

DRAG: Distilling RAG for SLMs from LLMs to Transfer Knowledge and Mitigate Hallucination via Evidence and Graph-based Distillation

Jennifer Chen^{*,‡,†}, Aidar Myrzakhan^{*,‡}, Yaxin Luo[‡], Hassaan Muhammad Khan^{‡,‡}

Sondos Mahmoud Bsharat[‡], Zhiqiang Shen^{‡,§}

[‡]VILA Lab, Mohamed bin Zayed University of AI

[†]McGill University

[‡]National University of Science and Technology

Abstract

Retrieval-Augmented Generation (RAG) methods have proven highly effective for tasks requiring factual consistency and robust knowledge retrieval. However, large-scale RAG systems consume significant computational resources and are prone to generating “hallucinated” content from Humans¹. In this work, we introduce DRAG, a novel framework for distilling RAG knowledge from large-scale Language Models (LLMs) into small LMs (SLMs). Our approach leverages evidence- and knowledge graph-based distillation, ensuring that the distilled model retains critical factual knowledge while significantly reducing model size and computational cost. By aligning the smaller model’s predictions with a structured knowledge graph and ranked evidence, DRAG effectively mitigates hallucinations and improves factual accuracy. We further present a case demonstrating how our framework mitigates user privacy risks and introduce a corresponding benchmark. Experimental evaluations on multiple benchmarks demonstrate that our method outperforms the prior competitive RAG methods like MiniRAG for SLMs by up to 27.7% using the same models, preserving high-level efficiency and reliability. With DRAG, we provide a practical and resource-efficient roadmap to deploying enhanced retrieval and generation capabilities in small-sized LLMs. Code is available at <https://github.com/VILA-Lab/DRAG>.

1 Introduction

The development of retrieval-augmented generation (RAG) frameworks has significantly advanced the capabilities of large language models (LLMs) by integrating external knowledge retrieval with generative capabilities. RAG models allow for dy-

namic retrieval of evidence, enhancing both factual accuracy and contextual relevance. However, these frameworks are computationally expensive by maintaining an up-to-date large-scale knowledge base and are primarily designed for large-scale LLMs, making them impractical for smaller LLMs deployed on resource-constrained environments. Furthermore, the hallucination problem, where the model generates plausible-sounding but factually incorrect information, remains a critical challenge even in advanced RAG systems. Addressing these issues is crucial for the effective utilization of LLMs in real-world applications.

In this work, we introduce DRAG (**Distilling RAG**), a novel approach aimed at transferring the knowledge and capabilities of large models to smaller LLMs while simultaneously mitigating hallucination through evidence-based distillation. Our method is motivated by the need to make RAG frameworks more accessible and efficient for smaller models without compromising their ability to retrieve and generate accurate information. By leveraging the retrieval process as a core component of distillation, DRAG provides a structured mechanism to teach smaller LLMs how to ground their outputs in external evidence.

The proposed DRAG framework employs a multi-stage paradigm as in Figure 1. First, it generates associated evidences and knowledge graphs based on the context to the input questions from a large RAG teacher model. Then, it distills the retrieval and generation of knowledge into a smaller student / target model. By aligning the student’s retrieval and generation processes with those of the teachers, DRAG ensures that the student model can effectively mimic the evidence-driven reasoning of the teacher. We further introduce an evidence-based privacy protection mechanism to reduce privacy issues, as an additional use case of our proposed framework. To achieve this, we construct a new benchmark dataset with information leakage. Then, we let

^{*}Equal contribution. Work done while Jennifer and Hassaan visiting VILA Lab at MBZUAI, supervised by Zhiqiang Shen. [§]Corresponding author.

¹Human incorrect answers can pollute RAG’s database.

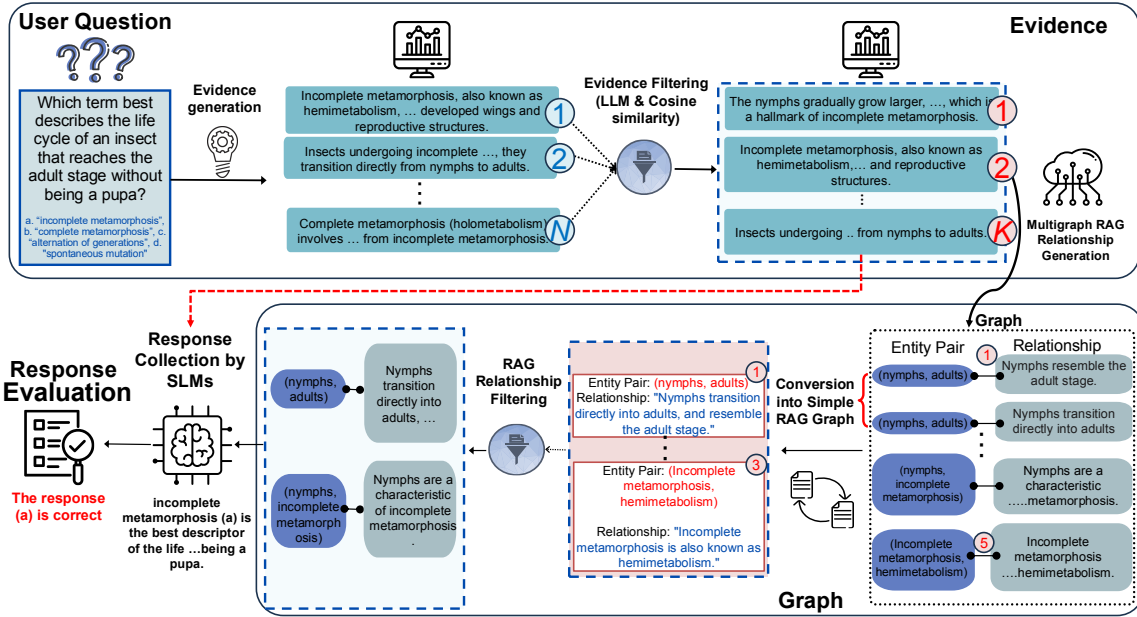


Figure 1: **Framework Overview of Our Evidence- and Graph-based RAG Distillation.** Given a user query (top-left), the approach first retrieves and filters evidence by collecting relevant text snippets. Then, these references are fed into relationship filtering and ranking using an LLM and cosine similarity to yield high-quality ordered references. The resulting multigraph RAG structure is then converted into a simplified RAG graph (bottom-right), distilling crucial relationships and factual context, also extracts key entity pairs and links (e.g., “nymphs→adults”). Finally, SLMs leverage this distilled information, mitigating hallucination and transferring knowledge effectively.

the local target model rephrase input questions before uploading them to a cloud-based large-scale teacher LLM to generate corresponding evidence and knowledge graphs. Finally, the target model utilizes these received evidence and knowledge graphs to produce more reliable and accurate answers.

To evaluate the effectiveness of DRAG, we conduct extensive experiments across various datasets and tasks. Our results demonstrate that DRAG significantly enhances the performance of SLMs in retrieval-augmented tasks by more than 20%, achieving results comparable to their larger teacher models. Moreover, DRAG consistently outperforms baseline RAG methods in mitigating hallucination, as evidenced by improved factual accuracy and reduced error in generated outputs. This advantage stems from the teacher LLM’s ability to generate more relevant and abstract evidence and knowledge graphs, making them easier for SLMs to interpret and utilize effectively. These findings highlight the potential of DRAG to bridge the gap between LLMs and SLMs in a retrieval-augmented setting.

The contributions of this paper are threefold:

- 1) We propose a novel evidence and knowledge graph-based distillation framework for transferring RAG capabilities and mitigating hallucination from large to small-scale LLMs.
- 2) We construct a privacy leakage benchmark

and introduce a privacy mitigation mechanism based on our framework that integrates evidence consistency to demonstrate the additional advantage and strong applicability of our approach.

3) We provide a comprehensive evaluation of DRAG on diverse tasks and datasets, as well as various teacher LLMs and student SLMs, showing its superior ability to balance efficiency, accuracy, and factual consistency.

In summary, by allowing SLMs to harness the strengths of RAG frameworks from a distillation scheme, DRAG opens new opportunities to deploy powerful and reliable LLMs in resource-constrained settings, offering a way for their wider adoption in real-world applications.

2 Related Work

RAG frameworks have been widely explored for tasks requiring factual accuracy and enhanced knowledge retrieval. The foundational work (Lewis et al., 2020) introduced the RAG model, which integrates dense neural retrievers with sequence-to-sequence language models, achieving state-of-the-art performance on knowledge-intensive tasks. Subsequent research has focused on refining the retrieval mechanisms within RAG frameworks (He et al., 2024; Yan et al., 2024a; Wang et al., 2023). For example, Active Retrieval Augmented Gen-

eration (Jiang et al., 2023) introduces a dynamic retrieval strategy that selects information based on the input query, thereby improving generation quality. Similarly, a unified active retrieval approach was proposed to employ multiple criteria for assessing the necessity of retrieval, optimizing the overall generation process (Cheng et al., 2024).

Incorporating structured knowledge into RAG has garnered significant interest in addressing hallucination and enhancing factual grounding. Graph-based methods, such as Graph RAG (Edge et al., 2024), construct an entity knowledge graph from source documents, enabling large language models to handle global questions over entire text corpora. This method enhances query-focused summarization by leveraging graph-based text indices, leading to substantial improvements in the comprehensiveness and diversity of generated answers. Therefore, other graph-based methods have been explored that utilize graph structures to improve both retrieval precision and generative coherence (Hu et al., 2024; Mao et al., 2024; Mavromatis and Karypis, 2024). Similarly, a framework was proposed to align retrieval processes with knowledge graph structures, improving logical consistency in generated responses (Ma et al., 2024).

