Say Less, Mean More: Leveraging Pragmatics in Retrieval-Augmented Generation

Haris Riaz
Department of Computer Science
University of Arizona
Tucson, AZ, 85721
hriaz@arizona.edu

Ellen Riloff
Department of Computer Science
University of Arizona
Tucson, AZ, 85721
riloff@cs.arizona.edu

Mihai Surdeanu
Department of Computer Science
University of Arizona
Tucson, AZ, 85721
msurdeanu@arizona.edu

Abstract

We propose a simple, unsupervised method that injects pragmatic principles in retrieval-augmented generation (RAG) frameworks such as Dense Passage Retrieval [9]. Our approach first identifies which sentences in a pool of documents retrieved by RAG are most relevant to the question at hand, cover all the topics addressed in the input question and no more, and then highlights these sentences in the documents before they are provided to the LLM. We show that this simple idea brings consistent improvements in experiments on three question answering tasks (ARC-Challenge, PubHealth and PopQA) using three different LLMs. It notably enhances accuracy by up to 19.7% compared to a conventional RAG system on PubHealth.

1 Introduction

Retrieval-augmented generation (RAG) [13] has emerged as a solution to the limited knowledge horizon of large language models (LLMs). RAG combines "pre-trained parametric and non-parametric memory for language generation," [13] with the non-parametric memory typically retrieved from large collections of documents. RAG has been shown to dramatically improve the performance of LLMs on various question-answering and reasoning tasks (see section 2). However, we argue that RAG often overwhelms the LLM with too much information, only some of which may be relevant to the task at hand. This contradicts Grice's four maxims of effective communication [4], which state that the information provided should be "as much as needed, and no more" and that it should be "as clear, as brief" as possible. The four maxims are enumerated as follows: (1) **Maxim of Quantity**: Provide as much information as needed, but no more; (2) **Maxim of Quality**: Be truthful; avoid giving information that is false or unsupported; (3) **Maxim of Relation**: Be relevant, sharing only information pertinent to the discussion; (4) **Maxim of Manner**: Be clear, brief, and orderly; avoid obscurity and ambiguity. While these maxims were originally formulated in the context of human communication, we argue that they are also applicable in a RAG setting.

We propose a simple, unsupervised method that injects pragmatics in any RAG framework. In particular, our method: (a) identifies which sentences in a pool of documents retrieved by RAG are

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

most relevant to the question at hand (maxim of relation), and cover all the topics addressed in the input question and no more (maxim of quantity and manner);[1] and (b) highlights these sentences in the documents before they are provided to the LLM. Table 1 shows an example of our method in action.

The contributions of our paper are:

(1) We introduce a strategy to introduce pragmatics into any RAG method such as Dense Passage Retrieval [9]. To our knowledge, we are the first to investigate the impact of pragmatics for RAG.

(2) We evaluate the contributions of pragmatics in RAG on three datasets: ARC-Challenge [2], PubHealth [12] and PopQA [14] and with three different LLMs: Mistral-7B-Instruct-v0.1, Alpaca-7B [18] and Llama2-7B-chat [19]. Our results indicate that pragmatics helps when the QA task involves single-hop or multi-hop reasoning. Our post-hoc analysis further shows that this approach fares well in extracting relevant evidence sentences when entities in the query and KB passages share causal relationships. However it also uncovers challenges related to handling negation cues and arithmetic reasoning in retrieval setups such as ours, where the model may fail to answer the query correctly even if a complete set of relevant evidences are retrieved. Furthermore, we find that for factoid QA tasks: if a set of ambiguous contexts are first retrieved for a given query where the query itself contains no information for disambiguating between these contexts, our approach may highlight irrelevant evidences, which can slightly degrade the LLM's QA performance.

2 Related Work

Table 1: Example of a multiple-choice question (MCQ) from the ARC-C dataset [2] together with a fragment of a supporting document retrieved, in which the relevant evidence is highlighted with "<evidence>" tokens by our pragmatics-inspired algorithm. This evidence highlighting allows the downstream LLM to identify the correct answer (option B).

