Logic Matters in Lightweight Hallucination Classification for RAG System

Anonymous EMNLP submission

Abstract

This paper presents a lightweight hallucination classifier specifically designed for Retrieval-Augmented Generation (RAG) systems. To address the inherent limitations of compact models in processing long-context information and performing multi-hop reasoning, our approach systematically analyzes the logical relationships among retrieved documents within the vector space. By capturing these geometric patterns through a novel feature extraction framework, the proposed classifier significantly enhances context-aware hallucination detection without requiring complex architectures or pre-training on datasets. Meanwhile, we find out that all the current benchmark datasets fail to fairly evaluate multi-hop reasoning. To alleviate this issue, we contribute the community a new datase called HotPotQA-derived, a hallucination dataset preserving separate retrieved texts and enabling comprehensive assessment of multi-hop reasoning capabilities. Experimental results on HotPotQA-derived and several open-source datasets demonstrate that our framework can achieve results comparable to or even surpassing those of large language models (LLMs) on the task of hallucination detection.

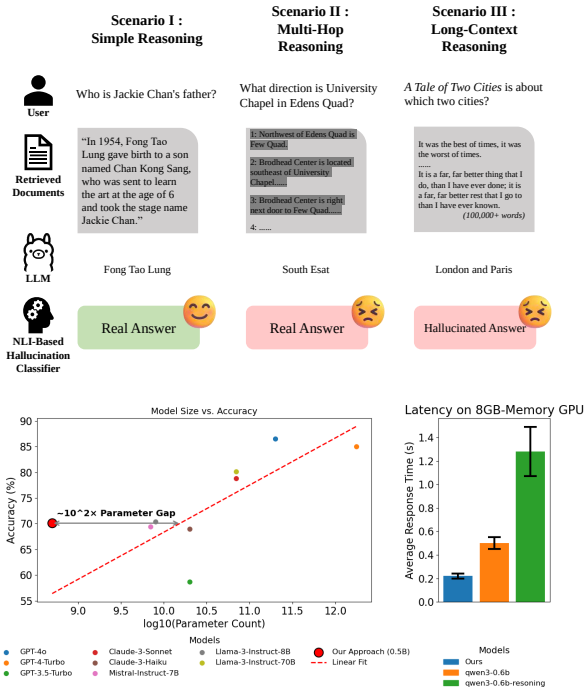

Figure 1: Upper: Dilemmas of NLI-Based Model in Multi-Hop and Long Context Reasoning. Lower: Comparison of our approach with various baselines in terms of size, accuracy, and latency.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has emerged as a powerful strategy for mitigating hallucinations in large language models (LLMs) by grounding their outputs in externally retrieved documents. However, under constrained computational budgets, the retrieval stage itself can introduce new hallucinations: for instance, when asked "Which Japanese city served as the imperial capital during the Heian period?", RAG may retrieve passages describing Kyoto's Heian Shrine and ancient palace grounds, yet the LLM hallucinates "Tokyo" and goes on to describe the Meiji-era Imperial Palace. Such errors arise because a compact model, faced with limited context, may lose track of the precise information in retrieved segments or misinterpret their logical relationship.

Hallucination detection in RAG systems follows two main paths. The first adapts the language model through fine-tuning on hallucination-annotated corpora, yielding strong in-domain performance. In real-world settings, however, developers often cannot access sufficiently large, domain-specific datasets due to privacy restrictions on proprietary content and the high latency of additional validation calls on local hardware. Another way is applying a natural language inference (NLI) model to score factual consistency. While this approach incurs minimal overhead, its effectiveness degrades sharply on tasks that demand understanding long

contexts or performing multi-hop reasoning, as shown in Figure 1. The reason is that the restricted parameter size and network structure of the NLI model cannot maintain a holistic perception of long contexts. Therefore, it does not have the ability to load all retrieved documents into one context window as LLM does, or maintain good recognition effect when a single text is too long.

Prior arts in this field have focused primarily on alleviating the challenges of long contexts. RAGAS (Es et al., 2025) and Provenance (Sankararaman et al., 2024) reduces per-call context length by weighting and classifying each retrieved document independently. A contemporaneous effort, grounded context retrieval (Gerner et al., 2025), further segments and filters individual passages to ease the burden on NLI models when judging factual consistency. Despite these advances, the complexity of the *logical relationship* across multiple retrieved documents remains elusive to lightweight methods, and existing approaches still struggle to detect hallucinations in multi-hop reasoning settings.

In response to these challenges, we introduce a compact, three-module framework specifically designed to bolster multi-hop hallucination detection under tight compute constraints. Our method addresses the limited reasoning capacity of small models by explicitly capturing inter-segment *logical relationship* information before consistency scoring. It comprises:

- **Long Context Segmentation**: splitting both retrieved passages and generated answers into concise, semantically coherent segments;
- **Logical Relation Capture**: embedding each segment into a shared vector space and constructing a segment graph whose weighted edges encode pairwise logical relationships;
- **Consistency Scoring**: grouping related segments via graph traversal and applying a similarity module plus an NLI classifier to each group, then aggregating scores into a global hallucination indicator.

By capturing segments with *logical relationship* scattered across different texts and efficiently combining them to supply the NLI classifier with clear, structured prior information, our framework significantly improves hallucination detection accuracy compared to previous lightweight methods.

To facilitate rigorous evaluation of multi-hop hallucination detection, we also contribute the community a new dataset called HotPotQA-derived, derived based on HotPotQA (Yang et al., 2018a). Existing benchmarks either provide only a single related passage or merge multiple sources into one long context—thus obscuring cross-document links; this naturally favors LLM-based classifiers that load an entire context into the window, and is potentially biased against other implementations. Distinctively, our dataset preserves separate retrieved documents, better reflecting real-world RAG pipelines and enabling comprehensive assessment of multi-hop reasoning capabilities.

