RASC: Retrieval-Augmentation with Synthetic Corrective Reasoning

Laurentiu Meirosu
Syntelesis AI Lab
laurentiu@syntelesis.ai

Abstract

Retrieval-Augmented Generation (RAG) extends large language models to specialised corpora, but it remains vulnerable to hallucination answers that seamlessly confabulate true evidence with unsupported claims triggered by irrelevant context [Huang et al., 2024b]. We introduce RASC (Retrieval-Augmentation with Synthetic Corrective reasoning), a supervised fine-tuning method that converts hallucination into training signal. For every question–answer pair in domain-specific datasets like PubMedQA and BioASQ, RASC synthesises (i) distractor passages, (ii) generates hallucination answers based on formal fallacy classes, and (iii) constructs chain-of-thought rationales that diagnose and correct the error using only verifiable evidence. Training on these "wrong then right" demonstrations conditions the model to locate golden passages, ignore distractors, and articulate faithful reasoning. Evaluated with RAGAS (faithfulness, answer relevance) and ROUGE, the RASC fine-tuned model, with only 1,000 generated samples, more than doubles grounding quality over a standard RAG baseline and surpasses conventional supervised fine-tuning across all datasets. These results present corrective chain-of-thought supervision as a lightweight yet effective method to significantly reduce domain-specific hallucinations in RAG systems.

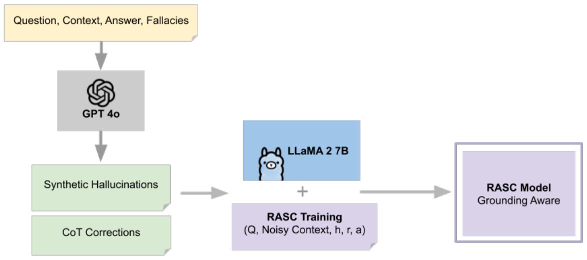


Figure 1: **Overview of our Supervised fine-tuning method with Synthetic Hallucinations and Corrections**: Using a higher model we generate the hallucinations and corrections for the relevant Question, Context and Answer pairs. Then we use instruction fine-tuning (IFT) on the lower model with reasoning steps on how to write the correction of the fallacies itself before answering the question. The result is a model that has learned to account for corrections via few-shot demonstrations and Chain-of-Thought (CoT) reasoning.

1 Introduction

Large-scale language models (LLMs) have demonstrated impressive general knowledge and emergent reasoning abilities, enabling strong performance on tasks ranging from open-domain question answering to code generation. Their general coverage, acquired through pre-training on large corpora, has made them indispensable tools across multiple industries. Yet, when these models are deployed in domain-specific settings such as biomedicine, finance or law their seemingly universal competence often degrades sharply. The deterioration is most visible when the model must answer questions whose solutions lie outside its pre-training distribution and inside a narrow, specialised set of documents.

Retrieval-Augmented Generation (RAG) has emerged as the de-facto strategy for bridging this gap [Lewis et al., 2020, Gao et al., 2024]. By retrieving relevant passages from an external corpus and concatenating them to the prompt as context, RAG enables an LLM to ground its generations in up-to-date, domain-specific evidence without retraining the entire model. In principle, this architecture should yield faithful and context-aware responses. However, in practice, RAG pipelines often suffer from a severe misalignment between the retrieved context and the final answer to the model, resulting in factual errors or hallucinations [Li et al., 2024]. These hallucinations manifest when the model (i) overlooks pertinent evidence, (ii) fabricates statements that are unsupported by any retrieved passage, or (iii) conflates distractor documents with genuinely relevant ones.

Hallucinations linger in RAG systems for three reasons: (1) Retriever noise dense retrievers [Karpukhin et al., 2020] still surface topically similar but non-answer chunks due to coarse boundaries and embedding collisions [Morris et al., 2023]; (2) Ungrounded training—pretraining treats all tokens alike, so models learn facts without learning to link claims to specific passages; without grounding-aware fine-tuning, they cannot reliably ignore distractors; (3) Weak incentives rarely rewards correct citation or penalises misuse of context, letting the model favour high-probability continuations over evidence fidelity. Thus, noisy retrieval, unsupported knowledge, and absent contrastive rewards combine to produce confident yet unsupported answers.