Recently, ranking-based methods like LambdaMART (Burgess, 2010) with RRF (Cormack et al., 2009) enhance RAG by refining retrieval and reducing hallucinations. However, their effectiveness is limited by small context windows and reliance on synthetic data (Anantha et al., 2023). To overcome this, long-context LLMs have been integrated to handle larger retrieval units, improving both retrieval and generation performance while reducing the retriever’s workload (Zhu et al., 2024). This integration has shown promising gains, particularly in tasks requiring deep contextual understanding (Xu et al., 2023).

Several efforts have preliminarily explored distillation techniques for RAG systems (Izacard and Grave, 2020; Jia et al., 2024; Bezerra and Weigang, 2025). For instance, LLMQuoter (Bezerra and Weigang, 2025) fine-tunes a model using Low-Rank Adaptation (LoRA) on a 15k-sample subset of HotpotQA and employs a *quote-first-then-answer* strategy. In contrast, our method is finetuning-free and more efficient. Our approach focuses on enhancing response quality, factual consistency, and retrieval accuracy by integrating evidence and knowledge graphs.

3 Method

3.1 Preliminaries

Naive RAG. Let \mathcal{X} denote the space of input queries and \mathcal{Y} the space of possible outputs. To enhance the generation process, RAG leverages a large external corpus $\mathcal{D} = \{d_1, d_2, \dots, d_{|\mathcal{D}|}\}$. Given an input query $\mathbf{x} \in \mathcal{X}$, the framework retrieves relevant documents \mathbf{d}_k from \mathcal{D} and uses these retrieved documents to condition the output generation. RAG decomposes conditional output distribution $p(\mathbf{y} | \mathbf{x})$. Formally, the output distribution is represented as:

$$p(\mathbf{y} | \mathbf{x}) = \sum_{\mathbf{d} \in \mathcal{D}} p_{\theta_r}(\mathbf{d} | \mathbf{x}) p_{\theta_g}(\mathbf{y} | \mathbf{x}, \mathbf{d}), \quad (1)$$

where $p_{\theta_r}(\mathbf{d} | \mathbf{x})$ is referred to as the retrieval distribution, which assigns a relevance score to each document \mathbf{d} given the input query \mathbf{x} , and $p_{\theta_g}(\mathbf{y} | \mathbf{x}, \mathbf{d})$ is the generation distribution, which generates the final output \mathbf{y} while attending to both the input \mathbf{x} and the retrieved document \mathbf{d} , where θ represents the model parameters.

Graph RAG. Let $\mathcal{G} = \mathcal{V}, \mathcal{E}$ be a knowledge graph, where each node $v \in \mathcal{V}$ denotes an entity (e.g., a concept or object) and each edge $(v_i, v_j) \in \mathcal{E}$ captures a relationship between entities v_i and v_j . Rather than retrieving individual documents \mathbf{d} from a corpus, Graph RAG seeks relevant *subgraphs* \mathbf{z} within \mathcal{G} to provide structured context for generation. A subgraph \mathbf{z} can be formed by selecting a subset of nodes (and their induced edges) that are topically or semantically related to the input query \mathbf{x} . Formally, the framework factorizes the conditional distribution $p(\mathbf{y} | \mathbf{x})$ as:

$$p(\mathbf{y} | \mathbf{x}) = \sum_{\mathbf{z} \subseteq \mathcal{G}} p_{\theta_r}(\mathbf{z} | \mathbf{x}) p_{\theta_g}(\mathbf{y} | \mathbf{x}, \mathbf{z}), \quad (2)$$

where $p_{\theta_r}(\mathbf{z} | \mathbf{x})$ is the retrieval distribution that assigns a relevance score to each subgraph \mathbf{z} , and $p_{\theta_g}(\mathbf{y} | \mathbf{x}, \mathbf{z})$ is the generation distribution conditioned on both the input \mathbf{x} and retrieved subgraph.

3.2 DRAG Framework Overview

In this work, we propose DRAG (Distilling RAG for SLMs) as a novel framework to transfer retrieval-augmented generation capabilities from large-scale LLMs to smaller, more efficient models. DRAG mitigates hallucination and enhances answer accuracy by leveraging evidence-based distillation. The overall procedure consists of four sequential steps: **1)** Evidence generation, **2)** RAG evidence ranking, **3)**

Algorithm 1: DRAG Framework

Step 1: Evidence Generation

Input: A question q ; a large-scale LLM $\mathcal{M}_{\text{large}}$; a small-scale LLM $\mathcal{M}_{\text{small}}$; number of evidences to generate N ; number of top relationships K

foreach question q do

```

1   Prompt  $\mathcal{M}_{\text{large}}$  to generate  $N$  textual evidences
    relevant to  $q$ .
     $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$ 

```

Step 3: Graph RAG Generation
foreach evidence $d_i \in \mathcal{D}_{\text{filtered}}$ do

```

6   Prompt  $\mathcal{M}_{\text{large}}$  to extract entity pairs and their
    relationships
     $\mathcal{R}_i = \{(a, b, r) \mid a, b \in \mathcal{V}, r \in \mathcal{E}\}$ ,
    where  $\mathcal{V}$  is the set of entities, and  $\mathcal{E}$  is the set
    of relationships.
7   Construct  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  by adding nodes  $a, b$ 
    and an edge labeled  $r$ .

```

Step 2: RAG Evidence Ranking
foreach generated evidence $d_i \in \mathcal{D}$ do

```

2    $\text{score}_i^{(\text{sim})} = \cos(\mathbf{e}_i, \mathbf{q})$ 
3    $\text{rank}_{\text{LLM}}(d_i)$  via  $\mathcal{M}_{\text{large}}$ 
4   // Compute combined ranking score
     $s_i = \text{score}_i^{(\text{sim})} + \text{rank}_{\text{LLM}}(d_i)$ .
5   Select top-ranked subset  $\mathcal{D}_{\text{filtered}} \subset \mathcal{D}$  based on  $s_i$ .

```

Step 4: Small-Scale LLM Evaluation

```

8   Select the top  $K$  evidences and relationships based
    on semantic and LLM-based scores.
9   Prompt  $\mathcal{M}_{\text{small}}$  with:
     $\{d_i \in \mathcal{D}_{\text{filtered}}\}$  and/or  $\{(a_j, b_j, r_j)\}_{j=1}^K$ 
    to generate the final answer:
     $\hat{y} \leftarrow \mathcal{M}_{\text{small}}(q, \text{context})$ .
10  return  $\hat{y}$  // Final answer from small LLM.

```

Graph RAG generation, and **4)** Small-scale LLM evaluation. A full paradigm of our framework is in Alg. 1. Each step is described in detail below.

Evidence Generation. Given an input question q , the first stage of DRAG involves eliciting a diverse set of potentially relevant facts from a large-scale language model $\mathcal{M}_{\text{large}}$. Our perspective here is that a well-trained LLM is a stronger and more efficient retriever for SLMs than the traditional “query encoder + document index based retriever”, especially given the relatively weaker target model. Specifically, we design a prompt (details in our appendix) for $\mathcal{M}_{\text{large}}$ to generate N distinct textual evidences: $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$. Each evidence d_i is intended to encapsulate a factual snippet or a useful piece of information that could help answer q . This step not only diversifies the candidate knowledge but also forms the basis for subsequent ranking and structured extraction processes.

RAG Evidence Ranking. Once evidence set \mathcal{D} is obtained, each evidence d_i is quantitatively evaluated to determine its relevance to the question q . This ranking process involves two key components: **Step-1. Semantic Similarity Score:** We compute a vector embedding using the sentence-transformers (Reimers and Gurevych, 2019) for both the evidence d_i and the query q , denoted by \mathbf{e}_i and \mathbf{q} respectively. The semantic similarity score is then calculated using cosine similarity:

$$\text{score}_i^{(\text{sim})} = \cos(\mathbf{e}_i, \mathbf{q}), \quad (3)$$

This score captures the latent semantic alignment between the evidence and the question.

Step-2. LLM-based Ranking Score: In parallel,

$\mathcal{M}_{\text{large}}$ is prompted to provide an intrinsic relevance ranking, denoted as $\text{rank}_{\text{LLM}}(d_i)$, for each d_i . This ranking leverages the vast internal knowledge of $\mathcal{M}_{\text{large}}$ to assess the contextual appropriateness of the evidence.

The combined ranking score s_i for each evidence d_i is computed as an equally weighted sum:

$$s_i = \text{score}_i^{(\text{sim})} + \text{rank}_{\text{LLM}}(d_i), \quad (4)$$

Following the computation of s_i for all evidences, we discard the lowest-scoring X evidences, where X refers to the $N - K$ portion specified and illustrated in Figure 1, and retain a filtered subset:

$$\mathcal{D}_{\text{filtered}} \subset \mathcal{D}. \quad (5)$$

This filtering ensures that only the most relevant evidences are carried forward.

Graph RAG Generation. In order to further structure the distilled knowledge, the filtered evidences $\mathcal{D}_{\text{filtered}}$ are transformed into a relational graph. For each evidence $d_i \in \mathcal{D}_{\text{filtered}}$, we prompt $\mathcal{M}_{\text{large}}$ to extract structured information in the form of entity relationships. Specifically, for each d_i , a set of relationship triples is extracted:

$$\mathcal{R}_i = \{(a, b, r) \mid a, b \in \mathcal{V}, r \in \mathcal{E}\}, \quad (6)$$

where \mathcal{V} represents the set of entities, \mathcal{E} represents the set of relationships. These triples are then used to construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in which nodes represent entities and edges (labeled by r) represent relationships. To focus on the most salient connections, a ranking procedure is applied to the extracted relationships, and the top K relationships are selected based on a combination of semantic

and LLM-based scores (similar to the evidence ranking). This graph-based representation serves as an additional structured context that enriches the evidence pool with inter-entity relationships.