Extensive experiments demonstrate that our 0.5B-parameter model, without any task-specific pretraining, surpasses much larger LLM baselines on both HotPotQA-derived (82.4% overall Accu.) and the RAGTruth benchmark (Niu et al., 2023) (72.2% overall F1). Moreover, its small-parameter-scale and low-latency (see Figure 1) inference—achieving response times faster than all baselines—makes it highly practical for deployment in resource-constrained environments.

2 Related Work

Several lines of research have addressed hallucination detection in RAG systems.

Fine-tuning on hallucination corpora. Lynx (Ravi et al., 2024) and RAG-HAT (Song et al., 2024) both improve hallucination detection by fine-tuning language models on annotated hallucination datasets. Lynx also introduces the HaluBench benchmark, which provides a rich set of test examples for measuring detection performance. However, these methods assume that all relevant evidence can be loaded into the model's context window at once and do not evaluate scenarios with multiple retrieved passages—a common case in real-world RAG deployments.

NLI-based consistency checking. The Provenance framework (Sankararaman et al., 2024) first proposed repurposing natural language inference (NLI) models to score factual consistency between generated text and source documents, RAGAS Faithfullness (Es et al., 2025)suggests a similar idea. More recently, the Grounded in Context approach (Gerner et al., 2025) highlights NLI's weaknesses on long, multi-hop contexts and attempts to mitigate them by segmenting text and pruning

non-factual statements to increase inference density. Despite these improvements, existing work does not explicitly model the latent logical relations across multiple segments of retrieved data.

Our contribution. Building on these prior efforts, we introduce an explicit segmentation-and-graph module that decomposes long texts and captures the logical affinities among segments via weighted edges. By reassembling clusters of related passages in topological order and applying NLI-based scoring, our framework complements NLI models' limited long-context reasoning and enables a lightweight, unpretrained 0.5 B–parameter model to achieve detection performance on par with much larger, pretrained LLM baselines.

3 Methodology

3.1 Problem Statement

Designing a lightweight hallucination detection framework for RAG systems entails overcoming three key challenges:

1. **Limited NLI Context Window.** Natural Language Inference models can only process a finite amount of text at once. When the total volume of retrieved passages exceeds the model's context window, essential facts may be omitted, leading to degraded consistency judgments. Therefore, we must devise a strategy to pack the most informative content into each NLI input segment without exceeding length constraints.

2. **Cross-passage Logical Dependencies.** Standard NLI approaches evaluate one passage against the hypothesis at a time, which ignores logical relationships spanning multiple retrieved documents. In many RAG scenarios, verifying a generated claim requires chaining evidence from distinct sources. A lightweight yet effective mechanism is needed to identify and encode these inter-passage dependencies before NLI scoring.

3. **Ambiguity in NLI Scoring.** A low NLI entailment score can arise either from a direct contradiction between hypothesis and premise or from a lack of relevance altogether. Retrieved passages that are unrelated to the generated answer can therefore produce misleadingly low scores, triggering false alarms. Our method must distinguish contradictory evidence from mere irrelevance, ensuring that only genuinely conflicting information drives hallucination detection.

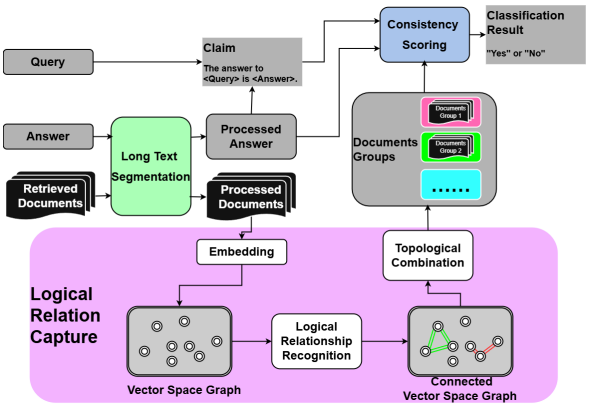

Figure 2: Hallucination Classifier Pipeline

Appendix A gives a few simple examples that reveal the boundaries of the NLI model's capabilities and the challenges it poses to the hallucination detection tasks.

3.2 Method

Our framework comprises three modules: Long Context Segmentation, Logical Relation Capture, and Consistency Scoring. Each module corresponds to one of the previously mentioned challenges respectively. A sketch of its pipeline can be found in Figure 2.

3.2.1 Long Context Segmentation

Let $\ell(s)$ denote the token length of sentence s. Define two thresholds: T_a for answers and T_d for documents. For any text X consisting of sentences $(s_1, \ldots, s_n)$, we define

$$\mathsf{Chunk}(X; T) = \{c_1, \ldots, c_m\}$$

such that each segment $c_k = (s_{i_k}, \ldots, s_{j_k})$ satisfies

$$\sum_{\ell=i_k}^{j_k} \ell(s_\ell) \le T, \quad \sum_{\ell=i_k}^{j_k+1} \ell(s_\ell) > T$$

(or, if $\ell(s_i) > T$, s_i is split into pieces of length at most T).

Answer chunking. Given answer A, if $\ell(A) > T_a$, compute

$$C_a = \mathsf{Chunk}(A; T_a).$$

Let $f(c) \in \{0, 1\}$ be a binary classifier indicating whether segment c is a factual statement. Reassem-

ble the high-density answer

$$A' = \bigoplus_{c \in C_a:\, f(c)=1} c.$$

Document chunking. Given retrieved document set $\{d\}$, for each d, we have:

$$C_d = \begin{cases} \mathsf{Chunk}(d; T_d), & \ell(d) > T_d, \\ \{d\}, & \text{otherwise.} \end{cases}$$

All chunks form $\mathcal{C} = \bigcup_d C_d$.

