DENSE-RAG: MEASURING AND IMPROVING CONTEXT UNDERSTANDING FOR CONSISTENT RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) has significantly advanced LLM performance in knowledge-intensive tasks. However, when LLMs misinterpret retrieved content, they often revert to pre-trained parametric knowledge or generate hallucinated responses, undermining RAG effectiveness. This highlights the need to assess LLMs' understanding of the retrieved context, since effective measurement is essential for evaluating and improving RAG performance. In this work we try to explore this problem by proposing DEgree-based uNcertainty with Semantically Equivalent contexts (DENSE), a training-free and model-agnostic method to evaluate LLM understanding of retrieved documents. DENSE constructs semantically equivalent context and introduces a degree-based entropy to quantify response semantic uncertainty. Building on DENSE, we further introduce DENSE-RAG, which includes two training-free DENSE-guided modules: adaptive semantic chunking and iterative context refinement. Extensive experiments on open-book QA datasets show that higher DENSE uncertainty correlates with lower QA performance, validating DENSE as a reliable indicator of LLM understanding measurement. DENSE-RAG also achieves performance competitive with state-of-the-art baselines approaches without introducing additional model or fine-tuning.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable success in NLP tasks. However, the reliance on parametric knowledge alone leads to knowledge cut-off issues and hallucination, making Retrieval-Augmented Generation (RAG) a crucial paradigm for knowledge-intensive tasks. In a RAG system, an information retrieval system fetches relevant documents from an external corpus, and LLMs generate the answer based on the retrieved documents. Recent research has advanced RAG by improving retrievers Shi et al. (2023); Lin et al. (2023b;a); Xu et al. (2024); Zhang et al. (2025) or through end-to-end fine-tuning of LLMs Yu et al. (2024); Izacard et al. (2023); Asai et al. (2023); Wang et al. (2024a); Huang et al. (2023). However, an essential research problem remains underexplored: how to measure LLM's understanding of the retrieved context? When LLMs fail to comprehend the input context, LLMs are observed to misinterpret contexts and make up unfaithful responses to the retrieved context Barnett et al. (2024); Song et al. (2024); Saad-Falcon et al. (2024). Consequently, the ability to assess an LLM's understanding of retrieved content represents a promising direction for evaluating and improving the reliability of RAG systems.

In this paper, we attempt to explore this problem through a semantic perspective. Consider two simple sentences, "*Bob is a physics teacher*" and "*Bob teaches physics*", which express the same meaning with different syntactic expressions. When asked "*What is Bob's occupation?*", a human reader would give the same answer regardless of which sentence is provided as context. This illustrates two key principles : (1) **the same semantics can be conveyed through different textual expressions;** and (2) **if an LLM truly understands the context, semantically equivalent inputs should yield semantically equivalent responses.**

Based on these insights, we propose **DE**gree-based u**N**certainty with **S**emantically **E**quivalent contexts (**DENSE**), for measuring LLM's understanding of contexts by evaluating the semantic consistency between multiple responses. DENSE is an **unsupervised**, **training-free** method that could be

applied to any LLMs. According to the principles, we construct semantically equivalent contexts and quantify the semantic uncertainty reflected in LLM responses across these contexts. Unlike prior approaches on non-retrieval settings Kuhn et al. (2023); Lin et al. (2024), DENSE captures semantic variation directly from response-level outputs and further enables fine-grained attribution of uncertainty to specific retrieved chunks. Experiments on open-book QA datasets show that LLMs perform significantly worse on questions with high DENSE uncertainty, demonstrating that DENSE provides a reliable indicator of contextual understanding.

On the basis of effective understanding measurement, we further leverage DENSE to enhance RAG performance. Specifically, we design two unsupervised modules: **Adaptive Semantic Chunking**, which leverages DENSE to trigger semantic chunking only under high uncertainty to improve intra-chunk semantic coherence, and **Iterative Context Refinement**, which incrementally supplements and reorganizes chunks guided by DENSE to enhance inter-chunk semantic completeness. Extensive experiments across diverse datasets and LLM backbones show that DENSE-RAG achieves competitive performance compared to state-of-the-art baselines, while offering a model-agnostic framework for both diagnosing and improving RAG systems.

Our contributions can be summarized as follows:

- We introduce **DENSE**, an unsupervised, training-free method to assess LLMs' understanding of retrieved contexts by measuring response uncertainty. Unlike prior work that primarily quantifies LLM inherent uncertainty, DENSE connect the presence of uncertainty to specific chunk, enabling targeted improvement to enhance RAG performance.
- We propose **DENSE-RAG**, which leverages DENSE to enhance RAG performance through two modules: Adaptive Semantic Chunking, which improves intra-chunk coherence under high uncertainty, and Iterative Context Refinement, which enhances inter-chunk completeness by reorganizing contexts in a DENSE-guided manner.
- We conduct extensive experiments on four open-book QA datasets with five LLMs of different scales, demonstrating that DENSE effectively evaluates LLM's understanding of contexts and is predictive of RAG performance. The proposed DENSE-RAG improves QA performance on challenging questions with high uncertainty, achieving competitive performance against state-of-the-art baselines, while maintaining flexibility and generality as a model-agnostic framework.

2 Related Work

In Retrieval-augmented generation, a retriever Karpukhin et al. (2020); Douze et al. (2024) is employed to obtain relevant document chunks from an external corpus, then LLM takes the retrieved context to generate replies Gao et al. (2023). Enhancing LLMs' understanding of retrieved documents to improve the overall alignment of the system remains a significant challenge. Some works improve retrievers to align the needs of LLMs Shi et al. (2023); Lin et al. (2023b;a) or add-on moderate-size models Xu et al. (2024); Zhang et al. (2025). Despite providing stronger retrievers, one potential approach is to finetune LLM in an end-to-end manner Yu et al. (2024); Izacard et al. (2023); Asai et al. (2023); Wang et al. (2024a); Huang et al. (2023); Yoran et al. (2023).

Enhancing the reliability of the generation by measuring the uncertainty in LLM responses has emerged as a promising direction Kuhn et al. (2023); Farquhar et al. (2024); Hou et al. (2024); Jiang et al. (2024). Semantic uncertainty was proposed to estimate uncertainty in natural language generation tasks in an unsupervised manner Kuhn et al. (2023). By quantifying the semantic differences among the responses, researchers can effectively measure the impact of hallucinations in LLMs Farquhar et al. (2024). Hou et al. (2024) proposed a method to decompose uncertainty by generating clarifications and ensembling. While many works discuss the uncertainty in LLMs in unsuperivsed scope Lin et al. (2024); Jiang et al. (2024), some works also try to identify uncertainty and improve the performance of LLMs in a supervised manner Kweon et al. (2025); Liu et al. (2024a); Arteaga et al. (2025). Several works have investigated how LLM uncertainty manifests in RAG settings. Dai et al. (2025) quantify the utility of retrieval by capturing LLM's internal belief in RAG scenarios. Hasegawa et al. (2024) measured certainty in retrieval and generation seperately through Rouge-L or the BERT score. Perez-Beltrachini & Lapata (2025) trained a passage utility model to predict the utility of each passage in the context of LLMs. However, these studies often focus on how to measure uncertainty within the system, while how to effectively link uncertainty to the

retrieved documents and leverage it to improve the performance of RAG systems remains largely underexplored.