In DRAG, converting evidence into a graph inevitably results in information loss. However, since some evidence pieces are quite long, directly processing them with the SLM would impose a significant computational burden. By utilizing a graph-based representation, we greatly reduce this overhead while preserving essential structured knowledge. To further optimize efficiency, we introduce a *simple graph aggregation approach*, as in Figure 1 and Appendix H, which merges pairs of the same entity into a unified graph representation. This further minimizes computational costs during SLM inference, making the process more efficient without compromising key relational information.

SLMs Evaluation. In the final step, the distilled and structured evidence is used to inform and boost a small-scale language model $\mathcal{M}_{\text{small}}$ to generate the final answer. The context provided to $\mathcal{M}_{\text{small}}$ can include:

- 1) The set of filtered evidences: $\{d_i \in \mathcal{D}_{\text{filtered}}\}$,
- 2) The top K relationship triples extracted from the graph: $\{(a_j, b_j, r_j)\}_{j=1}^K$.

These elements are concatenated with the original question q to form a comprehensive prompt. The small-scale model is then queried as follows:

$$\hat{y} \leftarrow \mathcal{M}_{\text{small}}(q, \text{context}). \quad (7)$$

where \hat{y} is the final answer. By conditioning on both unstructured evidences and structured relational information, $\mathcal{M}_{\text{small}}$ is better grounded in factual knowledge, thereby mitigating hallucination while maintaining computational efficiency.

3.3 Mitigating Privacy for Users

Another key advantage of our framework is its potential for privacy protection. Typically, when querying large-scale LLMs, local deployment is not feasible, requiring users to upload their private queries to cloud-based LLMs, raising privacy concerns. Our framework addresses this issue by enabling local SLMs to leverage the knowledge of large models while preserving user privacy.

With our approach, the local model first reformulates the query (much simpler than answering the query directly), stripping any sensitive information before sending it to the cloud-based model. The cloud model then retrieves relevant evidence and knowledge graphs, which are subsequently passed

back to the local model for final processing and response generation. This ensures that private data remains protected while still benefiting from the power of large-scale LLMs.

To evaluate the feasibility of this privacy-preserving solution by our DRAG framework, we construct a specialized dataset containing privacy-sensitive information. The dataset includes several key processing steps, with details provided in Appendix G. We test our method on this dataset, as shown in the experimental section, we observe significant improvements in privacy protection while maintaining high accuracy and efficiency.

3.4 Discussion

In DRAG, instead of generating answers directly from the teacher model, it solely provides evidence and knowledge graphs for the student model. This strategy offers several key advantages: 1) The teacher LLM is an extremely large, general-purpose model, while the student is domain-specialized, ensuring higher accuracy and efficiency during usage without unnecessary general knowledge. 2) The teacher LLM is usually heavy in size and deployed on the cloud, and the student is on local devices, it is simple to develop methods for preserving privacy using our proposed framework by transmitting only de-identified queries and structured knowledge instead of full responses, as we have introduced in Section 3.3.

4 Experiments

Datasets. The following benchmarks are used in our work: **ARC-Challenge** (Clark et al., 2018), **MedMCQA** (Pal et al., 2022), **GPQA** (Rein et al., 2024), **MMLU** (Hendrycks et al., 2020), **Open-LLM-Leaderboard** (Myrzakhan et al., 2024), **AVERITEC** (Schlichtkrull et al., 2023). More details of these datasets are provided in Appendix A.

4.1 Models and Experimental Settings

For our experiment we use a set of large-scale teacher models and small student models. The teacher models include GPT-4o (Hurst et al., 2024), DeepSeek-V3 (Liu et al., 2024), Gemini Flash 1.5 (Team et al., 2024a), Claude 3.5 Sonnet (Anthropic, 2024), and LLaMA-3.3-70B (Dubey et al., 2024). For student models, we use Gemma-2-2B-it (Team et al., 2024b), Phi-3.5-mini-instruct (Abdin et al., 2024), Qwen2.5-3B-Instruct (Yang et al., 2024), LLaMA-3.2-3B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024),

Framework	Backbone LLM	MedMCQA Acc	MMLU Acc	ARC-C Acc
Self-RAG (Asai et al., 2023)	SelfRAG-LLaMA-2-7B	–	–	67.3
CRAG (Yan et al., 2024b)	SelfRAG-LLaMA-2-7B	–	–	68.6
DRAG (Ours)	LLaMA-2-7B	72.4	71.2	86.2 _{↑17.6}
SimRAG (Xu et al., 2024)	Llama-3.1-8B-Instruct	–	67.5	81.4
DRAG (Ours)	Llama-3.1-8B-Instruct	74.2	75.7 _{↑8.2}	93.1 _{↑11.7}
MiniRAG (Fan et al., 2025)	GLM-edge-1.5B-chat	52.1	47.0	62.3
	Qwen2.5-3B-Instruct	58.3	70.9	67.7
	Llama-3.2-3B-Instruct	52.7	69.1	65.3
	Gemma-2-2B-it	48.5	57.3	68.6
	Phi-3.5-mini-instruct	61.1	72.7	82.7
DRAG (Ours)	GLM-edge-1.5B-chat	56.9 _{↑4.8}	69.0 _{↑22.0}	90.0 _{↑27.7}
	Qwen2.5-3B-Instruct	72.8 _{↑14.5}	73.8 _{↑2.9}	93.0 _{↑25.3}
	Llama-3.2-3B-Instruct	73.6 _{↑20.9}	74.4 _{↑5.3}	93.0 _{↑27.7}
	Gemma-2-2B-it	72.4 _{↑23.9}	71.2 _{↑13.9}	91.5 _{↑22.9}
	Phi-3.5-mini-instruct	74.4 _{↑13.3}	77.8 _{↑5.1}	94.1 _{↑11.4}

Table 1: **Comparison with other state-of-the-art RAG frameworks.** We compare DRAG (evidence-based) with prior approaches across multiple benchmarks and backbone LLMs. In the table, the “↑” indicates improvements over other methods under the same backbone and inference configuration.

LLaMA-3.1-8B-Instruct (Dubey et al., 2024), and Gemma-2-9B-it (Team et al., 2024b) covering a range of 2B to 9B parameters. We evaluate the performance of the student models in a zero-shot setting using the lm-evaluation-harness framework (Gao et al., 2024) on a 4×RTX 4090 GPUs setup.

4.2 Comparison with State-of-the-Arts

Table 1 compares DRAG against existing RAG frameworks on MedMCQA, MMLU, and ARC-C. Overall, DRAG consistently outperforms previous state-of-the-art methods and effectively boosts small language models (SLMs). For instance, on ARC-C, Self-RAG (Asai et al., 2023) and CRAG (Yan et al., 2024b) achieve 67.3% and 68.6%, respectively, while DRAG obtains up to 94.1%, exceeding them by +26.8% and +25.5%. SimRAG (Xu et al., 2024), based on the Llama-3.1-8B-Instruct backbone, achieves 67.5% and 81.4% on MMLU and ARC-C, respectively. By contrast, DRAG attains 75.7% and 93.1% with the same backbone, surpassing SimRAG by +8.2%, and +11.7%. Compared to MiniRAG (Fan et al., 2025), DRAG achieves notable gains of at most +23.9% on MedMCQA, +13.9% on MMLU, and +11.4% on ARC-C with the same SLMs backbones. These results confirm that DRAG delivers superior retrieval-augmented performance while substantially reducing hallucination and computational overhead.

4.3 Ablation Studies

Number of evidence/graph relations K . We analyze the impact of varying the number of generated/retrieved evidence on student SLM’s perfor-

mance, as in Table 2 and Figure 2. Our experiments show that ~ 15 evidence pieces generally provide the optimal cost-accuracy trade-off results, using fewer leads to insufficient knowledge, while using more introduces redundancy and slightly degrades performance due to increased noise. Since the graph representation is constructed from raw evidence, it naturally loses information but remains more concise and computationally efficient. Combining both evidence and graph might be beneficial, our results show that this scheme is redundant, yielding similar accuracy to using evidence alone, while incurring extra inference overhead.

Effects of different teacher LLMs. We investigate the effect of different teacher LLMs on student performance as in Table 3. Surprisingly, a more powerful teacher does not always lead to better student accuracy. Our experiments show that GPT-4o produces the best distillation results, outperforming all other models. The ranking of teacher models for student SLM is as follows: GPT-4o > Claude 3.5 Sonnet > DeepSeek V3 > Llama 3.3 70B > Gemini 1.5 Flash. These results indicate that the quality and consistency of generated evidence matter more than just using a more capable LLM. Certain models, such as Claude and DeepSeek V3, perform competitively, but their evidence generation may not be as structured or factually aligned as GPT-4o.