Highlighted evidence	[. . .] Bats are famous for using echolocation to hunt down their prey, using sonar sounds to capture them in the dark. Another reason for nocturnality is avoiding the heat of the day. **<evidence>This is especially true in arid biomes like deserts, where nocturnal behavior prevents creatures from losing precious water during the hot, dry daytime.</evidence>** This is an adaptation that enhances osmoregulation. One of the reasons that (cathemeral) lions prefer to hunt at night is to conserve water.
MCQ	Question: Many desert animals are only active at night. How does being active only at night most help them survive in a hot desert climate? **Choices**: A. They see insects that light up at night. B. They lose less water in the cool air. C. They find more plant food by moonlight. D. They absorb sunlight during sleep.

Since it was first proposed [13], RAG has become an essential arrow in the quiver of LLM tools. However, many of the proposed RAG approaches rely on supervised learning to jointly optimize the retrieval component and the LLM [13, 5, 21, inter alia] or to decide "when to retrieve" [1]. Instead, our approach is training free: it uses a set of unsupervised heuristics that approximate Grice's maxims (refer to Section 1).

Part of our method is similar to Active-RAG, which also reformulates the input query [8]. However, unlike Active-RAG, we use pragmatics to reformulate the input query and retrieve evidence for it, instead of relying on LLM probabilities.

Our work is also similar to [21] and [17], which also touch on pragmatics by reducing the quantity of text presented to the LLM through summarization. However, the method used in [21] is supervised. Furthermore, both of these methods exhibit considerably higher overhead compared to our proposed approach, which relies on simple yet robust heuristics.

Our method adopts a *pre-retrieval* reasoning approach that is complementary to post-retrieval reasoning approaches such as [20, 11], which reason after document retrieval. Further, we do not focus on reasoning about whether the retrieval was useful or not [6]. Instead we incorporate reasoning directly into retrieval, i.e., we first reason about the task, then retrieve following the simple technique described in [24]

Lastly, our work focuses on improving the utility of retrieved documents, somewhat similar to CRAG [23]. However, we do not improve utility by retrieving more documents (e.g., from a web search)

[1]We envision that the maxim of quality could be considered too by identifying factual statements [15]. We leave this for future work.

but rather by highlighting useful information already present in the current set of documents through pragmatics.

3 Approach: Combining Step-Back Reasoning With Pragmatic Retrieval

Conceptually, our approach is a simple plug-and-play extension that emphasizes important information in any standard RAG setup. In this paper, we apply our extension to a collection of documents retrieved by a dense passage retriever (DPR) [7].[2] We adapt the unsupervised iterative sentence retriever proposed by [22] to identify important sentences in the documents retrieved by RAG with DPR, as follows: **(1)** Given a query and associated passages retrieved by DPR, the query is first conjoined with a more abstract *step-back* version of itself created by a *step-back LLM* [24]. **(2)** In the first sentence retrieval iteration, this conjoined query is used to retrieve a set of relevant evidence sentences from the corresponding passages (see Eqs. 1 and 2). **(3)** In the next iteration(s), the query is reformulated to focus on *missing information*, i.e., query keywords not covered by the current set of retrieved evidence sentences (see Eq. 3) and the process repeats until all question phrases are covered. As such, this strategy implements Grice's maxims of relation (because the evidence sentences are relevant to the question), quantity, and manner (because we identify as many sentences as needed to cover the question and no more). By aggregating sets of retrieved evidence sentences across iterations, this retrieval strategy allows constructing *chains* of evidence sentences for a given query, which can extend dynamically until a parameter-free termination criteria is reached. Further, by varying the first evidence sentence in the top N[3] retrieved evidences, we can trivially extend this retriever to extract *parallel evidence chains*, each of varying lengths, to create a more diverse set of evidence sentences that support the query.

Lastly, we condition the generation of the Question Answering (QA) LLMs on the retrieved evidences, highlighted with special *evidence tokens*, embedded in their original DPR contexts, in order (see Table 1 for an example). We describe each of these stages in more detail below.

3.1 Step-Back Query Expansion

In this work, we employ *Step-Back Prompting* [24], a simple technique to integrate LLM driven reasoning into the retrieval process. A step-back prompt elicits from the LLM an abstract, higher-level question derived from the original query, encouraging higher-level reasoning about the problem.

