60 *	3.45 *
	Deg	0.37	0.16 *	0.13 *	0.10 *	0.07 *	0.05 *	0.20	0.18 *	0.10 *	0.10 *	0.07 *	0.19 *	0.48	0.05 *	0.07 *	0.06 *	0.05 *	0.03 *
	ECC	0.54	0.20 *	0.18 *	0.15 *	0.11 *	0.08 *	0.37	0.32 *	0.17 *	0.19 *	0.13 *	0.17 *	0.68	0.03 *	0.08 *	0.08 *	0.05 *	0.04 *
	EigV	2.83	1.49 *	1.40 *	1.32 *	1.23 *	1.37 *	2.04	1.65 *	1.36 *	1.39 *	1.25 *	1.41 *	4.18	1.08 *	1.16 *	1.17 *	1.11 *	1.08 *

Table 5: Average uncertainty values for various settings. Lighter colors indicate lower uncertainty. Statistically significant differences are compared to *No Doc* are marked with $*$.

eral real-world scenario. We obtain the gold documents for each query from the corpus and dataset annotated by (Cuconasu et al., 2024) [1], in which the gold documents are integrated with the original corpus. For evaluation, we use the test set containing 2,889 queries. TriviaQA consists of trivia questions sourced from the web (Jeong et al., 2024). To ensure a dataset size comparable to NQ-open, we randomly sample 3,000 queries from its development set. PopQA is an open-domain QA dataset designed to evaluate factual knowledge, particularly regarding long-tail entities. Constructed from 16 diverse relationship types in Wikidata, PopQA is originally a closed-book dataset comprising 14,000 QA pairs without gold document annotations. Consequently, following (Soudani et al., 2024), we consider the summary section of the corresponding Wikipedia page as the gold document. Since PopQA is entirely based on Wikipedia, we employ the same corpus for retrieval. To maintain consistency with the other datasets, we randomly select 3,000 samples from the test set. Additionally, we create a validation set for each dataset, comprising 10% of the test set (approximately 300 samples), which is used to compute calibration coefficients as described in Section 4. For NQ-open and TriviaQA, the validation set is sampled from the training set, whereas for PopQA, it is derived from the test set.

Language Models. In accordance with established baselines, we select two generative LLMs: Llama2-chat-7B and Mistral-7B. For inputs that are not augmented with retrieved documents, we employ the following template: `"Answer the question. Question: <question> Answer:"` For inputs augmented with retrieved documents, we utilize this template: `"You are given a question, and you MUST respond with an answer (max 10`

[1] Dataset: florin-hf/nq_open_gold

UE	NQ-open			TriviaQA			PopQA		
	BM25	Contriever	Doc$^+$	BM25	Contriever	Doc$^+$	BM25	Contriever	Doc$^+$
Axiom 1: Positively Consistent ↓									
PE	1.445 → 1.194 *	1.535 → 1.216 *	1.549 → 1.159 *	0.700 → 0.753 *	0.718 → 0.743 *	0.731 → 0.724 *	0.735 → 0.419 *	0.735 → 0.408 *	1.242 → 0.340 *
SE	4.656 → 3.933 *	4.756 → 3.907 *	4.800 → 3.823 *	3.644 → 3.412	3.664 → 3.424 *	3.738 → 3.388 *	3.781 → 3.205 *	3.791 → 3.158 *	4.682 → 3.113 *
PE+M	3.389 → 3.124	3.412 → 3.052 *	3.437 → 3.069 *	1.051 → 1.110 *	1.131 → 1.178 *	1.141 → 1.120	0.896 → 0.483 *	0.881 → 0.458 *	1.530 → 0.406 *
SE+M	6.640 → 5.778 *	6.705 → 5.632 *	6.740 → 5.667 *	4.142 → 3.832 *	4.212 → 3.898 *	4.293 → 3.824 *	4.102 → 3.286 *	4.091 → 3.248 *	5.146 → 3.173 *
EigV	2.030 → 1.270 *	2.129 → 1.189 *	2.166 → 1.112 *	1.622 → 1.318 *	1.617 → 1.234 *	1.679 → 1.254 *	1.951 → 1.166 *	2.025 → 1.143 *	4.074 → 1.078 *
ECC	0.479 → 0.149 *	0.538 → 0.120 *	0.557 → 0.071 *	0.346 → 0.228 *	0.338 → 0.169 *	0.367 → 0.180 *	0.417 → 0.110 *	0.426 → 0.094 *	0.710 → 0.055 *
Deg	0.227 → 0.084 *	0.262 → 0.061 *	0.270 → 0.035 *	0.144 → 0.087 *	0.142 → 0.066 *	0.155 → 0.067 *	0.220 → 0.048 *	0.230 → 0.043 *	0.496 → 0.022 *
Axiom 2: Negatively Consistent ↑									
PE	2.317 → 2.261	2.230 → 2.153	2.232 → 2.194	1.543 → 1.478	1.534 → 1.438	1.495 → 1.528	1.068 → 0.746	0.820 → 0.593	1.083 → 0.597
SE	5.626 → 4.989 *	5.515 → 4.848 *	5.572 → 4.841 *	4.715 → 4.460	4.672 → 4.291 *	4.897 → 4.638 *	4.163 → 3.548 *	4.104 → 3.381 *	4.388 → 4.107
PE+M	5.284 → 4.891 *	5.052 → 4.904	5.665 → 5.652	2.716 → 2.633	2.381 → 2.249	2.594 → 2.597	1.309 → 0.844	1.016 → 0.782	1.328 → 0.684
SE+M	8.566 → 7.579 *	8.377 → 7.471 *	8.914 → 7.962 *	5.978 → 5.521 *	5.737 → 5.170 *	6.109 → 5.733 *	4.599 → 3.700 *	4.481 → 3.610 *	4.764 → 4.221
EigV	2.410 → 1.694 *	2.454 → 1.375 *	2.340 → 1.216 *	2.147 → 1.802 *	2.271 → 1.700 *	2.654 → 2.508	2.453 → 1.338 *	2.088 → 1.274 *	2.758 → 1.910
ECC	0.564 → 0.302 *	0.600 → 0.240 *	0.542 → 0.166 *	0.554 → 0.382 *	0.561 → 0.331 *	0.617 → 0.600	0.541 → 0.197 *	0.477 → 0.152 *	0.503 → 0.443
Deg	0.304 → 0.172 *	0.314 → 0.113 *	0.299 → 0.069 *	0.274 → 0.194 *	0.294 → 0.186 *	0.353 → 0.325	0.286 → 0.101 *	0.228 → 0.073 *	0.343 → 0.254
Axiom 3: Positively Changed ↓									
PE	2.113 → 0.909 *	1.989 → 0.939 *	2.006 → 0.847 *	1.481 → 0.665 *	1.413 → 0.702 *	1.403 → 0.653 *	1.375 → 0.347 *	1.416 → 0.298 *	1.342 → 0.268 *
SE	5.606 → 3.589 *	5.459 → 3.589 *	5.500 → 3.544 *	4.970 → 3.347 *	4.966 → 3.469 *	4.972 → 3.287 *	4.889 → 3.015 *	5.091 → 3.013 *	4.884 → 3.051 *
PE+M	3.479 → 2.056 *	3.420 → 1.991 *	3.416 → 2.012 *	2.001 → 0.917 *	2.026 → 1.020 *	1.930 → 0.938 *	1.708 → 0.398 *	1.735 → 0.374 *	1.604 → 0.340 *
SE+M	7.268 → 4.703 *	7.069 → 4.616 *	7.101 → 4.637 *	5.790 → 3.648 *	5.804 → 3.825 *	5.760 → 3.579 *	5.514 → 3.072 *	5.681 → 3.082 *	5.379 → 3.099 *
EigV	3.692 → 1.220 *	3.561 → 1.182 *	3.551 → 1.159 *	3.588 → 1.245 *	3.625 → 1.346 *	3.650 → 1.277 *	4.131 → 1.139 *	4.733 → 1.114 *	4.449 → 1.102 *
ECC	0.756 → 0.144 *	0.701 → 0.111 *	0.714 → 0.115 *	0.801 → 0.163 *	0.807 → 0.218 *	0.810 → 0.179 *	0.790 → 0.085 *	0.823 → 0.081 *	0.780 → 0.072 *
Deg	0.507 → 0.065 *	0.484 → 0.057 *	0.488 → 0.051 *	0.497 → 0.076 *	0.502 → 0.093 *	0.504 → 0.079 *	0.547 → 0.044 *	0.588 → 0.035 *	0.544 → 0.032 *
Axiom 4: Negatively Changed ↑									
PE	1.609 → 1.695	1.621 → 1.635	1.598 → 1.688	0.945 → 1.325 *	0.889 → 1.364 *	1.034 → 1.396 *	0.933 → 0.636	1.006 → 0.558	1.252 → 0.463
SE	4.899 → 4.457 *	4.899 → 4.437 *	4.915 → 4.497	4.160 → 4.312	4.157 → 4.273	4.297 → 4.339	4.152 → 3.552 *	4.192 → 3.409 *	4.830 → 3.690 *
PE+M	3.446 → 3.653	3.522 → 3.692	3.465 → 4.158	1.566 → 2.123 *	1.306 → 1.946 *	1.486 → 2.178 *	1.164 → 0.714 *	1.298 → 0.748 *	1.689 → 0.747
SE+M	6.764 → 6.286 *	6.803 → 6.377 *	6.643 → 6.442	4.953 → 5.121	4.769 → 4.933	4.983 → 5.088	4.553 → 3.690 *	4.653 → 3.608 *	5.381 → 4.007 *
EigV	2.262 → 1.582 *	2.244 → 1.503 *	2.233 → 1.367 *	2.089 → 1.908	2.141 → 1.908	2.399 → 2.131	2.593 → 1.449 *	2.557 → 1.412 *	3.567 → 1.449 *
ECC	0.594 → 0.332 *	0.565 → 0.295 *	0.490 → 0.270 *	0.501 → 0.453	0.542 → 0.456	0.614 → 0.555	0.540 → 0.262 *	0.548 → 0.220 *	0.707 → 0.237 *
Deg	0.301 → 0.163 *	0.294 → 0.148 *	0.308 → 0.123 *	0.239 → 0.237	0.253 → 0.251	0.313 → 0.299	0.320 → 0.128 *	0.320 → 0.115 *	0.463 → 0.140 *