Computation comparison. To compare computation efficiency, we evaluate the average length of generated evidence versus knowledge graphs in Table 5. As expected, graph-based RAG significantly reduces computational costs, as the structured rep-

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	78.55	91.69	92.76	93.48	93.30	92.31	93.74	94.01	94.10	92.40	93.12	93.74	93.74
Qwen2.5-3B-Instruct	48.26	88.47	89.81	91.15	90.17	91.06	91.96	93.03	92.85	89.81	91.06	92.49	92.23
Llama-3.2-3B-Instruct	42.45	86.77	88.47	90.44	90.44	89.10	91.78	92.31	93.03	88.56	91.51	92.67	92.40
Llama-3.1-8B-Instruct	54.24	89.01	90.71	91.96	91.42	92.05	94.10	94.28	93.30	91.69	93.03	94.28	93.57
Qwen2.5-7B-Instruct	55.23	90.97	91.78	92.49	92.58	91.87	93.12	93.39	93.83	91.51	93.03	93.48	93.74
gemma-2-9b-it	63.27	92.58	93.30	93.74	94.10	93.74	94.28	94.73	94.46	93.30	94.19	94.28	94.37
gemma-2-2b-it	53.71	85.08	87.94	88.56	88.92	88.11	90.71	91.33	91.51	87.67	90.71	91.33	91.33

Table 2: **Comparison of results on the ARC-Challenge using GPT-4o as the teacher model.** The *Original* represents the baseline performance without RAG. *Evidence Only* uses ranked context textual evidences, *Graph Only* utilizes structured relationships, and *Graph and Evidence Combined* integrates both sources. Results are reported for different retrieval sizes (5, 10, 15, 20).

Target SLM	Original	GPT-4o			Llama 3.3 70b			Claude 3.5 Sonnet			Gemini 1.5 Flash			DeepSeek V3		
		Graph	Evide.	Comb.	Graph	Evide.	Comb.	Graph	Evide.	Comb.	Graph	Evide.	Comb.	Graph	Evide.	Comb.
Phi-3.5-mini-instruct	55.41	72.01	74.35	74.06	65.43	68.80	68.42	65.14	69.78	70.28	58.91	61.18	60.03	66.12	68.95	68.49
Qwen2.5-3B-Instruct	51.59	70.21	72.32	72.03	61.06	66.48	65.34	60.32	65.50	67.68	55.08	57.66	57.04	63.18	66.27	66.08
Llama-3.2-3B-Instruct	50.90	71.65	73.56	73.15	64.26	67.99	67.34	64.69	68.56	69.21	59.62	61.58	60.60	66.56	69.16	68.80
Llama-3.1-8B-Instruct	59.14	73.13	73.97	73.54	67.34	69.52	69.02	68.95	71.74	71.69	62.99	63.78	63.09	68.35	70.40	69.4
Qwen2.5-7B-Instruct	56.08	72.51	73.66	73.58	64.04	68.13	67.54	65.67	70.07	70.14	59.84	60.77	60.58	66.56	68.76	68.54
gemma-2-9b-it	56.83	72.79	74.13	73.73	66.39	69.33	69.23	69.38	71.34	71.48	61.34	61.65	61.32	68.06	69.59	69.42
gemma-2-2b-it	42.91	68.68	71.38	71.72	60.79	65.54	65.19	57.76	63.93	66.65	54.24	57.21	56.32	62.20	66.24	65.46

Table 3: Ablation comparison across various large-scale teacher models on MedMCQA.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	66.22	69.24	70.35	71.46	71.97	72.11	74.48	75.13	75.58	72.22	74.05	74.46	74.48
Qwen2.5-3B-Instruct	62.01	64.89	66.73	67.29	68.17	68.81	70.78	71.50	71.95	67.97	70.41	70.58	71.58
Llama-3.2-3B-Instruct	57.27	63.20	64.93	65.38	66.76	67.54	69.98	70.82	71.36	67.49	69.51	70.10	71.13
Llama-3.1-8B-Instruct	65.01	67.78	68.95	69.92	70.70	70.94	73.29	73.96	74.15	71.15	72.83	73.31	73.39
Qwen2.5-7B-Instruct	69.71	70.49	72.55	73.27	73.65	73.65	75.59	76.30	76.34	73.72	75.75	75.95	76.16
gemma-2-9b-it	69.73	71.25	73.35	73.82	74.68	73.31	75.30	76.32	76.49	73.74	75.91	76.16	76.49
gemma-2-2b-it	55.59	59.61	61.66	62.50	63.45	64.25	66.16	66.94	67.58	63.63	66.12	66.59	66.88

Table 4: Performance of using GPT-4o as the teacher on privacy protection task and benchmark.

resentation is much shorter than raw evidence while still preserving core relational information. Specifically, graph representations require significantly fewer tokens during inference, making them ideal for low-resource or real-time retrieval scenarios.

Category	Total Number	Average Length
Evidence	29,698,547	26.30
Graph	24,324,636	21.55 _{↓18.1%}

Table 5: Token statistics per evidence and graph.

4.4 Multi-choice QA

To assess the impact of DRAG on retrieval-augmented multiple-choice question answering (MCQA), we conduct extensive experiments across four benchmark datasets: ARC-C, MedMCQA, GPQA, and MMLU in Table 2 and Tables 8, 10, and 11 in Appendix, testing various students and with GPT-4o as the teacher model. The results demonstrate significant improvement across various student and teacher model architectures.

Among student models, Gemma-2-9b-it consistently achieves the strongest performance across benchmarks when paired with GPT-4o as the

teacher, reaching 94.73% on ARC-C, 77.80% on MMLU, 74.42% on MedMCQA, and 40.11% on GPQA with evidence-only distillation. This represents substantial improvements over baseline performance: 53.71%, 71.80%, 56.83%, and 34.80%, respectively. The Phi-3.5-mini-instruct model, despite its smaller size, shows surprisingly competitive results, particularly on ARC-C (94.10%) and MMLU (77.80%). In contrast, smaller variants like Gemma-2-2b-it and Llama-3.2-3B-Instruct consistently perform 3-5% lower than their larger counterparts, although still showing substantial improvements over their baselines. In particular, on MMLU, Gemma-2b-it improves from 56.80% to 71.16%, demonstrating effective knowledge transfer even to resource-constrained architectures.

4.5 Open-ended QA

To assess the effectiveness of DRAG on open-style questions, where it requires models to generate unconstrained, contextually appropriate responses rather than selecting from predefined choices. We used the Open-LLM Leaderboard (Myrzakhan

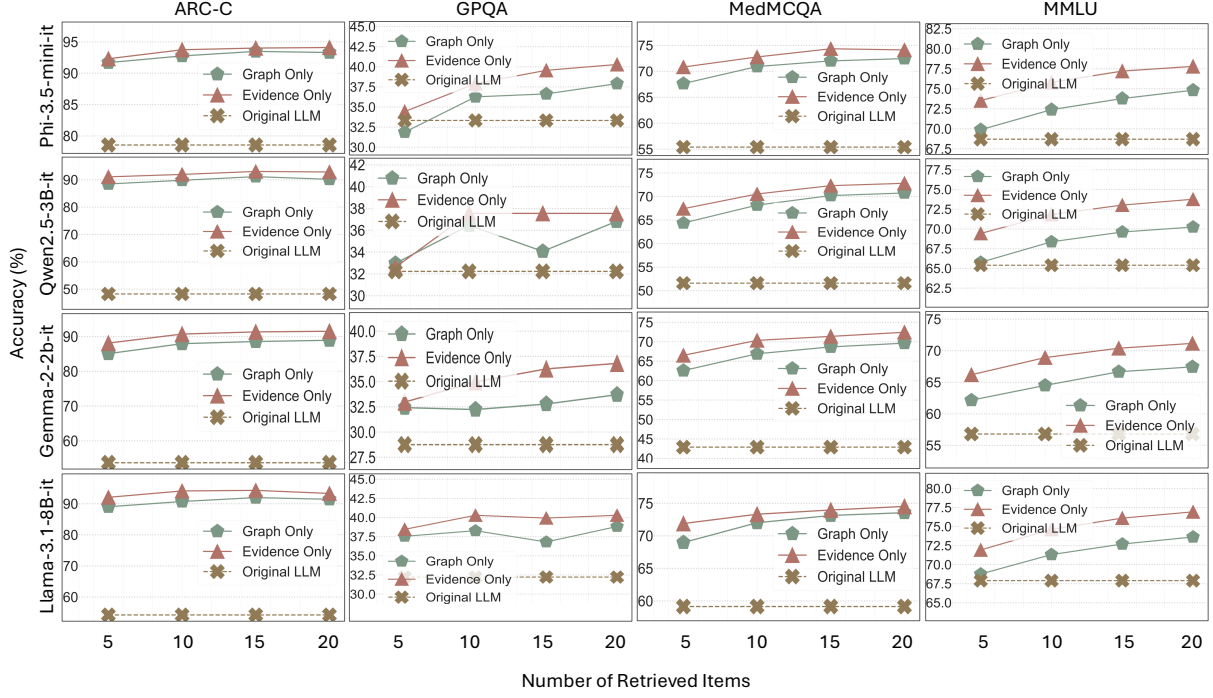


Figure 2: **Effect of retrieved graph-based and evidence-based RAG on multiple-choice QA tasks.** We evaluate different retrieval strategies: *Graph Only*, *Evidence Only*, and the *Original LLM* across four benchmarks (ARC-C, GPQA, MedMCQA, and MMLU) using various backbone models. The x-axis represents the number of retrieved items, while the y-axis denotes accuracy (%).

Open-LLM Leaderboard	Original	DRAG _G	DRAG _E	DRAG _C
Phi-3.5-mini-instruct	35.26	38.22 _{+2.96}	40.13 _{+4.87}	39.84 _{+4.58}
Qwen2.5-3B-Instruct	32.59	36.01 _{+3.42}	37.23 _{+4.64}	37.50 _{+4.91}
Llama-3.2-3B-Instruct	29.85	32.79 _{+2.94}	34.12 _{+4.27}	33.65 _{+3.80}
Llama-3.1-8B-Instruct	45.07	49.13 _{+4.06}	50.26 _{+5.19}	51.73 _{+6.66}
Qwen2.5-7B-Instruct	44.67	48.73 _{+4.06}	52.36 _{+7.69}	52.66 _{+7.99}
Gemma-2-9b-it	46.44	49.12 _{+2.68}	53.45 _{+7.01}	53.19 _{+6.75}
Gemma-2-2b-it	32.54	35.98 _{+3.44}	37.31 _{+4.77}	37.26 _{+4.72}

Table 6: Accuracy on Open-LLM leaderboard. DRAG_G, DRAG_E, and DRAG_C represent graph-based, evidence-based and combined configurations, respectively.

et al., 2024) for evaluation. Given the large scale of the Open-LLM Leaderboard, we considered computational cost and opted for GPT-4o-mini (OpenAI, 2024) as the teacher model, balancing efficiency with effective knowledge transfer.

