WHEN SEMANTICALLY EQUIVALENT CONTEXTS YIELD DIFFERENT ANSWERS: MEASURING CONTEXTUAL UNCERTAINTY TO IMPROVE RETRIEVAL-AUGMENTED QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

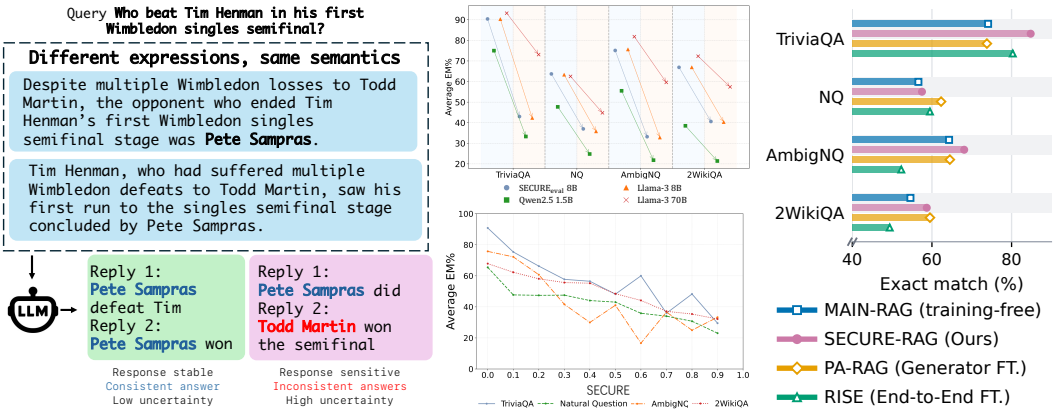


Figure 1: **Contextual uncertainty is a fundamental challenge in retrieval-based QA. Left**: Rephrasing retrieved evidence can change the answer. **Middle**: Higher SECURE scores are associated with lower QA accuracy in the evaluated groups. **Right**: Reported QA accuracy of SECURE-RAG and RAG baselines. Our training-free method achieves competing performance.

ABSTRACT

A fundamental challenge of retrieval-augmented generation (RAG) is that retrieving relevant evidence does not guarantee that an LLM uses it consistently or correctly. Current RAG systems lack a model-agnostic mechanism to identify such failures and use the diagnosis to improve QA performance. We introduce SECURE (SEmantic Context Uncertainty for REtrieval-based QA), a training-free method requiring only text input/output access to the generator. SECURE rephrases one retrieved chunk at a time, then measures the semantic consistency of the resulting answers across these intended meaning-preserving context variants. The resulting score estimates answer risk, while each single-chunk comparison exposes which edit changed the response, without requiring reference answers or model internals. Across four benchmarks, SECURE achieves state-of-the-art performance in uncertainty evaluation, using the same correctness criterion for all methods. We then close the diagnostic-to-improvement loop: SECURE-RAG uses this same signal to trigger adaptive semantic chunking and guide iterative context refinement. On the selected high-uncertainty questions, SECURE-RAG improves reported QA accuracy in 15 of 16 backbone–benchmark combinations. Together, these results show that evidence-rephrasing sensitivity can both rank answer risk and guide targeted RAG refinement in the evaluated settings.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Retrieval-augmented generation (RAG) supplies large language models (LLMs) with external evidence for knowledge-intensive questions. Improving retrieval and context processing Xu et al. (2024); Zhang et al. (2025), or fine-tuning the generator Yu et al. (2024), can improve the resulting answers. Yet retrieving relevant evidence does not ensure that the generator uses it consistently: the same question can receive different answers when the evidence is expressed in different words, as illustrated in Figure 1. This leaves a practical challenge for retrieval-based QA: **how can we detect answer uncertainty, and use that diagnosis to improve the context supplied to the model?**

Uncertainty in LLM generation has been widely studied. Prior work estimates uncertainty from the meanings of multiple generated answers Kuhn et al. (2023); Lin et al. (2024), and input perturbations have also been used to probe model uncertainty Gao et al. (2024). For retrieval-based applications, however, an uncertainty score alone offers limited guidance for improvement: **without a link to specific retrieved evidence, it is unclear which part of the context should be revised.** We therefore develop one diagnostic signal for two connected purposes: estimating answer risk and guiding targeted context refinement.

However, addressing this problem is non-trivial: it requires constructing meaning-preserving interventions, distinguishing genuine context sensitivity from rewrite noise, and converting black-box response changes into chunk-level actions without labeled answers. Therefore, we propose **SECURE** (**SE**mantic **C**ontext **U**ncertainty for **RE**trieval-based QA). SECURE is an **unsupervised, training-free** method that requires no labeled QA examples, reference answers, or updates to model parameters. SECURE generates an answer for the original context and for each single-chunk rewrite. It compares the meanings of these answers and uses their semantic relationships to compute a degree-based uncertainty score. Because each comparison is indexed by the chunk that was rewritten, the same evaluation also records which edits changed the answer. Thus SECURE produces both a question-level measure of response uncertainty and a chunk-indexed signal.

SECURE-RAG uses these two outputs as parts of one unsupervised, training-free process. Rephrasing sensitivity can arise from disrupted within-chunk coherence or unstable interactions across incomplete and distracting chunks; SECURE-RAG therefore targets these two failure modes through adaptive chunking and chunk-guided refinement. A high SECURE score triggers adaptive semantic chunking, which regroups sentences in the source documents for selected questions. The chunk-indexed response changes then guide iterative context refinement: the method tests the effect of removing uncertain chunks, prunes those whose removal leaves the answer semantically unchanged, and supplements the remaining context with related evidence. The score determines when to intervene, the chunk-indexed comparisons guide where to revise, and remeasurement together with the sufficiency check governs when to stop.

We evaluate the diagnostic and the RAG performance on four open-book QA benchmarks. SECURE improves AUROC over seven uncertainty baseline configurations, including SPUQ and SRE, on each dataset (Table 1). On the high-SECURE questions, the SECURE-RAG pipeline improves QA accuracy over the corresponding vanilla pipeline in 15 of 16 backbone–dataset combinations (Table 2). These results support using response sensitivity to evidence rephrasing both to rank answer risk and to guide context revision in the evaluated settings.

Our contributions are two connected parts:

- **A context-sensitive uncertainty metric.** SECURE measures semantic variation across answers to single-chunk evidence rewrites. Its degree-based score summarizes answer instability, while the single-chunk comparisons retain the information needed to locate response-sensitive evidence.
- **An unsupervised metric-guided RAG method.** SECURE-RAG uses SECURE to trigger adaptive chunking and guide iterative context refinement without task-specific fine-tuning. We evaluate the complete diagnostic-to-refinement process across four QA datasets and four generator backbones.

Our implementation is available anonymously.[1]

[1] https://anonymous.4open.science/r/SECURE-RAG-8B6C

2 Related Work

Retrieval-augmented QA systems. Retrieval-based QA combines retrieved evidence with LLM generation Karpukhin et al. (2020); Gao et al. (2023); Li et al. (2024c). Prior work improves retriever–generator alignment Shi et al. (2023); Lin et al. (2023b;a), adds context-processing modules Xu et al. (2024); Zhang et al. (2025), or fine-tunes RAG models Yu et al. (2024); Asai et al. (2023); Wang et al. (2024); Yoran et al. (2023). Knowledge-aware refinement improves retrieved contexts using external knowledge signals Li et al. (2025), while evidence condensation seeks to improve efficiency and reduce hallucinations Xu et al. (2025b). Evaluation frameworks assess RAG quality at the system level Liu et al. (2024a), including reference-free assessment of retrieval and generation Es et al. (2024). SECURE complements these directions by probing whether answer meaning remains stable when retrieved evidence is rephrased, without requiring reference answers or task-specific training.

Uncertainty estimation for reliable LLM generation. Semantic and graph-based estimators assess answer variability without reference labels Kuhn et al. (2023); Farquhar et al. (2024); Hou et al. (2024); Jiang et al. (2024); Lin et al. (2024), while supervised approaches learn reliability signals Kweon et al. (2025); Liu et al. (2024b); Arteaga et al. (2025). RAG-specific work studies retrieval utility, component-level confidence, passage utility, conformal answer sets, and contextual representation gaps Dai et al. (2025); Hasegawa et al. (2024); Perez-Beltrachini & Lapata (2025); Li et al. (2024a); Bakman et al. (2025). SPUQ perturbs inputs and samples outputs Gao et al. (2024); SRE reformulates questions Tong et al. (2025); Ensemble-of-Ensembles combines prompt perturbations with repeated sampling Ma et al. (2026). SECURE rephrases retrieved evidence one chunk at a time, producing a question-level uncertainty score and a chunk-indexed signal. SECURE-RAG uses this same signal to trigger re-chunking and guide context refinement, linking uncertainty estimation to context improvement.

3 Preliminaries

Given a question q, a retriever selects the top-k documents to form context $C = [c_1, c_2, ..., c_k]$, and an LLM f_θ generates response r:

$$r = f_\theta(q, C), \tag{1}$$

Following the distinction between meaning and surface form Kuhn et al. (2023), we formulate two assumptions about the semantics of retrieved context.

Assumption 1 (Semantic Projection). Given a question q and a retrieved context C, there exists a semantic space $\mathcal{S}$ and a semantic projection function:

$$\pi : \mathcal{Q} \times \mathcal{C} \to \mathcal{S}, \tag{2}$$

which maps the question-context pair (q, C) to its underlying semantic meaning $\pi(q, C) \in \mathcal{S}$.

Assumption 2 (Question-Conditioned Semantic Equivalence). For a fixed question q, two contexts C and C' are said to be *semantically equivalent with respect to* q if

$$\pi(q, C) = \pi(q, C'). \tag{3}$$

Under these assumptions, *semantic consistency* Rabinovich et al. (2023) requires answers to depend on contextual meaning rather than surface form:

Property 1 (Semantic Consistency in Retrieval-based QA). If an LLM f_θ reliably uses the retrieved context C to answer question q, then for any context C' such that

$$\pi(q, C) = \pi(q, C'),$$

the generated responses should be semantically equivalent:

$$\pi(f_\theta(q, C)) = \pi(f_\theta(q, C'))\,.$$

Property 1 describes idealized semantic consistency. Its violations (Figure 1, Left) motivate measuring context-induced response sensitivity through semantic variation in generated responses.