Table 6: Comparison of changes in average uncertainty values for Axioms 1–4 before (left) and after (right) applying RAG with Llama2-chat. Axioms are implemented using the *Reference-based* method. Colors green and deep red indicate significant changes aligning or conflicting with axioms, respectively. Color shallow red represents non-significant changes conflicting with axioms. Significance is marked by $*$.

`tokens) using either the provided document or your memorized knowledge. Document: <context> Question:<question> Answer:`". Although more sophisticated prompts were examined in preliminary experiments, the marginal improvement they offered relative to the simple template did not justify their use, particularly given the increased risk of model overfitting. Furthermore, following MARS (Bakman et al., 2024), we utilize the Huggingface library's "generate" function for model output generation. We designate the token "." as the "eos_token_id" to prevent the model from generating overly lengthy paragraphs in response to closed-book questions. We also set "num_beams" to 1, corresponding to greedy decoding.

Retrieval Models. We employ a suite of retrieval models to acquire relevant contexts for the RAG approach. The models utilized include BM25 (Robertson and Zaragoza, 2009), Contriever (Izacard et al., 2022), and a two-stage reranking system. In the two-stage configuration, BM25 is applied for initial retrieval, followed by re-ranking using a pre-trained cross-encoder model, specifically, `ms-marco-MiniLM-L-6-v2` from the `sentence-transformers` library. Additionally, we report results for two variations: Doc$^+$, in which the gold context is incorporated into the input prompt, and Doc$^-$, in which an irrelevant context is substituted. Although several methods exist to obtain irrelevant contexts (Zhao et al., 2024c), in our experiments, these are generated by randomly sampling a context from the corpus.

NLI Models. A NLI classifier takes a sequence pair (x_1, x_2) and outputs a label $y \in \{\mathit{Contradiction}, \mathit{Neutral}, \mathit{Entailment}\}$ with corresponding probabilities. The two sequences are concatenated with a separator token `[SEP]` before input. To study ordering effects, we consider both x_1`[SEP]`x_2 and x_2`[SEP]`x_1. In the reference-free

