

PRISM-RAG: Boosting RAG Factuality with Distractor Resilience and Strategized Reasoning

Mohammad Kachuee¹, Teja Gollapudi¹, Minseok Kim¹, Yin Huang¹, Kai Sun¹,
Xiao Yang¹, Jiaqi Wang¹, Nirav Shah¹, Yue Liu¹, Aaron Colak¹,
Anuj Kumar¹, Wen-tau Yih², Xin Luna Dong¹

¹Meta Reality Labs, ²Meta FAIR

Abstract

Retrieval-augmented generation (RAG) often falls short when retrieved context includes confusing semi-relevant passages, or when answering questions require deep contextual understanding and reasoning. We propose an efficient fine-tuning framework, called PRISM-RAG, that (i) trains the model with distractor-aware QA pairs mixing gold evidence with subtle distractor passages, and (ii) instills reasoning-centric habits that make the LLM plan, rationalize, and synthesize without relying on extensive human engineered instructions. Evaluated across 12 open-book RAG QA benchmarks spanning diverse application domains and scenarios, PRISM-RAG improves average factuality by 5.4%, outperforming state-of-the-art solutions.

1 Introduction

Factual question answering (QA) is an important application for large language models (LLMs). However, LLMs lack the necessary knowledge to answer questions that require current or external information not present in their parametric knowledge. To address this, *retrieval-augmented generation* (RAG) is commonly employed to incorporate grounding context into the prompt (Lewis et al., 2020). The grounding context is usually organized as reference documents fetched by search, semantic retrieval, or knowledge tools (Yang et al., 2024; Friel et al., 2024; Kachuee et al., 2025). In such settings, the LLM is instructed to utilize information presented in the references in conjunction with its parametric knowledge to deliver the most accurate responses (Huang et al., 2025).

Despite its intuitive appeal, RAG still struggles to deliver reliable answers. Prior studies show that appending large blocks of retrieved content, such as entire web pages or lengthy documents, can overwhelm the model and induce hallucinations (Fang et al., 2024). The deficit becomes acute for nu-

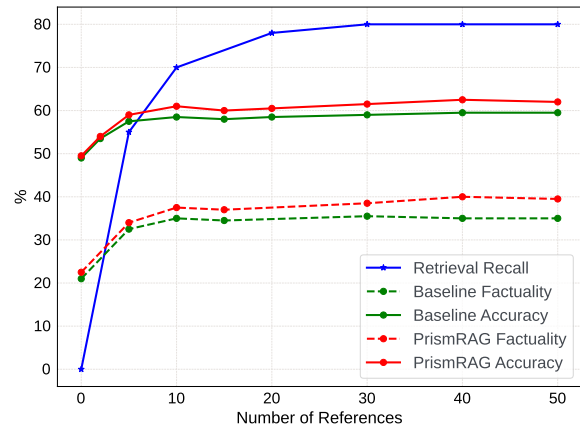


Figure 1: On the CRAG (Yang et al., 2024) benchmark, as we increase retrieval pages from 0 to 50, whereas the retrieval recall increases to 80%, answer factuality score (computed by accuracy minus hallucination rate) flattens out after 10 pages for Llama-3.1-70b-instruct. PRISM-RAG improves over baseline by 4.5%, allowing for further QA improvement as we retrieve more pages.

anced questions that demand synthesizing evidence scattered across several sources. As revealed in Figure 1, expanding the number of context webpages from 5 to 50 boosts retrieval recall by $\sim 25\%$, yet the factuality curve plateaus, underscoring the limited benefit of ever-larger context windows.

In this paper, we propose PRISM-RAG, an approach that enhances LLM’s capability in generating answers from multiple retrieval results. To mitigate the negative impact of semi-relevant content retrieved alongside useful passages, we develop methods for generating training data targeted at building robustness against such distractors. Furthermore, to address the reasoning challenges inherent in RAG QA—such as assessing the relevance of retrieved passages, resolving inconsistencies, and aggregating information—we fine-tune the LLM to strategize and deliberate before producing answers. Unlike approaches that rely heavily on extensive chain-of-thought (CoT) prompt engineering, our

method directly improves the model’s reasoning competence and generalization. To summarize, we make three contributions.

1. We propose PRISM-RAG, a fine-tuning approach that teaches LLMs to be robust against retrieval noises, and to plan and rationalize for improved answer generation.
2. We propose a training data generation framework that combines synthetic data generation with LLM-based verification, to generate high-quality training data in a scalable fashion.
3. We conducted comprehensive empirical study on 12 benchmarks, showing that PRISM-RAG improves over baseline by 5.4%, and outperforms state-of-the-art solutions. As illustrated in Figure 1, PRISM-RAG boosts factuality even when RAG recall plateaus, demonstrating the model’s ability to effectively leverage additional references without being distracted by the extra context.

2 Related Work

Since the early breakthroughs in LLMs, RAG QA has emerged as a significant application area. This approach offers a streamlined alternative to traditional web search and research for finding relevant answers (Lewis et al., 2020). Additionally, RAG QA systems facilitate the development of domain-specific applications, supporting use cases such as product user manuals, legal documents, or financial reports to name a few (Sadat et al., 2023; Hendrycks et al., 2021; Chen et al., 2021).

RAG QA is an active area of research with multiple dimensions, including information retrieval (Friel et al., 2024; Kachuee et al., 2025), LLM prompting (Asai et al., 2023; Semnani et al., 2023), LLM fine-tuning (Cai et al., 2024; Gekhman et al., 2024; Lin et al., 2024), evaluation (Yang et al., 2024), and guard-railing (Kim et al., 2024).

More relevant to this work is the research on enhancing the core capabilities of LLMs through fine-tuning, aimed at addressing task complexities. For instance, Fang et al. (2024) evaluated the impact of retrieval noise on the answer quality. They categorized the noise into three types: irrelevant, relevant, and counterfactual. Their findings indicate that relevant and counterfactual noises are most detrimental. RobustRAG (Xiang et al.) is another example which proposed building robustness by an isolate-then-aggregate strategy.

From another perspective, recent literature highlights CoT reasoning as a powerful capability for enhancing the response quality. Reasoning helps decompose complex problems into smaller but more manageable parts, facilitating advanced contextual understanding. The current CoT literature primarily focuses on reasoning for math, logic, and decision-making, while reasoning in RAG QA remains a less explored area (Wang and Zhou, 2024; Zheng et al., 2023; Phan et al., 2023; Zhang et al., 2024b). Wang et al. (2025) proposed leveraging multi-step reasoning to dynamically query, retrieve, and evaluate documents, thereby enhancing retrieval processes. Zhang et al. (2024a) introduced retrieval augmented fine-tuning (RAFT) and showed the value of fine-tuning on the RAG QA task, especially for CoT style responses. Another example is LLMQuoter (Bezerra and Weigang, 2025), a recent method that involves fine-tuning to produce quotes before generating the answer.

3 Problem Definition

We consider the problem of QA given a set of retrieved documents. Specifically, we focus on enhancing core capabilities of LLMs through fine-tuning to best leverage the internal knowledge, common sense, and reasoning capabilities in conjunction with a set of retrieved documents to produce the most factual answers. Formally, given a question Q , LLM parameters θ , a set of retrieved documents $D = \{d_1 \dots d_n\}$, and any relevant contextual information (e.g., user time and location) C , the generative model generates an answer A :

$$A = \mathcal{G}(Q, D, C|\theta), \quad (1)$$

where $\mathcal{G}$ is used to represent the generative process. The objectives are: (i) to produce an answer that addresses the question while being grounded on information present in D and C ; (ii) to refrain from answering the question if the available knowledge and context do not provide sufficient information.