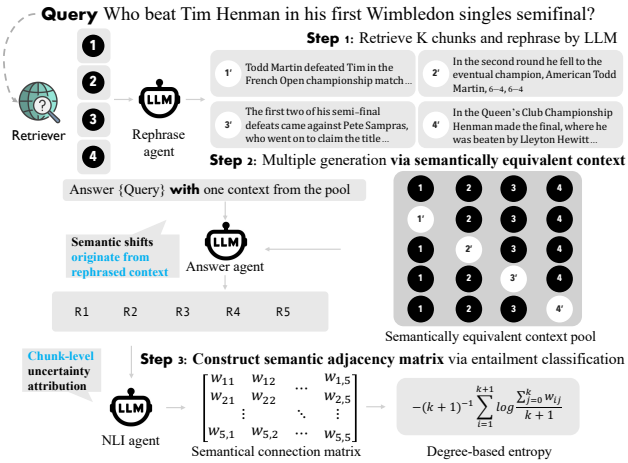


Figure 2: SECURE rephrases one retrieved chunk at a time, generates an answer for each context variant, and scores answer consistency. Each single-chunk comparison is indexed by the edited chunk; interpreting a response change assumes that the rewrite preserves question-relevant facts.

4 Evaluating Context-Induced Response Sensitivity

Semantic uncertainty commonly compares sampled answers under a fixed input (Kuhn et al., 2023; Lin et al., 2024), while input perturbations have also been used for uncertainty estimation (Gao et al., 2024). SECURE instead varies one retrieved chunk at a time, keeping the question and remaining evidence fixed, so that response sensitivity is indexed by the edited chunk. We introduce **SECURE** (**SE**mantic **C**ontext **U**ncertainty for **RE**trieval-based QA) to turn these chunk-indexed response changes into an uncertainty score and a signal for context refinement. SECURE (i) rephrases each retrieved chunk with the aim of preserving its meaning, (ii) generates an answer for each single-chunk variant with greedy decoding, and (iii) measures answer consistency with a degree-based score.

4.1 Semantically Equivalent Context Rephrasing

We construct context variants that differ in the wording of one retrieved chunk. Each single-chunk edit lets us associate a change in the answer with the edited chunk; this association depends on the rewrite retaining the question-relevant facts. For a retrieved context $C_0 = [c_1, \ldots, c_k]$, we obtain a rephrased c_i' for each chunk c_i:

$$c_i' = f_\theta(p_r(c_i)) \tag{4}$$

The rewriter receives only c_i and the meaning-preserving prompt p_r (Appendix A.1), without the question or other chunks. We then construct k single-edit contexts by replacing exactly one chunk at a time:

$$C_i[j] = \begin{cases} c_i', & \text{if } j = i \\ c_j, & \text{otherwise} \end{cases} \qquad \text{for } j = 1, \ldots, k \tag{5}$$

This yields $k{+}1$ contexts: the original C_0 and k intended meaning-preserving variants $C_1, \ldots, C_k$. Appendix C evaluates entailment in both directions between each original chunk and its rewrite. The LLM judge reports bidirectional entailment rates of 93.24–98.37%. These automatic checks support semantic preservation but do not guarantee preservation of every question-relevant fact.

Greedy decoding removes answer-sampling variation under a fixed input in our setup; variation across edited inputs may also reflect imperfect rewrites or judging. The fixed-context check is in Appendix B. Generating from $C_0, \ldots, C_k$ yields $R = \{r_0, \ldots, r_k\}$.

4.2 DEGREE-BASED SEMANTIC UNCERTAINTY

To retain response-level information associated with each edited chunk, we use Degree-based Semantic Entropy. Whereas semantic entropy clusters answers into semantic groups Kuhn et al. (2023), our formulation treats each response in R as a graph node and constructs a semantic adjacency matrix W from pairwise entailment scores Jiang et al. (2024); Lin et al. (2024):

$$w_{ij} = ((NLI(r_i, r_j) + NLI(r_j, r_i))/2)_{i,j\in[0,\ldots,k]}, \tag{6}$$

where a natural language inference (NLI) model classifies the relation between r_i and r_j as *entailment (1)* or *neutral (0)*; w_{ij} represents the link between the two responses. In SECURE, an LLM performs this classification using the prompt in Appendix A.2. After constructing W, we compute the degree-based semantic uncertainty as follows:

$$DSE(q, C) = -(k+1)^{-1} \sum_{i=0}^{k} \log \frac{D_i}{k+1}, \tag{7}$$

where $D_i = \sum_{j=0}^{k} w_{ij}$ is the weighted degree of response r_i. Higher degree indicates stronger aggregate semantic connections, and $D_i/(k+1)$ is its average pairwise score across all responses, including r_i itself. DSE averages negative log connectivity rates and equals empirical semantic entropy under the equivalence-class condition in Appendix D. Pseudocode appears in Appendix E.1.

4.3 SECURE: DISCRIMINATING CONTEXT-INDUCED SEMANTIC UNCERTAINTY

We next test whether SECURE's response-sensitivity score ranks answers by correctness. Property 1 motivates the expectation that reliable use of meaning-preserving context variants should yield semantically consistent responses; prior work further associates semantic consistency with answer correctness Kuhn et al. (2023). Concretely, we implement a RAG system on Llama3-8B with SECURE and conduct experiments on four open-book QA datasets. The details of dataset and backbone are discussed in Section 6.1 and Appendix F. We examine the relationship between SECURE and QA performance by grouping questions according to SECURE values. As shown in Figure 1(Middle-Bottom), average exact match (EM) consistently decreases as SECURE increases across all four datasets. This inverse relationship motivates using SECURE as a reference-free ranking signal for answer risk in the evaluated setting.

To test whether this degradation persists across model scales, we split questions into relatively *certain* and *uncertain* subsets using a fixed SECURE threshold (set to 0.2 for illustration). We then evaluate three LLMs of different scales on both subsets, as summarized in Figure 1(Middle-Top). The comparison between $\text{SECURE}_{\text{eval}}$ 8B and Llama-3 8B provides a check that the evaluation procedure itself does not substantially change aggregate QA accuracy. Across language models of different scales, QA performance consistently degrades on the uncertain subset, showing that the high-SECURE/low-accuracy association persists across the tested model capacities. A comprehensive quantitative comparison with existing uncertainty estimation methods is provided in Section 6.

5 IMPROVING CONTEXT QUALITY WITH SECURE

Section 4 shows that questions with high SECURE scores have lower QA accuracy in the evaluated RAG setting. We therefore reserve extra processing for high-uncertainty questions rather than revising every context. We use a fixed threshold to separate relatively *certain* and *uncertain* cases (SECURE ≤ 0.2 vs. SECURE > 0.2). For these uncertain cases, SECURE-RAG uses two training-free modules: adaptive semantic chunking changes how evidence is grouped, and iterative refinement prunes and supplements retrieved chunks.

5.1 ADAPTIVE SEMANTIC CHUNKING

Chunking strategies in retrieval-based QA face a fundamental trade-off: fixed-size chunking is efficient but often splits contextually related sentences, disrupting semantic coherence Gao et al. (2023); Finardi et al. (2024); semantic chunking, which groups sentences by embedding similarity, can

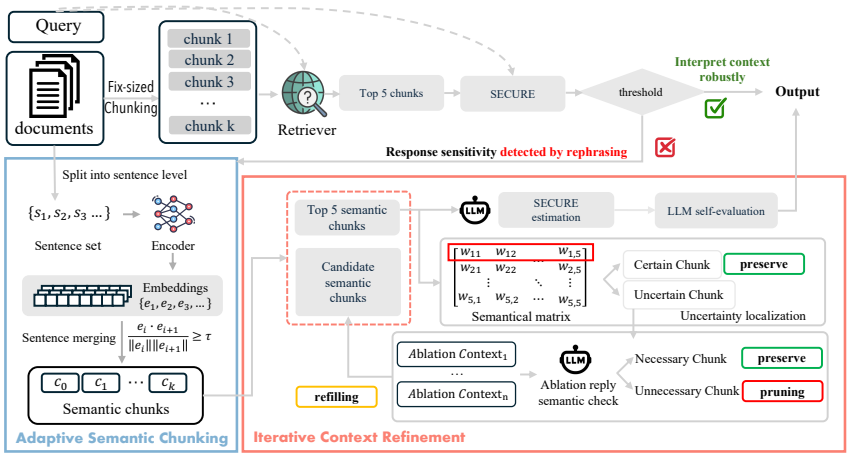


Figure 3: SECURE-RAG uses the uncertainty score to trigger adaptive semantic chunking and the chunk-indexed response signal to guide iterative context refinement.

better preserve semantics, but its computational cost frequently outweighs its performance gains Qu et al. (2024).

To address this trade-off, SECURE-RAG applies semantic chunking only to high-uncertainty questions, which are more likely to be sensitive to disrupted chunk coherence. As illustrated in Figure 3, we first apply fixed-size chunking, retrieve the top-k chunks, and compute SECURE. If q is certain, we return the original response; otherwise, we split the documents into sentences $d = \{s_1, s_2, \ldots\}$, encode each s_i as e_i, and merge s_{i+1} into chunk c_j if

$$\frac{e_i \cdot e_{i+1}}{\|e_i\|\|e_{i+1}\|} \geq \tau \quad \text{and} \quad |c_j \cup s_{i+1}| \leq T_{max}, \tag{8}$$

where τ is the similarity threshold and T_{max} is the maximum chunk length; otherwise, s_{i+1} starts c_{j+1}. This operation only regroups the original documents and adds no evidence. Pseudocode appears in Appendix E.2.

5.2 Iterative Context Refinement

To address incomplete or distracting interactions across chunks, iterative context refinement uses SECURE to label retrieved chunks as *certain*, *necessary*, or *unnecessary*, then retains certain chunks, removes unnecessary ones, and supplements the context with evidence guided by necessary chunks.

Localize the source of uncertainty. Since SECURE computes semantic adjacency W, and each modified context C_i differs from the original C_0 in exactly one rephrased chunk c_i', we can evaluate the impact of each c_i and classify it as *certain* or *uncertain*:

$$l_i^{ce} = \mathbb{1}_{\{w_{i,0}=1\}}, \tag{9}$$

where $\mathbb{1}$ is the indicator. We call c_i *certain* when $l_i^{ce} = 1$ and *uncertain* otherwise. These are operational response labels, not proof of reliable context use: an answer change may reflect either model error or a meaning-changing rewrite. An uncertain chunk may be relevant but insufficient, or weakly relevant and distracting. We distinguish these cases by masking each uncertain chunk:

$$ar_i = f_\theta(q, [c_1, ..., c_{i-1}, c_{i+1}, ..., c_k]), \tag{10}$$

where ar_i denotes the response when chunk c_i is absent. Then we employ the entailment in Equation 6 to evaluate whether chunk c_i is necessary:

$$l_i^{ne} = \mathbb{1}_{\{[NLI(r_0, ar_i) + NLI(ar_i, r_0)]/2 = 0\}}, \tag{11}$$

where r_0 is the response under the original context. For an uncertain chunk, a changed answer after removal receives the *necessary* label and an unchanged answer the *unnecessary* label. These labels capture response sensitivity, not independently verified relevance. Thus each c_i is *certain* ($l_i^{ce} = 1$), *necessary* ($l_i^{ce} = 0, l_i^{ne} = 1$), or *unnecessary* ($l_i^{ce} = 0, l_i^{ne} = 0$).

Table 1: Uncertainty evaluation performance on four QA benchmarks. All methods use the same correctness criterion.