UE	NQ-open			TriviaQA			PopQA		
	BM25	Contriever	Doc$^+$	BM25	Contriever	Doc$^+$	BM25	Contriever	Doc$^+$
Axiom 1: Positively Consistent ↓									
PE	1.620 → 1.332 *	1.538 → 1.288 *	1.544 → 1.232 *	0.531 → 0.483 *	0.549 → 0.494 *	0.570 → 0.456 *	0.893 → 0.673 *	0.886 → 0.638 *	1.368 → 0.419 *
SE	4.874 → 4.164 *	4.876 → 4.060 *	4.941 → 3.922 *	3.460 → 3.265 *	3.508 → 3.299 *	3.565 → 3.241 *	4.063 → 3.361 *	4.162 → 3.354 *	5.379 → 3.112 *
PE+M	3.682 → 3.283 *	3.380 → 3.188 *	3.395 → 3.080 *	0.943 → 0.903 *	0.987 → 0.948 *	1.010 → 0.869 *	1.220 → 0.942 *	1.195 → 0.836 *	2.115 → 0.615 *
SE+M	6.710 → 5.863 *	6.531 → 5.746 *	6.594 → 5.602 *	3.839 → 3.638 *	3.913 → 3.715 *	3.971 → 3.615 *	4.315 → 3.539 *	4.424 → 3.459 *	6.087 → 3.224 *
EigV	1.724 → 1.285 *	1.788 → 1.172 *	1.901 → 1.069 *	1.277 → 1.129 *	1.294 → 1.162 *	1.344 → 1.114 *	1.614 → 1.119 *	1.837 → 1.095 *	3.781 → 1.041 *
ECC	0.356 → 0.169 *	0.381 → 0.104 *	0.405 → 0.043 *	0.155 → 0.082 *	0.164 → 0.094 *	0.187 → 0.076 *	0.260 → 0.050 *	0.288 → 0.052 *	0.621 → 0.021 *
Deg	0.175 → 0.088 *	0.185 → 0.053 *	0.208 → 0.023 *	0.063 → 0.038 *	0.067 → 0.042 *	0.076 → 0.030 *	0.129 → 0.045 *	0.157 → 0.033 *	0.426 → 0.016 *
Axiom 2: Negatively Consistent ↑									
PE	2.460 → 2.303 *	2.353 → 2.321	2.377 → 2.374	1.512 → 1.397	1.226 → 1.228	1.477 → 1.421 *	0.933 → 0.589 *	0.804 → 0.450 *	1.196 → 0.570
SE	5.846 → 5.233 *	5.614 → 5.074 *	5.619 → 4.966 *	4.697 → 4.384 *	4.449 → 4.133 *	4.936 → 4.699 *	4.407 → 3.314 *	4.290 → 3.215 *	4.620 → 3.442
PE+M	6.014 → 5.908	5.523 → 5.664	5.783 → 5.920	2.917 → 2.797	2.260 → 2.313	2.777 → 2.833	1.376 → 0.901 *	1.230 → 0.762 *	1.631 → 0.686
SE+M	9.087 → 8.557 *	8.493 → 8.092	8.728 → 8.147	6.033 → 5.699	5.462 → 5.100 *	6.121 → 6.009	4.819 → 3.551 *	4.702 → 3.456 *	4.875 → 3.504
EigV	2.177 → 1.529 *	2.047 → 1.303 *	1.869 → 1.071 *	1.648 → 1.472	1.655 → 1.489	2.284 → 2.025 *	2.041 → 1.098 *	2.055 → 1.181 *	2.188 → 1.143
ECC	0.507 → 0.256 *	0.437 → 0.166 *	0.453 → 0.040 *	0.367 → 0.223 *	0.394 → 0.243 *	0.476 → 0.394 *	0.338 → 0.065 *	0.411 → 0.069 *	0.514 → 0.041
Deg	0.260 → 0.134 *	0.227 → 0.080 *	0.210 → 0.022 *	0.153 → 0.127	0.152 → 0.120	0.254 → 0.205 *	0.200 → 0.030 *	0.194 → 0.044 *	0.260 → 0.055
Axiom 3: Positively Changed ↓									
PE	1.972 → 1.038 *	1.972 → 1.120 *	2.020 → 1.097 *	1.492 → 0.531 *	1.446 → 0.515 *	1.452 → 0.510 *	1.837 → 0.734 *	1.727 → 0.566 *	1.458 → 0.403 *
SE	5.861 → 3.808 *	5.813 → 3.855 *	5.898 → 3.838 *	5.527 → 3.337 *	5.569 → 3.326 *	5.497 → 3.364 *	6.309 → 3.349 *	6.227 → 3.276 *	5.662 → 3.104 *
PE+M	3.917 → 2.599 *	4.061 → 2.810 *	4.063 → 2.762 *	2.544 → 0.959 *	2.545 → 1.033 *	2.480 → 0.927 *	2.935 → 0.970 *	2.686 → 0.867 *	2.244 → 0.594 *
SE+M	7.587 → 5.162 *	7.662 → 5.299 *	7.746 → 5.303 *	6.542 → 3.690 *	6.606 → 3.798 *	6.465 → 3.716 *	7.365 → 3.467 *	7.156 → 3.436 *	6.439 → 3.211 *
EigV	3.745 → 1.168 *	3.449 → 1.131 *	3.547 → 1.119 *	3.575 → 1.191 *	3.611 → 1.179 *	3.470 → 1.210 *	5.124 → 1.054 *	5.217 → 1.055 *	4.323 → 1.040 *
ECC	0.653 → 0.089 *	0.633 → 0.072 *	0.661 → 0.069 *	0.756 → 0.110 *	0.752 → 0.104 *	0.747 → 0.131 *	0.854 → 0.024 *	0.841 → 0.024 *	0.700 → 0.025 *
Deg	0.471 → 0.053 *	0.450 → 0.048 *	0.466 → 0.037 *	0.462 → 0.053 *	0.471 → 0.047 *	0.454 → 0.063 *	0.614 → 0.022 *	0.615 → 0.021 *	0.492 → 0.016 *
Axiom 4: Negatively Changed ↑									
PE	1.450 → 1.284 *	1.570 → 1.490	1.518 → 1.256 *	0.791 → 1.173 *	0.833 → 1.144 *	0.881 → 1.021	0.941 → 0.881	1.014 → 0.807	1.660 → 0.913 *
SE	4.957 → 4.252 *	5.039 → 4.543 *	4.775 → 4.116 *	4.173 → 4.356 *	4.212 → 4.319	4.392 → 4.174	4.569 → 3.875 *	4.739 → 3.709 *	5.853 → 3.783 *
PE+M	3.045 → 2.901	3.421 → 3.597	3.159 → 2.924	1.383 → 1.989 *	1.361 → 2.107 *	1.424 → 1.849	1.323 → 1.354	1.447 → 1.303	2.705 → 1.735 *
SE+M	6.368 → 5.630 *	6.674 → 6.349	6.181 → 5.549	4.743 → 5.076 *	4.720 → 5.081	4.954 → 4.796	4.958 → 4.241 *	5.184 → 4.062 *	6.835 → 4.446 *
EigV	2.087 → 1.415 *	2.115 → 1.497 *	1.906 → 1.375 *	1.850 → 1.593	1.944 → 1.710	2.103 → 1.594 *	2.522 → 1.200 *	2.565 → 1.222 *	4.209 → 1.159 *
ECC	0.440 → 0.208 *	0.438 → 0.238 *	0.351 → 0.151 *	0.362 → 0.293	0.378 → 0.323	0.420 → 0.298	0.437 → 0.091 *	0.492 → 0.119 *	0.700 → 0.068 *
Deg	0.243 → 0.138 *	0.252 → 0.149 *	0.222 → 0.109 *	0.175 → 0.183	0.190 → 0.203	0.233 → 0.180	0.259 → 0.091 *	0.280 → 0.092 *	0.479 → 0.065 *

Table 7: Comparison of changes in average uncertainty values for Axioms 1–4 before (left) and after (right) applying RAG with Mistral-v0.3. Axioms are implemented using the *Reference-based* method. Colors green and deep red indicate significant changes aligning or conflicting with axioms, respectively. Color shallow red represents non-significant changes conflicting with axioms. Significance is marked by $*$.

Unc.	NQ-open	TriviaQA	PopQA
PE	2.227 → 1.778 *	0.657 → 0.780 *	1.014 → 1.087
SE	5.453 → 4.964 *	3.570 → 3.892 *	3.976 → 4.021
PE+M	5.634 → 4.293 *	1.223 → 1.374 *	1.686 → 1.759
SE+M	8.543 → 7.216 *	4.089 → 4.463 *	4.310 → 4.521
EigV	1.696 → 1.637	1.256 → 1.496 *	1.215 → 1.452
ECC	0.357 → 0.335	0.154 → 0.300 *	0.059 → 0.362
Deg	0.160 → 0.206 *	0.056 → 0.128 *	0.093 → 0.140

Table 8: Comparison of changes in average uncertainty values for Axiom 5 before (left) and after(right) applying RAG with Mistral-v0.3. Color coding and significance markers follow those in Table 7.

setting (Section 3.3), if either order yields a contradiction, the input is labeled as such; otherwise, it is labeled as entailment. In Section 4.1, we use the maximum entailment probability from the two orders.

Calibration Function. We perform a grid search on the validation set of each dataset to determine the axiomatic coefficients (k_1, k_2, k_3, k_4) as described in Section 4. This grid search simultaneously pursues two objectives: satisfying the axioms and maximizing the overall AUROC. For the CTI method, the optimal coefficients are (0.05, 0.75, 0.20, 1.30); for the NLI and MiniCheck methods, the optimal coefficients are (0.05, 0.90, 0.05, 1.20) consistently across all datasets.

Computational Cost. We conducted all experiments using Nvidia A-100 GPUs with 40 GB of memory, accumulating approximately 250 GPU hours. Due to the substantial computational demands, all results presented are based on a single run.