4 Proposed Method

In this section, we present our fine-tuning approach to enhance LLM’s answer generation capability in presence of retrieval noises. Our approach is based on two key intuitions. First, although LLMs in general can decide the relevancy between the question and retrieved content, they may make mistakes for nuances such as for events with different dates and locations. We can enhance this capability with

tailored fine-tuning (Section 4.2). Second, RAG answer generation is reasoning heavy, deciding which retrieved passages are relevant, identifying inconsistencies among retrieved content, and aggregating information across passages when necessary. Thus, we shall fine tune LLM’s reasoning capability for generating better answers (Section 4.3). These two methods improve RAG summarization from two complementary and orthogonal dimensions.

As we generate fine-tuning data, we face a dilemma. On the one hand, relying on human annotators to produce high-quality data is expensive and not scalable; this is especially true since RAG retrieval results typically contain many pages of unstructured (e.g., long concatenated text chunks) or semi-structured (e.g., parsed tables) text, as well as structured data blobs (e.g., json output from knowledge graphs), not suited for human readers (Sun et al., 2025). On the other hand, synthetic training data may not exhibit high quality, and noisy labels may even encourage more hallucinations. We describe in this section a simple yet effective solution for automatically producing high-quality seed data (Section 4.1), and how we use LLMs to iteratively generate examples and check example quality (Section 4.2-4.3).

4.1 Seed QA Data

We start with generating seed question-answer-passage triplets. To scale up the process, we generate synthetic QA pairs from a given set of raw documents. Initially, we split documents into references, and then randomly select a “golden” reference to design a QA pair. Solving this reverse problem is a more straightforward task and can be accomplished using a reasonably strong LLM. This approach helps to produce high-quality data at scale, as the primary limiting factors would be access to raw content and compute.

We generate QA pairs using two main content sources. (a) Wiki Pages, content sampled from the English Wikipedia dataset¹; (b) Web Search, parsed web pages retrieved using web search queries in response to a set of internal factual questions (details in Appendix C).

4.2 Improving Resilience to Distractors

Recent studies have highlighted the detrimental impact of distractor content (Fang et al., 2024). To address this challenge, we propose fine-tuning

on distractor-augmented content. By incorporating semi-relevant or confusing information into the training data, our approach aims to enhance the model’s resilience to relevant retrieval noise thereby mitigating the risk of hallucinations and improving overall factuality.

Informed by observations from real-world RAG QA applications and based on the seed data from Section 4.1, we adopt a synthetic approach to generate targeted distractor content. This method allows for large-scale generation and precise control over the introduction of relevant noise on named-entities and temporal information. We hypothesize that while direct solutions, such as utilizing irrelevant passages from a text corpus (Zhang et al., 2024a) as distractor content add retrieval noise, they do not ensure that the noise is relevant. Through targeted synthetic generation of distractions via modifying the golden passage, we introduce retrieval noises that are specifically tailored to challenge named entities and temporal information, which our observations have shown to be particularly problematic for summarization models. As depicted in Figure 2, this process consists of three steps.

1. **Synthetic distractor generation:** Given a question, user context ground-truth answer, and its grounding passage, we synthetically generate a distractor passage and an open-ended question (prompt in Appendix G.8). Specifically, we conduct three steps:

- Identifying key entities, locations, and temporal expressions in the golden passage that are most relevant.
- Reformulating the original question into a more open-ended form, such that a plausible answer can be extracted from both the golden and distractor passages.
- Systematically altering the identified named-entities including person/object names, locations, and temporal information to create a distractor passage that is grammatically coherent, while being stylistically similar to the original.

2. **Critique the effectiveness of synthetic distractor passage:** We employ an evaluation prompt to assess the effectiveness of the generated distractor passages on a scale of 1 to 5 (details in Appendix G.7). This step is crucial in ensuring that the distractors are relevant, confusing, and well-formatted.

¹<https://dumps.wikimedia.org>

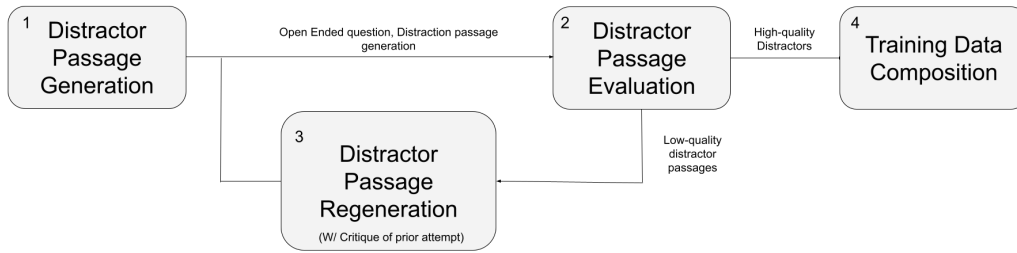


Figure 2: Overview of the synthetic distractor generation process.

3. **Train Data Composition:** For each sample, we iterate on the above mentioned steps until each sample scores 4 or higher or reaches five iterations. For samples that pass the quality bar, we compose the training instances by combining golden passage and the generated distractor passage as references for the task.

4.3 Improving Reasoning and Generalization

Numerous studies have highlighted the advantages of CoT fine-tuning, as opposed to directly training on the final answers (Kim et al., 2023). We believe this is particularly crucial for the QA task, where directly training on the final answer may increase risk of hallucinations due to a knowledge mismatch with the pre-training (Gekhman et al., 2024).

Nonetheless, there are two main considerations related to CoT fine-tuning. First, using human annotators to prepare reasoning data is challenging, as authoring rationales is not a well-defined task. Second, as reported in the literature (Cheng et al., 2024; Liu et al., 2024) and confirmed by our experiments (Appendix A), CoT instructions significantly impact the quality of the final answer. However, prompt engineering is resource-intensive, and optimizing instructions for each domain or benchmark is neither scalable nor optimal.

To address these challenges, we propose strategization as a meta-task aimed at reducing reliance on CoT instruction engineering while enhancing the performance. In this approach, the generative model is first tasked with producing a reasoning strategy that outlines the necessary thought steps. It then follows this strategy to generate the CoT and the final answer. We hypothesize this approach teaches the model “*how to think*” rather than “*what to think*”, which in turn improves its ability to tackle different problem settings by dynamically adjusting CoT steps. See Figure 3 for an example (additional examples in Appendix B). As illustrated in Figure 4, this process consists of:

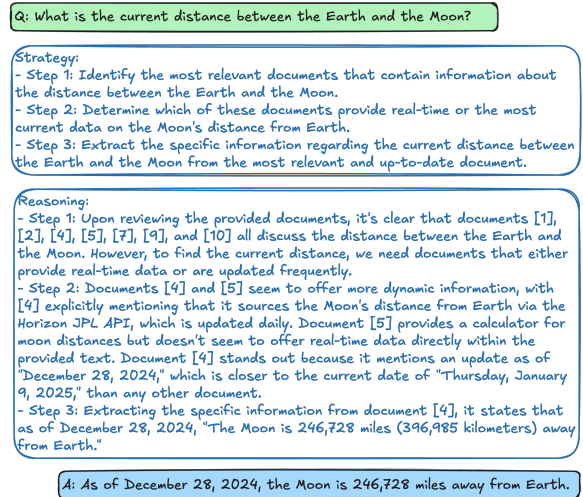


Figure 3: An example of strategization CoT inference process demonstrating the dynamic generation of CoT steps. Due to space limitations, reference documents are not shown here. The generated strategy outlines high-level reasoning steps to extract the relevant information from references, cite specific key pieces to support the asked question, and produce a final answer to the question. Subsequently, the reasoning follows the strategy steps to extract, quote, and summarize the final answer.