Uncertainty Measurements	TriviaQA		Natural Question		AmbigNQ		2WikiQA	
	AUROC↑	AUARC↑	AUROC↑	AUARC↑	AUROC↑	AUARC↑	AUROC↑	AUARC ↑
Discrete Semantic Entropy Farquhar et al. (2024)	64.07	88.72	61.33	72.29	60.40	75.94	57.28	60.80
U_{eigv} Lin et al. (2024)	66.64	89.05	64.16	73.08	62.90	76.01	57.37	60.43
U_{deg} Lin et al. (2024)	66.56	89.03	64.05	73.69	62.74	76.79	57.25	60.82
KLE heat Nikitin et al. (2024)	66.40	89.00	63.93	<u>73.89</u>	62.53	75.84	56.99	60.36
KLE deberta matern Nikitin et al. (2024)	63.97	89.11	61.98	72.51	62.33	<u>77.56</u>	56.78	61.05
SPUQ Gao et al. (2024)	<u>71.81</u>	<u>91.21</u>	<u>64.74</u>	67.53	65.21	77.05	<u>64.36</u>	64.04
SRE Tong et al. (2025)	69.82	89.83	59.55	60.48	<u>68.27</u>	76.55	63.87	<u>65.50</u>
SECURE*(ours)*	**75.63**	**92.52**	**67.96**	**75.10**	**69.94**	**79.54**	**65.74**	**66.38**

Iterative refinement. After categorizing all chunks, we refine the context with two steps: (i) pruning, which removes *unnecessary chunks*, and (ii) refilling, which adds new chunks most similar to the *necessary chunks* based on cosine similarity of embeddings. After each update, we recompute SECURE and stop only when DSE is below 0.2 and the context-sufficiency self-evaluation passes, matching Algorithm 3; neither condition alone certifies correctness. Otherwise, we return r_0 from the evaluated context with the lowest SECURE score. The prompt and pseudocode appear in Appendix A.3 and E.3.

6 Experiments

6.1 Experiment Setup

Datasets. We test SECURE and SECURE-based RAG modules on open-book QA datasets, which require free-form answers: TriviaQA Joshi et al. (2017), Natural Question Kwiatkowski et al. (2019), AmbigNQ Min et al. (2020) and 2WikiQA Ho et al. (2020). The first three are single-hop QA datasets, while 2WikiQA is a multi-hop QA dataset. We use Exact Match (EM) as the metric to evaluate QA performance. The detailed information is in Appendix F.

Implementation Details. We conduct experiments on Qwen-2.5 1.5B, Qwen-3 8B, Llama-3 8B, and Llama-3 70B, using the documents provided by each dataset as the retrieval corpus. A vanilla retrieval-based pipeline is built with recursive chunking (chunk size = 512, overlap size = 64) as the default strategy. For dense retrieval, we adopt UAE-Large-V1 as the encoder for both questions and documents, and use FAISS for indexing. Unless otherwise specified, the top-5 retrieved chunks are used as context in all experiments. We also conduct experiments with different numbers of chunks in Appendix J. Details of each component and the QA prompt are provided in Appendix A.4.

Following prior work Kuhn et al. (2023); Lin et al. (2024), we use the Area Under the Receiver Operating Characteristic curve (**AUROC**) and the Area Under the Accuracy-Rejection Curve (**AUARC**) Nadeem et al. (2009) to evaluate uncertainty. AUARC measures how well uncertainty estimation prioritizes the rejection of incorrect predictions, with higher values indicating that accuracy improves more rapidly as high-uncertainty samples are removed. For uncertainty ranking, we follow the ROUGE-L > 0.3 fuzzy correctness criterion used by semantic-uncertainty evaluation Kuhn et al. (2023), applying the same criterion to SECURE and all baselines. For QA performance, we use exact match (EM); further details appear in Appendix G.

6.2 Evaluation of Context-Induced Uncertainty

We compare SECURE with seven baseline configurations (Table 1): five semantic/graph-based sampling baselines, SPUQ Gao et al. (2024), and SRE Tong et al. (2025). SECURE uses greedy decoding to remove answer-sampling variation; for comparison with sampling-based methods, we evaluate at $T = 0.25$. It achieves the highest AUROC and AUARC on all four datasets. Relative to the strongest result in each column, SECURE improves AUROC by 3.82, 3.22, 1.67, and 1.38 points and AUARC by 1.31, 1.21, 1.98, and 0.88 points on TriviaQA, Natural Question, AmbigNQ, and 2WikiQA, respectively. These gains extend across both single-hop and multi-hop QA. High-temperature and additional evaluations, including summarization, appear in Appendices H.1 and H.

Table 2: QA accuracy on selected high-SECURE questions across four datasets and four backbones. The full chunking-plus-refinement pipeline improves over the corresponding vanilla pipeline in 15 of 16 cells; individual module effects vary. Chunking and Refinement denote the proposed adaptive semantic chunking and iterative context refinement modules.

Backbone	SECURE Component		TriviaQA	Natural Question	AmbigNQ	2WikiQA
	Chunking	Refinement				
Qwen-2.5 1.5B			33.28	24.79	21.82	**21.39**
	✓		35.27	27.00	24.09	21.19
	✓	✓	**36.36**	**27.08**	**24.55**	21.33
Qwen-3 8B			46.51	35.35	33.18	42.03
	✓		49.42	37.73	36.36	42.68
	✓	✓	**49.83**	**38.25**	**37.73**	**44.35**
Llama-3 8B			42.60	35.60	32.27	40.57
	✓		51.33	38.93	40.00	41.73
	✓	✓	**53.91**	**40.54**	**41.36**	**45.78**
Llama-3 70B			73.12	44.80	59.55	57.37
	✓		74.63	45.49	**61.36**	59.85
	✓	✓	**74.87**	**46.93**	60.00	**61.20**

Table 3: Results of SECURE-RAG and baselines. Ret. FT and Gen. FT indicate whether the retriever and generator of the method were fine-tuned, respectively. Results unavailable in public reports are marked as "–". **SECURE-RAG achieves state-of-the-art performance among methods without LLM fine-tuning.**

Method	Generator Model	Ret. FT	Gen. FT	TriviaQA	Natural Question	AmbigNQ	2WikiQA
ChatQA-1.5Liu et al. (2024c)	Llama3 8B	✓	✓	81.0	42.4	<u>65.5</u>	26.8
UncertaintyRAGLi et al. (2024d)	Llama2 13B	✓	✗	82.5	-	-	38.3
ERM4Shi et al. (2024)	GPT-3.5-turbo	✓	✗	-	52.7	53.5	46.8
Astute-RAGWang et al. (2025)	Claude 3.5 Sonnet	✗	✗	<u>84.5</u>	53.6	-	-
RankRAGYu et al. (2024)	Llama3 8B	✗	✓	82.9	50.6	-	31.4
MAIN-RAGChang et al. (2025)	Llama3 8B	✗	✗	74.1	56.6	64.3	54.6
PA-RAGWu et al. (2025)	Llama3 8B	✗	✓	73.8	**62.3**	64.5	**59.5**
RISEHe et al. (2025)	Llama3.1 8B	✓	✓	80.3	<u>59.5</u>	52.3	49.4
SECURE-RAG(*ours*)	Llama3 8B	✗	✗	**84.8**	57.5	**68.1**	<u>58.7</u>

6.3 Improving QA with SECURE Guidance

SECURE-RAG is effective on uncertain questions. Table 2 progressively adds adaptive semantic chunking and iterative context refinement. The full pipeline improves the reported score over the vanilla pipeline in 15 of 16 selected-subset comparisons; the Qwen-2.5 1.5B–2WikiQA cell is slightly lower (21.33 versus 21.39). The marginal effect of each module is not positive in every cell. For Llama-3 8B, it improves EM by 11.31, 4.94, 9.09, and 5.21 points across the four datasets; gains also persist for Qwen-3 8B and Llama-3 70B. The larger TriviaQA and AmbigNQ gains for Llama-3 8B suggest that context adaptation is especially useful when the vanilla pipeline is sensitive to retrieved evidence. Two cells qualify component-wise claims: the full pipeline is slightly below vanilla for Qwen-2.5 1.5B on 2WikiQA (21.33 versus 21.39), and on Llama-3 70B–AmbigNQ, chunking alone reaches 61.36 while the full pipeline reaches 60.00, still above vanilla at 59.55. Among baselines without LLM fine-tuning, SECURE-RAG achieves state-of-the-art overall performance. On the complete evaluation sets, its 8B model obtains EM scores of 84.8, 57.5, 68.1, and 58.7, ranking first in all four columns among the compared methods whose generators are not fine-tuned. This complements the high-SECURE subset results by showing gains on complete evaluation sets without parameter updates. Additional model scales and baseline comparisons appear in Appendix I.

SECURE-RAG uses text input/output access to the generator. For one uncertainty-evaluation pass with k retrieved chunks, pairwise semantic judging can require $O((k+1)^2)$ calls; the complete refinement cost also depends on the number of iterations and candidate chunks. Further complexity details appear in Appendix K.

6.4 Ablation study

Analysis of chunking strategies. Table 4 compares fixed-size recursive chunking at 256/32 and 128/16 with semantic chunking applied to all questions.

Table 4: Ablation results on certain/uncertain questions with different chunking strategies.

Chunking	TriviaQA	NQ	AmbigNQ	2WikiQA
512/64	**90.4**/43.0	**63.7**/36.0	**75.1**/33.2	**66.9**/40.6
256/32	88.3/47.7	57.6/33.6	68.3/33.2	61.1/41.7
128/16	87.7/48.9	53.6/33.2	53.6/31.7	58.8/41.4
Full semantic	87.6/**53.0**	59.3/**41.3**	68.2/39.5	65.9/41.5
Adaptive (ours)	90.3/52.4	63.4/**41.3**	75.0/**40.5**	**66.9**/**42.3**

Table 5: Ablation results for iterative context refinement on uncertain questions.

Refinement	TriviaQA	NQ	AmbigNQ	2WikiQA
w/o refine	52.4	41.3	40.5	42.3
only removing chunks	54.5	40.6	39.5	43.3
w/o self-evaluation	52.7	41.5	38.2	42.1
refill on certain chunks	55.7	41.8	**43.6**	44.6
Context-refiner	**56.3**	**42.2**	**43.6**	**46.6**

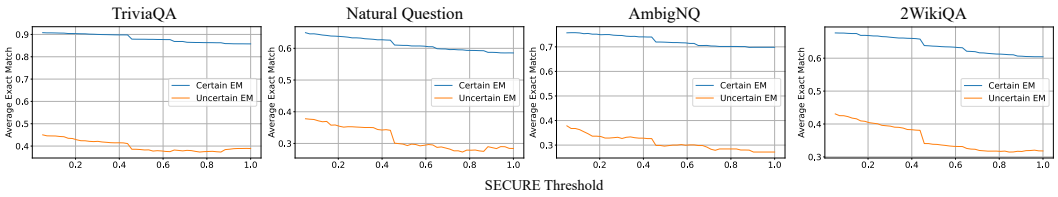


Figure 4: Average exact match on certain vs. uncertain questions under different SECURE thresholds, showing a consistent performance gap between the two groups.