We hypothesize that step-back queries, representing a more generalized query formulation, when utilized as initialization seeds for the iterative retrieval, will generate a more diverse yet still relevant set of candidate evidence sentences. For multiple-choice questions (MCQs), we generate step-back answer choices for each option, combining them with the step-back query to guide retrieval. This approach introduces an additional dimension of parallelism in constructing evidence chains for MCQs. The stepback prompts used for multi-hop reasoning are adapted from [24] (refer to Table 7 in Appendix A.4 for exemplars).

3.2 Parallel Iterative Evidence Retrieval

Computing an alignment score between queries & documents is a critical step in any retrieval system. Keeping in mind the Gricean maxim's of *quality* and *relation* (Section 1), which emphasize relevance and factual grounding, we leverage a principle similar to "late interaction" [10] & [16], where evidences are selected based on token-level similarities between queries and KB passages. We align query tokens with tokens from each sentence in the KB passages to construct evidence sentences, by selecting the most maximally similar token from the KB passage based on cosine similarity scores over dense embeddings[4] (Equation 1).

$$s(Q, P_j) = \sum_{i=1}^{|Q|} align(q_i, P_j) \tag{1}$$

[2]We use the same KB collection of documents as Self-RAG [1] and CRAG [23].

[3]In our experiments, we set $N = 3$.

[4]While [22] align tokens based on similarity over GloVe embeddings, we use sentence transformer embeddings: `https://huggingface.co/jinaai/jina-embeddings-v2-base-en`

$$align(q_i, P_j) = \max_{k=1}^{|P_j|} cosSim(q_i, p_k) \quad (2)$$

where q_i and p_k are the i^{th} and k^{th} terms of the query (Q) and evidence sentence (P_j) respectively.

Query reformulation is driven by remainder terms, defined as the set of query terms which have not yet been covered by the set of evidence sentences which were retrieved in the first i iterations of the multi-hop retriever (Equation 3):

$$Q_r(i) = t(Q) - \bigcup_{s_k \in S_i} t(s_k) \quad (3)$$

where $t(Q)$ represents the unique set of query terms, $t(s_k)$ represents the unique terms of the k^{th} evidence sentence in set S_i, which is the set of evidences retrieved in the i^{th} iteration of the retrieval process.

The notion of coverage here is based on soft matching alignment: a query term is considered to be included in the set of evidence terms if its cosine similarity with a evidence term is greater than M[5]. Note that the goal of query reformulation is to maximize the coverage of the query keywords by the retrieved chain of evidences, which aligns with the notion of the maxim of *quantity* (Section 1).

Ambiguous queries are mitigated by dynamically expanding the current query with terms from all previously retrieved evidence sentences if the number of uncovered terms in the query falls below T,[6] which also satisfies the last of Grice's maxims (maxim of *manner*).

Settings	**ARC-C**	**PubHealth**	**PopQA**
No Retrieval			
Mistral-7B-Instruct	**62.39** (+6.72%)	74.82 (+0.96%)	32.52 (-49.73%)
Alpaca-7B	34.02 (-17.43%)	43.25 (-7.78%)	30.24 (-53.04%)
Llama2-7B	40.94 (-9.78%)	68.02 (+10.57%)	23.73 (-64.07%)
DPR (No Evidence Highlighting)			
Mistral-7B-Instruct	58.46	74.11	64.69
Alpaca-7B	41.20	46.90	64.40
Llama2-7B-chat	45.38	61.52	**66.05**
DPR + Evidence Highlighting + No Step-back			
Mistral-7B-Instruct	59.23 (+1.32%)	76.04 (+2.60%)	63.90 (-1.22%)
Alpaca-7B	41.28 (+0.19%)	50.56 (+7.80%)	63.83 (-0.89%)
Llama2-7B-chat	47.44 (+4.54%)	62.64 (+1.82%)	65.98 (-0.10%)
DPR + Evidence Highlighting + Step-back			
Mistral-7B-Instruct	59.57 (+1.90%)	**76.14** (+2.74%)	64.19 (-0.77%)
Alpaca-7B	41.37 (+0.41%)	56.14 (+19.70%)	64.05 (-0.54%)
Llama2-7B-chat	47.95 (+5.66%)	66.40 (+7.94%)	65.76 (-0.43%)

Table 2: Our pragmatics driven RAG versus a Standard DPR RAG setup. **Bold** numbers indicate the best performance among all methods and LLMs for a specific dataset. Percentage changes relative to the *DPR without Evidence Highlighting* setting are shown in parentheses. Positive changes are highlighted in green, negative in red. In the *No Retrieval* setting, we do not retrieve any documents and test the LLM's parametric knowledge. *DPR (No Evidence Highlighting)* refers to the setting where we provide the top-K passages for each query to the LLM without highlighting any evidence sentences within those passages. In the *DPR + Evidence Highlighting + No Step-back* setting, we provide DPR passages annotated with highlighted evidences using "<evidence>" tokens. The *DPR + Evidence Highlighting + Step-back* setting extends the previous setting by introducing reformulated queries and answer choices using Step-back prompting.