3.2.2 Logical Relation Capture

For each chunk $c_i \in \mathcal{C}$, it can be embedded into $\mathbb{R}^d$:

$$v_i = E(c_i).$$

Let $N = |\mathcal{C}|$. We can compute pairwise distances

$$d_{ij} = \|v_i - v_j\|_2, \quad \mu = \frac{2}{N(N-1)} \sum_{i<j} d_{ij}.$$

For constant $\alpha > 0$, we then define initial edge set

$$E^{(0)} = \{(i,j) \mid d_{ij} \le \alpha\, \mu\}, \quad w_{ij}^{(0)} = d_{ij}.$$

On graph $G^{(0)} = (\{1, \ldots, N\}, E^{(0)}, w^{(0)})$, we compute all-pairs shortest paths P_{ij}. For each edge $e \in E^{(0)}$, let

$$f_e = \big|\{(i,j) : e \in P_{ij}\}\big|, \quad w_e^{(1)} = f_e.$$

We sort edges by $w^{(1)}$ descending. We also Initialize clusters as

$$\mathcal{S} = \{\{i\}\}_{i=1}^{N}.$$

For each edge $e = (i,j)$ in order, let $S_p, S_q \in \mathcal{S}$ be the clusters containing i, j. Denote

$$\tau(S) = \sum_{k \in S} \text{tokens}(k)$$

as the total token count of cluster S. If

$$\tau(S_p) + \tau(S_q) \le T_t,$$

we then merge S_p and S_q:

$$S_{\text{new}} = S_p \cup S_q, \quad \mathcal{S} \leftarrow (\mathcal{S} \setminus \{S_p, S_q\}) \cup \{S_{\text{new}}\}.$$

Finally, for each cluster $S_k \in \mathcal{S}$, we can form

$$D_k = \bigoplus_{i \in S_k} c_i.$$

3.2.3 Consistency Scoring

For each grouped document D_k and reassembled answer A':

$$r_k = R(D_k, A'), \quad \tilde{r}_k = \frac{r_k}{\sum_{j=1}^{K} r_j},$$

we construct claim H from query Q and A', and compute entailment probability

$$e_k = \text{NLI}(D_k, H).$$

With Scoring threshold T_s, we define the overall score as

$$S = \sum_{k=1}^{K} (\tilde{r}_k * e_k).$$

If $S > T_s$, the output is classified as *no-hallucination*, otherwise as *hallucination*.

3.3 Rationale

Logical Relation Capture can be expected to be effective for the following reasons.

Community-Bridge Intuition

The method is inspired by the Girvan-Newman algorithm (Słoczyński, 2020) in community discovery algorithms. The algorithm detects communities by iteratively removing edges with highest betweenness. We invert this insight: edges that appear on many shortest paths (i.e. have high betweenness f_e) serve as *bridges* between semantic communities. By ranking edges with

$$f_e = \big|\{(i,j) : e \in P_{ij}\}\big|,$$

we prioritize connections that link distinct clusters of chunks.

Filtering Irrelevant Chunks

An edge $(i,j) \notin E^{(0)}$ whenever $d_{ij} > \tau$, so only semantically related chunks satisfy $d_{ij} \le \tau$. This strict threshold prevents the merging of unrelated text segments.

Avoiding Redundant Merges

Because Euclidean distance respects the triangle inequality

$$d_{ij} \le d_{ik} + d_{kj},$$

intra-community edges within a dense cluster rarely lie on shortest paths between other nodes, yielding low f_e. Consequently, our procedure selects only true inter-community bridges, avoiding meaningless merges of highly similar segments.

Together, these properties ensure that Logical Relation Capture merges only those chunks that genuinely bridge separate semantic "communities," accurately reflecting the underlying logical relationships in the retrieved text.
Experiment data can demonstrate the effectiveness of the methodology, an example of which is given in Appendix B.

4 Experiment

4.1 Benchmark Datasets

We evaluate our framework on four benchmarks designed for hallucination detection with long-context and multi-hop reasoning. Detailed data size is listed in Appendix C:

RAGTruth (Niu et al., 2023) A large-scale dataset specifically for evaluating hallucinations in RAG systems. It comprises 2,965 query instances, each with six distinct LLM-generated answers, yielding a total of 17,790 responses. One answer per instance is reserved for testing, while the remaining five serve as training data.

HaluBench (Ravi et al., 2024) A comprehensive hallucination detection benchmark comprising 13,867 samples drawn from six different source corpora. Each instance includes a context paragraph, a question based on that context, an LLM-generated answer, and a binary label (PASS for faithful, FAIL for hallucination). HaluBench covers a variety of domains—general knowledge, reasoning, and specialized topics such as finance and healthcare—and includes particularly challenging "hard-to-detect" hallucinations that appear plausible but are contextually unfounded.

HaluEval (Li et al., 2023) This dataset contains two parts totaling 35,000 samples:

- *Manual subset* (5,000): drawn from 52,000 Alpaca-style prompts, with ChatGPT-generated responses; the 5,000 with lowest response similarity were annotated by 30 trained annotators for hallucination.
- *Automatic subset* (30,000): randomly sampled across QA, knowledge dialogue, and summarization tasks (10,000 each), then processed through a "sampling–filtering" pipeline using ChatGPT to generate diverse hallucinations and an instruction-enhanced filter to select the most challenging examples.

HotPotQA-derived Multi-hop Hallucination Benchmark Building on the multi-hop QA dataset HotPotQA (Yang et al., 2018b), we synthesized a specialized hallucination detection set to evaluate models' ability to spot inconsistencies across separate retrieved passages. Motivated by the need to benchmark multi-hop reasoning—which existing single-passage hallucination datasets cannot adequately assess—we generated three levels of "bridge" hallucinated answers ("bridge-easy," "bridge-medium," "bridge-hard"), yielding 14,282, 45,863, and 12,246 examples respectively (72,391 in total). During benchmarking, each question is paired, at random, with either its faithful (ground-truth) answer or one of its hallucinated variants; the classifier then predicts whether the answer is hallucinated. We report accuracy as the proportion of correct hallucination-vs-truth judgments over all examples. Prompts used for data generation and evaluation are put in Appendix E.