Certain questions benefit from larger chunks. In Table 4, smaller fixed-size chunks lower QA accuracy in the low-SECURE group but improve it in the high-SECURE group. This subgroup pattern motivates conditional chunking; the table does not isolate whether context length, relevance, or sentence boundaries drive the difference.

Uncertain questions benefit from semantic chunking. Semantic chunking improves uncertain questions but degrades certain ones, consistent with prior findings Qu et al. (2024); for already well-answered questions, it restricts information diversity. Our adaptive method therefore applies it only under high SECURE, improving uncertain questions while preserving fixed-size chunking for certain ones and helping explain mixed findings in prior work.

Analysis of iterative context refinement. Table 5 compares no refinement, pruning only, refinement without self-evaluation, and refilling from certain rather than necessary chunks. Pruning outperforms no refinement on TriviaQA and 2WikiQA, with only about a one-point drop on Natural Questions and AmbigNQ. This pattern is consistent with removing distracting context. Disabling self-evaluation lowers all four scores, supporting the gate. Refilling from certain chunks is second-best, but remains weaker than guidance from necessary chunks.

6.5 Hyperparameter Sensitivity Analysis

SECURE threshold. In Figure 4, the certain group consistently outperforms the uncertain group across all tested thresholds and datasets, so the SECURE–EM relationship is not tied to one cutoff. We use 0.2 by default, following Figure 1(Middle-Bottom). Lower thresholds route more questions through intervention; higher thresholds are more selective. Retrieval-scope and chunking-threshold analyses and case studies appear in Appendices J and L.

7 Conclusions

SECURE is a training-free, text-only diagnostic that estimates answer risk from sensitivity to meaning-preserving context rewrites. Higher scores are associated with lower QA accuracy across four benchmarks. SECURE-RAG uses the same signal for adaptive semantic chunking and iterative context refinement, achieving state-of-the-art overall performance among methods without LLM fine-tuning. Future work could reduce inference costs with smaller models and improve context interpretation through uncertainty-aware training.

AI USE STATEMENT

During manuscript revision, we used AI for language editing, literature search, interpretation of reported results, feedback on experimental design, checking and revising algorithm pseudocode, and preparing LaTeX and plotting scripts from existing results. The LLM-based rephrasing, semantic judging, and QA components of the research method are described in Sections 4 and 5 and Appendix G. The authors take responsibility for the final content, including AI-assisted text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 define SECURE and SECURE-RAG, with pseudocode in Algorithms 1–3. Appendices A.1–A.4 provide the prompts. Dataset splits, retrieval settings, model backbones, and hardware are described in Section 6.1 and Appendices F and G. Threshold sensitivity is reported in Section 6.5 and Appendix J; Appendix D states the assumptions and derivation relating DSE to empirical semantic entropy. The implementation is provided in our anonymous SECURE-RAG repository.[2]

ETHICS STATEMENT

This work uses only publicly available question-answering benchmarks and pretrained language models. It does not involve human participants, private or personally identifiable information, or deployment in high-stakes settings. We identify no ethical risks specific to the proposed uncertainty-estimation and context-refinement methods.

REFERENCES

Gabriel Y Arteaga, Thomas B Schön, and Nicolas Pielawski. Hallucination detection in llms: Fast and memory-efficient finetuned models. In *Northern Lights Deep Learning Conference*, pp. 1–15. PMLR, 2025.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

Yavuz Bakman, Sungmin Kang, Zhiqi Huang, Duygu Nur Yaldiz, Catarina G Belém, Chenyang Zhu, Anoop Kumar, Alfy Samuel, Salman Avestimehr, Daben Liu, et al. Uncertainty as feature gaps: Epistemic uncertainty quantification of llms in contextual question-answering. *arXiv preprint arXiv:2510.02671*, 2025.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W. Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. *arXiv preprint arXiv:2211.12717*, 2022.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, et al. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2607–2622, 2025.

Lu Dai, Yijie Xu, Jinhui Ye, Hao Liu, and Hui Xiong. Seper: Measure retrieval utility through the lens of semantic perplexity reduction. *arXiv preprint arXiv:2503.01478*, 2025.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[2] https://anonymous.4open.science/r/SECURE-RAG-8B6C

Paulo Finardi, Leonardo Avila, Rodrigo Castaldoni, Pedro Gengo, Celio Larcher, Marcos Piau, Pablo Costa, and Vinicius Caridá. The chronicles of rag: The retriever, the chunk and the generator. *arXiv preprint arXiv:2401.07883*, 2024.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2336–2346, 2024. doi: 10.18653/v1/2024.eacl-long.143. URL https://aclanthology.org/2024.eacl-long.143/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

Kento Hasegawa, Seira Hidano, and Kazuhide Fukushima. Rag certainty: Quantifying the certainty of context-based responses by llms. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 912–917. IEEE, 2024.

Bolei He, Xinran He, Mengke Chen, Xianwei Xue, Ying Zhu, and Zhen-Hua Ling. Rise: reasoning enhancement via iterative self-exploration in multi-hop question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14925–14948, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19023–19042, 2024.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong-Cheol Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050. Association for Computational Linguistics, 2024.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori Hashimoto. Graph-based uncertainty metrics for long-form language model generations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 32980–33006. Curran Associates, Inc., 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pp. 6769–6781, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Wonbin Kweon, Sanghwan Jang, SeongKu Kang, and Hwanjo Yu. Uncertainty quantification and decomposition for llm-based recommendation. In *Proceedings of the ACM on Web Conference 2025*, pp. 4889–4901, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. Traq: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3799–3821, 2024a.

Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, et al. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. *arXiv preprint arXiv:2410.13509*, 2024b.

Yongjian Li, HaoCheng Chu, Yukun Yan, Zhenghao Liu, Shi Yu, Zheni Zeng, Ruobing Wang, Sen Song, Zhiyuan Liu, and Maosong Sun. Kare-rag: Knowledge-aware refinement and enhancement for rag. *arXiv preprint arXiv:2506.02503*, 2025.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024c.

Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, et al. Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. *arXiv preprint arXiv:2410.02719*, 2024d.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6385–6400, 2023a.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023b.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Jintao Liu, Ruixue Ding, Linhao Zhang, Pengjun Xie, and Fie Huang. Cofe-rag: A comprehensive full-chain evaluation framework for retrieval-augmented generation with enhanced data diversity. *arXiv preprint arXiv:2410.12248*, 2024a.

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

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Advances in Neural Information Processing Systems*, 37:15416–15459, 2024c.

Wang Ma, Debarun Bhattacharjya, Junkyu Lee, Nhan Pham, Harsha Kokel, and Qiang Ji. Black-box uncertainty quantification for large language models via ensemble-of-ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL `https://openreview.net/pdf?id=jfq9DUrW90`.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*, pp. 65–81. PMLR, 2009.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Laura Perez-Beltrachini and Mirella Lapata. Uncertainty quantification in retrieval augmented question answering. *arXiv preprint arXiv:2502.18108*, 2025.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.

Renyi Qu, Ruixuan Tu, and Forrest Bao. Is semantic chunking worth the computational cost? *arXiv preprint arXiv:2410.13070*, 2024.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. Predicting question-answering performance of large language models through semantic consistency. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

Y Shi, X Zi, Z Shi, H Zhang, Q Wu, and M Xu. Enhancing retrieval and managing retrieval: A four-module synergy for improved quality and efficiency in rag systems. *ECAI 2024*, 2024.

Chaodong Tong, Qi Zhang, Lei Jiang, Yanbing Liu, Nannan Sun, and Wei Li. Semantic reformulation entropy for robust hallucination detection in QA tasks, 2025. URL `https://arxiv.org/abs/2509.17445`.

V Venktesh, Mandeep Rathee, and Avishek Anand. Sunar: Semantic uncertainty based neighborhood aware retrieval for complex qa. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5818–5835, 2025.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. Instructretro: instruction tuning post retrieval-augmented pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 51255–51272, 2024.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2025.

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. Pa-rag: Rag alignment via multi-perspective preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9091–9112, 2025.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024.

Ran Xu, Wenqi Shi, Yuchen Zhuang, Yue Yu, Joyce C Ho, Haoyu Wang, and Carl Yang. Collab-rag: Boosting retrieval-augmented generation for complex question answering via white-box and black-box llm collaboration. *arXiv preprint arXiv:2504.04915*, 2025a.

Zhichao Xu, Minheng Wang, Yawei Wang, Wenqian Ye, Yuntao Du, Yunpu Ma, and Yijun Tian. Recon: reasoning with condensation for efficient retrieval-augmented generation. *arXiv preprint arXiv:2510.10448*, 2025b.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184, 2024.

Jintao Zhang, Guoliang Li, and Jinyang Su. Sage: A framework of precise retrieval for rag. *arXiv preprint arXiv:2503.01713*, 2025.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, 2024.

APPENDIX

A PROMPT FORMATS

A.1 REPHRASING PROMPT

The format template of LLM inputs in building semantically equivalent contexts as follows:

```
User: Rewrite the following text at the syntactic level without
changing its meaning. Modify the sentence structure, but preserve
the original intent and semantic meaning. ONLY return the
rewritten content without any additional token. Text: {chunk}
```

A.2 NLI PROMPT

The format template of LLM inputs in evaluating semantic entailment as follows:

```
User: We are evaluating answers to the question {question} Here
are two possible answers:

Possible Answer 1: {response1}

Possible Answer 2: {response2}

Does Possible Answer 1 semantically entail Possible Answer 2?
Respond with ONLY entailment, contradiction, or neutral.
```

A.3 SELF-EVALUATION PROMPT

The format template of LLM inputs in performing self-evaluation on context as follows:

```
User: Context: {context}

Question: {question}

Does the context contain enough information to answer the
question? Only answer yes or no.
```

A.4 QA PROMPT

The format template of LLM inputs in asking questions as follows:

```
User: Answer question {query} based on provided context, ONLY
output a short answer with minimum words.

Context:{context}
```

B DISENTANGLEMENT OF INHERENT LLM RANDOMNESS

In previous works, uncertainty related metrics are employed under high temperature settings Kuhn et al. (2023); Lin et al. (2024). These works focus on quantifying the intrinsic stochasticity of LLMs as well as their hallucination behavior during question answering. In contrast, we adopt greedy decoding to disentangle the influence of LLM-intrinsic randomness on response variability. As a fixed-input check, we run Llama3-8B with greedy decoding on TriviaQA, NQ, AmbigNQ, and 2WikiQA. We then compute discrete semantic entropy using the official implementation provided by Kuhn et al. (2023). In all datasets, the measured uncertainty is exactly zero, confirming that our greedy decoding setup eliminates randomness-induced variability under a fixed input. The remaining response variability measured by SECURE arises across rephrased contexts. Lin et al. (2024) explored the relationship between decoding temperature and uncertainty estimation.