1. **Thought Generation:** In the initial thought generation pass, we use an instruction prompt that takes the question and a set of references, and produces a strategy that outlines the necessary steps. This strategy guides the reasoning process and leads to the final answer (see the prompt in Appendix G.6).
2. **Thought Evaluation:** Subsequently, we reason about the quality of the generated rationale, its thought process, and how it leads to the final answer. Then, we assign a score of 1 to 4, 4 indicating excellent reasoning that leads to the ground-truth answer (prompt in Appendix G.3). We consider any samples with score below 4 to require further iterations.
3. **Answer Evaluation:** We are exclusively interested in rationales that lead to the correct

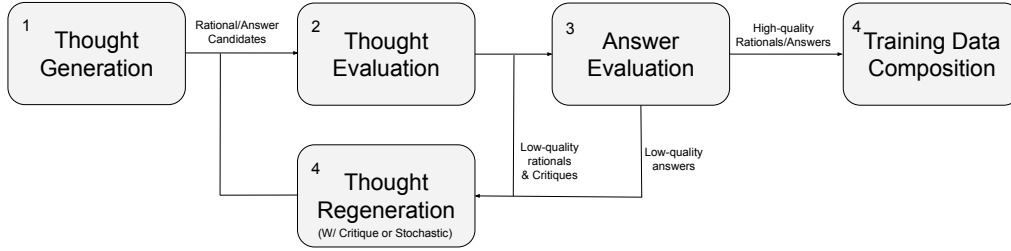


Figure 4: Overview of the iterative synthetic CoT generation process.

Method	Seed Data	Samples	Avg. Refs.
Distractor Resilience	Web Search	2,589	2
Dynamic Strategization	Web Search	2,674	9.1
Dynamic Strategization	Wiki Pages	5,079	4.6
Total		10,342	

Table 1: Breakdown of the final training data mix.

answer. For thoughts assessed as high-quality, we compare the candidate and the ground-truth answers to ensure the question is being fully answered and verify the factual consistency. We use a scoring system of 1 to 4 as the previous step (prompt in Appendix G.4).

4. **Thought Regeneration:** To regenerate thoughts, we consider a combination of stochastic generation and critique-based revision methods. For stochastic generation, we simply rerun the inference in step 1 (temperature=1.0 and top-p=0.90). However, for revision, we use analysis produced during the thought evaluation step as critique feedback (prompt in Appendix G.5).
5. **Train Data Composition:** We follow the iterative steps above until we reach a high-quality synthetic rationale, or exhaust a budget of 10 attempts. We use stochastic thought regeneration for the first 6 attempts, and then switch to critique-based revision for the rest.

4.4 Final training data

Table 1 presents the breakdown of the final training data, consisting of a mix of distractor resilience and dynamic strategization.

5 Experiments

5.1 Experiment Settings

5.1.1 Benchmarks

In our experiments, we employed a diverse set of 12 public RAG QA benchmarks, covering various question domains, answer formats, and reference

document types. Specifically, we used the web portion of the CRAG dataset (Yang et al., 2024), and filtered out the false-premise type of questions, resulting in a total of 643 samples. Furthermore, we utilized BGE embedding (Xiao et al., 2024) to rank references by semantic relevance to the question.

For the remaining 11, we leveraged pre-processed data from RAGBench (Friel et al., 2024), down-sampling each to 100 samples. These benchmarks cover relevant real-world application areas, including health (Jin et al., 2019; Möller et al., 2020), finance (Chen et al., 2021; Zhu et al., 2021), customer support (Sadat et al., 2023; Nandy et al., 2021; Malaviya et al., 2023), legal (Hendrycks et al., 2021), and general knowledge (Yang et al., 2018; Kamalloo et al., 2023; Malaviya et al., 2023).

5.1.2 Metrics

To evaluate the factual quality of generated answers, we follow metrics suggested by Yang et al. (2024), classifying each response as either accurate, hallucinated, or missing. Accurate responses fully and accurately answer the question. Hallucinated answers contain inaccurate or misleading information. Missing responses either refuse to answer or fail to completely address the question. To summarize the overall factuality, we use a *factuality score*, defined as the accuracy rate minus the hallucination rate. Factuality score provides a scalar measure of overall answer quality, where hallucinated answers are penalized twice as much as missing answers.

To evaluate factuality for CRAG, HotpotQA, MS Marco, FinQA, TAT-QA, HAGRID, and ExpertQA, we employed an LLM-as-judge to compare the ground-truth with the candidate answers. For other benchmarks, we utilized a fact-checking tool similar to VeriScore (Song et al., 2024), which is more suitable for long-form answers or scenarios with multiple potential answers.

Benchmark	Baseline	NaiveSFT	STaR (Zelikman et al., 2022)	LLMQuoter (Bezerra et al., 2025)	RAFT (Zhang et al., 2024a)	PRISM-RAG (This Work)
CRAG(Yang et al., 2024)	34.2%	27.8%	37.2%	34.4%	34.3%	39.2%
CovidQA(Möller et al., 2020)	80.0%	83.0%	76.0%	89.0%	90.0%	95.0%
DelucionQA(Sadat et al., 2023)	89.0%	90.0%	92.0%	89.0%	92.0	97.0%
Emanuel(Nandy et al., 2021)	92.0%	91.0%	91.0%	91.0%	92.0%	98.0%
ExpertQA(Malaviya et al., 2023)	83.0%	83.0%	84.0%	82.0%	83.0%	83.0%
FinQA(Chen et al., 2021)	83.0%	68.0%	72.0%	83.0%	75.0%	71.0%
HAGRID(Kamalloo et al., 2023)	89.0%	89.0%	83.0%	89.0%	87.0%	90.0%
HotpotQA(Yang et al., 2018)	93.0%	63.0%	58.0%	92.0%	90.0%	89.0%
MS Macro(Nguyen et al., 2016)	82.0%	76.0%	81.0%	81.0%	81.0%	82.0%
PubMedQA(Jin et al., 2019)	80.0%	78.0%	76.0%	77.0%	78.0%	90.0%
TAT-QA(Zhu et al., 2021)	77.0%	66.0%	66.0%	79.0%	90.0%	90.0%
TechQA(Castelli et al., 2019)	58.0%	62.0%	58.0%	75.0%	79.0%	82.0%
Avg.	78.4%	73.1%	72.9%	80.1%	80.9%	83.8%

Table 2: Comparison of RAG QA factuality scores for different approaches. PRISM-RAG improves average factuality by 5.4% over the baseline, and it outperforms SOTA such as STaR, LLMQuoter, and RAFT by $\sim 3\%$ – 11% .

5.1.3 Implementation

We used Llama-3.1-70b-instruct (Grattafiori et al., 2024) as the base model across all our experiments. For fine-tuning, we experimented with learning rate of 10^{-5} , computing loss over the entire assistant response (strategy, thought, and answer when present) but not using loss over the instruction prompt. For inference, we use a typical RAG prompt as in Appendix G.2 with generation temperature of 0.2 and top-p of 0.9.

Regarding the inference cost of producing the final synthetic train data, the iterative process for the distractor and CoT data took about 80,000 inference passes using an internal inference API. This translates to less than 48 hours of total wall clock time for our specific setup. Regarding the cost of SFT training, it took about 12 hours to post train the final model on a cluster of 64 NVIDIA H100 GPUs.

For comparisons with other work, to ensure fairness, we leveraged the same seed data and dataset size, while re-implementing their train data generation logic. Specifically, for: (a) **NaiveSFT**: Directly trained on final answers without strategization and CoT. (b) **STaR** (Zelikman et al., 2022): Implemented a rationalization chain for RAG QA using static CoT instructions. (c) **LLMQuoter** (Bezerra and Weigang, 2025): Implemented the same quote extraction instructions to produce training data. (d) **RAFT** (Zhang et al., 2024a): Reused their data generation code-base replacing closed-sourced model endpoints with LLaMA.

5.2 Results

Overall results: Table 2 presents a comparison of the proposed method (PRISM-RAG) with the baseline model and other related works in the literature. As shown in this Table, the proposed method demonstrates substantial improvements in factuality across benchmarks. Notably, PRISM-RAG achieves a factuality score improvement of 5% for the CRAG dataset, delivering the best results in 9 out of 12 benchmarks, and an overall macro-average gain of 5.4% over the baseline. From the average results, NaiveSFT and STaR regress over the baseline, while LLMQuoter and RAFT show more promise. This finding confirms the drawbacks of directly training on QA labels, while highlighting the benefits of training on intermediate tasks (e.g., quote extraction in LLMQuoter) and training for resilience (e.g., irrelevant references in RAFT). The breakdown of results is provided in Appendix D.