4 Results and Discussion

We evaluate our method on the test sets of ARC-Challenge, PubHealth & PopQA. We use the evaluation metrics used in self-RAG [1]. For closed-tasks (ARC-Challenge, PubHealth), the metrics

[5]In this work, we set $M = 0.98$.

[6]In this work, we set $T = 4$.

represent Accuracy. For the short-form generation task (PopQA), the metrics indicate performance based on whether gold answers are included in the model generations instead of strictly requiring exact matching. Table 2 shows that integrating pragmatic hints into RAG can enhance performance over a standard RAG system. With Mistral-7B-Instruct, we observe improvements over the DPR baseline of 1.90% on the ARC-Challenge dataset using *evidence highlighting + step-back reasoning* and 2.74% on PubHealth. Using Alpaca-7B, we observe a significant accuracy increase of up to 19.7% on PubHealth. Similarly, with Llama-2-7B-chat, we find that our approach helps it outperform the DPR baseline by 5.66% on ARC-Challenge and 7.94% on PubHealth, respectively. For both ARC-C and PubHealth, the "*DPR + Evidence Highlighting + Step-back reasoning*" setting consistently outperforms the "*Dense Passage Retrieval (DPR) (No Evidence Highlighting)*" setting and the "*DPR + Evidence Highlighting + No Step-back reasoning*" setting. It should be noted that in some cases, the model under the *No Retrieval* setting may achieve the best performance, suggesting possible data contamination [3] on the test set. However, our method improves over even the possibly contaminated LLM paired with Dense Passage Retrieval, which is a fair baseline in this case.

Dataset and Setting	**Llama-2–7B-chat**	**Alpaca-7B**	**Mistral-7B-Instruct**
ARC-C *(Evidences w/ Context)*	47.95	41.37	59.57
ARC-C *(Evidences w/o Context)*	47.69 (-0.54%)	38.03 (-8.07%)	58.29 (-2.14%)
PubHealth *(Evidences w/ Context)*	66.40	56.14	76.14
PubHealth *(Evidences w/o Context)*	54.82 (-17.44%)	49.34 (-12.11%)	62.23 (-18.27%)

Table 3: Performance of various models on ARC-C and PubHealth datasets when using highlighted evidences within their original context versus using highlighted evidences while discarding surrounding context. Percentage changes (decreases) are shown in parentheses relative to the full context setting. Using highlighted evidence without its surrounding context can significantly degrade the LLMs QA performance.

Our error analysis indicates that this method may help in answering single-hop and multi-hop cause-and-effect queries, especially when causal entities overlap with passage sentences, as seen in ARC-Challenge and PubHealth. However, it struggles with queries requiring arithmetic reasoning or manipulation of physical quantities, as such tasks test mathematical reasoning rather than factuality. The method also falls short in handling negation cues (e.g., double negation) and addressing hypothetical or counterfactual questions. In factoid QA tasks like PopQA, highlighted evidences slightly degrade performance compared to DPR, likely because such tasks rely more on the model's parametric knowledge. For instance, PopQA queries like "What is Joseph Weydemeyer's occupation?" often retrieve ambiguous contexts with multiple roles (e.g., military officer, politician), offering insufficient signals for disambiguation, thereby limiting the utility of highlighted evidence.

Maintaining full DPR context is useful: We conduct an experiment where we compare how dropping the context surrounding the highlighted evidence sentences versus keeping it affects QA performance. As shown in Table 3, on both ARC-C and PubHealth with three different LLMs, we find that just providing the highlighted evidence sentences without context can significantly degrade QA performance relative to the scenario where we highlight evidence while keeping the full, surrounding context.