Table 6: Bidirectional entailment between original and rephrased chunks. Ratios count pairs judged as entailing in both directions. We sample 1000 questions per dataset and 10 chunks from each question's retrieved documents.

Entailment methods	Bidirectional entailment ratio			
	TriviaQA	Natural Question	AmbigNQ	2WikiQA
LLM-as-a-judge	98.37%	95.95%	97.88%	93.24%
Deberta-as-a-judge	97.65%	98.28%	96.35%	98.40%

C Rephrased contexts semantic checking

To evaluate whether SECURE's LLM-based rephrasing preserves the semantics of the original chunks, we check rewritten chunks on the four QA datasets used in the main experiments. We randomly sample 1,000 questions from each dataset and 10 chunks from the retrieved documents for each question. We apply the LLM judge and DeBERTa-large-mnli separately to evaluate *both* entailment directions for each original/rephrased pair: $c_i \Rightarrow c_i'$ and $c_i' \Rightarrow c_i$. A pair passes the bidirectional check only when both directions are judged as entailment. Table 6 reports the corresponding pass rates: 93.24–98.37% for the LLM judge and 96.35–98.40% for DeBERTa.

The two directions address complementary risks. If $c_i \Rightarrow c_i'$ holds, the rewrite introduces no factual information beyond what the source entails; conversely, $c_i' \Rightarrow c_i$ rules out loss of information asserted by the source. Thus, true mutual entailment establishes semantic equivalence and excludes unsupported factual additions, even when the wording changes. This implication relies on correct entailment judgments: the reported automatic pass rates provide empirical evidence rather than a guarantee for every rewritten chunk.

The rephrasing model sees only one source chunk and the instruction in Appendix A.1; it receives neither the question nor other retrieved chunks. The prompt asks for syntactic changes while preserving meaning, so the intended transformation cannot draw on other chunks to add facts. This input restriction limits cross-chunk information mixing, while the bidirectional check assesses whether the rewrite preserves the source meaning. The high pass rates provide evidence of semantic preservation on the sampled pairs; they do not by themselves prove that rephrasing has a negligible effect on QA accuracy.

D Discussion of Semantic Entropy and SECURE

Semantic entropy. Semantic entropy is the Shannon entropy of the distribution over answer meanings Kuhn et al. (2023); Farquhar et al. (2024):

$$SE(q, C) = -\sum_{h \in H} p_h \log p_h, \qquad p_h = p(h \mid q, C), \tag{12}$$

where H contains semantic classes and $\sum_h p_h = 1$. For a finite collection of $N = k + 1$ responses, the empirical version uses $\widehat{p}_h = n_h/N$, where n_h counts responses in class h. Repeated answers remain separate observations. As in the main text, the pairwise semantic score is

$$E(r_i, r_j) = \frac{NLI(r_i, r_j) + NLI(r_j, r_i)}{2}, \tag{13}$$

with $w_{ij} = E(r_i, r_j)$. Mutual entailment is operational evidence of semantic equivalence, but automatic pairwise decisions need not be transitive.

Degree-based interpretation. SECURE uses $D_i = \sum_{j=0}^{N-1} w_{ij}$ and

$$DSE(q, C) = -\frac{1}{N} \sum_{i=0}^{N-1} \log \frac{D_i}{N}.$$

Thus, DSE is the average negative logarithm of normalized row degrees, or local connectivity rates. The quantities D_i/N generally do not sum to one over i, so they are not a categorical probability

Algorithm 1: SECURE

Input: Chunk set $\{c_1, c_2, c_3, \ldots, c_k\}$, question q, LLM f_θ, rephrase prompt p_r, QA prompt p_q, NLI agent
Output: degree-based uncertainty DSE, indexed replies R, semantic matrix W
$C_0 \leftarrow [c_1, \ldots, c_k]$; $r_0 \leftarrow f_\theta(p_q(q, C_0))$; $R \leftarrow [r_0]$
$W \leftarrow [0]_{k+1 \times k+1}$
foreach $c_i \in \{c_1, c_2, \ldots, c_k\}$ **do**
  $c_i' \leftarrow f_\theta(p_r(c_i))$
for $i \leftarrow 1$ **to** k **do**
  $C_i \leftarrow [c_1, \ldots, c_i', \ldots, c_k]$
  $r_i \leftarrow f_\theta(p_q(q, C_i))$
  $R \leftarrow \text{append}(R, r_i)$
foreach $r_i \in R$ **do**
  foreach $r_j \in R$ **do**
    $w_{ij} \leftarrow (NLI(r_i, r_j) + NLI(r_j, r_i))/2$
    $W[i, j] \leftarrow w_{ij}$
$DSE \leftarrow -(k+1)^{-1} \sum_{i=0}^{k} \log \frac{D_i}{k+1}$, $D_i = \sum_{j=0}^{k} w_{ij}$
return DSE, R, W

distribution over responses. We retain the name Degree-based Semantic Entropy for this score without identifying it with Shannon entropy on a general semantic graph.

Exact equality in the equivalence-class case. Suppose W is binary and reflexive, with $w_{ij} = 1$ exactly when r_i and r_j belong to the same semantic class. The graph is then a disjoint union of complete graphs, including self-links. Every response in class h has degree n_h, giving

$$DSE = -\frac{1}{N} \sum_{h \in H} n_h \log \frac{n_h}{N} = -\sum_{h \in H} \widehat{p}_h \log \widehat{p}_h = \widehat{SE}.$$

This equality concerns the same finite response collection; it does not equate context perturbations with sampling from a fixed-input output distribution. Weighted or nontransitive entailment graphs need not satisfy this equality.

Bounds and monotonicity. If $0 \leq w_{ij} \leq 1$ and $w_{ii} = 1$, then $1 \leq D_i \leq N$, implying $0 \leq DSE \leq \log N$. Complete agreement gives $D_i = N$ and $DSE = 0$; only self-links give $D_i = 1$ and $DSE = \log N$. The empirical semantic entropy has the same extrema, but shared extrema alone do not establish equivalence. At fixed N, increasing degrees cannot increase DSE. These bounds require reflexive self-links; a zero degree would make the logarithm undefined.

E ALGORITHMS

E.1 SECURE

We present the SECURE algorithm in Algorithm 1.

E.2 ADAPTIVE SEMANTIC CHUNKING

We present the adaptive semantic chunking Algorithm 2.

E.3 ITERATIVE CONTEXT REFINEMENT

Algorithm 3 retains the original-context answer r_0 from each SECURE evaluation; R is an indexed response list, including repeated answers. The visited set is initialized once and never reset. Each updated nonempty context is evaluated before testing candidate exhaustion. Early stopping requires both low DSE and a passing self-evaluation for that context. Otherwise, the algorithm returns the answer from the evaluated context with minimum DSE. A continued iteration either adds an unvisited chunk or removes a chunk; removed chunks are never reintroduced. Empty or unchanged contexts

Algorithm 2: Adaptive Semantic Chunking

Input: Document set $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, Similarity threshold τ, Maximum chunk length T_{max}
Output: Semantic chunk set $\mathcal{C}$
Initialize an empty list: $\mathcal{C} \leftarrow \emptyset$
foreach $d \in \mathcal{D}$ **do**
  $S \leftarrow \texttt{tokenize}(d)$ `// split documents into sentences`
  $\{e_1, e_2, e_3, ...\} \leftarrow [\texttt{Encode}(s) \mid s \in S]$ `// encode sentences into embeddings`
  $c \leftarrow [s_1]$ `// initialize current chunk`
  $T_{current} \leftarrow |s_1|$
  for $i \leftarrow 1$ **to** $|S| - 1$ **do**
    $sim \leftarrow \texttt{cosine_similarity}(e_i, e_{i+1})$
    $T_{next} \leftarrow |s_{i+1}|$
    if $sim > \tau$ ***and*** $T_{current} + T_{next} \leq T_{max}$ **then**
      $c \leftarrow c \cup \{s_{i+1}\}$
      $T_{current} \leftarrow T_{current} + T_{next}$
    else
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
      $c \leftarrow [s_{i+1}]$ `// initialize a new chunk`
      $T_{current} \leftarrow T_{next}$
  if c *is not empty* **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
return $\mathcal{C}$

terminate refinement, ensuring termination on a finite candidate pool. Cosine similarity is computed between chunk embeddings.

F DATASETS

We evaluate on the publicly labeled development splits described below. Some benchmarks withhold test answers from public release; for example, KILT releases test inputs without reference answers Petroni et al. (2020). The development splits support reproducible offline evaluation. SECURE and SECURE-RAG require no task-specific parameter training or reference answers for scoring and refinement. The default SECURE threshold is 0.2; Section 6.5 and Figure 4 report the QA accuracy of the certain and uncertain groups across thresholds. Dataset statistics appear in Table 7.

- **TriviaQA** Joshi et al. (2017) is a challenging QA dataset with supporting evidence documents and two question sources: Wikipedia and the Web. Following KILT Petroni et al. (2020), we use only Wikipedia questions with evidence documents from the *wikipedia-dev* split.
- **Natural Questions** Kwiatkowski et al. (2019) is a widely used QA dataset derived from Wikipedia. Its questions originate from Google searches, and the associated documents are Wikipedia pages. Following KILT Petroni et al. (2020), we use the *dev* questions for which at least one human annotator marked a short answer in the document.
- **AmbigNQ** Min et al. (2020) is constructed from NQ-OPEN questions using English Wikipedia as the evidence corpus. We use the *singleAnswer* questions from its *dev* split.
- **2WikiQA** Ho et al. (2020) is a multi-hop QA dataset designed to test relations between entities. Each question is paired with multiple evidence articles. We use the *dev* split.