Sensitivity to references: We conducted an experiment using the CRAG benchmark by limiting the number of references and measuring the impact on the final QA performance. From Figure 1 results, the proposed method consistently outperforms the baseline, with its margin of improvement increasing as more references are used. This demonstrates the effectiveness of PRISM-RAG in utilizing the retrieved documents (see breakdown charts in Appendix E).

Ablation study: Table 3 presents an ablation study that examines the impact of two major components of the proposed fine-tuning method: distractor resilience and dynamic strategization. From

this analysis, strategization contributes to increased accuracy and reduced hallucinations, while the distractor resilience task specifically aids in reducing hallucinations. The combination of both methods yields the best overall factuality results, highlighting their complementary roles and effectiveness.

Additional experiments for closed-book settings are provided in Appendix F.

Method	Accurate	Hallucinated	Missing	Factuality
Baseline	59.1%	24.9%	16.0%	34.2%
PRISM-RAG	62.1%	22.9%	15.1%	39.2%
-Distractor	59.3%	23.2%	17.6%	37.0%
-Strategization	62.4%	23.2%	12.3%	36.1%

Table 3: Ablation study using the CRAG dataset for the proposed fine-tuning method based on distractor resilience and dynamic strategization.

6 Conclusion

In this paper, we tackle key challenges in open-book question answering by bolstering the robustness of LLMs against distractors and enhancing their reasoning capabilities. Our approach involved fine-tuning LLMs with synthetic distractor data and introducing a novel strategization method focused on dynamic CoT step generation. We showed that leveraging the proposed iterative synthetic data generation and refinement is an effective method to produce train data with sufficient scale and complexity for finetuning LLMs to boost their RAG QA capability. Extensive evaluations across 12 public benchmarks demonstrated significant improvements in factual accuracy. These findings highlight the potential of our approach to advance the state-of-the-art, ultimately delivering a more factual and reliable QA system.

Limitations

While our proposed method shows significant improvements in factuality for open-book question answering, several limitations persist. First, the reliance on synthetically generated distractor data may not fully capture the complexity and variability of real-world distractors, potentially limiting the model’s robustness in diverse scenarios. Specifically, we acknowledge that the current method may not fully capture certain real-world complexities such as: HTML parsing artifacts/formatting inconsistencies, implicit contradictions requiring deep domain knowledge, and perspective differences in presenting similar events. Additionally, the use

of LLM-as-judge for factuality evaluation, both in our training data pipeline and benchmarking, presents its own challenges. For instance, it may exhibit bias towards the presence or absence of additional explanations, and its behavior can become less predictable when there is a mismatch between its internal parametric knowledge and the retrieved references.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- Yuri Façanha Bezerra and Li Weigang. 2025. Llmquote: Enhancing rag capabilities through efficient quote extraction from large contexts. *arXiv preprint arXiv:2501.05554*.
- Tianchi Cai, Zhiwen Tan, Xierui Song, Tao Sun, Jiyan Jiang, Yunqi Xu, Yinger Zhang, and Jinjie Gu. 2024. Forag: Factuality-optimized retrieval augmented generation for web-enhanced long-form question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 199–210.
- Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, and 1 others. 2019. The techqa dataset. *arXiv preprint arXiv:1911.02984*.
- Zhiyu Chen, Wenhua Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Chainlm: Empowering large language models with improved chain-of-thought prompting. *arXiv preprint arXiv:2403.14312*.
- Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *arXiv preprint arXiv:2405.20978*.
- Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: an expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.
- Yin Huang, Yifan Ethan Xu, Kai Sun, Vera Yan, Alicia Sun, Haidar Khan, Jimmy Nguyen, Mohammad Kachuee, Zhaojiang Lin, Yue Liu, and 1 others. 2025. Confqa: Answer only if you are confident. *arXiv preprint arXiv:2506.07309*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Mohammad Kachuee, Sarthak Ahuja, Vaibhav Kumar, Puyang Xu, and Xiaohu Liu. 2025. Improving tool retrieval by leveraging large language models for query generation. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 29–38.
- Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.
- Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.
- Siwon Kim, Shuyang Dai, Mohammad Kachuee, Shayan Ray, Tara Taghavi, and Sungroh Yoon. 2024. Groundial: Human-norm grounded safe dialog response generation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1582–1588.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *arXiv preprint arXiv:2405.01525*.
- Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. 2024. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.
- Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *arXiv preprint arXiv:2309.07852*.
- Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.
- Abhilash Nandy, Soumya Sharma, Shubham Madhaski, Kapil Sachdeva, Pawan Goyal, and Niloy Ganguly. 2021. Question answering over electronic devices: A new benchmark dataset and a multi-task learning based qa framework. *arXiv preprint arXiv:2109.05897*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.
- Du Phan, Matthew Douglas Hoffman, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Soutsov, Charles Sutton, Sharad Vikram, and Rif A Saurous. 2023. Training chain-of-thought via latent-variable inference. *Advances in Neural Information Processing Systems*, 36:72819–72841.
- Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering. *arXiv preprint arXiv:2312.05200*.
- Sina J Semnani, Violet Z Yao, Heidi C Zhang, and Monica S Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *arXiv preprint arXiv:2305.14292*.
- Yixiao Song, Yekyung Kim, and Mohit Iyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474.
- Kai Sun, Yin Huang, Srishti Mehra, Mohammad Kachuee, Xilun Chen, Renjie Tao, Zhaojiang Lin, Andrea Jessee, Nirav Shah, Alex Betty, Yue Liu, Anuj Kumar, Wen tau Yih, and Xin Luna Dong. 2025. Knowledge extraction on semi-structured content: Does it remain relevant for question answering in the era of llms? *Preprint*, arXiv:2509.25107.
- Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: how knowledgeable are large language models (llms)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.
- Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

- Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation. *arXiv preprint arXiv:2501.14342*.
- Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.
- Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust rag against retrieval corruption. In *ICML 2024 Next Generation of AI Safety Workshop*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, and 1 others. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.
- Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024a. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.
- Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024b. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356.
- Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Sensitivity to CoT Instructions

To study the sensitivity of RAG QA generated answers to prompt instructions, we experimented with three variations of the instruction prompt in G.2 (Prompt-A). For the first variant (Prompt-B), updated step 1 to directly ask for evaluating individual references, and consolidated step 2 and 3 with emphasis on saying “I don’t know” when provided information is insufficient. For the second variant (Prompt-C), we removed the emphasis on “I don’t know”. Finally, in the last variant (Prompt-D), we removed all CoT step instructions and just asked for thinking before answering the question. In these experiments, we used llama-3.3-70b-instruct (Grattafiori et al., 2024) as the generative model and a holdout set from the Web Pages (C) dataset as our test set.

Based on the comparison of results presented in Table 4, alight the base instructions are the same, and the only difference is CoT step instructions, we see significant sensitivity to the prompt variations. This is particularly important for the case of refusal answers as explicit step instruction on refusing with “I don’t know”, seem to severely increase the refusal rate, while a very similar instruction is provided in the task definition above CoT steps for the base prompt. Also, it is worth noting that having no instructions, and simply asking to “think before answer” has a competitive performance compared to engineered prompts for this specific task.

Instruction	Accuracy	Hallucination	Missing	Factuality
Prompt-A	69.9%	18.4%	11.7%	51.5%
Prompt-B	63.2%	15.5%	21.4%	47.7%
Prompt-C	69.2%	18.7%	12.1%	50.5%
Prompt-D	69.3%	19.6%	11.1%	49.7%

Table 4: Experiments to study the sensitivity of RAG QA factuality metrics to CoT step instructions.

B Strategization Examples

We present two additional examples for the strategization CoT process. In the first example, the reasoning steps are dynamically defined to focus on specific sections and entities in a legal document to extract the final summary. However, in the second example, the focus is on comparing and analyzing entities relented such as language and publication year for a book title before proving the final answer.