UE	NQ-open			TriviaQA			PopQA		
	BM25	**Contriever**	**Doc$^+$**	**BM25**	**Contriever**	**Doc$^+$**	**BM25**	**Contriever**	**Doc$^+$**
Axiom 1: Positively Consistent ↓									
PE	1.896 → 1.802	1.801 → 1.642	1.684 → 1.500 *	0.796 → 0.848 *	0.844 → 0.877 *	0.929 → 0.952 *	0.798 → 0.418	0.715 → 0.416 *	0.818 → 0.191
SE	5.174 → 4.524 *	5.071 → 4.344 *	4.957 → 4.145 *	3.779 → 3.533	3.823 → 3.569 *	4.019 → 3.725 *	3.869 → 3.260 *	3.805 → 3.152 *	3.700 → 3.053 *
PE+M	4.445 → 4.152 *	4.162 → 4.013 *	4.090 → 4.039	1.307 → 1.331 *	1.368 → 1.392	1.564 → 1.559	0.930 → 0.490	0.820 → 0.486 *	0.846 → 0.213 *
SE+M	7.716 → 6.855 *	7.483 → 6.619 *	7.380 → 6.577 *	4.422 → 4.062 *	4.495 → 4.142 *	4.783 → 4.381 *	4.185 → 3.354 *	4.090 → 3.265 *	3.909 → 3.080 *
EigV	2.248 → 1.451 *	2.264 → 1.236 *	2.183 → 1.105 *	1.656 → 1.375 *	1.704 → 1.296 *	1.913 → 1.583 *	2.088 → 1.215 *	2.030 → 1.175 *	2.126 → 1.145 *
ECC	0.546 → 0.224 *	0.583 → 0.163 *	0.548 → 0.080 *	0.353 → 0.237 *	0.369 → 0.187 *	0.422 → 0.289 *	0.447 → 0.133 *	0.432 → 0.093 *	0.405 → 0.071 *
Deg	0.264 → 0.123 *	0.277 → 0.075 *	0.262 → 0.032 *	0.153 → 0.097 *	0.161 → 0.081 *	0.201 → 0.134 *	0.222 → 0.058 *	0.218 → 0.051 *	0.236 → 0.040 *
Axiom 2: Negatively Consistent ↑									
PE	1.978 → 2.037	1.717 → 1.507	1.705 → 1.173 *	0.783 → 0.833	0.794 → 0.718	0.795 → 0.749	0.817 → 0.583	0.698 → 0.310	1.296 → 0.528
SE	5.499 → 5.108	5.039 → 4.210 *	5.034 → 3.845 *	3.707 → 3.740	3.744 → 3.521	3.897 → 3.568	3.570 → 3.238	3.698 → 3.119 *	4.233 → 2.986
PE+M	4.707 → 4.579	3.438 → 3.295	3.483 → 3.027	1.029 → 1.185	1.047 → 0.987	1.024 → 1.030	0.817 → 0.542	0.640 → 0.332	1.262 → 0.637
SE+M	8.217 → 7.579	6.970 → 6.014 *	6.959 → 5.525 *	4.103 → 4.149	4.198 → 3.894	4.351 → 3.892 *	3.771 → 3.321 *	3.854 → 3.181 *	4.458 → 3.029
EigV	2.563 → 2.233	2.610 → 1.464 *	2.236 → 1.192 *	1.811 → 1.537 *	1.804 → 1.426 *	1.970 → 1.632 *	1.911 → 1.217 *	2.015 → 1.175 *	2.998 → 1.214
ECC	0.580 → 0.403	0.612 → 0.294 *	0.626 → 0.169 *	0.419 → 0.359	0.390 → 0.283	0.450 → 0.319 *	0.406 → 0.142 *	0.473 → 0.145 *	0.667 → 0.058
Deg	0.308 → 0.255	0.331 → 0.125 *	0.269 → 0.063 *	0.177 → 0.141	0.173 → 0.111 *	0.215 → 0.139 *	0.217 → 0.077 *	0.217 → 0.052 *	0.407 → 0.093
Axiom 3: Positively Changed ↓									
PE	1.800 → 1.239 *	1.860 → 1.261 *	1.816 → 1.063 *	1.239 → 0.749 *	1.287 → 0.851 *	1.332 → 0.686 *	1.348 → 0.397 *	1.386 → 0.368 *	1.358 → 0.264 *
SE	5.575 → 4.025 *	5.603 → 4.055 *	5.685 → 3.742 *	4.773 → 3.511 *	4.908 → 3.641 *	5.029 → 3.312 *	5.092 → 3.161 *	5.203 → 3.135 *	4.987 → 3.050 *
PE+M	3.630 → 3.003 *	3.770 → 3.028 *	3.704 → 2.785 *	1.809 → 1.112 *	1.943 → 1.297 *	1.835 → 1.061 *	1.723 → 0.470 *	1.766 → 0.436 *	1.655 → 0.331 *
SE+M	7.504 → 5.709 *	7.565 → 5.705 *	7.693 → 5.285 *	5.504 → 3.907 *	5.740 → 4.134 *	5.771 → 3.681 *	5.691 → 3.262 *	5.782 → 3.228 *	5.506 → 3.102 *
EigV	3.693 → 1.335 *	3.822 → 1.321 *	3.947 → 1.149 *	3.377 → 1.332 *	3.626 → 1.411 *	3.840 → 1.281 *	4.772 → 1.222 *	5.100 → 1.197 *	4.622 → 1.102 *
ECC	0.762 → 0.203 *	0.760 → 0.207 *	0.817 → 0.098 *	0.738 → 0.213 *	0.796 → 0.261 *	0.845 → 0.162 *	0.814 → 0.135 *	0.855 → 0.125 *	0.806 → 0.065 *
Deg	0.494 → 0.105 *	0.517 → 0.100 *	0.538 → 0.048 *	0.460 → 0.097 *	0.494 → 0.121 *	0.525 → 0.076 *	0.593 → 0.069 *	0.630 → 0.059 *	0.569 → 0.032 *
Axiom 4: Negatively Changed ↑									
PE	2.027 → 1.829	2.245 → 1.342 *	2.423 → 1.386 *	1.139 → 1.017	1.017 → 0.911	1.427 → 1.067	1.248 → 0.874	1.600 → 0.543	1.964 → 0.223
SE	5.476 → 4.683 *	5.494 → 4.245 *	5.689 → 4.419 *	4.626 → 4.176 *	4.554 → 4.028 *	4.523 → 3.697 *	4.941 → 3.822 *	4.678 → 3.879 *	5.367 → 3.435 *
PE+M	3.922 → 3.817	3.501 → 2.822 *	4.112 → 3.021 *	1.649 → 1.611	1.465 → 1.570	1.646 → 1.313	1.634 → 1.153	1.784 → 0.621	2.302 → 0.339
SE+M	7.532 → 6.421 *	7.092 → 5.728 *	7.660 → 6.024 *	5.387 → 4.771 *	5.256 → 4.773 *	5.135 → 4.003 *	5.530 → 4.164 *	5.171 → 4.041 *	5.972 → 3.593 *
EigV	2.876 → 1.754 *	3.040 → 1.550 *	2.791 → 1.729 *	2.919 → 1.995 *	2.983 → 1.780 *	2.887 → 2.134	3.995 → 1.683 *	4.122 → 1.840 *	5.520 → 1.421 *
ECC	0.685 → 0.343 *	0.641 → 0.307 *	0.505 → 0.395	0.705 → 0.499 *	0.700 → 0.434 *	0.741 → 0.433 *	0.755 → 0.333 *	0.799 → 0.429 *	0.917 → 0.245 *
Deg	0.417 → 0.229 *	0.442 → 0.171 *	0.426 → 0.199 *	0.384 → 0.253 *	0.397 → 0.215 *	0.405 → 0.269	0.508 → 0.215 *	0.546 → 0.199 *	0.688 → 0.120 *

Table 9: Comparison of changes in average uncertainty values for Axioms 1–4 before (left) and after (right) applying RAG with Llama2-chat. Axioms are implemented using the *Reference-free* method. Colors green and deep red indicate significant changes aligning or conflicting with axioms, respectively. Color shallow red represents non-significant changes conflicting with axioms. Significance is marked by $*$.