As shown in Table 6, both graph-based and evidence-based distillation lead to significant improvements over the original model performances, where the original refers to student models without any distillation. Evidence-based setting provides the highest accuracy gains, with Gemma-2-9b-it improving from 46.44% to 53.45%, and Qwen2.5-7B-Instruct from 44.67% to 52.36%. These results highlight the importance of structured knowledge and retrieved evidence in enhancing open-style response generation. Graph-only distillation, while slightly less effective, still provides meaningful improvements, where Qwen2.5-3B-Instruct increases

from 32.59% to 36.01%. These results emphasize the importance of utilizing both structured and retrieved information to improve open-style response generation while demonstrating that smaller, cost-effective teacher models like GPT-4o-mini can still facilitate meaningful knowledge transfer.

4.6 Fact Verification

Model	Original	DRAG _G	DRAG _E	DRAG _C
BLOOM-7b	26	30.11 _{+4.11}	32.43 _{+6.43}	32.29 _{+6.29}
GPT-3.5-Turbo	29	40.98 _{+11.98}	49.10 _{+20.10}	45.63 _{+16.63}

Table 7: Performance on AVERITEC.

Following Schlichtkrull et al. (2023), BLOOM-7b and GPT-3.5-Turbo are used as students for fact verification benchmarking on AVERITEC dataset as in Table 7. For both models, evidence distillation yielded the strongest performance, with 32.43% for BLOOM-7b and 49.10% for GPT-3.5-Turbo. Evidence and graph combined distillation provided the second highest in both cases, with 32.29% for BLOOM-7b and 45.63% for GPT-3.5-Turbo.

4.7 Privacy Protection Evaluation

Our privacy-preserving framework effectively minimizes the risk of sensitive data exposure. We analyze the reduction in personally identifiable information (PII) before and after applying our SLM-based privacy filtering. Out of 15,090 injected PIIs,

only 649 remain post-processing, resulting in an overall reduction of 95.7%.

To understand the impact of privacy filtering on model accuracy, we evaluate performance on our MMLU-augmented dataset (see Appendix G). As shown in Table 4, our framework maintains strong performance across various student models, despite rigorous privacy filtering. The graph and evidence-based combined approach achieves the best results, with Gemma-2-9b-it increasing from 69.73% to 76.49% and Qwen2.5-7B-Instruct improving from 69.71% to 76.16%. Even smaller models like Gemma-2-2b-it show notable gains, rising by 11.29% from the baseline, demonstrating that privacy filtering does not significantly compromise the performance of DRAG. These findings confirm that our framework effectively mitigates privacy risks while preserving knowledge retrieval, ensuring high-quality LLM responses.

5 Conclusion

We presented DRAG, a novel approach that distills RAG knowledge into SLMs using evidence- and graph-guided distillation. By structuring knowledge extraction with ranked evidence and knowledge graphs, DRAG mitigates hallucinations while significantly reducing computational demands. Experimental results show that DRAG outperforms existing SoTAs like MiniRAG under similar constraints, preserving RAG’s benefits while enhancing efficiency. Our work offers a scalable, resource-efficient solution for deploying high-quality retrieval-augmented generation in small models, balancing factual consistency and computational efficiency in knowledge-intensive tasks.

Limitations

Despite its strong performance, DRAG has a few limitations that warrant further investigation: 1) *Knowledge retention trade-offs*. Our method successfully distills factual knowledge into smaller models, but some nuanced or implicit knowledge present in the teacher LLMs may be missing. This is especially relevant in creative, open-ended, or subjective tasks where explicit factual grounding is less defined. 2) *Computational overhead during distillation*. Although DRAG enables more efficient inference in SLMs, the distillation process itself requires significant computation, particularly when generating evidence rankings and graph-based knowledge representation. Future work

could explore optimizing this process to further reduce evidence generation costs. 3) In DRAG, when generating evidence, we aim to prevent data/answer leakage by instructing the model explicitly in the prompt with “do not give the answer away directly”. However, despite this precaution, there is still a potential risk of unintended leakage. This could raise concerns regarding the integrity of the distillation process. To mitigate this, we ensure that the generated evidence remains neutral, contextually relevant, and free from direct answer hints while still being informative for the target student model.

Ethics Statement

While DRAG minimizes hallucinations, its outputs must still be subject to critical evaluation in high-stakes applications such as legal, medical, or scientific domains. Human oversight remains crucial in ensuring that AI-generated content aligns with ethical and professional standards. Also, DRAG reduces hallucinations by aligning outputs with structured knowledge, but it remains susceptible to biases present in the teacher LLMs, knowledge graphs, and evidence. Biases inherent in training data may still propagate into distilled SLMs, necessitating continuous evaluation and mitigation strategies.

References

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Raviteja Anantha, Tharun Bethi, Danil Vodanik, and Srinivas Chappidi. 2023. Context tuning for retrieval augmented generation. *arXiv preprint arXiv:2312.05708*.
- Anthropic. 2024. [Claude 3.5 sonnet](#).
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.
- Yuri Façanha Bezerra and Li Weigang. 2025. Llmquoter: Enhancing rag capabilities through efficient quote extraction from large contexts. *arXiv preprint arXiv:2501.05554*.
- Christopher JC Burges. 2010. From ranknet to lambdamart: An overview. *Learning*, 11(23-581):81.

- Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Tianyu Fan, Jingyuan Wang, Xubin Ren, and Chao Huang. 2025. Minirag: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:2501.06713*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [The language model evaluation harness](#).
- Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. Grag: Graph retrieval-augmented generation. *arXiv preprint arXiv:2405.16506*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.
- Pengyue Jia, Derong Xu, Xiaopeng Li, Zhaocheng Du, Xiangyang Li, Xiangyu Zhao, Yichao Wang, Yuhao Wang, Huifeng Guo, and Ruiming Tang. 2024. Bridging relevance and reasoning: Rationale distillation in retrieval-augmented generation. *arXiv preprint arXiv:2412.08519*.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- 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. In *Advances in Neural Information Processing Systems*.
- Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *ACL*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2024. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. *arXiv preprint arXiv:2407.10805*.
- Qiheng Mao, Zemin Liu, Chenghao Liu, Zhuo Li, and Jianling Sun. 2024. Advancing graph representation learning with large language models: A comprehensive survey of techniques. *arXiv preprint arXiv:2402.05952*.
- Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.
- Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.
- OpenAI. 2024. [Gpt-4o mini: Advancing cost-efficient intelligence](#). Accessed: 2025-02-14.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*. PMLR.

- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, pages 3982–3992.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*.
- Wolfram Alpha. 2025. [Simple Graph](#). Accessed: 2025-02-14.
- Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.
- Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C Ho, Carl Yang, et al. 2024. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. *arXiv preprint arXiv:2410.17952*.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024a. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024b. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yun Zhu et al. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Appendix

A Datasets

The following benchmarks are used in our work, more details of them are as follows:

ARC-Challenge: The AI2 Reasoning Challenge (ARC) is comprised of questions pertaining to natural science (Clark et al., 2018). The benchmark is split into the Easy Set and the Challenge Set, we selected ARC-Challenge due to the fact that it consists of questions for which retrieval and co-occurrence methods both fail.

MedMCQA: It contains over 194k medical school entrance exam questions spanning over 21 medical subjects (Pal et al., 2022). This is a high-quality benchmark with real medical exam questions.

GPQA: This is the Google-Proof QA dataset, a collection of 448 multiple-choice questions written by experts in the natural sciences. What makes this dataset unique is that PhDs and PhD candidates in the corresponding domains have only reached 65% accuracy in the questions even with web access, hence making the questions "Google-proof" (Rein et al., 2024).

MMLU: The *Massive Multitask Language Understanding* benchmark is an evaluation dataset consisting of multiple-choice tasks covering a broad range of 57 subjects (Hendrycks et al., 2020). MMLU is used for its comprehensiveness and high popularity level.

Open-LLM-Leaderboard: To circumvent issues associated with multiple choice questions, such as preference towards certain options, we include an open-style benchmark in our evaluation that avoids the selection bias and random guessing problems (Myrzakhan et al., 2024).

AVERITEC: It measures LLM fact verification abilities by providing claims that LLMs either refute, support, claim not enough evidence, or claim conflicting evidence (Schlichtkrull et al., 2023).

B Additional Results and Ablations

B.1 More Ablation on Teacher Model

Tables 8, 12, 13, 14, and 15 illustrate the performance of various teacher models on the MedMCQA dataset. The results on MedMCQA demonstrate improvement across all small-scale models when incorporating additional context, whether through graph-based, evidence-based, or combined distillation approaches.

Under the GPT-4o teacher model, impressive improvements are seen. Phi-3.5-mini-instruct improves from 55.41% to 74.13% in the best configuration (20 evidence). Similarly, Qwen2.5-3B-Instruct experienced a 21.23% improvement, rising from 51.59% to 72.82%. Most notably, smaller models show dramatic improvements. For instance, the smaller gemma-2-2b-it, which originally scored 42.91%, achieves up to 72.44% in the 20-evidence experiment, representing a 29.52% increase.