3 PRELIMINARIES

To lay the groundwork for analyzing LLMs' understanding of retrieved content, we first introduce the RAG task formulation, semantic-related formulations and properties. Given a question q, a retriever fetches top-k documents from a knowledge base to construct context $C = [c_1, c_2, ..., c_k]$ that are most relevant to q. LLM f_θ is used to produce the response r:

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

where r could be a short phrase or sentence. Kuhn et al. (2023) discuss the meanings and forms of natural language, *"Although models' input is words, but for almost all applications we care about meanings"*. This observation underlines the central role of semantics in NLP tasks, we summarize the relationship between semantic meaning and retrieved context in RAG as two formulations.

Formulation 1: *For natural language context C and question q, there is a semantic space $\mathcal{S}$ and a mapping function π, that maps q, C to its underlying semantic $\pi(q, C)$ in semantic space $\mathcal{S}$.*

Formulation 2: *If there exists $C' \neq C$ such that $\pi(q, C') = \pi(q, C)$, then C' and C are semantically equivalent under question q in the semantic space $\mathcal{S}$.*

According to these formulations, two different textual contexts C and C' can yield the same semantics for a given question q. If a human reader or an LLM fully understands C, it should also produce a semantically equivalent response when given C'. This property is commonly referred as semantic consistency in prior work Rabinovich et al. (2023). Accordingly, we extend this notion of semantic consistency in the RAG setting and propose the following property:

Property 1: *If an LLM f_θ can understand C while answering question q. With $\pi(q, C) = \pi(q, C')$, the LLM's responses under C and C' should be semantically equivalent:*

$$\pi(q, C) = \pi(q, C') \rightarrow \pi(f_\theta(q, C)) = \pi(f_\theta(q, C')) \tag{2}$$

This property highlights that if an LLM produces diverse semantic under semantically equivalent contexts, the discrepancy signals a misalignment in its interpretation of input semantics.

4 EVALUATING LLM'S UNDERSTANDING OF RETRIEVED CONTEXT

Building on the aforementioned important property, we can examine whether the LLM has adequately understood the retrieved context by measuring its semantic variance under C and C'. We introduce DENSE (DEgree-based uNcertainty with Semantically Equivalent contexts), to evaluate LLM's understanding of contexts (Figure 1). DENSE consists of two main steps: we first construct semantically equivalent contexts through rephrasing and obtain responses via greedy decoding; then we propose a degree-based uncertainty measure to capture semantic variations across these responses, which further enables us to localize the specific chunks that contribute to the uncertainty.

4.1 SEMANTICALLY EQUIVALENT CONTEXT REPHRASING

To leverage Property 1, we need to construct semantically equivalent and textually diverse contexts C'. Specifically, we use an LLM to rephrase each retrieved chunk in isolation so that any semantic variation in the model's responses can be attributed to a single chunk without cross-chunk interference. For a retrieved context $C_0 = [c_1, \ldots, c_k]$, we obtain a rephrased c_i' for each chunk c_i:

$$c' = f_\theta(p_r(c)),\ where\ \pi(c') = \pi(c),\ and\ c' \neq c \tag{3}$$

p_r is the rephrasing prompt (see Appendix B.1). This yields a set of rephrased chunks $\{c_1', \ldots, c_k'\}$. We then construct k single-edit contexts by replacing exactly one chunk at a time:

$$C_i[j] = \begin{cases} c_i', & \text{if } j = i \\ c_j, & \text{otherwise} \end{cases} \quad \text{for } j = 1, \ldots, k \tag{4}$$

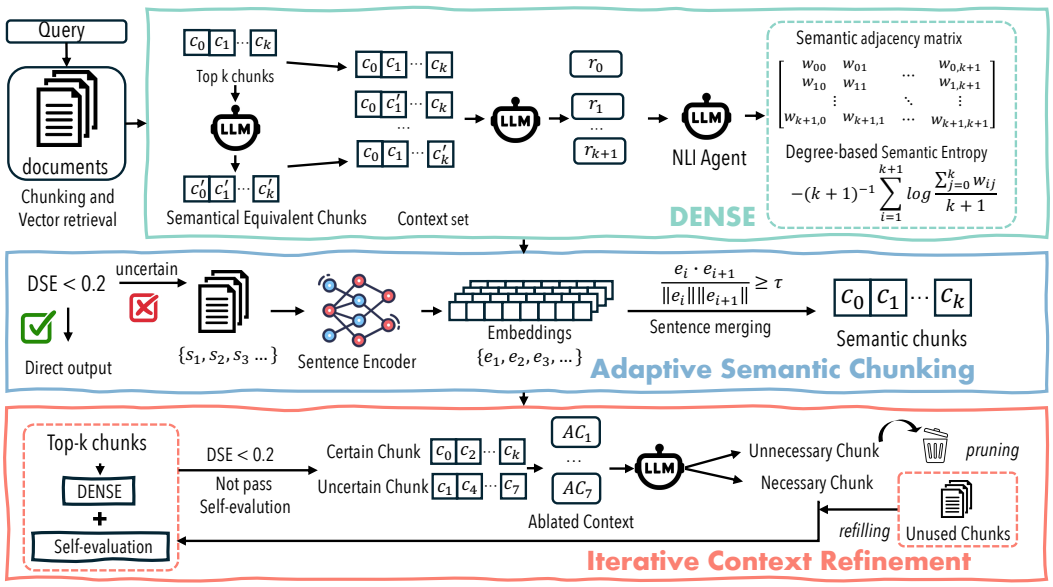

Figure 1: Overview of DENSE-RAG framework. DENSE evaluates the LLM's understanding of retrieved chunks. Adaptive semantic chunking improves intra-chunk coherence, and iterative context refinement enhances inter-chunk completeness.

In this manner, we obtain $k+1$ contexts in total: the original C_0 and k semantically equivalent variants $C_1, \ldots, C_k$. To ensure that rephrasing preserves the original semantics, we empirically compare QA results with and without rephrasing in Section 6.2, demonstrating the robustness of LLM rephrasing.

Before measuring semantic uncertainty in LLM responses, it is crucial to rule out randomness originating from the LLM itself. Prior work in non-retrieval settings studied the LLM inherent randomness by increasing the decoding temperature for diverse outputs Kuhn et al. (2023); Lin et al. (2024). In our setting, we disentangle this randomness by greedy decoding, ensuring that when the input is identical, the LLM always produces the same output. This guarantees that any semantic variation observed in our experiments arises solely from the LLM's understanding of rephrased contexts. Supporting experiments and detailed discussion are provided in Appendix C. Under this setup, generating with semantically equivalent contexts yields a reply set $R = \{r_0, r_1, \ldots, r_{k+1}\}$.