Q: What was the main purpose of Amendment XIV, which was adopted by the 1854 and 1855 Legislatures and approved by the people on May 23, 1855?

Strategy:

- Step 1: Identify the relevant document that contains information about Amendment XIV.
- Step 2: Locate the specific section or paragraph in the identified document that discusses Amendment XIV.
- Step 3: Read and understand the purpose of Amendment XIV as described in the document.
- Step 4: Summarize the main purpose of Amendment XIV based on the information found.

Reasoning:

- Step 1: The relevant document that contains information about Amendment XIV is ID:6, as it discusses amendments adopted after the convention, including Amendment XIV.
- Step 2: The specific section that discusses Amendment XIV is the part that lists the amendments adopted by the 1854 and 1855 Legislatures and approved by the people on May 23, 1855.
- Step 3: According to the document, Amendment XIV provided for plurality voting for all civil officers of the state, to ensure that there would be no repeat of the unsatisfactory situation where the legislature decided the 1850 gubernatorial election.
- Step 4: The main purpose of Amendment XIV was to introduce plurality voting for all civil officers of the state, preventing the legislature from deciding elections when no candidate received a majority of the votes.

A: Amendment XIV provided for plurality voting for all civil officers of the state, to ensure that there would be no repeat of the unsatisfactory situation where the legislature decided the 1850 gubernatorial election.

Q: In what year was Janine Chasseguet-Smirgel's book 'The ego ideal: a psychoanalytic essay on the malady of the ideal' first published in English?

Strategy:

- Step 1: Review the provided references to identify the relevant information about Janine Chasseguet-Smirgel's book 'The ego ideal: a psychoanalytic essay on the malady of the ideal'.
- Step 2: Look for the publication year of the book in English within the relevant references.
- Step 3: Verify if the book was originally published in another language and if so, determine the year of the original publication.
- Step 4: Confirm the correct publication year in English based on the information gathered.

Reasoning:

- Step 1: Upon reviewing the references, it appears that the most relevant information about the book 'The ego ideal: a psychoanalytic essay on the malady of the ideal' can be found in references [1] and [3].
- Step 2: In reference [1], it is mentioned that the book 'The ego ideal: a psychoanalytic essay on the malady of the ideal' was written by Janine Chasseguet-Smirgel, but the publication year in English is not explicitly stated. However, reference [3] provides a bibliography that includes the book with a publication note.
- Step 3: According to reference [3], the book 'The ego ideal: a psychoanalytic essay on the malady of the ideal' was originally published in French as 'La maladie d'idealité'. The English version is noted as being published by W.W. Norton in 1985, with a translation by Paul Barrows.
- Step 4: Based on the information from reference [3], it is clear that the English publication of 'The ego ideal: a psychoanalytic essay on the malady of the ideal' occurred in 1985.

A: 1985

Figure 5: Examples of strategization CoT inference process demonstrating the dynamic generation of CoT steps. Samples taken from the Wiki dataset. Reference documents are not shown here.

C Seed Data Generation

Here, we consider two main content sources:

- **Wiki Pages:** We sampled about 10K pages from the English Wikipedia data set². First, filtered very short pages of less than 500 words or 10 lines as well as any pages longer than 7000 words or more than 1000 lines. Then, we split each page to non-overlapping text chunks, randomized and shuffled to eventually have 2 to 15 references of size between

²<https://dumps.wikimedia.org>

250 to 1000 words. This process resulted in final sample size of about 5K RAG QA task samples.

- **Web Search:** To further diversify and target more challenging scenarios, we curated a collection of pages retrieved using web search queries in response to a set of internal knowledge related questions that are primarily focused on time-sensitive queries (e.g. news events, sports outcomes). For each query we retrieved up to 10 pages. Consequently, each page was converted from html to a plain text reference document that is truncated to 3000 words. For this data source, we also use and include relevant time and location (city name) of the request as additional contextual information.

To generate synthetic QA pairs for references extracted from the Wiki Pages, we randomly select a reference and leverage an instruction prompt focused on producing QA pairs that are grounded on the reference content. We found that emphasizing on question difficulty (e.g. college graduate level and setting a difficulty grade), listing a few common category of problematic questions to avoid (e.g., questions about user time/location), and providing exemplars significantly helps with the final data quality. Additionally, we instruct the model to reject designing QA when reference content is not readable, which is sometimes the case for the Web Search references due to html parse failures or web page retrieval/access problems. For the complete prompt, please refer to Appendix G.1. Regarding the LLM, we used llama-3.3-70b-instruct (Grattafiori et al., 2024)³ as we found it reasonably capable and cost-efficient for this use case.

D Breakdown of Results

Table 5 presents the breakdown results for the baseline and proposed method. We follow a similar metric design as suggested by Yang et al. (2024) to classify each response to one of accurate, hallucinated, or missing. Accurate responses fully and accurately answer the question. Hallucinated answers contain inaccurate and misleading facts. Missing is any answer that refuses to answer or fails to completely address the asked question. Finally,

to summarize the overall factuality, we use *factuality score*, defined as hallucination rate subtracted from accuracy. Note that the “Unverified” portion is applicable to benchmarks without provided ground-truth labels and cases where our factuality analysis tool was not able to verify correctness of all claims in the answer.

E CRAG Resilience Experiment

We conducted an experiment using the CRAG benchmark by limiting the number of reference documents and measuring the impact on the performance. From Figure 6, the proposed method consistently outperforms the baseline. For the accuracy and hallucination rates, the margin of improvement increases with the use of more references, showing the effectiveness of the proposed method to reject noises and leverage additional grounding content.

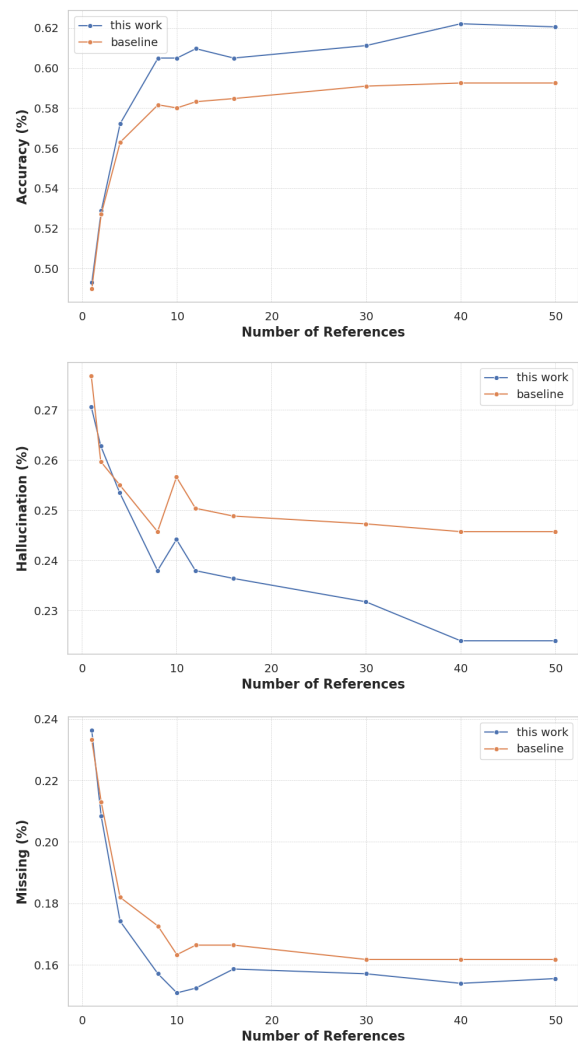


Figure 6: Impact of changing the number of reference documents on accuracy, hallucination, and missing rate for the CRAG (Yang et al., 2024) benchmark.