In this work, we hypothesise that Instruction Fine-Tuning (IFT) [Ouyang et al., 2022] can allow a large language model to recognise and avoid hallucinations by making it learn how to create and correct fallacies during training. To evaluate this claim, we introduce the RASC method (Retrieval-Augmented with Synthetic Corrections). This method teaches the model to generate an incorrect context-relevant response and then goes through an explicit reasoning to identify the error and produce the correct response. Unlike approaches that suppress generative freedom, this hypothesis asserts that exposing the model to its hallucinations and teaching it to correct them will train its representations to downweight distractors and better align the response to the given context.

2 Prior Literature

Retrieval-augmented generation (RAG) promises factual grounding yet still fails when large language models (LLMs) treat stray context as evidence. Shi et al. [2023] show this failure mode systematically in "Large Language Models Can Be Easily Distracted by Irrelevant Context". When a single, semantically plausible but answer-irrelevant sentence is inserted into GSM8K problems, accuracy collapses by up to 30 points and fewer than 18% of base questions are answered consistently across distractor variations. Their micro/macro analysis pinpoints two root causes we address in RASC: (i) retrievers surface distractors; (ii) the generator lacks an internal mechanism for separating "gold" from noise.

A second line of work tries to detect when the model simply does not know. "Don't Hallucinate, Abstain" [Feng et al., 2024] ensembles multiple LLMs, measuring answer disagreement as a proxy for knowledge gaps. While collaboration reduces blatant fabrications, it does not teach any single model why a candidate span is ungrounded. Related work on calibration [Kadavath et al., 2022] explores whether models can accurately assess their own uncertainty, but this does not directly address context-grounding failures.

In very low-resource settings, "Embedding Hallucination for Few-shot Language Fine-tuning" [Jian et al., 2022] shows that injecting short, noisy continuations during fine-tuning encourages the model to allocate separate sub-spaces to "real" and "fake" examples, improving robustness on GLUE [Wang et al., 2019]. RASC generalises this idea to the RAG setting: we synthesise context-conditioned hallucinations that mirror the ten formal-fallacy profiles most often observed in RAG output and couple them with guided correction. Experiments confirm that conditioning on the same retrieval pipeline is crucial. The gains reported for embedding hallucination alone vanish when distractors are semantically on-topic.

Recent evidence suggests that training on noisy retrieval results is as critical as evaluating on them. Zhang et al. [2024] introduce "RAFT: Retrieval-Augmented Fine-Tuning", showing that explicitly mixing golden and distractor documents during supervision sharpens a model's ability to disregard irrelevant context at test time. Their ablations reveal two findings highly pertinent to RASC: (i) a 4:1 ratio of distractors to gold passages at training maximises downstream RAG robustness, and (ii) omitting the gold passage in a controlled fraction of examples forces the model to memorise core facts when retrieval misses, further reducing hallucinations. RAFT therefore complements the diagnostic study of Shi et al. [2023]: while Shi quantifies how badly models fail in the presence of a single misleading sentence, RAFT demonstrates that exposing the model to such noise—paired with chain-of-thought rationales that highlight the correct evidence leads to sizeable gains on PubMedQA, HotPotQA [Yang et al., 2018], and Gorilla APIBench. RASC adopts this principle by fusing distractor rich contexts with labelled fallacy/correction pairs, extending RAFT's document-level noise curriculum to a finer-grained taxonomy of ten formal fallacies.