The pattern of stronger performance on smaller student models is consistent across Claude 3.5 Sonnet, Llama 3.3 70B, Gemini 1.5 Flash, and DeepSeek V3 as teachers. For gemma-2-2b-it, performance increases up to 23.05% for Llama (20 graph and evidence), up to 25.13% for Claude (20 graph and evidence), up to 14.70% for Gemini (20 evidence), and up to 23.5% for DeepSeek (20 evidence). Across the five teacher models, the best-performing augmentation leads to improvements between 14.7% and 29.5%.

Performance gains vary based on model size, with larger models like Llama-3.1-8B and gemma-2-9b-it showing relatively less improvement. Thus, while structured knowledge integration is beneficial across all student models, smaller models gain the most from these enhancements due to their initial performance limitations.

B.2 Comparison Across Various Student Models

Figure 3 presents extra results extending Table 1, further examining the impact of retrieval strategies for Llama-3.2-3B-it and Gemma-2-9B-it.

Consistently across all benchmarks, **Evidence Only** retrieval achieves the highest accuracy, reinforcing the importance of direct textual evidence in retrieval-augmented learning. The performance gap between evidence-only distillation and graph-only distillation is particularly noticeable in knowledge-intensive tasks such as MedMCQA and ARC-C, where factual precision is crucial. While graph-based retrieval provides some benefits, its improvements remain more limited, especially in benchmarks requiring extensive factual recall. The results suggest that direct evidence retrieval provides a more reliable source of knowledge for model reasoning, whereas graph-based retrieval alone may not be sufficient to bridge factual gaps. Additionally, increasing the number of retrieved items consistently enhances performance, though

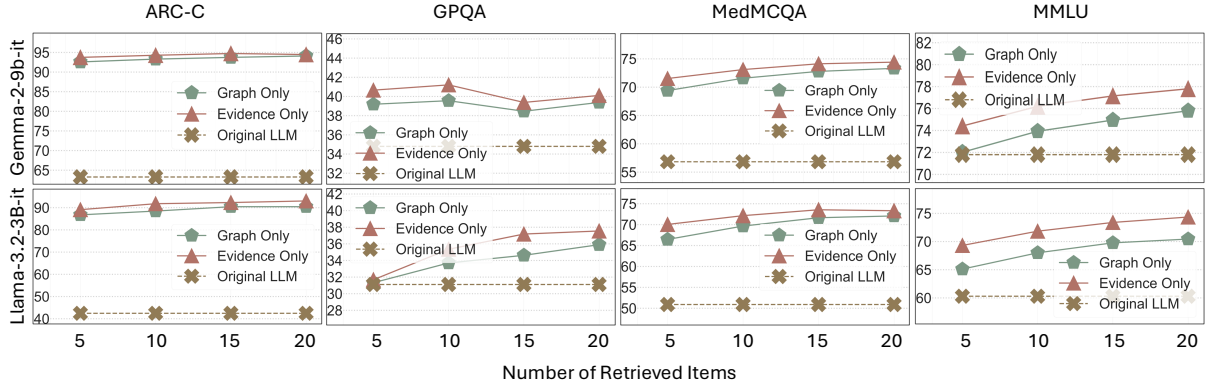


Figure 3: **More results on Retrieval Strategies.** We evaluate different retrieval strategies: *Graph Only*, *Evidence Only*, and *the Original LLM* across four benchmarks (ARC-C, GPQA, MedMCQA, and MMLU) extended on Llama3.2-3B-it and Gemma-2-9B-it benchmarks. In this figure, the x-axis represents the number of retrieved items, while the y-axis denotes accuracy (%).

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	55.41	67.65	70.93	72.01	72.46	70.83	72.77	74.35	74.13	69.73	72.36	74.06	74.06
Qwen2.5-3B-Instruct	51.59	64.38	68.18	70.21	70.76	67.42	70.55	72.32	72.82	66.75	69.95	72.03	72.13
Llama-3.2-3B-Instruct	50.90	66.46	69.66	71.65	72.08	70.05	72.15	73.56	73.32	69.47	71.62	73.15	73.27
Llama-3.1-8B-Instruct	59.14	68.97	72.01	73.13	73.56	71.89	73.32	73.97	74.52	71.58	72.96	73.54	74.18
Qwen2.5-7B-Instruct	56.08	67.44	70.43	72.51	72.53	70.48	72.56	73.66	74.01	69.54	72.24	73.58	73.68
gemma-2-9b-it	56.83	69.40	71.58	72.79	73.32	71.53	73.11	74.13	74.42	71.34	72.89	73.73	73.97
gemma-2-2b-it	42.91	62.59	66.91	68.68	69.64	66.51	70.36	71.38	72.44	65.62	69.35	71.72	72.03

Table 8: Comparison of results on the MedMCQA using GPT-4o as the teacher model.

Model	ARC-C	MedMCQA	GPQA	MMLU
GPT-4o	96.7	77.0	53.6	88.7
Claude 3.5 Sonnet	96.4	76.1	59.4	88.7
DeepSeek V3	95.3	74.3	59.1	87.1
Llama 3.3 70B-Instruct	95.1	72.7	50.5	86.0
Gemini 1.5 Flash	90.3	69.9	39.5	78.9

Table 9: Comparison of teacher models’ performance across ARC-Challenge, MedMCQA, GPQA, and MMLU benchmarks.

the gains diminish beyond 15 items. These findings highlight the effectiveness of evidence-based retrieval as the dominant strategy for boosting model performance in multiple-choice QA tasks.

C More Results on Open QA Dataset

Following OpenQA (Lin et al., 2018)², we provide more results on WebQuestions dataset³ in Table 16. Our framework consistently achieves much better performance than the original baseline method.

D More Ablations for Analyzing N

We provide additional ablation studies using larger $N=30, 40$, and 50 to show the effect and analysis

of N in Table 17. Our results indicate that increasing N further does not lead to significant performance improvement. This is intuitive, as only a limited amount of evidence is typically needed to answer a question, excessive input may introduce noise and confuse the SLM in identifying the most relevant information. This finding aligns with our main experiments, which show that performance generally peaks around 15 evidence pieces, while adding more often leads to redundancy and slight performance degradation due to noise.

E Upper Bound of Teachers’ Performance

Table 9 presents a comparison across different teacher models. The results demonstrate that our distillation framework enables student models to achieve performance remarkably close to this upper bound. On ARC-C and MedMCQA, the best student models, such as Gemma-2-9b-it and Llama-3.1-8B-Instruct, perform within 2.0–2.5% of GPT-4o, demonstrating minimal performance loss after distillation. On GPQA and MMLU, student models show competitive performance, with Gemma-2-9b-it reaching 77.8% on MMLU, demonstrating that our approach narrows the gap between student and

²<https://github.com/thunlp/OpenQA>.

³<https://github.com/brmson/dataset-factoid-webquestions>.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	33.33	31.87	36.26	36.63	37.91	34.43	37.91	39.56	40.29	33.52	35.90	36.08	39.01
Qwen2.5-3B-Instruct	32.23	32.97	36.45	34.07	36.81	32.60	37.55	37.55	37.55	33.70	38.10	37.91	37.91
Llama-3.2-3B-Instruct	31.11	31.32	33.70	34.62	35.90	31.68	35.35	37.18	37.55	31.14	35.71	36.45	35.53
Llama-3.1-8B-Instruct	32.23	37.55	38.28	36.81	38.83	38.46	40.29	39.93	40.29	37.36	39.56	40.84	40.11
Qwen2.5-7B-Instruct	31.50	34.43	38.64	38.64	40.11	38.28	40.66	41.21	42.12	37.91	41.76	41.94	42.49
gemma-2-9b-it	34.80	39.19	39.56	38.46	39.38	40.66	41.21	39.38	40.11	39.19	41.76	39.74	41.03
gemma-2-2b-it	28.75	32.42	32.23	32.78	33.70	32.97	34.98	36.26	36.81	33.33	35.16	35.71	37.55

Table 10: Comparison of results on the GPQA using GPT-4o as the teacher model.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	68.7	69.88	72.39	73.78	74.82	73.53	75.80	77.21	77.80	73.18	75.82	77.05	77.46
Qwen2.5-3B-Instruct	65.4	65.77	68.37	69.61	70.24	69.43	71.83	73.02	73.76	69.31	71.95	72.79	73.20
Llama-3.2-3B-Instruct	60.3	65.11	68.00	69.77	70.44	69.31	71.84	73.39	74.35	69.31	71.43	73.39	73.90
Llama-3.1-8B-Instruct	67.9	68.77	71.32	72.70	73.63	71.93	74.62	76.11	76.93	72.11	74.72	76.10	76.77
Qwen2.5-7B-Instruct	71.7	71.45	73.42	74.17	75.24	73.96	76.29	76.94	77.82	74.26	76.26	76.95	77.42
gemma-2-9b-it	71.8	72.02	73.93	74.95	75.79	74.41	76.20	77.15	77.80	74.51	76.25	77.24	77.55
gemma-2-2b-it	56.8	62.16	64.50	66.66	67.44	66.19	68.91	70.39	71.16	66.24	68.93	70.37	70.64

Table 11: Comparison of results on the MMLU using GPT-4o as the teacher model.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	55.41	59.19	62.25	65.43	67.01	64.88	68.04	68.80	68.97	63.11	67.08	68.42	68.71
Qwen2.5-3B-Instruct	51.59	53.79	58.31	61.06	62.85	60.22	65.14	66.48	67.13	59.53	63.59	65.34	64.93
Llama-3.2-3B-Instruct	50.90	58.19	61.44	64.26	65.79	64.52	67.34	67.99	68.78	63.09	66.53	67.34	68.11
Llama-3.1-8B-Instruct	59.14	63.52	65.29	67.34	68.71	66.60	69.04	69.52	69.81	66.12	68.42	69.02	69.57
Qwen2.5-7B-Instruct	56.08	58.26	61.32	64.04	65.79	63.90	67.68	68.13	68.54	64.57	67.20	67.54	68.28
gemma-2-9b-it	56.83	62.80	65.55	66.39	67.75	66.22	68.90	69.33	68.99	66.41	68.92	69.23	69.35
gemma-2-2b-it	42.91	52.38	58.09	60.79	62.44	58.74	64.31	65.54	66.36	58.52	63.76	65.19	65.96

Table 12: Comparison of results on the MedMCQA using Llama 3.3 70B as the teacher model.

teacher models. While GPT-4o represents the upper bound for performance, our results show that smaller models can reach highly competitive accuracy levels.