4.2 Experimental Settings

Our framework employs the following off-the-shelf models at different stages:

Vectara Hallucination Evaluation Model (HHEM-2.1-Open) (Bao et al., 2024): assesses factual consistency between generated text and reference documents; 60 M parameters.

Mixedbread AI mxbai-rerank-base-v2 (Lee et al., 2025): re-ranks candidate documents to improve retrieval relevance; 480 M parameters.

Sentence-transformers/all-MiniLM-L6-v (Reimers and Gurevych, 2019): encodes sentences or paragraphs into 384-dimensional dense vectors for semantic search and clustering; 20 M parameters.

We tune several key hyperparameters across different benchmarks: the maximum chunk length C, the segmentation threshold T_a and T_d, and the hallucination scoring threshold T_s. Preliminary experiments on held-out validation splits showed that optimal values vary by dataset, but the following configuration achieves robust performance across all tasks: $C = 256$, $T_a = T_d = 512$, $T_t = 1,024$, $T_s = 0.4$. All reported results use these settings unless otherwise noted.

4.3 Main Results

RAGTruth On the RAGTruth benchmark (Table 1), our method surpasses all baselines that were

Pre-trained	Method	Question Answering			Data-to-Text			Summarization			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
LLM Methods													
✓	Finetuned Llama-2-13B	61.6	76.3	68.2	85.4	91.0	88.1	64.0	54.9	59.1	76.9	80.7	78.7
✓	RAG-HAT	76.5	73.1	74.8	92.9	90.3	91.6	77.7	59.8	67.6	87.3	80.8	83.9
✓	Luna	37.8	80.0	51.3	64.9	91.2	75.9	40.0	76.5	52.5	52.7	86.1	65.4
✗	$\text{Prompt}_{\text{gpt-3.5-turbo}}$	18.8	84.4	30.8	65.1	95.5	77.4	23.4	89.2	37.1	37.1	92.3	52.9
✗	$\text{Prompt}_{\text{gpt-4-turbo}}$	33.2	90.6	45.6	64.3	100.0	78.3*	31.5	97.6	47.6	46.9	97.9	63.4
✗	$\text{SelfCheckGPT}_{\text{gpt-3.5-turbo}}$	35.0	58.0	43.7	68.2	82.8	74.8	31.1	56.5	40.1	49.7	71.9	58.8
✗	$\text{LMvLM}_{\text{gpt-4-turbo}}$	18.7	76.9	30.1	68.0	76.7	72.1	23.3	81.9	36.2	36.2	77.8	49.4
NLI Methods													
✗	Provenance	17.8	100.0	30.2	64.3	100.0	78.3*	23.8	81.4	36.8	36.2	96.0	52.6
✗	RAGAS Faithfulness	18.1	63.1	28.1	66.0	89.7	76.0	20.4	66.5	31.2	27.5	73.4	40.0
✗	Our Approach	89.8	82.1	85.8*	39.7	42.4	41.0	80.7	67.2	73.7*	75.8	69.0	72.2*

Table 1: Response-level hallucination detection on RAGTruth (Niu et al., 2023): comparison of our lightweight, zero-pretraining approach and several approaches presented in RAGTruth (Niu et al., 2023), Luna (Belyi et al., 2024), and RAG-HAT (Song et al., 2024). $*$ represents the best performance in the non-pre-training method, underline represents the best performance in all methods.

not pretrained on this dataset. It attains the best F1 scores in both the Question Answering (85.8%) and Summarization (73.7%) sub-tasks, confirming its strong generalization without dataset-specific training.

HaluBench As the parameter size of RAGAS (Es et al., 2025) was incorrectly recorded in Provenance (Sankararaman et al., 2024), we reimplemented its experiment, more details in Appendix D. On the HaluBench test set (Table 2), our method achieves an overall accuracy of 70.1%, outperforming any other NLI methods and the GPT-3.5-Turbo baseline (62.2%), matching the performance of Claude-3-Haiku and Llama-3-Instruct-8B . A regression analysis of model size versus accuracy for all non-pretrained methods demonstrates that our 0.5 B-parameter model delivers substantial efficiency gains with no loss in detection quality.

HaluEval As shown in Table 3, our approach achieves the highest accuracy on the summarization subset and the second-highest accuracy on QA, yielding overall performance on par with much larger baselines such as ChatGPT and Claude 2, despite requiring no additional fine-tuning.

HotPotQA-derived We compare our model to GPT-3.5-Turbo (OpenAI, 2023), GPT-4o (OpenAI and et al., 2024), Qwen3-0.6B (Alibaba Cloud, 2025), a model with similar size with our approach, and implemented Provenance (Sankararaman et al., 2024) on the HotPotQA-derived hallucination set. Enabling the reasoning mode of Qwen3 increases model response time by more than threefold. Under this configuration, Qwen3-0.6B's la-

Model	Overall Accuracy (%)
LLM Methods	
gpt-4o	87.9
gpt-4-turbo	86.0
gpt-3.5-turbo	62.2
LYNX (70B)	88.4
Llama-3-Instruct-70B	87.0
Claude-3-Sonnet	84.5
LYNX (8B)	85.7
Llama-3-Instruct-8B	83.1
Mistral-Instruct-7B	78.3
Claude-3-Haiku	68.9
NLI Methods	
RAGAS Faithfullness	56.9
Provenance	65.6
Our Approach (0.5B)	**70.1**

Table 2: Overall Accuracy(%) on HaluBench (Ravi et al., 2024), compared with some approaches presented in Provenance (Sankararaman et al., 2024).