4.2 Degree-based Semantic Uncertainty

Given reply set R under semantically equivalent contexts, the next step is to assess the semantic variation in the LLM responses. To this end, we propose Degree-based Semantic Entropy, an effective method to quantify semantic uncertainty across these responses.

Semantic entropy was introduced by Kuhn et al. (2023), which measures uncertainty by clustering responses into semantic groups. However, since each generation in our setting is conditioned on a rephrased context, we need to further identify semantic variations introduced at the chunk level. In this case, semantic entropy becomes inadequate for capturing such fine-grained uncertainty.

We therefore propose degree-based semantic entropy, which avoids clustering and computes entropy directly over each response. Instead of clustering, we treat each response as a graph node, and construct a semantic adjacency matrix W using the entailment scores between multiple responses Jiang et al. (2024); Lin et al. (2024):

$$w_{ij} = ((NLI(r_i, r_j) + NLI(r_j, r_i))/2)_{i,j \in [0,\ldots,k]}, \tag{5}$$

where a natural language inference model(NLI) is used to classify whether r_i and r_j are *entailment(1)* or *neutral(0)*, w_{ij} represents the link between two responses. In DENSE, an LLM is employed to make the classification and the prompt is in Appendix B.2. After constructing W, we compute the

degree-based semantic uncertainty as follows:

$$DSE(q, C) = -(k+1)^{-1} \sum_{i=1}^{k+1} \log \frac{D_i}{k+1}, \tag{6}$$

where $D_i = \sum_{j=0}^{k} w_{ij}$, is the degree of response r_i. $\frac{D_i}{k+1}$ represents the average link strength of response r_i with respect to all other responses. Degree-based semantic entropy represents a non-clustered variant of semantic entropy, we discuss the relationship of DSE and semantic entropy in Appendix D and provide the pseudocode of the DENSE in Appendix E.1.

We empirically verify that DENSE effectively reflects LLMs' understanding of retrieved contexts, where higher DENSE scores consistently correlate with worse QA performance. The experimental setup, results, and corresponding discussions are presented in Section 6.

5 Improving Context Quality with DENSE Guidance

In the previous section, we proposed DENSE as an indicator of the LLM's ability to understand retrieved contexts. Building on this foundation, we move beyond measurement and leverage DENSE to improve context quality. Since higher uncertainty indicates worse performance, we categorize questions into **certain** ($DENSE \leq 0.2$) and **uncertain** ($DENSE > 0.2$). We propose two **model-agnostic, training-free** modules that enhance intra-chunk semantic consistency and inter-chunk completeness, thereby improving LLM performance on uncertain questions.

5.1 Adaptive Semantic Chunking

Chunking strategies in RAG 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 better preserve semantic but its computational cost frequently outweighs its performance gains Qu et al. (2024).

To address this trade-off, we design a DENSE-driven adaptation mechanism that selectively applies semantic chunking only when high uncertainty is detected under fixed-size chunking. The intuition is that uncertain questions are more likely to suffer from semantic inconsistencies in the retrieved chunks, and thus benefit more from semantic chunking. As illustrated in Figure 1, given a question q, we first run fixed-size chunking, use the top-k chunks for DENSE evaluation, and classify q as certain or uncertain. If q is a certain question, we directly output the response. Otherwise, we split the documents into sentences $d = \{s_1, s_2, s_3, ...\}$, encode each sentence s_i into a vector e_i, and iteratively merge s_{i+1} into chunk c_j if

$$\frac{e_i \cdot e_{i+1}}{\|e_i\|\|e_{i+1}\|} \geq \tau \; and \; |c_j \cup s_{i+1}| \leq T_{max}, \tag{7}$$

where τ is the similarity threshold and T_{max} is the maximum token length for a chunk. If the condition is not satisfied, s_{i+1} starts a new chunk c_{j+1}. The pseudocode of adaptive chunking is provided in Appendix E.2. We evaluate the effectiveness of Adaptive Chunking in Section 6. Our adaptive mechanism preserves the strong performance of fixed-size chunking on certain cases while selectively leveraging semantic chunking to enhance QA performance on uncertain ones.

5.2 Iterative Context Refinement

Semantic inconsistency can occur both within and across chunks. For complex questions, it is often the case that an individual retrieved chunk is insufficient to answer the query, highlighting the need for better inter-chunk coherence. To address this, we propose a DENSE based iterative context refinement module. The module evaluates retrieved chunks using DENSE and categorizes them into three types—certain, necessary, and unnecessary, based on their semantic contribution when serving as context. Then the module refine context by retaining certain chunks, removing unnecessary ones, and supplementing the context with new chunks guided by the necessary ones.

Localize the source of uncertainty. Since DENSE 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*:

$$l_i^{ce} = \mathbb{1}_{\{w_{i,0}=1\}}, \tag{8}$$

where $\mathbb{1}$ denotes the indicator which equals 1 if $w_{i,0} = 1$ and 0 otherwise. If $l_i^{ce} = 1$, chunk c_i is classified as a *certain chunk*, indicating that even after rephrasing, the corresponding response r_i remains semantically consistent with the original response r_0. This suggests that the LLM's understanding of c_i is robust and unaffected by rephrasing. Conversely, an *uncertain chunk* indicates that the LLM fails to correctly capture the intended semantics when that chunk is rephrased.

After distinguishing between certain and uncertain chunks, we further analysis the uncertain ones. Semantic uncertainty in LLM responses under a rephrased chunk can stem from two different scenarios: (i) the chunk is topically relevant but lacks the answer, leading to uncertainty due to incomplete information, or (ii) the chunk is weakly relevant(maybe total irrelevant) and contains noise, which misleads the model and introduces spurious uncertainty. To differentiate these two scenarios, we perform an ablation generation by masking each uncertain chunk:

$$ar_i = f_\theta(q, [c_1, ..., c_{i-1}, c_{i+1}, ..., c_k]), \tag{9}$$

where ar_i denotes the response when chunk c_i is absent. Then we employ the entailment in Equation 5 to evaluate whether chunk c_i is necessary:

$$l_i^{ne} = \mathbb{1}_{\{[NLI(r_0,ar_i)+NLI(ar_i,r0)]/2=0\}}, \tag{10}$$

where r_0 is the response under original context. If removing an *uncertain chunk* changes the model's answer, it implies that the chunk carries critical information, and we classify it as a *necessary chunk*. Conversely, if its removal does not affect the answer, the chunk is *unnecessary*, as it is either irrelevant or redundant. Through this process, DENSE together with ablation allows us to classify each chunk c_i into three types: *certain* ($l_i^{ce} = 1$), *necessary* ($l_i^{ce} = 0, l_i^{ne} = 1$), and *unnecessary* ($l_i^{ce} = 0, l_i^{ne} = 0$).