F Prompt Used to Produce Evidence

The following system prompt and user prompt are used for evidence generation:

System Prompt:

"You are an assistant in charge of generating factual evidence statements that aid in solving the provided question. Provide only the evidence statements with no additional remarks. Do not give the answer away directly."

User Prompt:

"Generate N evidences that pertain to answering the following question: $\{q\}$ "

This process is repeated for every task within each of the tested benchmarks.

G Details of Privacy Sensitive Benchmark Construction

To demonstrate the effectiveness of our DRAG framework in preventing the leakage of sensitive information while maintaining answer quality, we constructed a benchmark based on the MMLU dataset. Our approach begins with a diverse sample of 5,000 questions randomly drawn from MMLU. To simulate realistic privacy risks such as those involving the unintentional exposure of personally identifiable information (PII), we use GPT-4o to augment these questions by injecting synthetic yet realistic PII (e.g., fabricated names, email addresses, and affiliations). To mitigate these risks, we use SLM to detect and remove any sensitive information. This redaction process preserves the original semantic content of each query while ensuring that only privacy-safe inputs are forwarded to our DRAG framework. The cleaned questions are then processed by DRAG, which retrieves relevant external knowledge before generating high-quality answers. The overall process is summarized as follows:

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	20
Phi-3.5-mini-instruct	55.41	60.51	64.07	65.14	66.15	65.03	68.78	69.78	70.26	66.24	70.16	70.28	70.86
Qwen2.5-3B-Instruct	51.59	55.73	58.47	60.32	61.22	60.29	64.38	65.50	65.55	62.40	66.39	67.68	68.35
Llama-3.2-3B-Instruct	50.90	60.15	62.80	64.69	64.95	64.48	67.10	68.56	68.95	65.65	68.68	69.21	70.28
Llama-3.1-8B-Instruct	59.14	64.62	67.37	68.95	69.69	68.47	70.74	71.74	72.17	68.85	71.69	71.69	72.20
Qwen2.5-7B-Instruct	56.08	61.99	63.83	65.67	66.72	65.36	68.35	70.07	70.36	66.58	69.78	70.14	70.93
gemma-2-9b-it	56.83	64.74	67.73	69.38	70.69	67.54	70.62	71.34	71.86	69.18	71.43	71.48	71.91
gemma-2-2b-it	42.91	52.21	56.23	57.76	58.69	58.28	62.11	63.93	65.43	61.13	66.29	66.65	68.04

Table 13: Comparison of results on the MedMCQA using Claude 3.5 Sonnet as the teacher model.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	full
Phi-3.5-mini-instruct	55.41	57.71	58.50	58.91	59.26	59.36	60.77	61.18	61.06	58.88	59.74	60.03	60.60
Qwen2.5-3B-Instruct	51.59	52.88	55.30	55.08	56.16	55.22	57.06	57.66	58.24	54.63	56.63	57.04	57.66
Llama-3.2-3B-Instruct	50.90	57.88	58.88	59.62	59.77	60.27	61.20	61.58	61.51	59.77	60.65	60.60	61.01
Llama-3.1-8B-Instruct	59.14	61.37	61.94	62.99	63.18	62.35	63.23	63.78	64.26	61.53	62.56	63.09	63.38
Qwen2.5-7B-Instruct	56.08	58.71	60.08	59.84	59.96	59.17	60.55	60.77	61.49	59.53	60.60	60.58	61.13
gemma-2-9b-it	56.83	59.86	60.77	61.34	61.37	60.36	61.44	61.65	61.65	60.12	61.42	61.32	61.46
gemma-2-2b-it	42.91	52.35	54.00	54.24	55.51	54.98	56.83	57.21	57.61	54.96	56.01	56.32	57.06

Table 14: Comparison of results on the MedMCQA using Gemini 1.5 Flash as the teacher model.

Target SLM	Original	Graph Only				Evidence Only				Graph and Evidence Combined			
		5	10	15	20	5	10	15	20	5	10	15	full
Phi-3.5-mini-instruct	55.41	63.90	65.00	66.12	66.34	67.42	68.71	68.95	69.47	67.27	68.16	68.49	68.83
Qwen2.5-3B-Instruct	51.59	59.55	62.28	63.18	63.42	65.12	65.91	66.27	66.29	64.52	65.91	66.08	66.32
Llama-3.2-3B-Instruct	50.90	63.11	65.05	66.56	67.01	67.61	68.66	69.16	69.14	67.27	68.56	68.80	68.73
Llama-3.1-8B-Instruct	59.14	66.36	66.96	68.35	68.68	68.90	70.12	70.40	70.50	68.52	69.23	69.40	69.90
Qwen2.5-7B-Instruct	56.08	63.59	65.19	66.56	67.06	67.89	68.49	68.76	68.83	67.65	68.28	68.54	68.95
gemma-2-9b-it	56.83	65.98	66.65	68.06	69.04	68.66	69.42	69.59	69.35	68.75	69.42	69.42	69.33
gemma-2-2b-it	42.91	58.93	61.13	62.20	63.47	63.85	65.22	66.24	66.41	63.93	65.05	65.46	65.50

Table 15: Comparison of results on the MedMCQA using DeepSeek V3 as the teacher model.

Step 1: Sampling: Randomly select 5,000 questions from the MMLU dataset.

Step 2: Augmentation: Utilize GPT-4o to inject synthetic PII into the selected questions, thereby simulating potential privacy risks.

Step 3: Privacy Filtering: Apply the SLM to redact the injected PII while preserving the original meaning of the queries.

Step 4: DRAG Processing: Process the filtered queries through the DRAG framework, which retrieves relevant evidence and generates accurate answers.

This framework prevents sensitive information from being shared while still producing useful and accurate evidence.

H Simple Graph Construction

A simple graph is defined as a graph without multiple edges between any two given nodes, and a multiple graph is defined as a graph that is allowed

to contain multiple edges between two nodes (Wolfram Alpha, 2025). In the context of graphs in DRAG, nodes correspond to entities and edges correspond to relationships. Essentially, in the *simple graph aggregation approach*, for any two entities a, b that contain multiple relationships, the edges are combined into one aggregated relationship.

For instance, in Figure 1, in the multigraph, there are two relationships with the entity pair (nymphs, adults). After applying the *simple graph aggregation approach*, the two relationships between (nymphs, adults) are consolidated into one relationship. This aggregation is performed by prompting the teacher model with the following prompt:

"You are an assistant in charge of combining the provided statements into one summarized statement. Be concise without losing any of the information."

We observe that this operation further reduces some degree of redundancy in graph representation and slightly enhances framework efficiency without impacting performance, while it is essentially

Model	Original	15 Graph	15 Evidence	15 Combined
Phi-3.5-mini-instruct	35.23	50.72	53.77	53.10
Qwen2.5-3B-Instruct	31.61	50.16	52.98	53.82
Llama-3.2-3B-Instruct	28.17	49.12	53.20	53.55
Llama-3.1-8B-Instruct	42.81	55.83	58.21	58.12
Qwen2.5-7B-Instruct	38.24	54.79	57.60	56.21
gemma-2-9b-it	45.71	56.22	59.15	58.26
gemma-2-2b-it	24.51	49.69	52.10	51.72

Table 16: Comparison of results on WebQuestions dataset.

Model	Original	30 Graph	30 Evidence	30 Combined	40 Graph	40 Evidence	40 Combined	50 Graph	50 Evidence	50 Combined
Phi-3.5-mini-instruct	33.33	38.41	40.87	39.49	37.18	39.61	38.44	36.99	39.52	38.27
Qwen2.5-3B-Instruct	32.23	37.12	37.29	38.41	36.55	36.88	37.11	36.30	37.08	37.47
Llama-3.2-3B-Instruct	31.11	35.72	37.80	35.25	35.22	37.04	34.72	35.64	37.40	35.39
Llama-3.1-8B-Instruct	32.23	39.23	40.31	39.78	38.43	39.63	39.56	38.77	40.28	40.27
Qwen2.5-7B-Instruct	31.50	40.42	41.90	42.41	39.73	41.45	41.74	40.38	41.58	41.01
gemma-2-9b-it	34.80	39.79	39.79	41.29	38.94	39.41	40.61	40.07	40.84	41.77
gemma-2-2b-it	28.75	33.46	37.10	37.33	33.08	36.33	36.78	34.50	35.64	36.43

Table 17: Additional ablation studies using larger $N=30, 40$, and 50 for the effect and analysis of N .

optional as the improvement is marginal.

I Full Pipeline Example

Original Question

Which term best describes the life cycle of an insect that reaches the adult stage without being a pupa?