Evaluating Quality of Highlighted Evidence: We also conduct a human evaluation of the quality of evidence highlighting for a sample of 40 questions, 20 of which are sampled from ARC-Challenge and 20 of which are sampled from the PubHealth dataset. We score each highlighted evidence according to the following scale: **0 (bad)**, **0.5 (medium)** and **1 (good)**. Overall, 60% to 70% of highlighted evidences were rated at least "medium" by the human evaluator across both datasets. See Appendix A.3 for examples of 'good', 'medium' and 'bad' evidence sentences. We include more examples of low quality retrieved evidences in Appendix A.2. Please refer to Appendix A.5 for details of the prompts used and other experimental details.

5 Conclusion

We introduce a simple unsupervised method that injects pragmatic principles into retrieval-augmented generation (RAG) frameworks such as Dense Passage Retrieval. Our approach identifies and highlights sentences within retrieved documents that are most relevant to the question, ensuring they cover all the topics addressed without introducing extraneous information. By providing these highlighted sentences to large language models, we show that we can improve the accuracy of retrieval.

Acknowledgments and Disclosure of Funding

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.

[3] Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. *CoRR*, abs/2308.08493, 2023.

[4] Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.

[5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020.

[6] Shayekh Bin Islam, Md Asib Rahman, KSM Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[7] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2021.

[8] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation, 2023.

[9] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[10] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert, 2020.

[11] Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models, 2023.

[12] Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*, 2020.

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

[14] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*, 2022.

[15] Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. Neural models of factuality. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction, 2022.

[17] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *Proceedings of the International Conference on Machine Learning*, 2024.

[18] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[20] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, 2023.

[21] Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. In *Proceedings of the International Conference on Machine Learning*, 2024.

[22] Vikas Yadav, Steven Bethard, and Mihai Surdeanu. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering, 2020.

[23] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

[24] Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models, 2024.

A Appendix / supplemental material

A.1 Limitations

This study investigates the effectiveness of pragmatics in enhancing Retrieval Augmented Generation (RAG) systems. Our evaluation, however, is limited to a comparison against a standard Dense Passage Retriever (DPR) baseline. The proposed method has potential for integration with more sophisticated RAG systems, such as those developed by [1], [21], [17]. Our assessment encompasses three datasets, but a more comprehensive evaluation would involve a broader range of single-hop and multi-hop tasks, as well as a wider array of advanced RAG systems to validate the consistency of improvements. While we hypothesize that our retrieved & highlighted evidences constitute "shallow chains of thought" which are faithfully utilized by the Large Language Model in its generations, this assertion remains to be formally validated through rigorous analysis.

Table 4: Examples of Good, Medium and Bad Highlighted Evidences

Category	Examples of Evidences
Good Evidence	**Question:** A certain atom has 20 electrons, 21 neutrons, and 20 protons. What is the atomic mass of the atom? **Highlighted Evidence**: - "Mass number Mass number The mass number (symbol "A", from the German word "Atomgewicht" (atomic weight), also called atomic mass number or nucleon number, is the total number of protons and neutrons (together known as nucleons) in an atomic nucleus." - "The modern form of the whole number rule is that the atomic mass of a given elemental isotope is approximately the mass number (number of protons plus neutrons) times an atomic mass unit (approximate mass of a proton, neutron, or hydrogen-1 atom)."
Medium Evidence	**Question:** A law in Japan makes it illegal for citizens of that country to be fat **Highlighted Evidence**: - "Japan implemented the 'metabo' law which included the measurement of waist sizes in 2008 in attempt to overcome increasing obesity rates." - "The New York Times wrote: To reach its goals of shrinking the overweight population by 10 percent over the next four years and 25 percent over the next seven years, the government will impose financial penalties on companies and local governments that fail to meet specific targets." - "In January 2008, Japan passed the "Metabo Law," named after metabolic syndrome, a cluster of conditions - increased blood pressure, a high blood sugar level, excess body fat around the waist and abnormal cholesterol levels - that occurring together can increase the risk of heart disease, stroke and diabetes, Snopes.com reported". - "The law requires models to have a minimum body mass index to work and if an image was photoshopped to make the model appear thinner, it must have a warning."
Bad Evidence	**Question:** Ted Cruz Says Democrats are embracing abortion up until the moment of birth and even, horrifically, after that **Highlighted Evidence**: - "In January 2016, Cruz announced his P̈ro-Lifers for Cruzc̈oalition, chaired by Tony Perkins; co-chairs include Troy Newman, who has previously stated that the government has a responsibility to execute abortion doctors "in order to expunge bloodguilt ["sic"] from the land and people."" - "Kamala Harris refutes ridiculous Republican claims about Democrats abortion views: Or if you would prefer:" - "In the mid-1990s, Moynihan was one of the Democrats to support the ban on the procedure known as partial-birth abortion.".