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. The refinement proceeds iteratively and after each update, we recompute DENSE and stop once either (a) DENSE falls below 0.2, indicating LLM understands context with certainty, or (b) the LLM evaluates the context as sufficient in self-evaluation. If neither condition is met after all candidate chunks are explored, we fall back to the subset of chunks yielding the lowest DENSE score. The self-evaluation prompt is described in Appendix B.3 and the complete pseudocode is provided in Appendix E.3.

6 Experiments

In this section, we conduct experiments to demonstrate the effectiveness of DENSE-RAG and analyze the contributions of each component. First we validate DENSE as an indicator of contextual understanding in Section 6.2. In Section 6.3, we evaluate how DENSE-RAG improves QA performance on uncertain questions across different LLM backbones as well as comparing with other baselines. Section 6.4 presents ablation studies to examine the design choices of adaptive semantic chunking and iterative context refinement. We also include additional robustness demonstration, sensitivity analyses and case studies in Appendix J and Appendix L.

6.1 Experiment Setup

Datasets. We test our methods on open-book QA datasets, which require free-form answers: TriviaQA Joshi et al. (2017), Natutal 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, Llama-3.1 8B, and Llama-3.1 70B, using the documents provided by each dataset as the retrieval corpus. A vanilla RAG pipeline is built with recursive chunking (chunk size = 512) 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 number of chunks in Appendix J. Details of each component and the QA prompt are provided in Appendix B.4 and G.

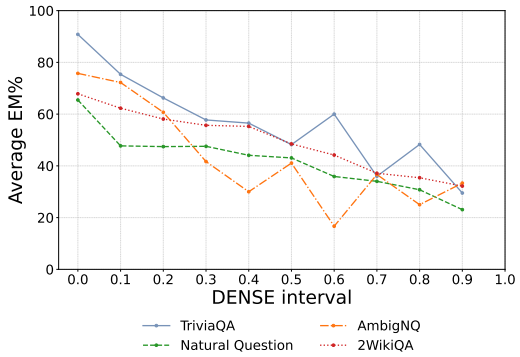

Figure 2: Average EM across different DENSE intervals. On four Open-book QA tasks, average EM decreases as DENSE increases.

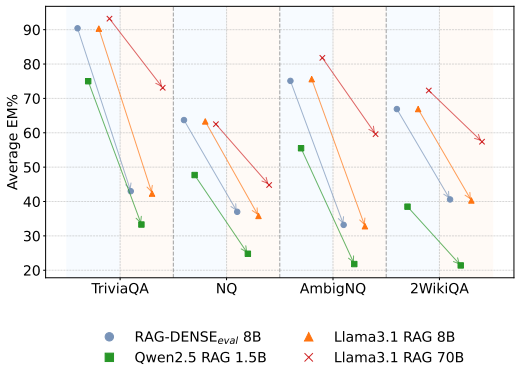

Figure 3: Exace match on certain (blue: DENSE ≤ 0.2) and uncertain (ivory: DENSE > 0.2) questions across LLM backbones of different scales.

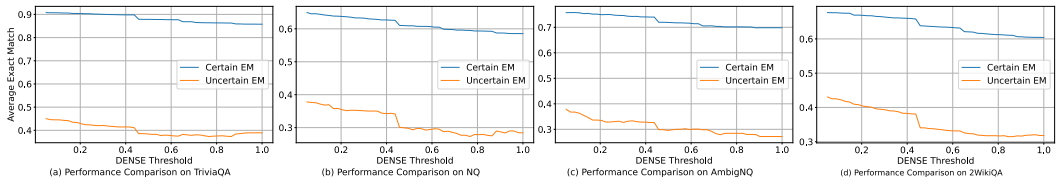

Figure 4: Average exact match on certain vs. uncertain questions under different DENSE thresholds, showing a consistent performance gap between the two groups.

6.2 DENSE AS A MEASURE OF CONTEXT UNDERSTANDING

We first demonstrate that DENSE is an effective way to quantify LLM's understanding of retrieved context. Followed prior work Kuhn et al. (2023) settings, when the LLM understands the semantically equivalent contexts, the responses tend to be *more consistent*, and are more likely to be *correct*. We compute DENSE on a Llama-3.1 8B vanilla RAG, and present the average exact match within different DENSE intervals in Figure 2. The results show that average exact match decreases as DENSE increases, confirming that higher semantic uncertainty corresponds to lower QA accuracy and that **DENSE provides an effective unsupervised measure of LLMs' contextual understanding**.

To verify that performance drops occur across various LLM backbones, we conduct RAG experiments on Qwen-2.5 1.5B, Llama-3.1 8B and Llama-3.1 70B without DENSE, comparing their performances on certain and uncertain questions in Figure 3. The consistent performance drop confirms that DENSE provides a reliable measurement. The comparison between RAG-DENSE$_{eval}$ 8B and Llama3.1 RAG 8B in Figure 3 shows that rephrasing in DENSE has negligible impact on QA performance, which verifies that our rephrasing process does not cause semantic drift in the chunks. Additional results, including experiments on summarization datasets, are presented in Appendix H.

6.2.1 ROBUSTNESS OF DENSE UNDER DIFFERENT THRESHOLDS

We evaluate the robustness of DENSE by testing multiple thresholds for separating certain/uncertain questions, as shown in Figure 4. Across all thresholds, the performance gap between the two groups remains significant, confirming that DENSE measurement is stable and effective. For our main experiments, we adopt 0.2 as the default threshold, supported by the observation from Figure 2 that across all datasets, the average exact match decreases monotonically as DENSE increases when $DENSE \leq 0.2$. In different applications, the choice of threshold can be flexible. In domains requiring higher certainty, such as healthcare or law, a lower threshold enforces more certain outputs but classifies more questions as uncertain, triggering chunking and refinement more frequently. Higher thresholds reduce computation at the cost of tolerating greater semantic variability.

Table 1: Experimental result on **uncertain questions** on 4 datasets. Each component shows improvement across LLMs of different scales. Chunking and Refinement denotes the proposed adaptive semantic chunking and iterative context refinement modules.

Backbone	DENSE 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-3.1 8B			42.26	35.78	32.27	40.30
	✓		52.41	41.31	40.45	42.34
	✓	✓	**56.32**	**42.16**	**43.63**	**46.55**
Llama-3.1 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**

6.3 DENSE-RAG QA PERFORMANCE

DENSE-RAG is effective on uncertain questions. To evaluate the effectiveness of DENSE-RAG, we progressively incorporate adaptive semantic chunking and iterative context refinement into RAG pipeline. The results on uncertain questions are summarized in Table 1. We observe consistent gains as each module is added, with the full DENSE-RAG achieving improvements across uncertain question. For Qwen-2.5 1.5B, the limited parameter size makes it inherently difficult to handle multi-hop reasoning, which is also evident when compared with other LLMs. In contrast, Llama-3.1 70B is already very strong, so the gain on AmbigNQ is marginal.