³<https://www.llama.com/models/llama-3/>

		Correct	Hallucination	Refusal	Unverified	Factuality Score
Baseline	CRAG	59.1%	24.9%	16.0%	0.0%	34.2%
	CovidQA	85.0%	5.0%	8.0%	2.0%	80.0%
	DelucionQA	93.0%	4.0%	3.0%	0.0%	89.0%
	Emanual	93.0%	1.0%	2.0%	4.0%	92.0%
	ExpertQA	86.0%	3.0%	7.0%	4.0%	83.0%
	FinQA	89.0%	6.0%	2.0%	3.0%	83.0%
	HAGRID	93.0%	4.0%	1.0%	2.0%	89.0%
	HotpotQA	96.0%	3.0%	0.0%	1.0%	93.0%
	MS Macro	90.0%	8.0%	1.0%	1.0%	82.0%
	PubMedQA	85.0%	5.0%	7.0%	3.0%	80.0%
	TAT-QA	87.0%	10.0%	2.0%	1.0%	77.0%
	TechQA	61.0%	3.0%	21.0%	15.0%	58.0%
This Work	CRAG	62.1%	22.9%	15.1%	0.0%	39.2%
	CovidQA	96.0%	1.0%	2.0%	1.0%	95.0%
	DelucionQA	98.0%	1.0%	1.0%	0.0%	97.0%
	Emanual	98.0%	0.0%	2.0%	0.0%	98.0%
	ExpertQA	84.0%	1.0%	13.0%	2.0%	83.0%
	FinQA	81.0%	10.0%	3.0%	6.0%	71.0%
	HAGRID	91.0%	1.0%	7.0%	1.0%	90.0%
	HotpotQA	92.0%	3.0%	0.0%	5.0%	89.0%
	MS Macro	89.0%	7.0%	1.0%	3.0%	82.0%
	PubMedQA	92.0%	2.0%	6.0%	0.0%	90.0%
	TAT-QA	94.0%	4.0%	2.0%	0.0%	90.0%
	TechQA	86.0%	4.0%	7.0%	3.0%	82.0%

Table 5: Breakdown of factuality benchmark results.

F Evaluation on Closed-Book QA

F.1 Experiment Settings

To further demonstrate the effectiveness of our proposed approach, we conducted an additional evaluation on closed-book factuality QA benchmarks. These benchmarks contain short answers for given questions and do not have access to retrieved documents, allowing us to isolate and showcase the enhancement in reasoning capability afforded by our method. We describe their details in the following section.

F.1.1 Benchmarks

- **CRAG** (Yang et al., 2024) is designed to evaluate a model’s RAG capability. For the comparisons made in this section, we have used the same web summarization split from main text but without any reference documents.
- **SimpleQA** (Wei et al., 2024) comprises 4,326 short QA pairs that cover a wide range of topics, including history, technology, science, and entertainment.

- **Head-to-Tail** (Sun et al., 2022) contains simple QA pairs generated from a general knowledge graph, DBPedia, and an internet movie database, IMDb. For each dataset, we sampled 200 entities respectively from head, torso, and tail entities which are divided by entity popularity following (Sun et al., 2023), resulting in 1,200 QA pairs.

F.1.2 Metrics and Implementation

For short-form factuality QA, we use accuracy, missing, and hallucination following (Yang et al., 2024) followed by factuality score described in the main text.

Identical to the evaluation in the main text, we compared llama-3.1-70b-instruct (Grattafiori et al., 2024) and its fine-tuned version using our proposed approach. We note that the recipe for fine-tuning and inference is also the same with the setting used in the main text.

F.2 Results

As shown in Table 6, our proposed method demonstrates significant factuality gains across all datasets, with the exception of DBpedia for which we see a marginal regression. Notably, we observe a substantial reduction in hallucination rates across all four datasets. These findings suggest that the proposed approach can effectively enhance a model’s factuality capabilities and generalization to even in closed-book settings, despite being trained on samples that utilize search documents.

Model	CRAG ^a (closed-book)			
	Acc (Rec)	Miss.	Hall.	Fac.
Baseline	58.7	15.6	25.7	33.0
This Work	55.3	22.6	22.1	33.2
Model	SimpleQA (closed-book)			
	Acc (Rec)	Miss.	Hall.	Fac.
Baseline	20.0	44.1	35.9	-15.8
This Work	14.0	64.0	22.0	-8.0
Model	IMDB (closed-book)			
	Acc (Rec)	Miss.	Hall.	Fac.
Baseline	44.8	34.2	21.0	23.8
This Work	40.7	47.7	11.7	29.0
Model	DBpedia (closed-book)			
	Acc (Rec)	Miss.	Hall.	Fac.
Baseline	52.0	22.0	26.0	26.0
This Work	41.0	43.0	16.0	25.0

^a CRAG benchmark here is different from the main text as reference documents are omitted for the close-book task.

Table 6: Closed-book factuality benchmark results.

G Prompts

G.1 Synthetic QA Generation

G.2 Vanilla RAG QA CoT Prompt

G.3 Thought Evaluation Prompt

G.4 Answer Evaluation Prompt

You are a helpful assistant. Always follow the provided instructions and generate outputs in valid json format without any extra information. Generate a question and answer pair based on the Provided Content below.

Requirements:

- You must ground your question and answer to the Provided Content
- The question should be selected to resemble what a curious college graduate would ask an intelligent conversational system. From the difficulty level of 1 to 10, aim for an 8.
- The answer should be fully and directly grounded on the Provided Content. Never use any information other than what is available in the Provided Content to generate the question and answer.
- Never generate a question that is asking for the current time, date, or location.
- The question should not be too general or vague. When applicable, include specific entities, names, times, locations, events, and keywords in the question.
- The question and answer must be grammatically correct and be conversationally natural.
- A good question should be meaningful and provides enough context. For example, questions like "when was this updated?" do not provide enough context and are not meaningful, but "when was the start date for world war two?" is clear and meaningful.
- Always return in json format with two keys: "question" and "answer". If the Provided Content is not readable, you may set the value corresponding to the question and answer keys to "N/A".

Examples:

Here are some examples of questions types to consider:

1. How old is Obama?
2. What was the name of the first president of the United States?
3. How is the weather in Seattle this weekend?
4. What is the population of China?
5. Is there a movie theater nearby?
6. What time is high tide tonight in Santa Cruz, CA?
7. Who is leading in the election between Trump and Kamala Harris?
8. What is the dodgers current score?
9. When does daylight saving time start?
10. Any updates on Morgan Freeman's health?
11. What are the main ingredients in a margarita?
12. Why is chocolate bad for dogs?

Provided Content:

<reference content>

Figure 7: Prompt used for synthetic QA generation given reference content.

For this task you are asked to answer a question (Question). Please provide factually accurate, direct, and clear responses. Do not ask any clarification questions or ask for additional information. To make sure the right facts are being considered, you should always ground your responses to the provided references below (References). If the references are not relevant to the question or do not provide the right information, you can respond with an apology rather than fabricating facts.

References:

<Reference Documents>

Question:

<Question>

Now let's think step by step:

Step 1: summarize what the question is asking, and what are the specific key pieces of information that are needed to answer the question.

Step 2: analyze the provided references, one by one. Identify the relevant information that can be used to answer the question. Pay close attention to the entities, names, times, locations, events, and keywords that are relevant to the question. If the question is related to (or implies) the current user location and/or time, you must consider that in finding the relevant information and answering the question.

Step 3: based on Step 1 and Step 2, you must provide an answer that directly addresses the question and is fully grounded on the provided references.

Now, provide your reasoning steps in a few sentences followed by the final answer to the question as a new line starting with "## Answer:".

Figure 8: Prompt used for answer and thought generation given the question and reference content.

For this task you are given a question (Question), a set of references (References) as well as a reasoning (Reasoning) and an answer (Answer) to the question.