Models	QA	Dialogue	Summarization	General
LLM Methods				
ChatGPT	62.59	72.40	58.53	79.44
Claude 2	69.78	64.73	57.75	75.00
Claude	67.60	64.83	53.76	73.88
Davinci002	60.05	60.81	47.77	80.42
Davinci003	49.65	68.37	48.07	80.40
GPT-3	49.21	50.02	51.23	72.72
Llama 2	49.60	43.99	49.55	20.46
ChatGLM	47.93	44.41	48.57	30.92
Falcon	39.66	29.08	42.71	18.98
Vicuna	60.34	46.35	45.62	19.48
Alpaca	6.68	17.55	20.63	9.54
NLI Methods				
RAGAS Faithfullness	61.01	52.79	51.20	53.64
Provenance	67.48	62.97	62.27	56.70
Our Approach	68.49	60.01	59.84	**58.10**

Table 3: Accuracy (%) on HaluEval (Li et al., 2023) across different task types, compared with some approaches presented in Provenance (Sankararaman et al., 2024).

Method	Easy	Medium	Hard	Overall
LLM Methods				
gpt-3.5-turbo	42.3	44.1	43.9	43.7
gpt-4o	86.0	80.1	79.8	81.2
qwen3-0.6b	48.9	48.8	49.1	48.9
qwen3-0.6b-reasoning	69.6	67.0	66.1	67.4
NLI Methods				
RAGAS Faithfulness	55.1	49.9	50.2	51.0
Provenance	54.5	52.2	52.4	52.7
Our Approach	80.6	82.9	82.7	**82.4**

Table 4: Accuracy (%) on the HotPotQA-derived dataset across different task levels.

Module Configuration			Group	Accuracy (%)
A (LC)	B (LR)	C (CS)		
+	+	+	Full Model	82.7*
−	+	+	w/o A	78.2 (-4.5)
+	−	+	w/o B	63.1 (-19.6)
+	+	−	w/o C	59.4 (-23.3)
−	−	+	w/o A+B	55.0 (-27.7)
+	−	−	w/o B+C	51.9 (-30.8)
−	+	−	w/o A+C	52.1 (-30.6)

Table 5: Ablation study results on the HotPotQA-derived *bridge-hard* subset. Modules: (A) Long Context Segmentation (LC), (B) Logical Relation Capture (LR), (C) Consistency Scoring (CS).

tency matches that of other ∼7B-parameter models; accordingly, we include both the reasoning-disabled and reasoning-enabled variants of Qwen3-0.6B alongside in our comparison. As shown in Table 4, our method attains the highest overall accuracy across all baselines. Moreover, accuracy remains stable or even increases from `bridge-easy` through `bridge-hard`, showing that our Logical Relation Capture module effectively models multi-segment logical dependencies and ceases to be the primary bottleneck in hallucination detection.

Latency Evaluation To assess runtime performance under constrained hardware or high-concurrent-stress environment, we measured average response times on an NVIDIA RTX 4060 (8 GB) using HotPotQA-derived. Our model achieves an average latency of 0.22s, vs. 0.50s for `qwen3-0.6b` and 1.28s for `qwen3-0.6b-resoning`, verifying its efficiency under high-concurrency or resource-limited cases.

Remarks In all above comparisons, certain LLM methods may outperform our method. Nonetheless, these LLM methods could not be applicable under high-concurrency or resource-limited scenarios. Distinctively, our novel NLI framework can achieve results comparable to or even surpassing those of LLMs with the minimum latency.

4.4 Ablation Study

We quantify the individual contributions of our three modules on the HotPotQA-derived *bridge-hard* subset. The full model achieves the highest accuracy; Table 5 reports the accuracy after removing each component in turn.

Analysis. Removing *Long Context Segmentation* causes a 4.5-point drop, demonstrating the importance of chunking for capturing critical information. Omitting *Logical Relation Capture* leads to an 19.6-point decrease, confirming that inter-passage reasoning is essential for multi-hop questions. Finally, using equal weights in lieu of our combined relevance–entailment scoring reduces accuracy by 23.3 points. By slightly decreasing the determination threshold T_s, the accuracy gained a slight rebound, but still could not exceed 65%, which is a huge difference from full model. This highlights the necessity of balancing both signals in *Consistency Scoring*.

4.5 Parameter Sensitivity Analysis

We assess the robustness of our method on the HotPotQA-derived *bridge-hard* subset by varying three key hyperparameters and measuring overall accuracy. All results are summarized in Figure 3, which shows these hyperpameters are overall less sensitive.

Chunk Size C. Figure 3(a) plots $A(C)$ for $C \in \{64, 128, 256, 512\}$. Accuracy remains above 75% for $C \geq 128$, but drops by more than 5 points at $C = 64$, indicating that overly fine segmentation fragments essential evidence and degrades multi-hop reasoning.

Segmentation Thresholds. Due to the relevance of the three thresholds T_a, T_d, T_t in the method, we set $T_a = T_d = \frac{1}{2}T_t$. Figure 3(b) shows $A(T)$ for $T \in \{256, 384, 512, 768, 1024\}$. Performance is stable for $384 \leq T \leq 1024$, yet falls below 75% at $T = 256$. This asymmetry demonstrates that over-segmentation (too low T) is a more critical failure mode than handling longer contexts.

Decision Threshold T_s. Figure 3(c) reports $A(T_s)$ for $T_s \in [0.1, 0.9]$. A broad plateau around 80% for $0.3 \leq T_s \leq 0.7$ confirms that our com-

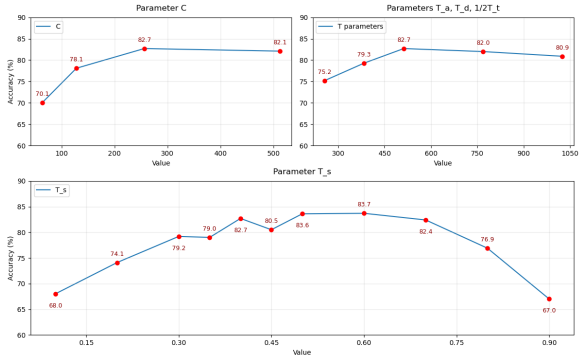

Figure 3: Results on the HotPotQA-derived *bridge-hard* subset with different hyperparameters.

bined relevance–entailment scoring is highly discriminative and robust to threshold selection.