DENSE-RAG achieves competitive performance against SOTA RAG. We compare DENSE-RAG with state-of-the-art baselines in Table 2. At the 8B scale, DENSE-RAG achieves performance comparable to finetuned systems such as RankRAG Yu et al. (2024). Although RAG-DDR outperforms DENSE on TriviaQA, it is an end-to-end trained framework, whereas DENSE-RAG requires no additional training and can be flexibly integrated into diverse RAG applications. Results on all five backbones and more baseline comparisons are provided in Appendix I.

6.4 ABLATION STUDY

To better understand the contributions of individual designs in DENSE-RAG, we conduct a set of ablation studies and comparative experiments. We focus on two main aspects: (i) the impact of different chunking strategies on performance under certain and uncertain questions, and (ii) the effectiveness of the iterative context refinement and its key components under varying configurations.

6.4.1 ANALYSIS OF CHUNKING STRATEGIES FROM UNCERTAINTY PERSPECTIVE

Chunk size introduces a natural trade-off: larger chunks preserve more information, while smaller ones reduce noise Zhang et al. (2025). Although semantic chunking has been proposed to improve coherence, prior work reports inconsistent gains compared to fixed-size chunking Qu et al. (2024). To examine this issue from an uncertainty perspective, we compare: (1) fixed-size recursive chunking with 256/32 and 128/16 settings, and (2) universal semantic chunking. Experiment results on certain and uncertain questions are shown in Table 3, which leads to the following interesting findings:

Uncertain questions benefit from smaller chunks. Reducing chunk size improves performance on uncertain questions but reduces accuracy on certain ones (Table 3). For certain questions, larger chunks maintain robustness by providing sufficient context despite added noise. In contrast, for uncertain questions, smaller chunks help filter irrelevant information, yielding marginal gains.

Semantic chunking works on uncertain questions. Semantic chunking improves uncertain questions but degrades certain ones, consistent with prior findings Qu et al. (2024). For questions already well-answered, semantic chunking restricts information diversity and limits performance.

Table 2: Comparison of DENSE-RAG with baselines. Ret. FT and Gen. FT indicate whether the retriever and generator of the method were fine-tuned, respectively.

Method	Generator Model	Ret. FT	Gen. FT	TriviaQA	Natural Question	AmbigNQ	2WikiQA
RePlug-LSR (few-shot)Shi et al. (2023)	Codex 175B	✓	✗	77.3	45.5	-	-
ChatQA-1.5Liu et al. (2024b)	Llama3 8B	✓	✓	81.0	42.4	-	26.8
UncertaintyRAGLi et al. (2024b)	Llama2 13B	✓	✗	82.5	-	-	38.3
ERM4Shi et al. (2024)	GPT-3.5-turbo	✓	✗	-	52.7	53.5	46.8
Astute-RAGWang et al. (2024b)	Claude 3.5 Sonnet	✗	✗	84.5	53.6	-	-
RankRAGYu et al. (2024)	Llama3 70B	✗	✓	85.6	54.2	-	38.2
RAG-DDRLi et al. (2024a)	Llama3 8B	✓	✓	89.6	52.1	-	-
DENSE-RAG	Llama3 8B	✗	✗	84.8	57.5	68.1	58.7
	Llama3.1 70B	✗	✗	**90.3**	**58.2**	**76.8**	**69.3**

Table 3: Experiment result on certain/uncertain questions with various chunk strategy.

Chunking	TriviaQA	Natural Question	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 4: Experiment result of iterative context refinement on uncertain questions.

Refinement	TriviaQA	Natural Question	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**

Our adaptive method applies semantic merging only when DENSE indicates high uncertainty, thereby improving uncertain question performance while preserving the advantages of fixed-size chunking on certain ones. Beyond the empirical gains, this observation provides an uncertainty-based explanation for the controversial effectiveness of semantic chunking reported in previous work.

6.4.2 Analysis of iterative context refinement

We further analyze iterative context refinement through the following settings: (1) no refinement after DENSE, (2) only removing unnecessary chunks, (3) disabling the self-evaluation condition, and (4) refilling based on certain chunks instead of necessary chunks. Table 4 summarizes the results.

Notably, removing unnecessary chunks outperforms the baseline (w/o refine) on TriviaQA and 2WikiQA, with only a minor drop (1%) on Natural Questions and AmbigNQ. This confirms that DENSE effectively identifies and filters irrelevant documents. Disabling self-evaluation leads to consistent drops, showing its usefulness in preventing contexts from generating consistently incorrect responses. Refilling based on certain chunks performs second-best, suggesting that adding information similar to certain chunks can indeed improve QA performance, but the gains are limited compared to refilling guided by necessary chunks.

7 Conclusions

In this work, we explore a fundamental problem in RAG: how to assess whether LLMs understand the retrieved context. We introduced DENSE, a training-free and model-agnostic method that quantifies semantic uncertainty through responses generated under semantically equivalent contexts. Our analysis shows that higher DENSE values consistently correspond to worse performance, validating its effectiveness as an unsupervised measure of contextual understanding. Building on this insight, we designed two modules—Adaptive Semantic Chunking and Iterative Context Refinement—to enhance both intra-chunk semantic coherence and inter-chunk semantic completeness for uncertain questions. Extensive experiments across multiple datasets and backbones demonstrate that DENSE-RAG delivers competitive or superior performance compared to state-of-the-art methods, while requiring no additional training.

While DENSE eliminates the need for training or access to model internals, it introduces extra inference calls. However, such interactions are necessary in a strictly black-box setting where no auxiliary models are introduced. The detailed time complexity is discussed in Appendix K. Future work could explore adaptive integration of smaller models for simpler tasks to reduce inference costs. Beyond our method on improving context quality, another promising direction for future work is to enhance LLMs' ability to interpret retrieved texts, for example by incorporating uncertainty-aware training objectives during pretraining or finetuning, which may further strengthen QA performance.

Ethics statement. Our work focuses on contributions to Retrieval-Augmented Generation (RAG) and does not involve human subjects, private data, or personally identifiable information. All experiments are conducted on publicly available open-book QA datasets, following their respective licenses and intended use.

Reproducibility statement. We have made extensive efforts to ensure the reproducibility of our work. The full implementation of our methods, along with detailed instructions for running the experiments, is provided in the anonymous link as well as uploaded Supplementary Materials.

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.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pp. 194–199, 2024.

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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 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.

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.

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.

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.

Jie Huang, Wei Ping, Peng Xu, Mohammad Shoeybi, Kevin Chen-Chuan Chang, and Bryan Catanzaro. Raven: In-context learning with retrieval-augmented encoder-decoder language models. *arXiv preprint arXiv:2308.07922*, 2023.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

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.

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*, 2024a.

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*, 2024b.

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.

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

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, 2024b.

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

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.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. 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. 338–354, 2024.

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.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. Rag-hat: A hallucination-aware tuning pipeline for llm in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1548–1558, 2024.

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, 2024a.

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. *arXiv preprint arXiv:2410.07176*, 2024b.

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*, 2025.