Finally, "Can Rationalization Improve Robustness?" [Chen et al., 2022] tests whether training models to output free-text rationales shields them from adversarial noise. Gold rationales improve human alignment but paradoxically decrease robustness: models latch onto the rationale template itself instead of grounding in supporting facts. This finding aligns with concerns about unfaithful chain-of-thought reasoning [Turpin et al., 2023, Lanham et al., 2023], where models may generate plausible-sounding but disconnected explanations. RASC departs in two ways: (i) rationales are generated after a hallucination is induced, ensuring they explicitly negate a known error, and (ii) rationales are evaluated with RAGAS–Faithfulness so only evidence-grounded traces survive curriculum filtering. Recent work on improving CoT faithfulness [Paul et al., 2024, Lyu et al., 2023] provides additional motivation for our correction-based approach.

3 Method

We now formalise the RASC training objective, describing how corrective examples provide gradient signal that teaches the model to distinguish gold evidence from distractors. We begin by establishing notation, then contrast the standard supervised fine-tuning loss with the RASC combined loss, and finally explain why conditioning corrections on hallucinations yields more robust learning.

3.1 Notation

Throughout this paper, we denote a question as q, the gold passage containing the correct evidence as c_g, and the set of $k = 3$ distractor passages as $\mathcal{D} = \{d_1, d_2, d_3\}$. The ground-truth answer is written a^*.

A key component of RASC is the noisy context $\tilde{c} = \text{shuffle}(c_g, d_1, d_2, d_3)$, which interleaves the gold passage with distractors in random order. This simulates the realistic retrieval setting where relevant and irrelevant chunks appear together without clear demarcation. We also introduce $\hat{a}_h$ to denote a synthetically generated hallucinated answer, and r to denote the corrective chain-of-thought rationale that identifies and fixes the hallucination. The use of hard negative examples during training draws on principles from contrastive learning [Robinson et al., 2021, Chuang et al., 2020].

3.2 Standard SFT Loss

Standard supervised fine-tuning optimises on clean question-context-answer triples, minimising the negative log-likelihood of the correct answer given only the gold passage:

$$\mathcal{L}_{\text{SFT}} = -\log p_\theta(a^* \mid q, c_g) \tag{1}$$

This formulation assumes that all provided context is relevant—an assumption that breaks down in real RAG deployments where retrievers inevitably surface some off-topic material. Because the model never encounters distractors during training, it cannot learn to discriminate gold from noise, leaving it vulnerable to hallucination when irrelevant passages appear at inference time.

3.3 RASC Loss

RASC addresses this limitation by fine-tuning on a combined objective that incorporates both noisy context and explicit correction supervision. The total loss comprises two terms:

$$\mathcal{L}_{\text{RASC}} = \mathcal{L}_{\text{ans}} + \lambda\, \mathcal{L}_{\text{corr}} \tag{2}$$

where λ controls the relative weight of the correction term. The answer loss trains the model to produce correct responses despite the presence of distractors:

$$\mathcal{L}_{\text{ans}} = -\log p_\theta(a^* \mid q, \tilde{c}) \tag{3}$$

The correction loss trains the model to generate a rationale that diagnoses errors in the hallucinated answer:

$$\mathcal{L}_{\text{corr}} = -\log p_\theta(r \mid q, \tilde{c}, \hat{a}_h) \tag{4}$$

Two key differences separate RASC from standard SFT. First, the use of noisy context $\tilde{c}$ forces the model to extract correct answers even when distractors are present, rather than assuming all context is trustworthy. Second, the correction loss conditions on the hallucinated answer $\hat{a}_h$ before generating the rationale r, teaching the model to diagnose specific errors rather than merely producing generic explanations.

3.4 Why Corrections Help

The correction loss provides gradient signal that standard answer-only training cannot deliver. When the model learns to predict rationale tokens such as "*The second paragraph is irrelevant because it discusses a different study...*", gradients flow back through the attention mechanism [Vaswani et al., 2017] and reduce the weights assigned to distractor positions.

Intuitively, to predict *why* something is wrong, the model must first learn *what* to ignore. This creates an implicit curriculum [Bengio et al., 2009]: the model cannot minimise $\mathcal{L}_{\text{corr}}$ without developing internal representations that distinguish gold evidence from noise. Standard SFT, by contrast, never forces this discrimination because the training context contains only relevant material. Recent surveys on self-correction in LLMs [Kamoi et al., 2024, Huang et al., 2024a] suggest that while intrinsic self-correction without external feedback often fails, training-time correction with explicit supervision—as in RASC—can be effective.