G IMPLEMENTATION DETAILS

Input-side uncertainty baselines. In addition to the output-sampling baselines in Table 1, we include SPUQ Gao et al. (2024) and Semantic Reformulation Entropy (SRE) Tong et al. (2025). SPUQ generates input perturbations, samples outputs for the perturbed inputs, and aggregates their consistency into an uncertainty estimate. SRE reformulates the question and estimates entropy over

Algorithm 3: Iterative Context Refinement

Input: Question q, finite candidate $Chunk\ Set$, initial context size $1 \leq k \leq |Chunk\ Set|$, LLM f_θ, self-evaluation prompt p_e
Output: One answer $Answer_{best}$

$C \leftarrow$ initial top-k chunks for q; $Visited\ Set \leftarrow \emptyset$;
$DSE_{min} \leftarrow \infty$; $Answer_{best} \leftarrow \perp$;
while true **do**
  $Visited\ Set \leftarrow Visited\ Set \cup C$; $m \leftarrow |C|$;
  $DSE, R, W \leftarrow SECURE(q, C)$; $r_0 \leftarrow R[0]$;
  if $Answer_{best} = \perp$ or $DSE < DSE_{min}$ **then**
    $Answer_{best} \leftarrow r_0$; $DSE_{min} \leftarrow DSE$;
  $Flag_{eval} \leftarrow f_\theta(p_e(q, C))$;
  if $DSE < 0.2$ and $Flag_{eval} = 1$ **then**
    return r_0
  if $Visited\ Set = Chunk\ Set$ **then**
    break;
  $C_{previous} \leftarrow C$;
  $Uncertain\ Set \leftarrow \{c_i \in C : W_{i0} \neq 1,\ i = 1, \ldots, m\}$;
  $Necessary\ Set \leftarrow \emptyset$; $Unnecessary\ Set \leftarrow \emptyset$;
  foreach $c_i \in Uncertain\ Set$ **do**
    $ar_i \leftarrow f_\theta(q, C \setminus \{c_i\})$;
    $e_i \leftarrow [NLI(r_0, ar_i) + NLI(ar_i, r_0)]/2$;
    if $e_i = 0$ **then**
      $Necessary\ Set \leftarrow Necessary\ Set \cup \{c_i\}$;
    else
      $Unnecessary\ Set \leftarrow Unnecessary\ Set \cup \{c_i\}$;
  $C \leftarrow C \setminus Unnecessary\ Set$;
  foreach $c_i \in Necessary\ Set$ **do**
    $Candidates \leftarrow Chunk\ Set \setminus Visited\ Set$;
    if $Candidates = \emptyset$ **then**
      break;
    $best_chunk \leftarrow \arg\max_{c_j \in Candidates} \text{cosine_similarity}(c_i, c_j)$;
    $C \leftarrow C \cup \{best_chunk\}$; $Visited\ Set \leftarrow Visited\ Set \cup \{best_chunk\}$;
  if $C = \emptyset$ or $C = C_{previous}$ **then**
    break;
return $Answer_{best}$

Table 7: Dataset statistics. Questions are divided into certain and uncertain groups by SECURE using Llama-3.1 8B.

Datasets	No. valid questions	No. certain questions	No. uncertain questions
TriviaQA	7928	6726	1202
NQ	4289	3115	1174
AmbigNQ	1000	780	220
2WikiQA	12576	7663	4913

answer groups obtained through hybrid semantic clustering. These baselines extend the comparison to input-side variation; SECURE instead rephrases one retrieved chunk at a time and retains the link between each answer variant and the edited chunk.

Role of rephrasing. Each SECURE variant changes one retrieved chunk, retaining a correspondence between response variation and an identifiable context edit. The degree-based score summarizes this variation, and SECURE-RAG uses the score and chunk indices to guide context refinement. Our contribution couples this context-sensitive uncertainty measurement with its refinement procedure. Table 1 evaluates the complete uncertainty estimator.

We implement a naive retrieval-based QA framework Gao et al. (2023) on Qwen-2.5 1.5B, Qwen-3 8B, Llama-3 8B and Llama-3 70B as backbones. We use the documents provided by each dataset as

Table 8: AUROC and AUARC under different temperature settings. Under high-temperature sampling, answer variability increasingly reflects decoding randomness, reducing the interpretability of uncertainty scores for retrieval-based question answering.

Uncertainty Measurements	TriviaQA		Natural Question		AmbigNQ		2WikiQA	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
greedy decoding								
SECURE$_{deberta}$*(ours)*	66.94	91.02	59.42	73.69	58.01	74.89	53.49	61.42
temperature=0.50								
Discrete Semantic Entropy Farquhar et al. (2024)	71.15	90.56	65.59	74.46	66.61	78.79	61.50	63.57
U$_{eigv}$ Lin et al. (2024)	74.32	91.90	**69.35**	**77.08**	68.79	77.47	61.97	63.93
U$_{deg}$ Lin et al. (2024)	74.24	91.53	69.28	76.82	68.64	77.89	61.84	63.44
KLE heat$_{t=0.1}$ Nikitin et al. (2024)	73.87	91.48	69.18	76.79	68.04	77.78	61.41	63.72
KLE deberta matern$_{\kappa=3.0,\nu=3.0}$ Nikitin et al. (2024)	68.84	90.72	64.57	75.09	63.25	77.95	60.17	63.54
SECURE$_{llm}$*(ours)*	**77.53**	**93.06**	69.23	76.71	**72.02**	**81.28**	**67.28**	**68.08**
temperature=1.0								
Discrete Semantic Entropy Farquhar et al. (2024)	78.82	93.23	72.29	79.15	72.19	82.67	67.18	67.61
U$_{eigv}$ Lin et al. (2024)	**81.75**	93.87	**75.12**	81.49	**75.50**	82.90	68.01	68.12
U$_{deg}$ Lin et al. (2024)	81.69	93.78	75.00	81.35	74.98	**83.25**	67.78	67.86
KLE heat$_{t=0.1}$ Nikitin et al. (2024)	81.30	93.72	74.84	**81.65**	74.31	82.50	67.29	67.67
KLE deberta matern$_{\kappa=3.0,\nu=3.0}$ Nikitin et al. (2024)	71.62	91.78	67.73	77.86	66.18	79.47	62.86	66.28
SECURE$_{llm}$*(ours)*	80.20	**93.92**	70.92	77.00	74.73	81.68	**70.05**	**70.25**
temperature=3.0								
Discrete Semantic Entropy Farquhar et al. (2024)	84.11	90.60	**77.36**	**69.78**	80.23	79.15	73.15	56.86
U$_{eigv}$ Lin et al. (2024)	80.42	89.55	67.35	64.45	75.73	76.91	70.06	55.50
U$_{deg}$ Lin et al. (2024)	84.00	90.87	72.45	67.63	79.81	79.12	72.51	57.21
KLE heat$_{t=0.1}$ Nikitin et al. (2024)	**85.74**	**91.44**	75.31	69.40	**81.72**	**80.08**	**73.89**	**58.03**
KLE deberta matern$_{\kappa=3.0,\nu=3.0}$ Nikitin et al. (2024)	59.95	76.50	58.22	58.16	58.34	66.54	56.16	45.01
SECURE$_{llm}$*(ours)*	78.58	87.33	72.33	65.14	75.39	71.07	70.33	53.93
temperature=5.0								
Discrete Semantic Entropy Farquhar et al. (2024)	**68.08**	**49.21**	61.22	20.51	62.74	24.73	71.31	21.45
U$_{eigv}$ Lin et al. (2024)	65.06	47.14	56.51	18.19	60.62	23.32	71.93	22.05
U$_{deg}$ Lin et al. (2024)	66.53	48.20	58.07	18.89	62.03	24.20	72.30	22.30
KLE heat$_{t=0.1}$ Nikitin et al. (2024)	67.88	48.96	59.31	19.48	63.28	24.84	**72.33**	22.24
KLE deberta matern$_{\kappa=3.0,\nu=3.0}$ Nikitin et al. (2024)	51.15	34.49	48.02	14.74	49.13	18.44	51.92	12.33
SECURE$_{llm}$*(ours)*	66.94	46.45	**62.25**	**20.99**	**68.96**	**27.69**	71.10	**22.53**
temperature=7.0								
Discrete Semantic Entropy Farquhar et al. (2024)	64.67	36.77	59.68	13.86	63.77	18.77	68.32	16.89
U$_{eigv}$ Lin et al. (2024)	64.31	36.27	57.58	12.91	62.35	18.29	71.12	18.38
U$_{deg}$ Lin et al. (2024)	64.90	36.66	58.30	13.30	63.92	18.66	70.94	18.31
KLE heat$_{t=0.1}$ Nikitin et al. (2024)	65.51	36.81	58.73	13.51	64.16	**18.78**	70.23	17.97
KLE deberta matern$_{\kappa=3.0,\nu=3.0}$ Nikitin et al. (2024)	52.33	27.63	49.98	10.88	49.47	12.23	52.94	10.83
SECURE$_{llm}$*(ours)*	**66.43**	**38.83**	**61.91**	**15.22**	**66.13**	18.04	**74.03**	**20.76**

the retrieval corpus and employ recursive chunking in LangChain[3]. The chunk size is 512 and the overlap is 64. For document retrieval, we use UAE-Large-V1 to encode both questions and document chunks; it is one of the top zero-shot embedding models on the MTEB English v2 leaderboard[4]. We then use FAISS[5] to build the dense index. The default initial context contains the top-5 retrieved chunks. Iterative refinement can inspect additional candidates and change the context size, so the initial retrieval count does not imply equal total evidence exposure or inference budgets. We discuss the impact of different values of top-k on the model's performance in Appendix J. At the generation stage, we use the simple prompt described in Appendix A.4.

In adaptive semantic chunking, we set the merge threshold to $\tau = 0.6$ and use the same encoder for chunk and sentence embeddings. The maximum chunk length is $T_{max} = 512$, consistent with recursive chunking. The NLI agent and the model used to construct meaning-preserving context variants use the same LLM as the generation stage. For the 1.5B and 8B SECURE-RAG models, a single NVIDIA RTX 3090 GPU is sufficient for embedding and inference. For the 70B model, we use two NVIDIA A100 80GB GPUs.

Effective uncertainty measures should reflect the reliability of LLM responses, such that higher uncertainty is associated with a higher likelihood of incorrect generations. Following prior work Kuhn et al. (2023); Band et al. (2022), we evaluate the quality of uncertainty estimates by using them to predict whether a generated answer is correct, and report the Area Under the Receiver Operating Characteristic curve (AUROC).

While AUROC is widely used, it has two notable limitations: it only applies to binary correctness labels, and its absolute values are often difficult to interpret. Therefore, we additionally adopt the Area Under the Accuracy–Rejection Curve (AUARC) Nadeem et al. (2009) as a complementary evaluation metric. For further discussion of AUARC in uncertainty evaluation, see Lin et al. (2024).

[3]https://python.langchain.com/docs/concepts/text_splitters/

[4]https://huggingface.co/spaces/mteb/leaderboard

[5]https://ai.meta.com/tools/faiss/

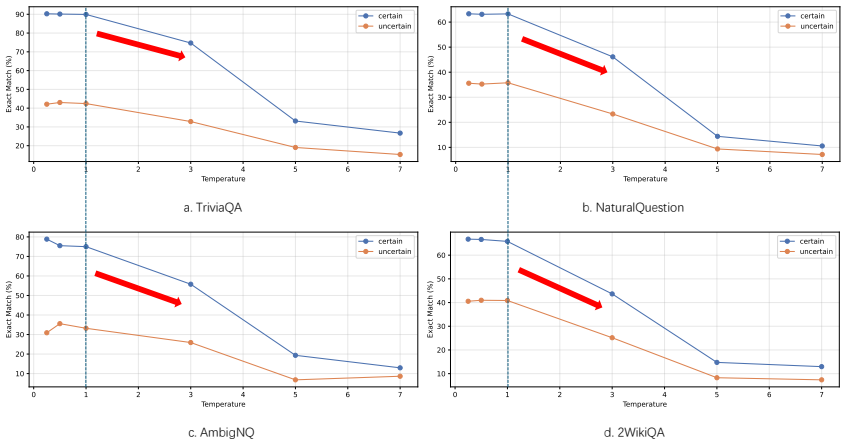


Figure 5: RAG QA performance at different sampling temperatures. Exact-match accuracy drops sharply above temperature 1 in the evaluated setting.