Your task is to evaluate the reasoning process. Use the following guidelines to evaluate the reasoning process and assign a score between 1 to 4 to the provided reasoning:

- Score 1: The reasoning is not correct. It is not related to the question and the references, or it is a simple repetition of them with no thought process.
- Score 2: The reasoning is related to the question and the references, but it is not really helpful in answering the question. It is not clear how the reasoning leads to the answer.
- Score 3: The reasoning is related to the question and the references, and it is helpful in answering the question. It may partially help answering the question but there are gaps in the reasoning and parts of the question are not addressed.
- Score 4: The reasoning is related to the question and the references, and it is helpful in answering the question. It provides a clear and complete thought process that leads to the answer.

References:

<Reference Documents>

Question:

<Question>

Reasoning:

<Thought>

Answer:

<Answer>

First, explain your assessment of the Reasoning based on the criteria above in a few sentences followed by the final score in a new line starting with "## Score:" followed by the score value.

Figure 9: Prompt used for evaluation of the thought quality.

For this task you are given a question (Question), a reference answer (Reference Answer), and a candidate answer (Candidate Answer).

Your task is to evaluate if the Candidate Answer is fully answers the question while being consistent to the information presented in the Reference Answer.

Use the following guidelines to evaluate the reasoning process and assign a score between 1 to 4 to the candidate answer:

- Score 1: The candidate answer does not provide any information to answer the question. It is a simple repetition of the question, a refusal to answer, or providing irrelevant information.
- Score 2: The the candidate answer is not consistent with the reference answer and there are major points of contradiction.
- Score 3: The the candidate answer partially answers the question and is consistent with the reference answer. It may have some minor points of contradiction but it is consistent for the most part.
- Score 4: The the candidate answer fully answers the question and is consistent with the reference answer.

Note: the reference answer may provide additional information that is not needed to answer the question. In such cases, the candidate answer does not have to contain such extra information to be considered consistent or complete, it simply has to provide a clear and complete answer to the original question while not contradicting the reference answer.

Examples:

Question: How tall is the Eiffel Tower?

Reference Answer: The Eiffel Tower in located in Paris and is about 984 feet tall.

Candidate Answer: It is 984 feet tall.

Reasoning: The candidate answer fully answers the question and is consistent with the reference answer. Extra information about the location of the Eiffel Tower is not needed to answer the question.

Score: 4

Question: What is the name of largest waterfall in the world and where is it located?

Reference Answer: The largest waterfall in the world is Angel Falls in Venezuela and is about 3,212 feet tall.

Candidate Answer: The largest waterfall in the world is Angel Falls.

Reasoning: The candidate answer parially answers the question and is consistent with the reference answer. However, it does not provide the location of the waterfall.

Score: 3

Question: Which country had most gold medals in the 2022 Winter Olympics?

Reference Answer: The country with most gold medals in the 2022 Winter Olympics was Canada

Candidate Answer: United States had most gold medals in that Olympics.

Reasoning: The candidate answer provides a different country name which is not consistent with the reference answer.

Score: 2

Question: What is the capital of France?

Reference Answer: The capital of France is Paris.

Candidate Answer: I don't know.

Reasoning: The candidate answer does not provide any information to answer the question. It is refusing to answer the question.

Score: 1

Figure 10: Part 1: Prompt used for evaluation of the answer quality.

Question: Who maintains traffic signals in the unincorporated areas of Collier County?

Reference Answer: Collier County operates and maintains 283 traffic signals in the unincorporated areas of the County, which generally includes areas north of Pine Ridge Road and east of Goodlette-Frank Road.

Candidate Answer: Collier County.

Reasoning: The candidate answer is consistent with the reference answer and provides the specific information needed to answer the question.

Score: 4

Question: What is included in the Footlong Quarter Pound Coney Combo at Sonic Drive-in? Reference Answer: The Provided Content does not specify the details of the Footlong Quarter Pound Coney Combo, but it is mentioned as one of the menu items.

Candidate Answer: Sorry, the provided references do not contain the necessary information to answer the question. Reasoning: The candidate answer does not provide any information to answer the question and instead saying the information is not available.

Score: 1

Question: What is the best-selling album of The Crystal Method in the United States? Reference Answer: The album Vegas, which has sold more than one million copies in the United States, certifying it platinum.

Candidate Answer: Vegas

Reasoning: The candidate answer is directly answering the question and is mentioning the same album name as the reference answer.

Score: 4

Question: What is the main street in Downtown Naples, Florida's Paradise Coast?

Reference Answer: Fifth Avenue South is Downtown's de facto Main Street, a one-mile stretch from 9th Street west to the beach.

Candidate Answer: Fifth Avenue South

Reasoning: The candidate answer is directly answering the question and is mentioning the same album name as the reference answer. While the reference answer provides additional information about the street those are not needed to answer the question.

Score: 4

Question:

<Question>

Reference Answer:

<Answer>

Candidate Answer:

<Candidate Answer>

First, explain your assessment of the Candidate Answer based on the criteria above in a few sentences followed by the final score in a new line starting with "## Score:" followed by the score value.

Figure 11: Part 2: Prompt used for evaluation of the answer quality.

G.5 Critique-Based Thought Revision Prompt

G.6 Strategization RAG QA CoT Prompt

G.7 Distractor Critique Prompt

The distraction critique process involves evaluating each distractor passage based on three key scores on a scale of 1-5:

1. **Relevance Score:** Measures how relevant the distractor passage is to the open-ended question, golden passage, location, and time.
2. **Distraction Score:** Assesses quality of the passage in its ability to provide relevant retrieval noise.
3. **Format Score:** Evaluates the similarity in text length and format between the distractor passage and the original passage.

The critique process provides constructive feedback on each score, highlighting areas where the distractor passage excels or falls short as a distractor. This feedback is then incorporated into the distractor passage generation step to refine the next iteration of distraction generation.

G.8 Distractor generation prompt

<Previous Thought Generation Steps as Context Instruction-Assistant Turns>

Your reasoning and answer can be further improved. Below is a critique of your previous reasoning and answer. You should use this critique to improve your reasoning and answer.

Critique: <Critique of Previous CoT>

Now let's think step by step while considering the provided critique above:

Step 1: summarize what is the question is asking, and what are the specific key pieces of information that are needed to answer the question.

Step 2: analyze the provided references, one by one. Identify the relevant information that can be used to answer the question. Pay close attention to the entities, names, times, locations, events, and keywords that are relevant to the question. If the question is related to (or implies) the current user location and/or time, you must consider that in finding the relevant information and answering the question.

Step 3: based on Step 1 and Step 2, you must provide an answer that directly address the question and is fully grounded on the provided references.

Now, provide your reasoning steps in a few sentences followed by the final answer to the question as a new line starting with "## Answer:"

Figure 12: Prompt used for thought regeneration given previous generated thought and evaluation critique.

For this task you are asked to answer a question (Question). Please provide factually accurate, direct, and clear responses. Do not ask any clarification questions or ask for additional information. To make sure the right facts are being considered, you should always ground your responses to the provided references below (References). If the references are not relevant to the question or do not provide the right information, you can respond with an apology rather than fabricating facts.

References:

<Reference Documents>

Question:

<Question>

Before answering the question, take a step back and carefully think about the best strategy to answer the question. Produce an outline for the reasoning steps that you can take to find the best answer. Then, use the outline to think step by step.

Use the following template to strategize your reasoning steps, reason step by step, and provide the final answer:

Strategy:

- Step 1: *** instructions for step 1 ***
- ...
- Step N: *** instructions for step N ***

Reasoning:

- Step 1: *** reasoning corresponding to step 1 in the strategy ***
- ...
- Step N: *** reasoning corresponding to step N in the strategy ***

Answer: <final answer to the question>

Figure 13: Prompt used for CoT strategization to solve the RAG QA task.

You are an intelligent assistant with expertise in linguistics. Always follow the provided instructions and generate outputs in valid json format without any extra information.

User

Given a question, answer, passage, location, user-time, distraction-passage and distraction passage's answer:

1. The goal of the distraction is to provide conflating information from the original passage with respect to the user's question and details, such that it would prompt a human to take a closer look at fine details of both of the passages before coming to an answer. If they glance superficially at a distraction passage, that should feel like a reasonable answer, but when juxtaposed against the original passage, only the original passage should lead to the 'answer'.