3.5 Training Data Generation

For each gold triple (q, c_g, a^*) in our datasets, we use GPT-4o to generate the additional components needed for RASC training. First, we generate three distractor passages $\mathcal{D}$ that are topically related to the question but do not contain the answer. These simulate the retriever noise that occurs in real deployments. Second, we generate a hallucinated answer $\hat{a}_h$ that exemplifies one of ten catalogued fallacy types, such as fabricating unsupported claims or misattributing evidence from distractors. Third, we generate a corrective rationale r that identifies the specific error in $\hat{a}_h$ and reasons step-by-step toward the correct answer a^*. This approach to synthetic data generation for fine-tuning builds on recent work demonstrating the effectiveness of LLM-generated training data [Wang et al., 2023].

The resulting training instances thus contain the full "wrong then right" demonstration: the model sees the question, the noisy context mixing gold and distractors, an example of what a hallucinated response looks like, and a detailed correction that models faithful reasoning.

4 Data

To evaluate the faithfulness of our fine-tuning method, we curate two domain-specific and one general-domain publicly available datasets.

4.1 Domain-Specific Benchmarks

PubMedQA [Jin et al., 2019] provides expert-annotated questions whose answers must be inferred from PubMed abstracts. Because each abstract contains highly technical findings, hallucinations surface whenever a model reads tangential sentences as evidence. PubMedQA therefore serves as our primary test bed for fact-grounding under narrow scientific discourse.

BioASQ [Krithara et al., 2023] extends this setting with thousands of "exact" and "summary" questions paired with gold PubMed snippets. The larger scale enables reliable measurement of false attribution rates and the impact of distractor passages injected by the retriever. Recent BioASQ challenges [Nentidis et al., 2024] have seen increasing use of RAG-based approaches, underscoring the relevance of our evaluation setting.

4.2 General-Domain Readability and Context-Use

SQuAD v1 [Rajpurkar et al., 2016] contains crowd-sourced QA pairs over Wikipedia paragraphs. Unlike the biomedical sets, SQuAD passages are written for a lay audience, allowing us to measure (i) whether corrective Chain-of-Thought harms fluency, and (ii) how precisely the model anchors answers to the immediate span that contains the ground truth.

5 Model

5.1 Baselines

We evaluate two reference systems on the identical LLaMA-2 7B backbone, each fed with a concatenated string $\langle$question, context$\rangle$. **B-Zero** is the untouched Meta checkpoint, showing how a general-purpose model performs when given a context. **B-SFT** represents the mainstream recipe for supervised RAG: the model is fine-tuned on (q, c_g, a^*) with $\mathcal{L}_{\text{SFT}}$.

5.2 Proposed Model: RASC

RASC adds synthetic hallucinations with correction-rich chain-of-thought reasoning (see Figure 2). For every gold instance in PubMedQA,

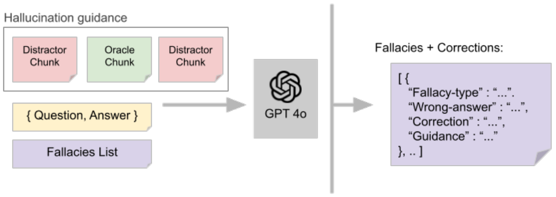
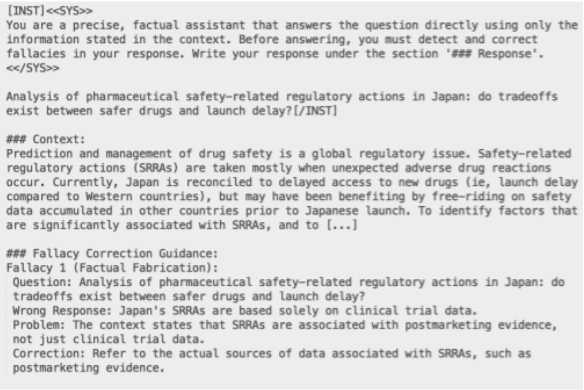