4.6 Discussion

Our results reveal a clear division in where the proposed lightweight framework excels versus where large pretrained models retain the upper hand:

Strength in Strong-Retrieval + Multi-Hop Reasoning On benchmarks such as RAGTruth and the HotPotQA-derived *bridge-hard* subset, the key challenge is to locate and aggregate evidence spread across multiple retrieved passages. Here, our segmentation and graph-based Logical Relation Capture modules directly address the need for multi-hop inference. By explicitly modeling inter-chunk dependencies and combining them before NLI scoring, our 0.5 B-parameter system consistently outperforms even GPT-4o on F1 (85.82 % vs. 83.5 % in QA) and accuracy in complex bridge-hard scenarios. This demonstrates that when retrieval is high-quality and reasoning chains are required, a compact graph + NLI architecture can surpass much larger end-to-end models.

Limitations on Common-Sense and Specialized Knowledge In contrast, on datasets like HaluBench and HaluEval's manual subset—where hallucination errors often hinge on subtle domain facts (e.g. finance, healthcare) or general world knowledge—GPT-4o and other billion-scale pretrained models maintain a significant lead. These tasks demand expansive factual memory and nuanced commonsense reasoning that our purely retrieval-driven pipeline cannot fully provide. Although our framework attains competitive accuracy (70.1% on HaluBench), it does so by relying solely on externally retrieved text, lacking the innate knowledge embedded in a large pretrained model.

Implications and Future Directions These findings suggest a hybrid strategy: leverage lightweight, interpretable graph-based modules for scenarios where high-recall retrieval and explicit multi-hop reasoning are paramount, but integrate or augment with pretrained knowledge sources when domain-specific or commonsense inferences are required. Future work will explore dynamic fusion of internal model priors with external retrieval graphs, and targeted fine-tuning of the NLI component on specialized data to bridge this gap.

5 Conclusion

We have presented a novel, lightweight hallucination detection framework for RAG systems, built around three modular components—Long Context Segmentation, Logical Relation Capture, and Consistency Scoring—and operating on a mere 0.5 B-parameter backbone without any task-specific pretraining. Our extensive evaluation demonstrates four key strengths:

High Parameter Efficiency. On HaluBench, our compact model matches or exceeds the accuracy of models tens of times larger, confirming that careful segmentation and lightweight inference can replace brute-force scale.

Cross-Domain Robustness. Without per-task tuning or fine-tuning, we achieve top-tier performance on the diverse tasks in HaluEval, underscoring the framework's ability to generalize across domains such as QA and summarization.

Zero-Pretraining Generalization. On RAGTruth (Niu et al., 2023), we outperform all baselines that did not see the test data during training, delivering state-of-the-art F1 scores in both QA and summarization sub-tasks and illustrating true out-of-the-box applicability.

Effective Multi-Hop Reasoning. On the HotPotQA-derived, our Logical Relation Capture module enables accurate detection of cross-passage hallucinations, surpassing both similarly sized and larger baselines on multi-context questions.

Together, these results show that our method not only achieves strong detection quality but also combines efficiency, privacy-preserving zero-pretraining, and robust logical reasoning. Future work will explore adaptive thresholding, dynamic graph construction, and integration of more powerful inference engines to further enhance cross-domain and ultra-long-context performance.

Limitation

Coarse-Grained Logical Relations. Our Logical Relation Capture module identifies and merges chunks purely based on aggregate graph-based connectivity, treating all inferred relations as equivalent. It does not distinguish finer-grained logical types such as causation, temporality, or coordination. As a result, the concatenated text segments may not follow a natural or causally coherent order, potentially impeding accurate hallucination detection.

Dependence on NLI Model Accuracy. Our framework's upper bound is constrained by the underlying NLI model's strengths and weaknesses. While models like HHEM excel at verifying explicit premises against hypotheses, they struggle with commonsense inferences or specialized domain knowledge that is not directly stated in the text. Elevating overall detection performance will require more capable inference modules—such as transformer-based multi-task NLI architectures—or targeted fine-tuning on domain-specific and commonsense-augmented corpora.

References

Alibaba Cloud. 2025. Qwen3 model documentation. Accessed: 2025-05-11.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2024. Luna: An evaluation foundation model to catch language model hallucinations with high accuracy and low cost.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2025. Ragas: Automated evaluation of retrieval augmented generation.

Assaf Gerner, Netta Madvil, Nadav Barak, Alex Zaikman, Jonatan Liberman, Liron Hamra, Rotem Brazilay, Shay Tsadok, Yaron Friedman, Neal Harow, Noam Bresler, Shir Chorev, and Philip Tannor. 2025. Grounded in context: Retrieval-based method for hallucination detection.

Sean Lee, Rui Huang, Aamir Shakir, and Julius Lipp. 2025. Baked-in brilliance: Reranking meets rl with mxbai-rerank-v2.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

OpenAI. 2023. Openai gpt-3.5 turbo.

OpenAI and : Aaron Hurst et al. 2024. Gpt-4o system card.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. 2024. Provenance: A light-weight fact-checker for retrieval augmented LLM generation output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1305–1313, Miami, Florida, US. Association for Computational Linguistics.

Adam Słoczyński. 2020. An overview of algorithms for community detection in networks.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. 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*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Premise	Hypothesis	NLI Score
CityA is a capital city.	The capital of CountryA is CityA.	0.0297
CityA is a city of CountryA.	The capital of CountryA is CityA.	0.1169
CityA is a capital city. CityA is a city of CountryA.	The capital of CountryA is CityA.	0.6322

Table 6: NLI scoring examples:how combining multiple premises improves entailment scores (cross-passage logical dependencies).