2. The goal of distraction-passage is such that when a human is provided distraction-passage and question, they can come to an answer different from the original answer which is retrieved from the passage. Open-question, distraction-answer, distraction-passage should be a slight modification of the original passage question and answer.

3. Open ended question and Distraction-answer is provided along with distraction-passage. Distraction-answer is the answer a human came up with, when provided with just the distraction-passage and the open ended question rather than the original passage and original question. Use that to guide your scoring.

score on the scale of 1 to 5 on Scores:

1.relevance-score:

Measures:

a.how relevant the distraction-passage is to the given question, answer, passage, location, user-time. The distraction passage is required to be relevant to the user question, such that when one or two details are omitted from the user question, the distraction passage answers the question sufficiently.

2.distraction: Measures:

a. how much of a distraction the passage is to a user who asked the question when provided with both passage and distraction-passage.

b. It ensures that when looking at distraction-passage alone, it would lead to a fully different answer.

3.format:

Measures:

a. how similar in text length and format, the distraction passage is, with respect to the original passage.

b. We require the distraction-passage to be of the same length as the original passage in terms of the number of words, or else the humans can easily distinguish based on the length differences.

c. Ensure format in terms of new lines, spaces etc are similar for original and distraction passage. If there are no new lines, humans can easily distinguish based on format. Penalize omission of new lines if it exists in original but not in distraction.

d. Compute the number of words in original vs distraction-passage and penalize if the difference is noticeable. Penalize if fewer words are present in distraction passage.

d. Penalize format difference, if original has new lines, the distraction should have the same.

e. Ensure if the original has tables, multiple tabs or new lines, the distraction has the same.

- Your thought-process field should contain constructive criticism which helps drive improvement. The distraction generation process will display your feedback, so that they can utilize it to provide a better distraction-passage. - Your thought process should clearly mention score and reasoning for each type of score and critique for each.

- Think it through step by step and provide a detailed explanation for each score in detail and what all measures are satisfied or missing. Eg: If new lines are missing, it falls under category c,d,e of format.

- Output in Json with following fields 'relevance-score', 'distraction-score', 'format-score', 'thought-process'.

question : {question}

answer : {answer}

user-time: {user-time}

location: {location}

passage :{passage}

open ended question: {open-ended-question}

distraction passage: {distraction}

distraction passage's answer: {distraction-answer}

Assistant:

Figure 14: Critique prompt used for synthetic distractor generation

You are an intelligent assistant with expertise in linguistics. Always follow the provided instructions and generate outputs in valid json format without any extra information.

User :

Think it through step by step:

1. Given a question, answer, passage, location and user-time identify relevant-named-entities, date times, locations in the passage based on the question and answer, such that modifying the relevant named entities will result in a new passage that can cause confusion to the user if they didn't have enough context.
2. You might also be provided with 'prior-distraction-passage', 'prior-distractor-rejecting-reason'. That logs your generation for the same question and answer in the previous turn. Use that information to guide and improve for this round of distraction passage generation.
3. Generate an open ended question "open-ended-question" by modifying provided question such that the answer provided answers the new open ended question.
4. Now using question, answer, user time, location, modify the passage and generate a new passage by modifying in the named entities such that
 - a. The new passage is relevant to the existing passage.
 - b. The new passage is grammatically coherent.
 - c. For the "open-ended-question", both provided and your generated passage are relevant.
 - d. The distraction-passage should have a similar number of characters as the original passage and similar format.
5. Score your confidence (from 1 to 5) that the generated passage will satisfy condition 4.

Follow the requirements.

Requirements:

- You must generate the passage based on the user question, location, answer, user-time .
- The generated distraction should be of similar length to the original passage and with similar special characters such as /n, / t. Do not reduce the total number of words.
- Think it through step by step and provide a detailed explanation in the "thought steps" field. Output in Json with following fields 'open-ended-question', 'thought-steps', 'distracting-named-entities', 'distracting-passage', 'score', 'reason' as json fields in your output.

Think it through step by step:

1. Given a question, answer, passage, location and user-time identify relevant-named-entities, locations and time information in the passage based on the question and answer, such that modifying the relevant named entities, location or date time will result in a new passage that can cause confusion to the user if they didn't have enough context.
2. You might also be provided with 'prior-distraction-passage', 'prior-distractor-rejecting-reason'. This provides information on your prior distraction generation for the same question and answer. Use that information along with the 'prior-distractor-rejecting-reason' to analyse and improve the distraction generation and address the gaps.
3. Generate an open ended question "open-ended-question" by modifying the provided question such that the answer provided answers the new open ended question.
4. Now using question, answer, user time, location, modify the passage and generate a new passage by modifying in the named entities and/or location and/or time such that
 - a. The new passage is relevant to the existing passage.
 - b. The new passage is grammatically coherent.
 - c. For the "open-ended-question", both provided and your generated passage are relevant.
 - d. The distraction-passage should have a similar number of characters as the original passage and similar format.
5. Pick what kind of named entities or location or time change would cause distraction based on user question, opened question , passage and details.
 - a. Eg: For a query about sporting events today, changing date time of the event will be a distraction as today (datetime) is a key component of the question.
 - b. Eg: For a question about actors in a movie, changing the movie name slightly and actors name will be a good distraction.
5. Score your confidence (from 1 to 5) that the generated passage will satisfy condition 4.

Figure 15: Part 1: Prompt used for generating distractor content.

```

## Examples:

Example 1 (Named entity change):

question: How hard was the stunt in Mission Impossible 7? passage: Mission Impossible 7 was a very dangerous movie. The most dangerous scene was the one where Tom Cruise jumped off a building and held onto a rope. The director was very careful with the stunt and made sure that Tom Cruise was safe."

relevant-named-entities, location, time : Mission Impossible, 7, building open-ended-question: What was the most dangerous scene in Mission Impossible?

distracting-named-entities: Mission Impossible, water

distracting-passage: Mission Impossible was a very dangerous movie. The most dangerous scene was the one where Tom Cruise jumped into water and held his breath. The director was very careful with the stunt and made sure that Tom Cruise was safe."

reason: open-ended-question is slight modification of the user question by omitting context that the movie was the 7th Mission Impossible movie. Distracting passage modified both the movie number and the stunt performed. While this would be a correct answer for any mission impossible movie. It's not the correct answer for mission impossible 7, making this the right distractor passage, while being of similar characters and length as the original passage.


Example 2 (date time change ):

question: Who are the Washington Commanders playing today?

user-time: 4pm, Thursday, January 9, 2025.

passage: Title: Washington Commanders vs. San Francisco 49ers - FOX Sports Snippet: Game-time: January 9, 2025, 5pm PST, Venue: Washington D.C, odds: Washington Commanders to win 63relevant-named-entities,location, time : January 9, 2025 5 pm PST, Washington Commanders, San Francisco 49ers

open-ended-question: Who are Washington Commanders playing?

distracting-named-entities, location, time : January 10, 2025, Washington Commanders, New Orleans Saints

distracting-passage: Title: Washington Commanders vs. New Orleans Saints - FOX Sports Snippet: Game-time: January 10, 2025, 5pm PST, Venue: Washington D.C, odds: Washington Commanders to win 52% probability, Game-a: 0-0 , Game-status: Not-started

reason: open-ended-question made it ambiguous by removing information that the user is asking about today's game (January 9th). The distracting-passage talks about the Washington Commanders vs New Orleans Saints game scheduled on the 10th instead, thus it's a slight modification of the original passage, while answering open ended questions correctly and being a distraction, while being of similar characters and length as the original passage.


## question : {question}
## answer : {answer}
## user-time: {time}
## location: {location}
## passage :{passage}
## prior-distraction-passage: {prior-distraction-passage}
## prior-distractor-rejecting-reason: {prior-distractor-rejecting-reason}

### Assistant:

```

Figure 16: Part 2: Prompt used for generating distractor content.