Figure 2: **Overview of our Synthetic Data Generation pipeline:** First, the context is composed from a combination of the context and distractor chunks. Second, a list of relevant fallacies is composed for the model to hallucinate and correct against. Third, the instruction is given on how to construct the wrong answer, correction, and the guidance. The result is a list of hallucinations and corrections demonstrations tailored for each question, context, and answer pair.



```
[INST]<<SYS>>
You are a precise, factual assistant that answers the question directly using only the
information stated in the context. Before answering, you must detect and correct
fallacies in your response. Write your response under the section '### Response'.
<</SYS>>

Analysis of pharmaceutical safety-related regulatory actions in Japan: do tradeoffs
exist between safer drugs and launch delay?[/INST]

### Context:
Prediction and management of drug safety is a global regulatory issue. Safety-related
regulatory actions (SRRAs) are taken mostly when unexpected adverse drug reactions
occur. Currently, Japan is reconciled to delayed access to new drugs (ie, launch delay
compared to Western countries), but may have been benefiting by free-riding on safety
data accumulated in other countries prior to Japanese launch. To identify factors that
are significantly associated with SRRAs, and to [...]

### Fallacy Correction Guidance:
Fallacy 1 (Factual Fabrication):
 Question: Analysis of pharmaceutical safety-related regulatory actions in Japan: do
 tradeoffs exist between safer drugs and launch delay?
 Wrong Response: Japan's SRRAs are based solely on clinical trial data.
 Problem: The context states that SRRAs are associated with postmarketing evidence,
 not just clinical trial data.
 Correction: Refer to the actual sources of data associated with SRRAs, such as
 postmarketing evidence.
```

Figure 3: **RASC prompt that guides the model on how to correct fallacies via Chain-of-Thought (CoT).** The detailed Fallacies Correction Guide includes the original Question, example of the Wrong Response, the explanation of the Problem and the Correction needed based on the given Question and Context.

BioASQ, and SQuAD, we sample three distractor passages and prompt GPT-4o, conditioned on ten catalogued fallacy tags, to produce:

- hallucinated answers that exemplify the fallacy, and
- a step-by-step corrective rationale that points out each error and provides guidance for the correction

The resulting training instances, each containing the question, noisy context, hallucinated answer, and correction, are combined with the gold answer. Fine-tuning minimises the combined RASC objective: cross-entropy on the correct answer tokens and teacher-forced loss over the corrective chain-of-thought.

5.3 Training and Inference Protocol

Fine-tuning is performed on a subset of the datasets (1,000 data points) with QLoRA [Dettmers et al., 2023]: 4-bit quantisation for the frozen base weights and rank-16 LoRA [Hu et al., 2022] adapters for trainable updates. Optimisation uses AdamW [Loshchilov and Hutter, 2019] with a learning rate of 2×10^{-5} on mini-batches of 128 sequence pairs. We set $\lambda = 0.5$ to balance the answer and correction loss terms. We use a standard 80/20 train/test split for evaluation.

6 Experimental Design

To isolate how corrective Chain-of-Thought (CoT) fine-tuning affects Retrieval-Augmented Generation, we organise evaluation on the subset of PubMedQA, BioASQ and SQuAD questions for which the raw checkpoint (B-Zero) scored < 0.40 on the RAGAS-Faithfulness metric. This threshold is chosen because a score below 0.40 indicates that fewer than half of the claims in the generated answer are supported by the retrieved context, representing cases where the model demonstrably struggles with grounding and is prone to hallucination.

6.1 Synthetic Corrections Supervised Fine-tuning

We fine-tune LLaMA-2 7B with the RASC synthetic corrections but add no specific prompt engineering at inference. The model sees training instances containing question, noisy context, hallucinated answer, and correction during training, learns to downweight distractors, and is evaluated with the same retrieval bundle as B-Zero. This quantifies the standalone benefit of fallacy-aware CoT supervision.