UE	NQ-open					TriviaQA					PopQA				
	A1 (%)	**A2** (%)	**A3** (%)	**A4** (%)	**AUROC**	**A1** (%)	**A2** (%)	**A3** (%)	**A4** (%)	**AUROC**	**A1** (%)	**A2** (%)	**A3** (%)	**A4** (%)	**AUROC**
PE	69.77	42.95	80.54	45.21	64.40	63.88	39.05	87.24	61.30	79.14	70.57	22.22	89.17	44.23	65.72
+CTI	67.91	40.39	78.60	48.85	64.82	65.67	37.87	86.01	64.04	79.90	72.91	14.82	89.63	46.80	67.14
+NLI	70.81	42.31	80.54	52.81	66.50	65.88	42.01	86.42	64.04	79.78	74.23	24.69	93.09	46.80	68.65
+MCH	78.05	32.05	85.41	49.18	67.15	67.53	44.38	87.24	64.38	**80.25**	74.22	17.28	91.01	48.72	68.63
SE	76.40	32.69	90.54	37.62	65.66	65.88	31.36	91.36	51.71	78.83	74.22	16.05	95.39	33.97	68.53
+CTI	74.12	31.41	87.57	42.57	65.13	53.22	40.23	85.60	59.93	76.28	70.57	12.35	93.55	37.18	67.10
+NLI	70.39	42.95	87.03	49.51	66.92	51.57	48.52	86.83	62.67	77.01	69.01	25.93	94.24	39.74	70.53
+MCH	78.47	30.13	89.73	42.57	**69.66**	68.60	45.56	89.30	56.16	77.65	80.21	12.35	94.01	42.31	70.10
EigV	54.66	14.10	85.40	27.72	63.03	19.24	23.08	85.19	41.44	72.25	41.15	6.17	93.08	26.92	66.35
+CTI	71.22	24.36	87.30	37.95	65.06	47.93	47.34	88.89	60.27	74.79	68.23	16.05	93.55	39.74	65.23
+NLI	69.98	38.46	88.65	41.91	67.29	48.64	50.89	87.24	60.27	74.48	65.88	32.10	95.16	39.10	68.40
+MCH	77.85	28.85	92.16	36.63	68.45	68.81	45.56	90.95	53.43	75.81	83.07	12.35	96.31	37.18	70.39
ECC	53.00	13.46	81.62	26.07	62.87	18.31	14.79	78.60	35.62	71.72	40.62	4.93	92.16	23.08	66.28
+CTI	72.05	29.48	86.48	39.60	66.76	47.13	50.29	82.71	60.95	76.22	67.45	18.52	94.24	37.82	68.36
+NLI	70.18	39.74	87.29	43.23	67.59	48.35	50.29	84.36	62.32	75.77	65.88	29.63	95.62	36.53	69.91
+MCH	79.08	32.05	90.81	37.29	68.80	68.74	43.19	90.12	53.08	77.67	81.77	11.11	96.08	36.53	**72.71**

Table 10: Percentage of samples passing the axioms before and after calibration for Contriever with Mistral-v0.3. The results show that as the number of samples passing the axioms increases, the AUROC also improves.

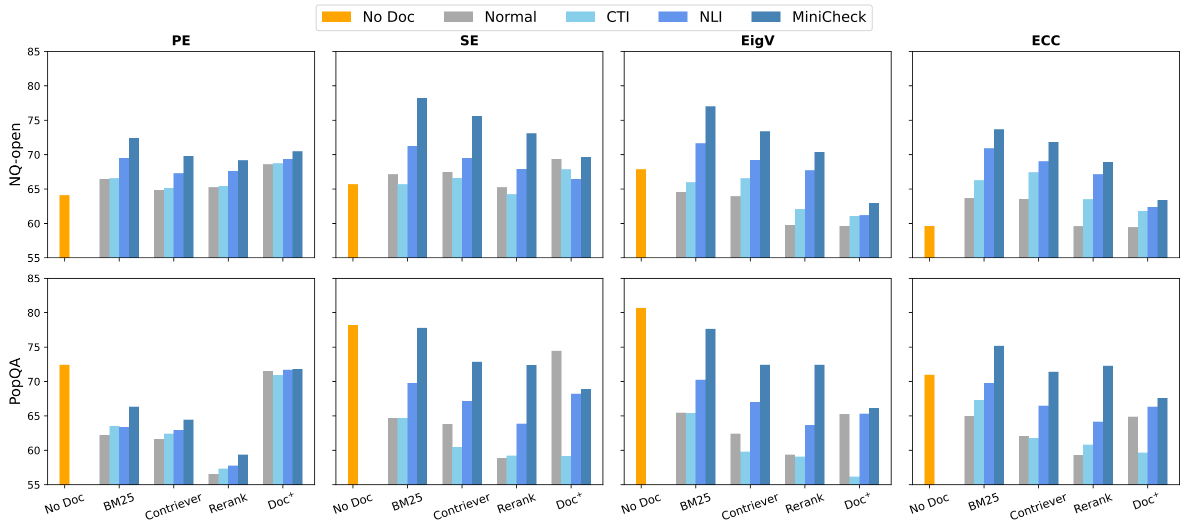

Figure 5: Comparison of AUROC between the no-RAG and calibrated RAG settings for Llama2-chat for NQ-open and PopQA datasets. AUROC improves significantly, either surpassing the no-RAG setting or reducing the gap between them.