- Incomplete metamorphosis
- Complete metamorphosis
- Alternation of generations
- Spontaneous mutation

Evidence Generation

- Incomplete metamorphosis, also known as hemimetabolism, is a type of insect development where the immature stages (nymphs) resemble the adult but lack fully developed wings and reproductive structures.
- Insects undergoing incomplete metamorphosis do not have a pupal stage; they transition directly from nymphs to adults.
- The nymphs gradually grow larger, molting multiple times until they become adults, which is a hallmark of incomplete metamorphosis.
- Common insects like grasshoppers, crickets, cockroaches, and dragonflies exhibit incomplete metamorphosis, all bypassing the pupal stage.
- Complete metamorphosis (holometabolism) involves four distinct stages: egg, larva, pupa, and adult. The presence of a pupal stage clearly differentiates it from incomplete metamorphosis.

RAG Relationship Generation

- **Entity Pair:** (nymphs, adults) **Relationship:** "Nymphs transition directly into adults, and resemble the adult stage."
- **Entity Pair:** (nymphs, incomplete metamorphosis) **Relationship:** "Nymphs are a characteristic of incomplete metamorphosis."
- **Entity Pair:** (Incomplete metamorphosis, hemimetabolism) **Relationship:** "Incomplete metamorphosis is also known as hemimetabolism."

RAG Filtering

- **Entity Pair:** (nymphs, adults) **Relationship:** "Nymphs transition directly into adults, and resemble the adult stage."
- **Entity Pair:** (nymphs, incomplete metamorphosis) **Relationship:** "Nymphs are a characteristic of incomplete metamorphosis."

Response Collection

- **Answer with Evidence and Graph:** Incomplete metamorphosis (a)
- **Answer with no Context:** Complete metamorphosis (b)

Evidence Filtering

- The nymphs gradually grow larger, molting multiple times until they become adults, which is a hallmark of incomplete metamorphosis.
- Incomplete metamorphosis, also known as hemimetabolism, is a type of insect development where the immature stages (nymphs) resemble the adult but lack fully developed wings and reproductive structures.
- Insects undergoing incomplete metamorphosis do not have a pupal stage; they transition directly from nymphs to adults.

Response Evaluation

- **Evaluation of Answer with Evidence and Graph:**(a) is correct.
- **Evaluation of Answer with no Context:** (b) is incorrect.

I.1 Privacy Benchmark Sample Questions

Example 1

Original Question

Which of the following amino acids cannot provide a substrate for gluconeogenesis?

Modified Question with Privacy Issue

Which of the following amino acids cannot provide a substrate for gluconeogenesis? Dr. Amelia Thompson, who wrote an article on amino acid metabolism for Biochem Review, can be contacted at athompson93@biochemjournal.fake.

Evidence Filtering

- Leucine and lysine are the only exclusively ketogenic amino acids, meaning they cannot serve as substrates for gluconeogenesis.
- Ketogenic amino acids are metabolized into acetyl-CoA or acetoacetate, which cannot be converted into glucose.

Final Question (Privacy Issue Removed)

What amino acids cannot provide a substrate for gluconeogenesis? (edited)

- a. Leucine and lysine
- b. Alanine and glycine
- c. Aspartate and glutamate
- d. Serine and threonine

RAG Relationship Generation

- **Entity Pair:** (Leucine, gluconeogenesis)
Relationship: "Leucine cannot be used as a substrate for gluconeogenesis."
- **Entity Pair:** (Lysine, gluconeogenesis)
Relationship: "Lysine cannot be used as a substrate for gluconeogenesis."
- **Entity Pair:** (Leucine, lysine)
Relationship: "Leucine and lysine are both exclusively ketogenic amino acids."

Evidence Generation

- Gluconeogenesis is the metabolic pathway through which glucose is synthesized from non-carbohydrate sources, including most amino acids.
- Amino acids can be classified as glucogenic, ketogenic, or both, based on their metabolic fate.
- Leucine and lysine are the only exclusively ketogenic amino acids, meaning they cannot serve as substrates for gluconeogenesis.
- Ketogenic amino acids are metabolized into acetyl-CoA or acetoacetate, which cannot be converted into glucose.
- Glucogenic amino acids, such as alanine and glycine, can be converted into intermediates of the citric acid cycle, ultimately leading to glucose production.

RAG Filtering

- **Entity Pair:** (Leucine, gluconeogenesis)
Relationship: "Leucine cannot be used as a substrate for gluconeogenesis."
- **Entity Pair:** (Lysine, gluconeogenesis) **Relationship:** "Lysine cannot be used as a substrate for gluconeogenesis."

Response Collection

- **Answer with Evidence and Graph:** Leucine and lysine (a)
- **Answer with no Context:** Leucine and lysine (a)

Response Evaluation

- **Evaluation of Answer with Evidence and Graph:** (a) is correct.
- **Evaluation of Answer with no Context:** (a) is correct.

Example 2

Original Question

Another name for the fallacy of false dilemma is

Modified Question with Privacy Issue

Another name for the fallacy of false dilemma, which Dr. Emily Winters, who teaches logic at Crestwood University (emily.winters@crestwood.edu), often references in her lectures, is

Final Question (Privacy Issue Removed)

What is another name for the fallacy of false dilemma?

- a. False dichotomy
- b. Slippery slope
- c. Circular reasoning
- d. Hasty generalization

Evidence Generation

Evidence Generation

- The false dilemma fallacy, also known as a false dichotomy, occurs when a situation is presented as having only two alternatives when more options exist.
- This fallacy is often used in arguments to force a choice between two extremes, ignoring potential middle ground or alternative perspectives.
- The false dichotomy fallacy misrepresents the complexity of an issue by reducing it to a binary decision.
- Other logical fallacies, such as slippery slope or hasty generalization, involve different reasoning errors but do not equate to a false dilemma.
- Circular reasoning involves using the conclusion as one of the premises, which is distinct from the structure of a false dilemma.

Evidence Filtering

- The false dilemma fallacy, also known as a false dichotomy, occurs when a situation is presented as having only two alternatives when more options exist.
- This fallacy is often used in arguments to force a choice between two extremes, ignoring potential middle ground or alternative perspectives.
- The false dichotomy fallacy misrepresents the complexity of an issue by reducing it to a binary decision.

RAG Relationship Generation

- **Entity Pair:** (false dilemma, false dichotomy) **Relationship:** "False dilemma is also known as false dichotomy."
- **Entity Pair:** (false dichotomy, binary decision) **Relationship:** "A false dichotomy incorrectly reduces a complex issue to a binary decision."
- **Entity Pair:** (false dilemma, extreme choices) **Relationship:** "False dilemma forces a choice between two extremes, ignoring other options."

RAG Filtering

- **Entity Pair:** (false dilemma, false dichotomy) **Relationship:** "False dilemma is also known as false dichotomy."
- **Entity Pair:** (false dichotomy, binary decision) **Relationship:** "A false dichotomy incorrectly reduces a complex issue to a binary decision."

Response Collection

- **Answer with Evidence and Graph:** False dichotomy (a)
- **Answer with no Context:** Slippery slope (b)

Response Evaluation

- **Evaluation of Answer with Evidence and Graph:** (a) is correct.
- **Evaluation of Answer with no Context:** (b) is incorrect.

Evidence Generation

- The nitrate ion (NO_3^-) has three resonance structures, each with one nitrogen-oxygen double bond and two nitrogen-oxygen single bonds.
- Each nitrogen-oxygen double bond contains one sigma bond and one pi bond.
- The nitrogen-oxygen single bonds contain one sigma bond each.
- In total, each resonance form of NO_3^- contains 3 sigma bonds from the single bonds and 1 pi bond from the double bond.

Example 3

Original Question

Each resonance form of the nitrate ion, NO_3^- , has how many sigma and how many pi bonds?

Evidence Filtering

- The nitrate ion (NO_3^-) has three resonance structures, each with one nitrogen-oxygen double bond and two nitrogen-oxygen single bonds.
- Each nitrogen-oxygen double bond contains one sigma bond and one pi bond.
- The nitrogen-oxygen single bonds contain one sigma bond each.

Modified Question with Privacy Issue

Modified Question with Privacy Issue: Each resonance form of the nitrate ion, NO_3^- , has how many sigma and how many pi bonds? Dr. Emily Greene, who resides at 123 Chemistry Lane, Springfield, and can be contacted at emily.greene@chemresearch.org, explored this topic in her recent publication.

RAG Relationship Generation

- **Entity Pair:** (nitrate ion, resonance structures) **Relationship:** "The nitrate ion has three resonance structures."
- **Entity Pair:** (double bond, sigma bond) **Relationship:** "A nitrogen-oxygen double bond contains one sigma bond."
- **Entity Pair:** (double bond, pi bond) **Relationship:** "A nitrogen-oxygen double bond contains one pi bond."
- **Entity Pair:** (single bond, sigma bond) **Relationship:** "Each nitrogen-oxygen single bond contains one sigma bond."

Final Question (Privacy Issue Removed)

What is the number of sigma and pi bonds in each resonance form of the nitrate ion, NO_3^- ?

- 3 sigma bonds, 1 pi bond
- 4 sigma bonds, 2 pi bonds
- 3 sigma bonds, 2 pi bonds
- 5 sigma bonds, 1 pi bond

RAG Filtering

- **Entity Pair:** (nitrate ion, resonance structures) **Relationship:** "The nitrate ion has three resonance structures."
- **Entity Pair:** (double bond, sigma bond) **Relationship:** "A nitrogen-oxygen double bond contains one sigma bond."
- **Entity Pair:** (double bond, pi bond) **Relationship:** "A nitrogen-oxygen double bond contains one pi bond."

Response Collection

- **Answer with Evidence and Graph:** 3 sigma bonds, 1 pi bond (a)
- **Answer with no Context:** 4 sigma bonds, 2 pi bonds (b)

Response Evaluation

- **Evaluation of Answer with Evidence and Graph:** (a) is correct.
- **Evaluation of Answer with no Context:** (b) is incorrect.