6.2 Evaluation Protocol

We assess every model on the subset—questions where the raw LLaMA-2 7B scored $<$ 0.40 on RAGAS—using two complementary families of metrics.

RAGAS Faithfulness. For each predicted answer we concatenate it with the retrieved passages and submit the pair to the GPT-4 model supplied by the RAGAS toolkit [Es et al., 2024]. The verifier decomposes the answer into single claims, checks each claim for textual entailment [Bowman et al., 2015] against the evidence, and returns a continuous score in [0, 1]: 1.0 when all claims are supported, 0.0 when none are. Because the computation is claim-level and context-aware, the metric is sensitive to hallucinations, yet agnostic to surface paraphrase, making it the primary measure for grounding quality. This approach aligns with recent work on fine-grained factuality evaluation [Min et al., 2023].

ROUGE-1 / ROUGE-L F1. These reference-based measures quantify the lexical fidelity to the gold answer. ROUGE-1 counts unigram overlap, capturing content words regardless of order, while ROUGE-L computes the longest-common-subsequence ratio, rewarding correct phrases and fluency. Both are reported as F-scores. Although ROUGE cannot detect hallucinations in supporting sentences, it remains the de-facto indicator of answer adequacy and is complementary to RAGAS.

All models are evaluated with the same question, context pairs and identical decoding settings. Any differences in RAGAS or ROUGE therefore stem solely from the fine-tuning data, ensuring a clean attribution of gains.

7 Results

Across the full evaluation suite (PubMedQA, BioASQ, and SQuAD datasets) RASC is the only model that clears the two-fold bar of (i) lexical fidelity (ROUGE-1/-L) and (ii) grounding quality (all RAGAS metrics). Averaged over datasets, RASC lifts the Faithfulness score from 0.233 (B-Zero) and 0.188 (B-SFT) to 0.395 (see Table 1). ROUGE follows the same pattern but on lower absolute values, climbing from 0.196 to 0.219 F-points, indicating that better grounding does not come at the expense of surface accuracy (see Tables 2 and 3).

Model	**PubMedQA**	**BioASQ**	**SQuAD**
B-Zero	0.233	0.228	0.189
B-SFT	0.188	0.217	0.221
RASC	**0.395**	**0.311**	**0.446**

Table 1: **RAGAS Faithfulness Score.** RASC improves faithfulness significantly in all domain-specific datasets. Across PubMedQA, BioASQ and SQuAD, we find that RASC fine-tuning with correction guidance demonstrations improves how factually consistent the response is with the given context. We compare our model with LLaMA-2 7B fine-tuned on the same datasets.

Model	**PubMedQA**	**BioASQ**	**SQuAD**
B-Zero	0.196	0.186	0.106
B-SFT	0.176	**0.238**	0.136
RASC	**0.219**	0.206	**0.224**

Table 2: **ROUGE-L F1 Score.** RASC improves significantly on structural hallucinations (distortions on relationships, context, or logical flow) on the SQuAD dataset, and only slightly on the more complex answer requirements. In our experiment, this shows the limitations of the ROUGE metric in measuring complex hallucinations.

Model	**PubMedQA**	**BioASQ**	**SQuAD**
B-Zero	0.279	0.189	0.139
B-SFT	0.239	0.271	0.341
RASC	**0.356**	**0.286**	**0.500**

Table 3: **ROUGE-1 Precision Score.** RASC improves significantly on additive hallucinations (missing words from the context) where this metric can properly assess this value. For example, RASC scores high on SQuAD dataset where answers are shorter and do not rely on inference, while BioASQ relies on lengthier answers with a high degree of inference from the context.