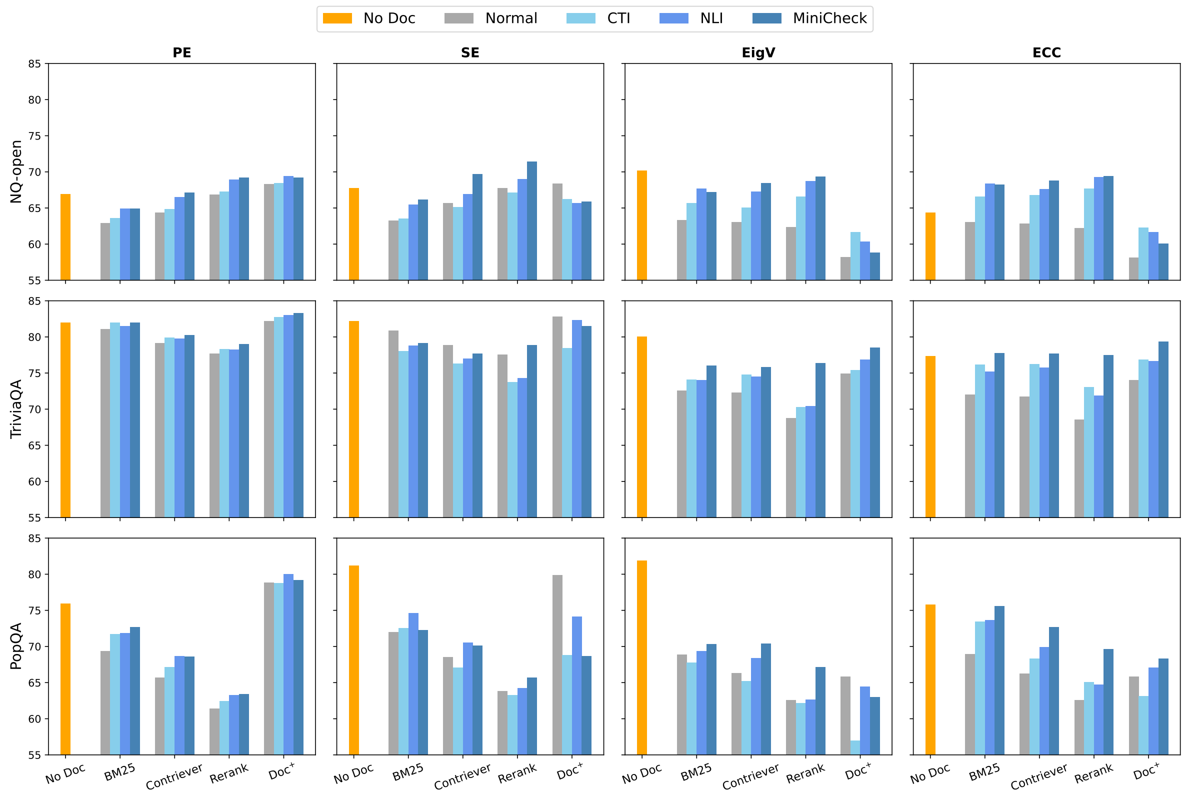

Figure 6: Comparison of AUROC between the no-RAG and calibrated RAG settings for Mistral-v0.3. AUROC improves significantly, either surpassing the no-RAG setting or reducing the gap between them.

UE	NQ-open			TriviaQA			PopQA		
	BM25	**Contriever**	**Doc^{+}**	**BM25**	**Contriever**	**Doc^{+}**	**BM25**	**Contriever**	**Doc^{+}**
Axiom 1: Positively Consistent ↓									
PE	55.357 → 55.357	60.194 → 61.489	61.793 → 61.598	41.454 → 43.124	45.532 → 46.002	45.853 → 46.401	62.369 → 68.293	66.189 → 69.628	79.511 → 81.040
SE	66.964 → 70.982	77.346 → 77.023	79.337 → 83.041	47.446 → 55.894	50.141 → 56.538	52.191 → 58.059	69.338 → 79.443	71.920 → 78.510	88.073 → 86.544
PE+M	57.589 → 57.589	61.165 → 63.430	62.378 → 62.378	44.008 → 44.499	43.744 → 45.720	46.870 → 48.044	62.021 → 68.293	69.628 → 73.639	81.040 → 82.875
SE+M	64.286 → 68.304	76.375 → 76.375	72.904 → 73.099	47.348 → 55.599	48.354 → 57.008	52.504 → 58.059	69.686 → 80.488	73.639 → 79.656	89.602 → 88.073
EigV	58.036 → 71.875	65.372 → 77.346	69.981 → 84.795	35.265 → 64.047	37.159 → 66.886	38.498 → 57.199	53.659 → 83.275	55.301 → 80.802	81.346 → 91.743
ECC	54.464 → 72.321	61.489 → 75.728	68.226 → 84.016	32.122 → 62.279	34.243 → 65.475	34.977 → 55.634	50.523 → 80.836	52.436 → 78.797	77.064 → 88.379
Deg	55.804 → 57.589	64.401 → 65.049	70.565 → 70.565	34.283 → 35.069	36.595 → 37.065	37.950 → 38.419	54.007 → 55.401	55.014 → 55.874	81.346 → 81.346
Axiom 2: Negatively Consistent ↑									
PE	46.237 → 44.086	47.853 → 44.172	47.059 → 44.118	52.299 → 50.575	43.781 → 43.781	56.477 → 56.218	49.275 → 44.928	42.466 → 39.726	57.143 → 57.143
SE	34.409 → 31.183	33.742 → 26.380	31.618 → 28.676	42.529 → 41.379	35.821 → 39.303	45.078 → 50.777	34.783 → 21.739	31.507 → 26.027	42.857 → 28.571
PE+M	39.247 → 39.247	42.945 → 38.037	47.794 → 46.324	49.425 → 47.701	41.791 → 41.791	52.332 → 53.886	44.928 → 43.478	43.836 → 42.466	57.143 → 57.143
SE+M	31.720 → 30.108	31.288 → 26.380	35.294 → 36.029	41.379 → 37.356	35.323 → 38.308	44.301 → 51.295	33.333 → 17.391	30.137 → 27.397	42.857 → 28.571
EigV	19.355 → 31.720	12.883 → 20.245	5.147 → 26.471	29.885 → 34.483	24.378 → 30.846	37.047 → 50.259	15.942 → 13.043	6.849 → 19.178	42.857 → 28.571
ECC	14.516 → 37.097	9.816 → 23.313	5.882 → 30.147	19.540 → 35.057	14.428 → 31.841	21.503 → 58.031	10.145 → 20.290	6.849 → 23.288	28.571 → 28.571
Deg	20.968 → 20.968	17.178 → 15.951	5.147 → 6.618	29.885 → 31.034	24.378 → 24.876	36.788 → 42.487	13.043 → 13.043	12.329 → 12.329	57.143 → 57.143
Axiom 3: Positively Changed ↓									
PE	82.215 → 81.544	77.346 → 76.430	82.557 → 81.541	73.402 → 72.634	70.256 → 69.231	74.870 → 73.830	82.331 → 83.083	87.572 → 87.954	84.314 → 84.540
SE	93.289 → 93.289	91.533 → 89.703	93.057 → 91.194	86.445 → 83.632	84.615 → 79.744	88.042 → 83.882	93.233 → 90.226	94.073 → 91.205	92.534 → 88.235
PE+M	81.544 → 79.866	77.574 → 76.888	80.271 → 78.493	76.982 → 77.749	73.590 → 72.308	80.069 → 77.296	88.346 → 87.594	90.822 → 90.440	84.389 → 84.691
SE+M	90.604 → 88.591	88.787 → 85.812	88.654 → 86.198	86.957 → 84.143	84.359 → 80.000	88.562 → 84.922	93.609 → 92.857	94.455 → 92.543	93.439 → 89.668
EigV	90.604 → 91.611	88.558 → 90.389	89.077 → 91.025	86.189 → 85.166	86.154 → 86.154	83.709 → 86.308	91.353 → 90.977	92.925 → 93.499	86.652 → 89.367
ECC	82.886 → 87.919	83.066 → 87.185	82.557 → 86.622	79.028 → 80.563	73.590 → 77.692	75.390 → 80.243	86.466 → 89.850	87.380 → 90.822	82.730 → 86.652
Deg	90.604 → 90.940	87.414 → 87.643	89.331 → 89.670	85.934 → 86.189	86.410 → 85.128	85.442 → 85.789	91.353 → 90.977	92.543 → 92.543	86.576 → 86.501
Axiom 4: Negatively Changed ↑									
PE	51.136 → 52.273	51.163 → 53.876	49.231 → 50.769	66.944 → 66.389	66.879 → 68.471	66.372 → 63.717	42.045 → 39.773	38.168 → 38.168	27.586 → 27.586
SE	36.080 → 40.625	36.047 → 40.310	44.615 → 40.000	55.556 → 58.889	54.777 → 56.688	52.212 → 57.522	31.818 → 36.364	29.008 → 26.718	25.287 → 22.989
PE+M	47.727 → 49.716	50.388 → 53.876	50.769 → 56.923	63.333 → 64.722	66.242 → 65.287	64.602 → 65.487	38.636 → 38.636	32.061 → 31.298	26.437 → 27.586
SE+M	38.636 → 41.193	40.698 → 42.636	41.538 → 49.231	55.278 → 57.500	53.503 → 56.369	53.097 → 55.752	31.250 → 33.523	28.244 → 24.427	24.138 → 20.690
EigV	24.432 → 34.091	24.419 → 35.271	16.923 → 18.462	38.333 → 51.944	39.172 → 48.408	38.938 → 51.327	21.591 → 35.795	20.611 → 29.771	8.046 → 12.644
ECC	19.602 → 39.205	18.992 → 37.984	16.923 → 30.769	30.556 → 57.500	30.892 → 53.185	26.549 → 60.177	18.182 → 44.318	18.321 → 34.351	8.046 → 19.540
Deg	25.284 → 26.989	24.806 → 27.132	20.000 → 23.077	42.500 → 45.278	42.357 → 44.904	42.478 → 44.248	22.727 → 23.864	19.084 → 19.084	11.494 → 11.494

Table 11: Changes in the percentage of samples that satisfy the axioms before and after calibration for Llama2-chat. The relation function $\mathcal{R}$ is implemented using CTI.