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 THE USE OF LARGE LANGUAGE MODELS (LLMS)

During the preparation of this manuscript, we employed tools such as GPT and Grammarly for language polishing and for assisting in literature search. We emphasize that no part of this work relies on unverified or irresponsible LLM-generated content, and the authors take full responsibility for all contents of the paper.

B PROMPT FORMATS

B.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}

LLM: ...
```

B.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.

LLM: ...
```

B.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.

LLM: ...
```

B.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}

LLM: ...
```

C 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. To prove the disentanglement works, we run experiment using Llama3-8B on TriviaQA, NQ, AmbigNQ and 2WiKiQA under greedy decoding and fixed context. 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 successfully eliminates randomness-induced variability from the LLM itself. And the observed uncertainty in DENSE is originated from rephrased contexts. Lin et al. (2024) explored the relationship between decoding temperature and uncertainty estimation. For more details on this topic, we refer readers to their work.

D DISCUSSION OF SEMANTIC ENTROPY AND DENSE

Kuhn et al. (2023) define the semantic entropy to measure the uncertainty of LLM's responses:

$$SE(q, C) \approx -|H|^{-1} \sum_{i=1}^{|H|} \log p(h|C), \tag{11}$$

where h is a semantic cluster belongs to $H = \{h_1, h_2...\}$, $p(h|x)$ estimates a categorical distribution over the cluster meanings. The cluster is computed by bi-directional entailment E_{r_i,r_j}. A natural language inference model (NLI) is used to classify whether r_i and r_j are *entailment(1)* or *neutral(0)*:

$$E_{r_i,r_j} = (NLI(r_i, r_j) + NLI(r_j, r_i))/2, \tag{12}$$

a Deberta-large model is employed to make the classification. r_i and r_j are clustered together into one semantic cluster h when $E_{r_i,r_j} = 1$.

We discuss the relationship of semantic entropy and the degree-based semantic entropy in DENSE from three perspectives:

Monotonicity: Both formulations decrease monotonically as semantic consistency among responses increases. In semantic entropy, diverse semantic lead to more clusters, and $p(H|x)$ decrease, the entropy increase. In our method, diverse semantic lead to smaller D, the entropy also increase.

Range Analysis: We now show that the two formulations share the same value range by analyzing two extreme cases. In the ideal case, where all responses are semantically equivalent, and there will be only one cluster in equation 11:

$$SE(x) \approx -1^{-1} \log 1 = 0 \tag{13}$$

And in our method, the D of all response will be $k + 1$:

$$DSE(x) \approx -(k+1)^{-1} \sum_{i=0}^{k+1} \log 1 = 0 \tag{14}$$

In the ideal case, two equations are equivalent. In the worst case, where all response shows distinct semantic, there are $k + 1$ clusters in equation 11 and each $p(C_i|x) = \frac{1}{k+1}$:

$$SE(x) \approx -(k+1)^{-1} \sum_{i=1}^{k+1} \log \frac{1}{k+1} = -(k+1)^{-1}(k+1) log(k+1)^{-1} = \log(k+1) \tag{15}$$

In our method, the $D = 1$ for all response, since every response only semantically equal to itself:

$$DSE(x) \approx -(k+1)^{-1} \sum_{i=1}^{k+1} \log \frac{1}{k+1} = \log(k+1) \tag{16}$$

Hence, two methods have a value range of $[0, log(k + 1)]$, and have the same value of both ideal case and the worst case.

Information-Theoretic Interpretation. From an information-theoretic perspective, semantic uncertainty is derived from Shannon entropy over the distribution $\{p(h_i|x)\}$, where h_i denotes a semantic cluster of responses. Intuitively, $p(h_i|x)$ represents the probability that a randomly sampled response falls into cluster h_i.

Our method avoids the explicit clustering step by directly considering the semantic graph, where each node is a response and edge weights represent pairwise semantic entailment. The degree D_i measures how semantically connected a response is to all others — in other words, it approximates how many responses are semantically similar to r_i.

E ALGORITHMS

E.1 DENSE

We present the DENSE algorithm in Algorithm 1.

Algorithm 1: DENSE

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 semantic entropy DSE, reply set R, semantic matrix W

$R \leftarrow \emptyset$
$W \leftarrow [0]_{k+1 \times k+1}$
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**
  $C_i \leftarrow [c_1, \ldots, c_i', \ldots, c_k]$
  $r_i \leftarrow f_\theta(p_q(q, C_i))$
  $R \leftarrow R \cup \{r_i\}$
foreach $r_i \in R$ **do**
  foreach $r_j \in R$ **do**
    $w_{ij} \leftarrow (NLI(r_i, r_j) + NLI(r_j, r_i))/2$
    $W[i, j] \leftarrow w_{ij}$
$DSE \leftarrow -(k+1)^{-1} \sum_{i=1}^{k+1} \log \frac{D_i}{k+1},\ where\ D_i = \sum_{j=0}^{k} w_{ij}$
return DSE, R, W

E.2 ADAPTIVE SEMANTIC CHUNKING

We present the adaptive semantic chunking Algorithm 2.

E.3 ITERATIVE CONTEXT REFINEMENT

We present the Iterative Context Refinement algorithm in Algorithm 3.

F DATASETS

We describe the open-book QA dataset here. Since all proposed methods requires no training, we only use the *dev* sets for evaluation. The statistics of each dataset is shown in Table 5:

- **TriviaQA** Joshi et al. (2017) is a challenging QA dataset that provinding evidence documents. There are two types of questions: Wikipedia and Web. We follow KILT benchmark, only consider Wikipedia cases Petroni et al. (2020) with evidence documents. We use the *wikipedia-dev* set in experiments.
- **Natural Question** Kwiatkowski et al. (2019) is a common-used QA dataset, which is extracted from Wikipedia. The questions are constructed from Google search engine and the provided documents are corresponding Wikipedia pages. We follow KILT benchmark Petroni et al. (2020) and only consider questions for which at least one human annotator has marked a short answer in the documents. We only use the *dev* set in experiments.

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 $\mathcal{C}$

- **AmbigNQ** Min et al. (2020) is a QA dataset proposed in AmbigQA, which is constructed using prompt questions from NQ-OPEN and English Wikipedia as the evidence corpus. In our task, we consider the *singleAnswer* questions in *dev* subset in AmbigNQ.
- **2WikiQA** Ho et al. (2020) is a multi-hop QA dataset, which is designed to test the relationship between two entities. In 2WikiQA, multiple evidence articles are provided for one question. We use the *dev* set in our experiments.

G Implementation Details

We implement a naive RAG framework Gao et al. (2023) on Qwen-2.5 1.5B, Qwen-3 8B, Llama-3 8B, Llama-3.1 8B and Llama-3.1 70B as backbones. We use the documents provided within dataset as the retrieval corpus and employ recursive chunking in Langchain[1]. The chunk size is set to 512 and chunk overlap is set to 64. For document retrieval, we use UAE-Large-V1 as the encoder for both questions and document chunks, which is one of the best zero-shot embedding models in MTEB (eng, v2) leaderboard[2]. Then we employ FAISS[3] to build dense index. To ensure a fair comparison with other baselines and demonstrate that the improvements of our method stem from better document understanding rather than an increased number of documents, we limit the retrieval to the top 5 documents—consistent with the minimum retrieval setting used in most RAG studies. We discuss the impact of different values of top-k on the model's performance in Appendix J. In generation stage, we use a simple prompt which is described in Appendix B.4.