8 Analysis

The raw B-Zero model fails on two fronts: it copies distractor facts into its narrative and it phrases answers in a style that mismatches the dataset annotations, yielding both low Faithfulness and modest ROUGE. B-SFT standard supervised fine-tuning on gold triples helps stylistic alignment [Zhou et al., 2024] but hardly budges grounding: the model has learned what a plausible answer looks like yet still cannot tell signal from noise in the retrieved bundle.

RASC succeeds because the “wrong then right” training structure forces explicit engagement with errors. The model cannot minimise $\mathcal{L}_{\text{corr}}$ without learning which context to ignore. This finding aligns with recent work showing that self-refinement approaches [Madaan et al., 2023] can be effective when combined with appropriate training signals.

These findings demonstrate that merely exposing a model to retrieved context is insufficient for robust reasoning; explicit adversarial CoT supervision, as embodied in RASC, is necessary to teach the model to produce domain-appropriate answers and to avoid misalignment with the context.

9 Conclusion

RASC offers a principled method for minimising hallucinations in RAG. By synthesising representative fallacies for each question-context pair and pairing every flawed answer with an explicit chain-of-thought correction, the method trains the model to recognise spurious cues, avoid them, and reason its way back to evidence-based answers.

Two design choices are central: (i) exposing the model to a balanced mix of relevant passages and distractors, and (ii) supervising with side-by-side hallucination/correction demonstrations that encode ten recurrent fallacy classes. Experiments on PubMedQA, BioASQ, and SQuAD confirm that these choices translate into substantial gains in RAG settings: increasing faithfulness far beyond base LLaMA-2 + RAG and standard supervised fine-tuning, while simultaneously improving ROUGE.

Taken together, the results underline RASC's promise as a lightweight yet effective defence in reducing domain-specific hallucinations where faithfulness to the context plays a key role.

Known Project Limitations

Domain over-fitting and generic degradation. Because RASC fine-tunes on corpora that pair specialised passages with fallacy-specific corrections, the model becomes highly sensitised to detecting and expunging unsupported content. Preliminary probes on open-domain benchmarks (e.g., TriviaQA [Joshi et al., 2017]) reveal a drop in answer accuracy and an increased tendency to hedge with phrases such as "based on the provided context". This suggests that the corrective bias learned for domain material can suppress otherwise valid inferences when no retrieval context is present.

Stylistic rigidity. RASC's training signal tightly couples factual fidelity with a didactic chain-of-thought style. While this improves transparency, it also tends to standardise phrasing; responses are more formal, and less adaptable to conversational registers than those produced by standard RAG systems.

Expanded context footprint. Each RASC training instance contains the question, gold passage, distractor chunks, a hallucinated answer, and a corrective rationale. At inference time the model consequently expects longer prompts (question + retrieved evidence + self-generated reasoning), consuming substantially more tokens than conventional RAG. This issue is particularly relevant given findings on how models use long contexts [Liu et al., 2024].

Authorship Statement

This paper, titled "RASC: Retrieval-Augmentation with Synthetic Corrective Reasoning" was solely authored by me, Laurentiu Meirosu, and contains original research and writing. I am responsible for all aspects of this work, which include but are not limited to:

- Conceptualisation and design of the study;
- Collection, analysis, and interpretation of data;
- Drafting and revising the manuscript critically for important intellectual content;
- Final approval of the version to be published

I affirm that this work is entirely my own and that no part of this paper has been plagiarised. All sources and aids used have been appropriately cited, and this manuscript has not been submitted elsewhere.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? *arXiv preprint arXiv:2204.11790*.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jie Huang, Shibo Simon Gu, Le Hosseini, Weijia Zhao, Yangfeng Wu, Jiawei Zhou, and Eric Hovy. 2024a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Embedding hallucination for few-shot language fine-tuning. *arXiv preprint arXiv:2205.01307*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2567–2577.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1–22.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Yihan Li et al. 2024. Mitigating hallucination in large language models (llms): An application-oriented survey on rag, reasoning, and agentic systems. *arXiv preprint arXiv:2510.24476*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, et al. 2024. Overview of bioasq 2024: The twelfth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Conference and Labs of the Evaluation Forum (CLEF)*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.