UE	NQ-open			TriviaQA			PopQA		
	BM25	**Contriever**	**Doc$^+$**	**BM25**	**Contriever**	**Doc$^+$**	**BM25**	**Contriever**	**Doc$^+$**
Axiom 1: Positively Consistent ↓									
PE	55.357 → 60.714	60.194 → 66.019	61.793 → 66.667	41.454 → 44.695	45.532 → 48.730	45.853 → 47.966	62.369 → 63.415	66.189 → 68.481	79.511 → 80.122
SE	66.964 → 73.661	77.346 → 80.259	79.337 → 79.922	47.446 → 58.743	50.141 → 58.231	52.191 → 57.433	69.338 → 73.868	71.920 → 74.785	88.073 → 81.040
PE+M	57.589 → 62.054	61.165 → 67.961	62.378 → 67.057	44.008 → 46.660	43.744 → 48.260	46.870 → 49.687	62.021 → 63.763	69.628 → 72.206	81.040 → 82.263
SE+M	64.286 → 70.089	76.375 → 77.023	72.904 → 74.269	47.348 → 58.350	48.354 → 58.325	52.504 → 57.825	69.686 → 76.655	73.639 → 75.072	89.602 → 84.709
EigV	58.036 → 71.429	65.372 → 82.201	69.981 → 86.160	35.265 → 59.136	37.159 → 60.960	38.498 → 59.077	53.659 → 73.868	55.301 → 74.785	81.346 → 85.933
ECC	54.464 → 70.982	61.489 → 78.641	68.226 → 84.795	32.122 → 55.403	34.243 → 58.043	34.977 → 55.399	50.523 → 72.474	52.436 → 71.347	77.064 → 84.404
Deg	55.804 → 56.250	64.401 → 64.401	70.565 → 70.955	34.283 → 35.069	36.595 → 36.877	37.950 → 38.185	54.007 → 55.052	55.014 → 57.307	81.346 → 81.040
Axiom 2: Negatively Consistent ↑									
PE	46.237 → 47.312	47.853 → 47.239	47.059 → 46.324	52.299 → 54.598	43.781 → 45.274	56.477 → 58.549	49.275 → 49.275	42.466 → 41.096	57.143 → 57.143
SE	34.409 → 38.172	33.742 → 38.037	31.618 → 31.618	42.529 → 48.276	35.821 → 43.284	45.078 → 56.218	34.783 → 33.333	31.507 → 32.877	42.857 → 57.143
PE+M	39.247 → 43.011	42.945 → 41.718	47.794 → 50.735	49.425 → 54.023	41.791 → 41.294	52.332 → 56.218	44.928 → 46.377	43.836 → 45.205	57.143 → 57.143
SE+M	31.720 → 38.710	31.288 → 36.196	35.294 → 33.824	41.379 → 45.977	35.323 → 39.801	44.301 → 55.440	33.333 → 30.435	30.137 → 32.877	42.857 → 42.857
EigV	19.355 → 35.484	12.883 → 26.380	5.147 → 20.588	29.885 → 46.552	24.378 → 38.806	37.047 → 58.290	15.942 → 26.087	6.849 → 32.877	42.857 → 42.857
ECC	14.516 → 43.011	9.816 → 32.515	5.882 → 25.000	19.540 → 55.172	14.428 → 42.786	21.503 → 76.425	10.145 → 34.783	6.849 → 32.877	28.571 → 57.143
Deg	20.968 → 23.656	17.178 → 17.791	5.147 → 8.824	29.885 → 34.483	24.378 → 26.866	36.788 → 45.596	13.043 → 13.043	12.329 → 13.699	57.143 → 57.143
Axiom 3: Positively Changed ↓									
PE	82.215 → 84.228	77.346 → 77.574	82.557 → 81.964	73.402 → 73.913	70.256 → 70.513	74.870 → 74.003	82.331 → 84.586	87.572 → 88.145	84.314 → 84.615
SE	93.289 → 88.591	91.533 → 86.270	93.057 → 86.113	86.445 → 84.910	84.615 → 80.513	88.042 → 84.749	93.233 → 91.729	94.073 → 92.161	92.534 → 87.029
PE+M	81.544 → 85.235	77.574 → 79.863	80.271 → 80.610	76.982 → 77.238	73.590 → 72.821	80.069 → 78.683	88.346 → 88.346	90.822 → 90.057	84.389 → 85.143
SE+M	90.604 → 87.248	88.787 → 84.211	88.654 → 82.557	86.957 → 84.655	84.359 → 81.026	88.562 → 86.482	93.609 → 92.105	94.455 → 93.690	93.439 → 88.235
EigV	90.604 → 92.617	88.558 → 91.533	89.077 → 90.517	86.189 → 87.724	86.154 → 87.436	83.709 → 86.655	91.353 → 93.609	92.925 → 95.602	86.652 → 90.875
ECC	82.886 → 88.255	83.066 → 87.185	82.557 → 86.791	79.028 → 84.655	73.590 → 77.436	75.390 → 78.163	86.466 → 91.353	87.380 → 92.161	82.730 → 88.537
Deg	90.604 → 89.933	87.414 → 86.270	89.331 → 89.162	85.934 → 86.189	86.410 → 86.154	85.442 → 84.749	91.353 → 91.353	92.543 → 92.352	86.576 → 86.652
Axiom 4: Negatively Changed ↑									
PE	51.136 → 56.250	51.163 → 55.426	49.231 → 58.462	66.944 → 68.611	66.879 → 68.790	66.372 → 69.027	42.045 → 42.614	38.168 → 41.221	27.586 → 31.034
SE	36.080 → 49.432	36.047 → 50.000	44.615 → 52.308	55.556 → 65.000	54.777 → 64.013	52.212 → 64.602	31.818 → 42.045	29.008 → 41.985	25.287 → 31.034
PE+M	47.727 → 52.273	50.388 → 56.977	50.769 → 55.385	63.333 → 66.667	66.242 → 67.834	64.602 → 67.257	38.636 → 38.636	32.061 → 35.115	26.437 → 29.885
SE+M	38.636 → 51.136	40.698 → 53.488	41.538 → 56.923	55.278 → 62.778	53.503 → 64.013	53.097 → 61.947	31.250 → 38.068	28.244 → 39.695	24.138 → 29.885
EigV	24.432 → 35.795	24.419 → 36.047	16.923 → 33.846	38.333 → 57.500	39.172 → 53.185	38.938 → 53.982	21.591 → 36.932	20.611 → 38.931	8.046 → 18.391
ECC	19.602 → 43.466	18.992 → 42.636	16.923 → 36.923	30.556 → 65.556	30.892 → 59.873	26.549 → 65.487	18.182 → 46.591	18.321 → 42.748	8.046 → 25.287
Deg	25.284 → 29.545	24.806 → 27.907	20.000 → 24.615	42.500 → 47.222	42.357 → 48.089	42.478 → 48.673	22.727 → 24.432	19.084 → 21.374	11.494 → 16.092

Table 12: Changes in the percentage of samples that satisfy the axioms before and after calibration for Llama2-chat. The relation function $\mathcal{R}$ is implemented using NLI.