In adaptive semantic chunking, we set the merge threshold $\tau = 0.6$ and use the same encoder in embeddin chunks as the sentence encoder. The maximum chunk length $T_{max} = 512$, consistent with recursive chunking. The NLI agent and the LLM for semantically equivalent context construction are the LLM used in generation stage. For the 1.5B and 8B DENSE-RAG, a single NVIDIA 3090 GPU is enough for embedding and inference. For the 70B DENSE-RAG, we use 2 NVIDIA A100 80GB GPUs for embedding and inference.

[1] https://python.langchain.com/docs/concepts/text_splitters/
[2] https://huggingface.co/spaces/mteb/leaderboard
[3] https://ai.meta.com/tools/faiss/

Algorithm 3: Iterative Context Refinement

Input: Question q, $Chunk\ Set$, k, LLM f_θ, self evaluation prompt p_e,
Output: $Answer_{best}$

Initialization:
$C \leftarrow$ initial top-k chunks for q
$DSE_{min} \leftarrow \infty$, $Flag_{current} \leftarrow 0$
while $Visited\ Set \neq Chunk\ Set$ **do**
  $Visited\ Set \leftarrow Visisted\ Set \cup C$
  $DSE, R, W \leftarrow DENSE(q, C)$
  $Flag_{eval} \leftarrow f_\theta(p_e(q, C))$
  if $DSE < DSE_{min}$ **then**
    if $Flag_{current} = 0 \vee Flag_{eval} = 1$ **then**
      $Answer_{best} \leftarrow R$
      $DSE_{min} \leftarrow DSE$
      $Flag_{current} = Flag_{eval}$
  if $DSE_{min} < 0.2\ \ \&\ \ Flag_{current} = 1$ **then**
    Return: $Answer_{best}$;
  $Visited\ Set \leftarrow \emptyset$, $Certain\ Set \leftarrow \emptyset$, $Uncertain\ Set \leftarrow \emptyset$, $Necessary\ Set \leftarrow \emptyset$, $Unnecessary\ Set \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** k **do**
    if $W_{i0} = 1\ \ \&\ \ W_{0i} = 1$ **then**
      $Certain\ Set \leftarrow Certain\ Set \cup \{c_i\}$;
    else
      $Uncertain\ Set \leftarrow Uncertain\ Set \cup \{c_i\}$;
  for $c_i \in Uncertain\ Set$ **do**
    $AC_i \leftarrow [c_0, ..., c_{i-1}, c_{i+1}, ...]$
    $ar_i \leftarrow f_\theta(q, AC_i)$
    if $E(r_0, ar_i) = 1$ **then**
      $Unnecessary\ Set \leftarrow Unnecessary\ Set \cup \{c_i\}$
    else
      $Necessary\ Set \leftarrow Necessary\ Set \cup \{c_i\}$
  for c_i *in* $Unnecessary\ Set$ **do**
    $C \leftarrow C \setminus \{c_i\}$;
  for c_i *in* $Necessary\ Set$ **do**
    $max\_sim \leftarrow -\infty$
    $best\_chunk \leftarrow \emptyset$
    for c_j *in* $Chunk\ Set \setminus Visited\ Set$ **do**
      $sim \leftarrow cosine\_similarity(c_i, c_j)$;
      if $sim > max\_sim$ **then**
        $max\_sim \leftarrow sim\_score$;
        $best\_chunk \leftarrow c_j$;
    $C \leftarrow C \cup \{best\_chunk\}$, $Visited\ Set \leftarrow Visited\ Set \cup best\_chunk$
Return: $Answer_{best}$

Table 5: Dataset Statistics. The certain and uncertain questions are devided by DENSE in Llama-3.1 8B.

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

H MORE DENSE EVALUATION EXPERIMENTAL RESULTS

Here we show the performance of RAG on certain and uncertain questions when using different LLM in open-book QA in Table 6. Regardless of model size, all LLMs exhibit a significant performance

Table 6: RAG performance on **certain** and **uncertain** questions. The EM% drop on uncertain questions to certain one is reported after the EM on ucnertain question. The split of certain questions and uncertain questions is according to DENSE$_{eval}$.

Task	TriviaQA		Natural Question		AmbigNQ		2WikiQA	
	certain	uncertain	certain	uncertain	certain	uncertain	certain	uncertain
Llama3.1 8B DENSE$_{eval}$	90.4	43.0(47.4↓)	63.7	37.0(26.7↓)	75.1	33.2(41.9↓)	66.9	40.6(26.3↓)
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↓)
Llama3.1 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↓)
Llama3.1 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 7: Experiment result on CNN/DailyMail summarization.

DENSE range	(,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,)
Ave RougeL/10^{-3}	90.47	86.97	86.84	82.40	86.29	81.48	78.56	76.54	80.36	81.48	76.43

drop on uncertain questions. While the drop is mitigated for larger models such as the llama3.1-70B, it still remains around 20%. This demonstrates that DENSE generalizes well across language models of different scales.

H.1 Experiments on Summarization Tasks

We focuses on the open-book QA task, but as a typical free-form language generation task, we also explore its effectiveness on summarization tasks. We add a simple verification experiment on CNN/DailyMail 3.0.0 test set. We build a RAG pipeline for summarization and compute the DENSE score, the RougeL and DENSE score relationship is shown in Table 7. As shown in the figure, higher uncertainty is correlated with lower RougeL score, showing DENSE's potential in measuring LLM's understanding in summarization tasks. 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 DENSE method.

I Compare with more baselines

We show extended comparison in Table 8. We consider following sota RAG methods: Astute RAG Wang et al. (2024b), RePlug Shi et al. (2023), Adaptive-RAG Jeong et al. (2024), shi et al. Shi et al. (2024), LongRAG Zhao et al. (2024), RAG-DDR Li et al. (2024a), ChatQA-1.5 Liu et al. (2024b), RankRAG Yu et al. (2024) and UncertaintyRAG Li et al. (2024b). Among these baselines, some baselines like Yu et al. (2024) employ finetuned LLMs to further optimize the retrieved context;