H Additional SECURE Evaluation Results

Table 9 reports RAG performance on certain and uncertain questions using different LLMs for open-book QA. All tested LLMs perform worse on uncertain questions. The gap narrows for larger models such as Llama-3.1 70B but remains around 20%, showing that the SECURE-based split remains informative across the tested model scales.

H.1 Experiments in high-temperature sampling

In addition to the low-temperature and greedy decoding evaluations reported in the main paper, we further examine the behavior of SECURE and other uncertainty estimating baselines under higher sampling temperatures. The results are shown in Table 8. Figure 5 shows how QA performance changes as temperature increases. At higher sampling temperatures, answer accuracy and uncertainty-ranking performance change in the reported experiments. These data indicate temperature sensitivity, but do not disentangle sampling noise from changes in evidence use. We therefore keep the low-temperature setting as the main comparison and report the high-temperature results as a stress test.

H.2 Experiments on Summarization Tasks

We focus on open-book QA, but we also conduct a preliminary evaluation of SECURE on summarization, a free-form generation task. We build a RAG summarization pipeline on the CNN/DailyMail 3.0.0 test set and compute SECURE scores. Table 10 reports the relationship between ROUGE-L and SECURE. Higher uncertainty is correlated with lower ROUGE-L, suggesting that SECURE may also serve as a response-stability signal for summarization. This verification experiment on CNN/DailyMail is preliminary, as we directly applied the method originally designed for open-book QA. Summarization presents different challenges compared to open-book QA, such as how to formulate effective queries. One important direction is to improve summarization performance according to the proposed SECURE method.

I Comparison with Additional Baselines

Table 11 provides an extended comparison with state-of-the-art RAG methods: Adaptive-RAG Jeong et al. (2024), Astute-RAG Wang et al. (2025), RePlug Shi et al. (2023), LongRAG Zhao et al. (2024), ERM4 Shi et al. (2024), UncertaintyRAG Li et al. (2024d), RAG-DDR Li et al. (2024b), ChatQA-1.5 Liu et al. (2024c), RankRAG Yu et al. (2024), Collab-RAG Xu et al. (2025a), MAIN-RAG Chang et al. (2025), PA-RAG Wu et al. (2025), SUNAR Venktesh et al. (2025), and RISE He et al. (2025). Methods such as RankRAG Yu et al. (2024) fine-tune the LLM to optimize retrieved context; we

Table 9: RAG performance on **certain** and **uncertain** questions. The EM drop from certain to uncertain questions is shown in parentheses after the uncertain-group EM. SECURE$_{eval}$ defines the two groups.

Task	TriviaQA		Natural Question		AmbigNQ		2WikiQA	
	certain	uncertain	certain	uncertain	certain	uncertain	certain	uncertain
SECURE$_{eval}$ Llama3 8B	90.4	43.0(47.4↓)	63.7	36.0(27.7↓)	75.1	33.2(41.9↓)	66.9	40.6(26.3↓)
SECURE$_{deberta}$ Llama3 8B	90.3	43.7 (46.6↓)	63.8	36.0(27.8↓)	75.1	34.1(41.0↓)	67.0	41.2(25.8↓)
Qwen2.5 1.5B	75.0	33.3(41.7↓)	47.7	24.8(22.9↓)	55.5	21.8(33.7↓)	38.5	21.4(17.1↓)
Llama3 8B	90.3	42.3(48.0↓)	63.3	35.8(27.5↓)	75.6	32.8(42.8↓)	66.9	40.3(26.6↓)
Llama3 70B	93.2	73.1(20.1↓)	62.5	44.8(17.7↓)	81.8	59.6(22.2↓)	72.3	57.4(14.9↓)

Table 10: Summarization results on CNN/DailyMail.

SECURE range	[0,0.1)	[0.1, 0.2)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1.0)	[1.0, ∞)
Average ROUGE-L/10^{-3}	90.47	86.97	86.84	82.40	86.29	81.48	78.56	76.54	80.36	81.48	76.43

mark these as Gen. FT. Approaches that train additional retrievers, encoders, or policy models are marked as Ret. FT. Collab-RAG Xu et al. (2025a) uses GPT-4o as an LLM reader during retrieval and generation.

J Additional Sensitivity Analysis

Number of retrieved documents. To examine SECURE with different numbers of retrieved chunks, we vary the context size with and without SECURE. As shown in Figure 7, both methods perform better on uncertain questions as more chunks are used. SECURE-RAG consistently outperforms RAG without SECURE across all datasets and tested context sizes.

Adaptive chunking threshold τ. We evaluate different adaptive chunking thresholds in Figure 6. As the sentence-merge threshold increases, performance on uncertain questions declines. Although semantic similarity can improve chunking, overly strict merge conditions may reduce overall effectiveness.

Different NLI backbones in SECURE. To assess whether SECURE depends on a particular NLI backbone, we replace the LLM-based entailment judge with a supervised model, DeBERTa-large-MNLI. We evaluate the DeBERTa-based SECURE using the same AUROC and AUARC metrics in Table 8. SECURE maintains consistently strong performance across datasets under both NLI backbones, indicating that the method is not tied to LLM-specific behaviors and generalizes well when using a small supervised NLI model. This robustness suggests that SECURE captures intrinsic semantic uncertainty rather than artifacts of any particular entailment model.

K Discussion of complexity

One SECURE evaluation with k chunks uses k rewrites, $k+1$ QA generations, and $O((k+1)^2)$ directional NLI judgments. These are model-invocation counts, not wall-clock bounds: token lengths, batching, and model choice affect the cost of each call. A smaller NLI model may be used for judging Kuhn et al. (2023).

For refinement, let M be the number of SECURE evaluations, k_t the context size at evaluation t, and $K = \max_t k_t$. Each round additionally uses at most k_t ablated QA generations, $2k_t$ directional necessity judgments, and one self-evaluation. Hence the full run requires

$$O\left(\sum_{t=1}^{M}(k_t+1)^2\right) \subseteq O\big(M(K+1)^2\big)$$

model invocations. The initial top-k size does not by itself bound later context sizes. If context sizes are bounded by k, the bound becomes $O(Mk^2)$; when $M = \Theta(k)$, this is $O(k^3)$, not $O(k^2)$. The fraction of questions triggering refinement affects average workload, not this per-query worst-case bound.

Candidate selection and embedding have additional costs. With N_c candidate chunks and precomputed d-dimensional embeddings, exhaustive cosine-similarity scans require at most $O(MKN_cd)$

Table 11: Results of our methods and baselines on four datasets. The best results are in **bold**, and the second best are underlined. Results unavailable in public reports are marked as "-".

Method	Generator Model	Ret. FT	Gen. FT	TriviaQA	Natural Question	AmbigNQ	2WikiQA
Adaptive-RAGJeong et al. (2024)	FLAN-T5-XL 3B	✓	✗	52.2	37.8	-	40.6
Astute-RAG Wang et al. (2025)	Claude 3.5 Sonnet	✗	✗	84.5	53.6	-	-
RePlug-LSR (few-shot)Shi et al. (2023)	Codex 175B	✓	✗	77.3	45.5	-	-
LongRAGZhao et al. (2024)	GLM4 32B	✗	✓	-	-	-	57.2
SUNARVenktesh et al. (2025)	GPT-3.5-turbo	✗	✗	-	-	-	54.8
ERM4Shi et al. (2024)	GPT-3.5-turbo	✓	✗	-	52.7	53.5	46.8
UncertaintyRAGLi et al. (2024d)	Vicuna 7B	✓	✗	85.0	-	-	29.9
UncertaintyRAGLi et al. (2024d)	Llama2 13B	✓	✗	82.5	-	-	38.3
RAG-DDRLi et al. (2024b)	Llama3 8B	✓	✓	89.6	52.1	-	-
ChatQA-1.5Liu et al. (2024c)	Llama3 8B	✓	✓	81.0	42.4	65.5	26.8
ChatQA-1.5Liu et al. (2024c)	Llama3 70B	✓	✓	85.6	47.0	-	34.9
RankRAGYu et al. (2024)	Llama3 8B	✗	✓	82.9	50.6	-	31.4
RankRAGYu et al. (2024)	Llama3 70B	✗	✓	86.5	54.2	-	38.2
Collab-RAGXu et al. (2025a)	Qwen2.5 3B	✗	✓	-	-	-	67.0
Collab-RAGXu et al. (2025a)	Llama3.1 8B	✗	✓	-	-	-	67.2
MAIN-RAGChang et al. (2025)	Llama3 8B	✗	✗	74.1	56.6	64.3	54.6
PA-RAGWu et al. (2025)	Llama3 8B	✗	✓	73.8	**62.3**	64.5	59.5
RISEHe et al. (2025)	Llama3.1 8B	✓	✓	80.3	59.5	52.3	49.4
SECURE-RAG	Qwen2.5 1.5B	✗	✗	69.8	42.4	50.3	31.5
	Qwen3 8B	✗	✗	82.4	53.5	65.2	54.2
	Llama3 8B	✗	✗	84.8	57.5	68.1	58.7
	Llama3 70B	✗	✗	**90.3**	58.2	**76.8**	**69.3**

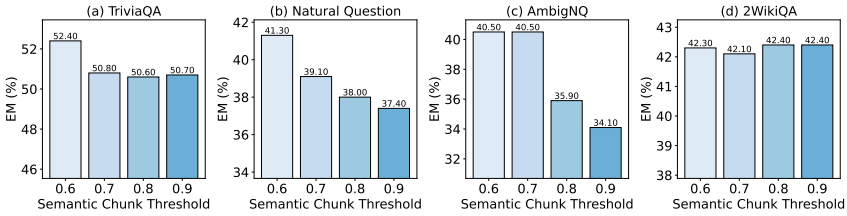


Figure 6: Performance on uncertain questions under different adaptive chunking merge thresholds.

arithmetic operations. Semantic chunking also incurs sentence-embedding costs that depend on document length and batching. These costs are separate from the model-invocation count.

In practice, on Llama-3 with A100 GPUs, computing SECURE adds about one second per query, while the full SECURE-RAG pipeline averages less than two seconds for certain cases and about four seconds for uncertain cases. **Inference cost and deployment trade-off.** Repeated inference is part of SECURE-RAG's iterative design: uncertainty estimation and context refinement add rewriting, generation, and semantic-judging calls. This unsupervised, model-agnostic design enables context adaptation without task-specific parameter training while requiring only text input/output access to the generator. The method can therefore operate through a black-box generator interface without access to logits, gradients, or model weights. Our results characterize the QA gains of this complete training-free pipeline with its additional inference computation. Comparisons at a fixed end-to-end inference budget remain outside the scope of this evaluation.