Premise	Hypothesis	NLI Score
I am a dog.	I am a cat.	0.0230
I am a dog.	You are a cat.	0.0223

Table 7: NLI scoring examples:low scores caused by contradiction versus irrelevance.

A Challenge of NLI model

Tables 6 and 7 reflect the restrictions of the existing NLI model in terms of factual scoring.

B Logical Relation Capture Example

The following example comes from HotpotQA-derived dataset.

```
Query "What football position did the manager of
↪  Tianjin Quanjian previously hold?"
Answer “Captain”
Label “True Answer”
```

The corresponding retrieved documents with their consistent scores and relevant scores are listed in Table 8. In the baseline approach using the NLI model, each text corresponds to a very low consistent score since no single text supports the answer. The final decision threshold is 0.115 and this answer is incorrectly categorized as a hallucinated answer.

After being processed by the Logical Relaton Capture module, the retrieved documents that has a logical relation is recombined and the results are listed in Table 9. The final decision score is 0.457, which is greater than the set threshold of 0.4, at which point the answer is categorized as correct.

C Experiment Dataset

See table 10.

D RAGAS Faithfullness

Specifically, the RAGAS baseline that emerged from the experiments used the method FaithfulnesswithHHEM. Although it used LLM to generate the results during the run, it is closer to an NLI method, both in terms of the method design and the final basis of identification. Meanwhile, the accuracy of the method is not sensitive to LLM. Therefore he was categorized as an NLI method in all subsequent experiments. To be fair, the method will not include a discussion of model size and accuracy.

E Prompts

E.1 Data Generation

In the data generation session, we use the `deepseek-v3` model with TEMPERATURE set to 0. The prompts below are shown in typewriter font, with literal “\n” preserved.

System Prompt:

```
"You are an expert in generating subtly
↪  hallucinated answers."\n
"Your task is to create responses that appear
↪  credible at first glance, but contain"\n
"verifiable factual errors when cross-checked
↪  with the provided golden answer and
↪  context."\n
```

User Prompt:

```
"Question: <question>\n"
"Golden Answer: <answer>\n"
"Context: <context>\n"
"Generate a plausible but factually incorrect
↪  answer that:\n"
"1. Maintains grammatical correctness\n"
"2. Contains subtle factual inconsistencies\n"
"3. Presents logical reasoning flaws\n"
"4. Includes inaccurate numerical/data
↪  references\n"
"5. If Golden Answer answers a question briefly
↪  with a noun or phrase, you should do the
↪  same\n"
```

E.2 Data Evaluation

In the data evaluation session, we give the following PROMPT for the test model and set TEMPERATURE = 0

System Prompt:

```
"You are an expert in verifying hallucination. "
"Please judge if the hallucination exists in the
↪  answer of query given contexts. "
"If hallucination exists, print 'Yes'. Otherwise,
↪  print 'No'."
```

Text	consistent Score	relevant Score
Tianjin Quanjian F.C. () is a professional Chinese football club that currently participates in the Chinese Super League division under licence from the Chinese Football Association (CFA). The team is based in Tianjin and their home stadium is the Haihe Educational Football Stadium that has a seating capacity of 30,000. Their current owners are Quanjian Nature Medicine who officially took over the club on 7 July 2015.	0.008477402850985527	7.5000
Tianjin Haihe Education Park Stadium is a multi-purpose stadium in Tianjin, China. It is currently used mostly for football matches of Tianjin Quanjian. They drew the highest average home attendance in the 2016 China League One (12,165), followed by Guizhou Hengfeng Zhicheng (11,089), Dalian Yifang (10,806) and Shenzhen FC (10,152). The stadium opened in 2011.	0.011761926114559174	3.1250
Zhang Lu (;born 6 September 1987 in Tianjin) is a Chinese footballer who currently plays for Tianjin Quanjian in the Chinese Super League.	0.053623996675014496	2.9375
Li Xingcan (Chinese: ***; born 23 July 1987 in Tianjin) is a Chinese football player who currently plays for Chinese Super League side Tianjin Quanjian.	0.055118858814239500	3.8125
Parma Associazione Calcio regained its respect following a lacklustre Serie A and Champions League performance the year before. Under new coach Cesare Prandelli, Parma played an offensive 4–3–3 formation, in which new offensive signings Adrian Mutu and Adriano starred. Both made up for the departure of Marco Di Vaio to Juventus. Mutu scored 18 goals from the left wing, and Parma accepted a multimillion-pound offer from Chelsea in the summer, which meant the Romanian international only spent a year at the club. Also impressing were goalkeeper Sébastien Frey and young centre-halves Matteo Ferrari and Daniele Bonera, who proved to be acceptable replacements for departed captain Fabio Cannavaro, who had joined Inter in late August 2002.	0.078507773578166960	0.3750
Axel Laurent Angel Lambert Witsel (born 12 January 1989) is a Belgian professional footballer who plays for Chinese club Tianjin Quanjian. During his play for the Belgium national team, he came into the first team as a right-winger, and can also play attacking midfielder, though his natural position is as a central midfielder.	0.067316725850105290	7.1875
Fabio Cannavaro, (] ; born 13 September 1973) is an Italian former professional footballer and current manager of Chinese club Tianjin Quanjian.	0.364537507295608500	8.6875
Quanjian Group Co., Ltd. () is a Chinese herbal medicine company based in Tianjin. The group is the parent company of Quanjian Nature Medicine Technology Development Co., Ltd. () for about 75.36% stake. Quanjian Group is the parent company of Dalian Quanjian F.C., which the group owned 80% stake of the women football club.	0.004174184985458851	2.4375
Tianjin Quanjian F.C. is a professional Chinese football club that currently participates in the Chinese Super League division under licence from the Chinese Football Association (CFA). The team is based in Tianjin and their home stadium is the Haihe Educational Football Stadium that has a seating capacity of 30,000. Their current owners are Quanjian Nature Medicine who officially took over the club on 7 July 2015.	0.007749349344521761	7.4375
Tianjin Tuanbo Football Stadium is a professional football stadium in Tianjin, China. It hosts the home matches of Tianjin Quanjian F.C. of the China League One. The stadium holds 22,320 spectators and opened in 2012.	0.012596431188285350	2.8125

Table 8: Retrieved documents in sample, with respective consistent score and relevant score.