Table 8: Results of our methods and baselines on 4 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. (2024b)	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
ERM4Shi et al. (2024)	GPT-3.5-turbo	✓	✗	-	52.7	53.5	46.8
UncertaintyRAGLi et al. (2024b)	Vicuna 7B	✓	✗	85.0	-	-	29.9
UncertaintyRAGLi et al. (2024b)	Llama2 13B	✓	✗	82.5	-	-	38.3
RAG-DDRLi et al. (2024a)	Llama3 8B	✓	✓	<u>89.6</u>	52.1	-	-
ChatQA-1.5Liu et al. (2024b)	Llama3 8B	✓	✓	81.0	42.4	-	26.8
ChatQA-1.5Liu et al. (2024b)	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. (2025)	Qwen2.5 3B	✗	✗	-	-	-	67.0
Collab-RAGXu et al. (2025)	Llama3.1 8B	✗	✗	-	-	-	<u>67.2</u>
DENSE-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	<u>68.1</u>	58.7
	Llama3.1 8B	✗	✗	85.1	<u>57.8</u>	67.9	59.1
	Llama3.1 70B	✗	✗	**90.3**	**58.2**	**76.8**	**69.3**

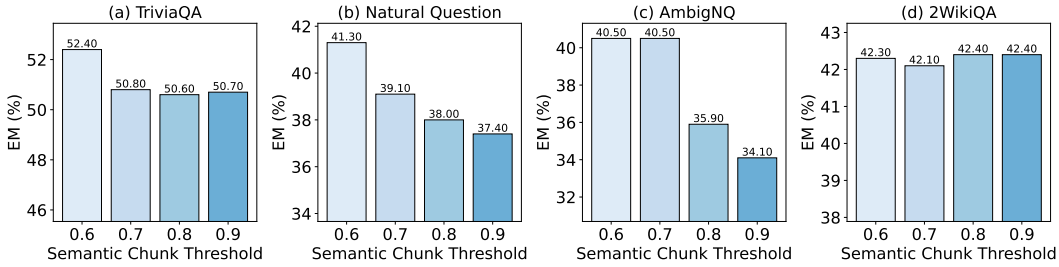

Figure 5: Performance comparison on uncertain questions when using different Adaptive chunking merging thresholds.

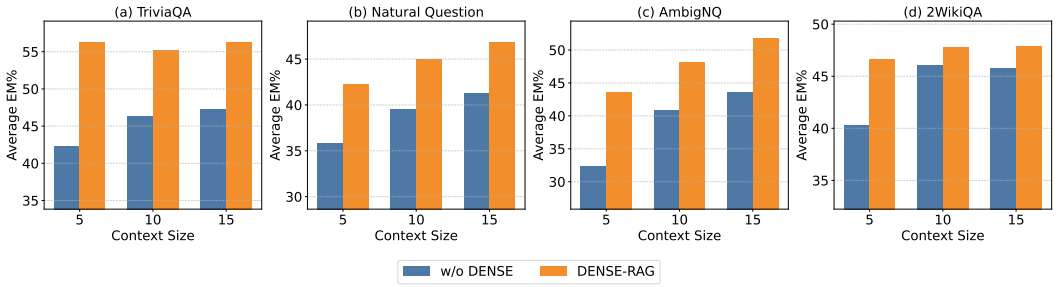

Figure 6: Performance comparison on uncertain questions when using different context size. Under different context sizes, DENSE-RAG demonstrates significant improvements.

in the table, we mark them simply as Gen. FT. Only approaches that introduce additional trained components such as retrievers, encoders, or policy models, are regarded as using trained retrievers. For Collab-RAG Xu et al. (2025), it utilize GPT4o as an LLM reader during the retrieval and generation. It is worth noting that some baselines employ different retrieval settings, such as retrieving a larger number of documents or searching over Wikipedia rather than the dataset-provided corpus; their results are thus reported for reference only. For fair comparison, our method uniformly uses the top-5 retrieved chunks (the minimum number adopted in most prior work) as context and performs retrieval strictly over the dataset-provided knowledge base.

J Additional Sensitivity Analysis

Number of retrieved documents. To discuss the performance of DENSE with different numbers of retrieved chunks, we conduct experiments with different chunk numbers with and without DENSE. As shown in Figure 6, when more chunks are utilized, two methods have better performance on uncertain questions. And DENSE-RAG consistently outperforms RAG w/o DENSE across all datasets and various chunk quantity settings, highlighting the robustness of our approach.

Adaptive chunking threshold τ We run experiments with different adaptive chunking threshold and show the result in Figure 5. As the sentence merge threshold of adaptive chunking increases, performance on uncertain questions significantly declines. This indicates that while leveraging semantic similarity for chunking can enhance performance, overly strict merge conditions may instead lead to a drop in overall effectiveness.

K Discussion of Complexity

As an unsupervised method applicable to any black-box LLM, DENSE requires multiple model calls during measurement. The computational complexity remains bounded by $O(k^2)$, where each LLM inference step (e.g., response generation or entailment check) is treated as $O(1)$. The main steps include: (1) rephrasing k retrieved chunks ($O(k)$); (2) generating responses under rephrased contexts ($O(k)$); and (3) pairwise entailment comparisons ($O(k^2)$). Notably, the entailment component can be efficiently accelerated using lightweight NLI models such as DeBERTa Kuhn et al. (2023), instead

Table 9: 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 DENSE	**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
DENSE-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

of relying on repeated LLM calls. Compared with DSE Kuhn et al. (2023), the only additional cost introduced by DENSE is the rephrasing step.

For semantic chunking, we embed sentences using a compact encoder rather than an LLM, so the overhead is negligible relative to the $O(k^2)$ entailment computations. In the refinement stage, if r chunks are updated, the process involves $O(r)$ generations and $O(r^2)$ entailment checks, giving an overall complexity of $O(r^2)$ (where r < k).

In practice, on Llama-3 with A100 GPUs, computing DENSE adds about one second per query, while the full DENSE-RAG pipeline averages six seconds. This additional cost is modest relative to standard RAG inference and is a reasonable trade-off for enabling reliable, training-free uncertainty measurement and context enhancement in a model-agnostic way.

L CASE STUDIES

We display the case studies of DENSE-RAG on uncertian questions in TriviaQA, Natural Question, AmbigNQ and 2WikiQA. We display the retrieved chunks and LLM responses with/without DENSE in Table 9, Table 10, Table 11 and Table 12. The red text is distractors, and the green text contains evidence for the correct answer. From all the cases, we can observe that these uncertain questions are often accompanied by distracting texts, which can easily mislead the LLM. Take 2WikiQA as an example: LLM needs to first identify the author of Sunday Papers and then locate the awards won by that author. By using DENSE, we effectively identify the chunks that are relevant but do not contain the answer, allowing the chunk refiner to pinpoint the chunk that holds the correct information.

Table 10: 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 DENSE	**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.
DENSE-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 11: Case study on uncertain questions in AmbigNQ.

Question Id: 3008542401703695479	
Q: **Who played robin on the original batman series?** A: Burt Ward	
w/o DENSE	**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
DENSE-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 12: Case study on uncertain questions in 2WikiQA.

Question Id: 0f1ac1a00bdb11eba7f7acde48001122	
Q: **What is the award that the performer of song Sunday Papers earned?** A: Grammy	
w/o DENSE	**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 bandThe 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 award that the performer of song "Sunday Papers" earned is none.
DENSE-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: Joe Jackson earned a Grammy Award nomination.