L CASE STUDIES

We present SECURE-RAG case studies on uncertain questions from TriviaQA, Natural Questions, AmbigNQ, and 2WikiQA. Tables 12–15 show retrieved chunks and LLM responses with and without SECURE-RAG. Red text marks distractors, and green text marks evidence supporting the correct answer. These cases contain distracting passages that can mislead the generator. In the 2WikiQA example, the model must first identify the performer of "Sunday Papers" and then determine the award for which the performer was nominated. SECURE-RAG identifies response-sensitive chunks and guides the refiner toward the evidence containing the answer.

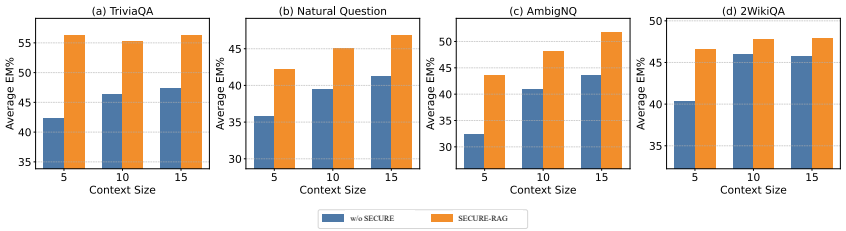


Figure 7: Performance on uncertain questions with different context sizes. SECURE-RAG improves performance across the tested sizes.

Table 12: Case study on uncertain questions in TriviaQA.

Question Id: tc_1693	
Q: **Who beat Tim Henman in his first Wimbledon singles semifinal?** A: Pete Sampras	
w/o SECURE	**Chunk 1**: Timothy Henry "Tim" Henman (born 6 September 1974) is a retired English professional tennis player. Henman played a serve-and-volley style of tennis... **Chunk 2**: In the second round he succumbed to the eventual champion American Todd Martin, 6–4, 6–4. Henman received a wildcard for the Manchester Open, where he lost in the first round to American Alex O'Brien... **Chunk 3**: At the time of his retirement, Henman had already committed to playing a Charity Exhibition at London's Royal Albert Hall during the Seniors Tennis Event The Blackrock Masters in December 2007... **Chunk 4**: Then breaking his opponent's serve twice in a row to win the final set 7–5 and beat reigning French Open champion Yevgeny Kafelnikov in the first round at Wimbledon, going on to reach the quarter finals before losing to Todd Martin... **Chunk 5**: He reached the second round after defeating German Martin Sinner, and in Nottingham he reached the quarter-finals, his first quarter-final in the ATP tour. His success in these tournaments increased his ranking from 272nd to 219th. LLM response: Todd Martin
SECURE-RAG	**Chunk 1**: Timothy Henry "Tim" Henman (born 6 September 1974) is a retired English professional tennis player. Henman played a serve-and-volley style of tennis... **Chunk 2**: In 2000 he reached the fourth round and in 1996, 1997, 2003 and 2004 he lost in the quarter-finals. The first two of those semi-final losses were to Pete Sampras, who went on to win the title on both occasions... **Chunk 3**: Then breaking his opponent's serve twice in a row to win the final set 7–5 and beat reigning French Open champion Yevgeny Kafelnikov in the first round at Wimbledon, going on to reach the quarter finals before losing to Todd Martin... **Chunk 4**: On the grass at Queen's Club Championship Henman reached the final, where he was defeated in straight sets by Australian Lleyton Hewitt... **Chunk 5**: However, Henman's winning streak did not last long, and in the second round he met Sampras, and was defeated 6–2, 6–3, 7–6... LLM response: Pete Sampras

Table 13: Case study on uncertain questions in Natural Question.

Question Id: -1580637763990313083	
Q: **When did muhammad ali win an olympic gold medal?** A: 1960	
w/o SECURE	**Chunk 1**: Ali said in his 1975 autobiography that shortly after his return from the Rome Olympics, he threw his gold medal into the Ohio River after he and a friend were refused service at a "whites - only" restaurant and fought with a white gang... **Chunk 2**: Ali received a replacement medal at a basketball intermission during the 1996 Olympics in Atlanta, where he lit the torch to start the gam professional boxing early career... **Chunk 3**: On September 13, 1999, Ali was named "Kentucky Athlete of the Century" by the Kentucky Athletic Hall of Fame in ceremonies at the Galt House East. On January 8, 2001, Muhammad Ali was presented with the Presidential Citizens Medal by President Bill Clinton. **Chunk 4**: In 1996, he had the honor of lighting the flame at the 1996 Summer Olympics in Atlanta, Georgia. Ali 's bout with Parkinson 's led to a gradual decline in his health... **Chunk 5**: W. Bush embraces Muhammad Ali after presenting him with the Presidential Medal of Freedom on November 9, 2005, during ceremonies at the White House. Muhammad Ali defeated every top heavyweight in his era, which has been called the golden age of heavyweight boxing... LLM response: Muhammad Ali did not win an Olympic gold medal.
SECURE-RAG	**Chunk 1**: On September 13, 1999, Ali was named "Kentucky Athlete of the Century" by the Kentucky Athletic Hall of Fame in ceremonies at the Galt House East... **Chunk 2**: Legacy As Mrs. Lonnie Ali looks on, President George W. Bush embraces Muhammad Ali after presenting him with the Presidential Medal of Freedom on November 9, 2005, during ceremonies at the White House... **Chunk 3**: Muhammad Ali defeated every top heavyweight in his era, which has been called the golden age of heavyweight boxing. Ali was named "Fighter of the Year" by The Ring magazine more times than any other fighter, and was involved in more Ring "Fight of the Year" bouts than any other fighter. **Chunk 4**: At age 18, he won a gold medal in the light heavyweight division at the 1960 Summer Olympics in Rome and turned professional later that year. At age 22 in 1964, he won the WBA, WBC, and lineal heavyweight titles from Sonny Liston in a major upset. **Chunk 5**: In time, Muhammad Ali Boulevard – and Ali himself – came to be well accepted in his hometown. In 1993, the Associated Press reported that Ali was tied with Babe Ruth as the most recognized athlete, out of over 800 dead or living athletes, in America. LLM response: 1960

Table 14: Case study on uncertain questions in AmbigNQ.

Question Id: 3008542401703695479	
Q: **Who played robin on the original batman series?** A: Burt Ward	
w/o SECURE	**Chunk 1**: Robin (Dick Grayson) was portrayed by Douglas Croft and Johnny Duncan, respectively, in the 1943 and 1949 fifteen chapter Batman serials. Burt Ward played him in the 1966–1968 Batman television series and the related 1966 film... **Chunk 2**: Batman is a 1960s American live action television series, based on the DC comic book character of the same name. It stars Adam West as Bruce Wayne / Batman and Burt Ward as Dick Grayson / Robin – two crime-fighting heroes who defend Gotham City from a variety of arch villains. **Chunk 3**: Lewis Wilson as Bruce Wayne / Batman, Douglas Croft as Richard "Dick" Grayson / Robin J. Carrol Naish as Dr. Tito Daka/Prince Daka * Shirley Patterson as Linda Page * William Austin as Alfred Pennyworth... **Chunk 4**: Burt Ward as Dick Grayson / Robin: Batman's sidekick, dubbed the "Boy Wonder": a high school student noted for his recurring interjections in the form of "Holy _, Batman!" **Chunk 5**: Lewis Wilson as Batman and Douglas Croft as Robin: Both actors and their stunt doubles lacked the "style and grace" of either the comic characters they were portraying or their equivalents at Republic Pictures. LLM response: Douglas Croft
SECURE-RAG	**Chunk 1**: His portrayal is based mainly on Burt Ward's Dick Grayson. Collected editions See also Homosexuality in the Batman franchise * List of exclamations by Robin... **Chunk 2**: Portrayals Robin (Dick Grayson) was portrayed by Douglas Croft and Johnny Duncan, respectively, in the 1943 and 1949 fifteen chapter Batman serials... **Chunk 3**: Batman is a 1960s American live action television series, based on the DC comic book character of the same name. It stars Adam West as Bruce Wayne / Batman and Burt Ward as Dick Grayson / Robin – two crime-fighting heroes who defend Gotham City from a variety of arch villains... **Chunk 4**: The entire first Batman serial was re-released theatrically in 1965 as An Evening with Batman and Robin, and proved very popular (some theatres showed the chapters as a Saturday matinee). Its success inspired the action-comedy lampoon series Batman (and its 1966 theatrical feature film spin-off) starring Adam West and Burt Ward. **Chunk 5**: * Burt Ward as Dick Grayson / Robin: Batman's sidekick, dubbed the "Boy Wonder": a high school student noted for his recurring interjections in the form of "Holy _, Batman!" LLM response: Burt Ward

Table 15: Case study on uncertain questions in 2WikiQA.

Question Id: 0f1ac1a00bdb11eba7f7acde48001122	
Q: **For what award was the performer of "Sunday Papers" nominated?** A: Grammy Award	
w/o SECURE	**Chunk 1**: "Sunday Papers" is a song written and performed by British new wave musician Joe Jackson. It was released on his debut album, "Look Sharp!" Written as a critique of the British press, "Sunday Papers" features mocking lyrics and reggae-inspired music... **Chunk 2**: Caspar Babypants is the stage name of children's music artist Chris Ballew, who is also widely known as the singer of The Presidents of the United States of America. **Chunk 3**: Dáithí Sproule (born 23 May 1950) is a guitarist and singer of traditional Irish music. His niece is the singer Claire Sproule. **Chunk 4**: David Ian "Joe" Jackson (born 11 August 1954) is an English musician and singer-songwriter. Having spent years studying music and playing clubs, Jackson scored a hit with his first release, "Is She Really Going Out with Him?", in 1979. **Chunk 5**: "O Valencia!" is the fifth single by the indie rock band The Decemberists and the first released from their fourth studio album, "The Crane Wife." The music was written by The Decemberists and the lyrics by Colin Meloy. LLM response: The performer was not nominated for an award.
SECURE-RAG	**Chunk 1**: Written as a critique of the British press, "Sunday Papers" features mocking lyrics and reggae inspired music. "Sunday Papers" was released as a single in the UK as the follow-up to his single... **Chunk 2**: "Sunday Papers" is a song written and performed by British new wave musician Joe Jackson. It was released on his debut album, "Look Sharp!" **Chunk 3**: Caspar Babypants is the stage name of children's music artist Chris Ballew, who is also widely known as the singer of The Presidents of the United States of America." **Chunk 4**: Dáithí Sproule (born 23 May 1950) is a guitarist and singer of traditional Irish music. His niece is the singer Claire Sproule. **Chunk 5**: He is associated with the 1980s Second British Invasion of the US. He has also composed classical music. He has recorded 19 studio albums and received 5 Grammy Award nominations. LLM response: Grammy Award