User Prompt:

```
f"Query: <query>\n"
f"Answer: <answer>\n"
f"Context: <context>"
```

F License and Terms for Artifacts

This appendix summarizes the licenses under which we use and distribute external and derived artifacts in this work.

HotPotQA-derived dataset. Our experiments build upon the HotPotQA dataset (Yang et al., 2018b), which is released under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). All questions, contexts, and annotations we derived retain this license; any redistribution of our HotPotQA-derived subset must comply with CC BY-SA 4.0.

HaluBench and HaluEval benchmarks. These benchmarks are constructed by the authors using passages from Wikipedia (licensed under Creative Commons Attribution-ShareAlike 3.0 Unported, CC BY-SA 3.0).

Pretrained models. We evaluate against `qwen3-0.6b` and its reasoning variant as provided by Qwen Foundation. The models are distributed under the Apache License 2.0. Use or redistribution of these model weights must comply with Apache 2.0 terms.

Implementation code and scripts. All code for segmentation, graph construction, clustering, NLI scoring, and evaluation (including the latency measurement scripts) is released under the Apache License 2.0. The repository will contain a `LICENSE` file with the full text of the Apache 2.0 license.

Third-party libraries. We utilize open-source Python packages, including `transformers` (Apache 2.0), `matplotlib` (PSF License), and `adjustText` (MIT License). Users must adhere to each library's respective license for any reuse or modification.

Text	consistent Score	relevant Score
Tianjin Quanjian F.C. () is a professional Chinese football club that currently participates in the Chinese Super League division under licence from the Chinese Football Association (CFA). The team is based in Tianjin and their home stadium is the Haihe Educational Football Stadium that has a seating capacity of 30,000. Their current owners are Quanjian Nature Medicine who officially took over the club on 7 July 2015. Tianjin Tuanbo Football Stadium is a professional football stadium in Tianjin, China. It hosts the home matches of Tianjin Quanjian F.C. of the China League One. The stadium holds 22,320 spectators and opened in 2012.	0.009305392391979694	6.625
Tianjin Haihe Education Park Stadium is a multi-purpose stadium in Tianjin, China. It is currently used mostly for football matches of Tianjin Quanjian. They drew the highest average home attendance in the 2016 China League One (12,165), followed by Guizhou Hengfeng Zhicheng (11,089), Dalian Yifang (10,806) and Shenzhen FC (10,152). The stadium opened in 2011.	0.011761926114559174	3.1250
Zhang Lu (;born 6 September 1987 in Tianjin) is a Chinese footballer who currently plays for Tianjin Quanjian in the Chinese Super League.	0.053623996675014496	2.9375
Li Xingcan (Chinese: ***; born 23 July 1987 in Tianjin) is a Chinese football player who currently plays for Chinese Super League side Tianjin Quanjian.	0.055118858814239500	3.8125
Axel Laurent Angel Lambert Witsel (born 12 January 1989) is a Belgian professional footballer who plays for Chinese club Tianjin Quanjian. During his play for the Belgium national team, he came into the first team as a right-winger, and can also play attacking midfielder, though his natural position is as a central midfielder.	0.067316725850105290	7.1875
Fabio Cannavaro, (] ; born 13 September 1973) is an Italian former professional footballer and current manager of Chinese club Tianjin Quanjian. Parma Associazione Calcio regained its respect following a lacklustre Serie A and Champions League performance the year before. Under new coach Cesare Prandelli, Parma played an offensive 4–3–3 formation, in which new offensive signings Adrian Mutu and Adriano starred. Both made up for the departure of Marco Di Vaio to Juventus. Mutu scored 18 goals from the left wing, and Parma accepted a multimillion-pound offer from Chelsea in the summer, which meant the Romanian international only spent a year at the club. Also impressing were goalkeeper Sébastien Frey and young centre-halves Matteo Ferrari and Daniele Bonera, who proved to be acceptable replacements for departed captain Fabio Cannavaro, who had joined Inter in late August 2002	0.8418206512928009	8.6875
Quanjian Group Co., Ltd. () is a Chinese herbal medicine company based in Tianjin. The group is the parent company of Quanjian Nature Medicine Technology Development Co., Ltd. () for about 75.36% stake. Quanjian Group is the parent company of Dalian Quanjian F.C., which the group owned 80% stake of the women football club.	0.004174184985458851	2.4375
Tianjin Quanjian F.C. is a professional Chinese football club that currently participates in the Chinese Super League division under licence from the Chinese Football Association (CFA). The team is based in Tianjin and their home stadium is the Haihe Educational Football Stadium that has a seating capacity of 30,000. Their current owners are Quanjian Nature Medicine who officially took over the club on 7 July 2015.	0.007749349344521761	7.4375

Table 9: Processed documents of sample, with respective consistent score and relevant score.

Test Set	**Subset**	**# Samples**
HaluBench	—	13,867
HaluEval	Manual	4,507
	Automatic (QA)	10,000
	Automatic (Dialogue)	10,000
	Automatic (Summarization)	10,000
RAGTruth Test	QA	989
	Data-to-Text Writing	1,033
	Summarization	943
HotPotQA-derived	bridge-easy	14,282
	bridge-medium	45,863
	bridge-hard	12,246
Total	—	123,730

Table 10: Sizes of test sets and their subsets used in our experiments.