UE	NQ-open			TriviaQA			PopQA		
	BM25	**Contriever**	**Doc**$^{+}$	**BM25**	**Contriever**	**Doc**$^{+}$	**BM25**	**Contriever**	**Doc**$^{+}$
Axiom 1: Positively Consistent ↓									
PE	55.357 → 70.089	60.194 → 75.081	61.793 → 75.634	41.454 → 48.134	45.532 → 51.364	45.853 → 51.174	62.369 → 68.293	66.189 → 69.341	79.511 → 81.040
SE	66.964 → 78.571	77.346 → 86.408	79.337 → 90.058	47.446 → 72.299	50.141 → 73.283	52.191 → 73.083	69.338 → 81.533	71.920 → 77.937	88.073 → 86.544
PE+M	57.589 → 68.750	61.165 → 78.317	62.378 → 76.023	44.008 → 48.723	43.744 → 51.646	46.870 → 52.034	62.021 → 70.035	69.628 → 71.347	81.040 → 81.957
SE+M	64.286 → 75.446	76.375 → 86.731	72.904 → 87.329	47.348 → 71.709	48.354 → 72.907	52.504 → 72.457	69.686 → 82.230	73.639 → 78.223	89.602 → 87.156
EigV	58.036 → 77.679	65.372 → 87.702	69.981 → 93.177	35.265 → 70.432	37.159 → 74.882	38.498 → 74.257	53.659 → 82.927	55.301 → 83.095	81.346 → 95.413
ECC	54.464 → 76.786	61.489 → 86.084	68.226 → 92.398	32.122 → 66.306	34.243 → 72.437	34.977 → 70.736	50.523 → 80.488	52.436 → 79.370	77.064 → 93.272
Deg	55.804 → 57.143	64.401 → 65.372	70.565 → 70.175	34.283 → 35.560	36.595 → 37.535	37.950 → 39.202	54.007 → 54.704	55.014 → 55.587	81.346 → 81.346
Axiom 2: Negatively Consistent ↑									
PE	46.237 → 52.151	47.853 → 39.877	47.059 → 38.971	52.299 → 57.471	43.781 → 49.254	56.477 → 60.622	49.275 → 46.377	42.466 → 39.726	57.143 → 57.143
SE	34.409 → 40.323	33.742 → 34.969	31.618 → 27.941	42.529 → 54.023	35.821 → 49.751	45.078 → 56.995	34.783 → 37.681	31.507 → 31.507	42.857 → 71.429
PE+M	39.247 → 46.774	42.945 → 34.969	47.794 → 40.441	49.425 → 58.621	41.791 → 48.756	52.332 → 59.585	44.928 → 43.478	43.836 → 39.726	57.143 → 71.429
SE+M	31.720 → 44.086	31.288 → 34.969	35.294 → 30.882	41.379 → 51.149	35.323 → 48.756	44.301 → 57.254	33.333 → 34.783	30.137 → 31.507	42.857 → 71.429
EigV	19.355 → 31.183	12.883 → 24.540	5.147 → 18.382	29.885 → 44.253	24.378 → 40.299	37.047 → 52.073	15.942 → 21.739	6.849 → 24.658	42.857 → 42.857
ECC	14.516 → 36.022	9.816 → 26.994	5.882 → 21.324	19.540 → 49.425	14.428 → 41.294	21.503 → 65.026	10.145 → 31.884	6.849 → 21.918	28.571 → 57.143
Deg	20.968 → 26.882	17.178 → 18.405	5.147 → 9.559	29.885 → 37.931	24.378 → 30.846	36.788 → 50.000	13.043 → 15.942	12.329 → 13.699	57.143 → 57.143
Axiom 3: Positively Changed ↓									
PE	82.215 → 91.946	77.346 → 83.982	82.557 → 84.589	73.402 → 76.726	70.256 → 74.103	74.870 → 74.697	82.331 → 86.842	87.572 → 89.484	84.314 → 84.691
SE	93.289 → 93.960	91.533 → 90.847	93.057 → 89.331	86.445 → 88.491	84.615 → 82.821	88.042 → 84.575	93.233 → 94.361	94.073 → 94.073	92.534 → 89.216
PE+M	81.544 → 91.275	77.574 → 84.439	80.271 → 83.065	76.982 → 79.028	73.590 → 75.385	80.069 → 79.029	88.346 → 89.850	90.822 → 91.396	84.389 → 85.143
SE+M	90.604 → 93.289	88.787 → 90.847	88.654 → 87.214	86.957 → 89.258	84.359 → 83.846	88.562 → 85.442	93.609 → 95.113	94.455 → 94.264	93.439 → 90.121
EigV	90.604 → 94.295	88.558 → 93.822	89.077 → 91.871	86.189 → 90.026	86.154 → 90.000	83.709 → 89.081	91.353 → 96.241	92.925 → 96.750	86.652 → 94.646
ECC	82.886 → 89.933	83.066 → 89.931	82.557 → 88.400	79.028 → 87.724	73.590 → 82.308	75.390 → 84.749	86.466 → 93.985	87.380 → 94.837	82.730 → 92.911
Deg	90.604 → 89.933	87.414 → 86.499	89.331 → 89.331	85.934 → 86.701	86.410 → 85.128	85.442 → 83.882	91.353 → 91.729	92.543 → 92.352	86.576 → 86.275
Axiom 4: Negatively Changed ↑									
PE	51.136 → 55.682	51.163 → 51.550	49.231 → 58.462	66.944 → 69.722	66.879 → 69.745	66.372 → 63.717	42.045 → 40.909	38.168 → 39.695	27.586 → 32.184
SE	36.080 → 48.011	36.047 → 46.512	44.615 → 53.846	55.556 → 65.833	54.777 → 67.197	52.212 → 63.717	31.818 → 46.023	29.008 → 41.221	25.287 → 36.782
PE+M	47.727 → 51.420	50.388 → 50.775	50.769 → 61.538	63.333 → 69.167	66.242 → 67.834	64.602 → 65.487	38.636 → 38.068	32.061 → 36.641	26.437 → 31.034
SE+M	38.636 → 50.568	40.698 → 48.450	41.538 → 56.923	55.278 → 62.778	53.503 → 66.879	53.097 → 64.602	31.250 → 43.182	28.244 → 39.695	24.138 → 34.483
EigV	24.432 → 35.227	24.419 → 34.496	16.923 → 32.308	38.333 → 55.278	39.172 → 55.414	38.938 → 53.982	21.591 → 34.091	20.611 → 32.824	8.046 → 17.241
ECC	19.602 → 42.330	18.992 → 39.535	16.923 → 33.846	30.556 → 61.389	30.892 → 58.917	26.549 → 61.062	18.182 → 41.477	18.321 → 35.878	8.046 → 21.839
Deg	25.284 → 29.830	24.806 → 28.295	20.000 → 26.154	42.500 → 49.167	42.357 → 49.363	42.478 → 50.442	22.727 → 26.136	19.084 → 22.137	11.494 → 19.540

Table 13: Changes in the percentage of samples that satisfy the axioms before and after calibration for Llama2-chat. The relation function $\mathcal{R}$ is implemented using MiniCheck.