Category	Frequency (ARC-Challenge)	Frequency (PubHealth)
Bad (0)	6	8
Medium (0.5)	10	4
Good (1)	4	8

Table 5: Highlighted Evidence Quality Scores for 20 randomly sampled queries from the ARC-Challenge and PubHealth datasets. The frequencies represent the number of instances falling into each quality category for the highlighted evidence in both datasets.

A.2 Errors in Evidence Highlighting

In Table 6, we include some examples of retrieved evidences from the ARC-C dataset that do not help the model to deal with specific tasks, especially those which requiring modeling negation and arithmetic reasoning.

A.3 Evaluating Quality of Highlighted Evidences:

We categorize highlighted evidence as "bad" (score: 0) when it includes completely irrelevant sentences or sentences within contexts that are somewhat related to the query but fail to provide any meaningful support in addressing it. In the case of fact-checking datasets like PubHealth, we also

Table 6: Examples of low-quality evidences retrieved for certain questions.

Dataset	**Examples of Low Quality Evidences**
ARC-Challenge	**Question:** Scott filled a tray with juice and put it in a freezer. The next day, Scott opened the freezer. How did the juice most likely change? **Evidence:** - Most recently, Scott produced the documentary film "Apple Pushers" with Joe Cross (filmmaker) juicer and a generator. - However, in March 1996, 70,000 Juice Tiger juicers, 9% of its models, were recalled after 14 injury incidents were reported.
ARC-Challenge	**Question:** A physicist wants to determine the speed a car must reach to jump over a ramp. The physicist conducts three trials. In trials two and three, the speed of the car is increased by 20 miles per hour. What is the physicist investigating when he changes the speed? **Evidence:** - Objects in motion often have variations in speed (a car might travel along a street at 50 km/h, slow to 0 km/h, and then reach 30 km/h). - Preparing an object for g-tolerance (not getting damaged when subjected to a Alfred E. Perlman control the car's speed. - Hence the round-trip time on traveler clocks will be $\Delta\tau = 4\left(\frac{c}{\alpha}\right)\cosh(\gamma)$.
ARC-Challenge	**Question:** Human activities affect the natural environment in many ways. Which action would have a positive effect on the natural environment? **Evidence:** - This environment encompasses the interaction of all living species, climate, weather and natural resources that affect human survival and economic activity. - For instance, the actions of the United States Army Corps of engineers, which threatened ecosystems within the Oklawaha River valley in Florida, and the numerous problems associated with preserving Pacific Coast Redwood communities, are utilized as case studies to elucidate the impact of human activity on the environment. - Humans have contributed to the extinction of many plants and animals.

classify highlighted evidence as "bad" if it appears to support a claim but overlooks negations in the surrounding context that would ultimately refute the claim.

Highlighted evidence is categorized as "medium" (score: 0.5) when it consists of sentences situated in relevant contexts that may allow the correct answer to be inferred indirectly in some instances but lack the direct or explicit support needed to effectively answer the query.

Highlighted evidence is categorized as "good" (score: 1) when it includes a sufficient number of sentences that directly address the query while ensuring no confounding factors (e.g., negations in the surrounding context) are overlooked.

Examples of 'Good', 'Medium' and 'Bad' evidences are shown in Table 4. The distribution of highlighted evidence quality scores assigned by a human evaluator are shown in Table 5 for ARC-C and PubHealth datasets. For ARC-Challenge, 14 out of 20 highlighted evidence sentences were rated as 'medium' to 'good,' with half receiving a 'medium' rating. Similarly, for PubHealth, 12 out of 20 highlighted evidence sentences were rated as 'medium' to 'good.' by the human evaluator.

A.4 Step-Back Reasoning Examples

Please refer to Table 7 for examples of original queries and the more abstract *Step-back* questions elicited from those queries.

A.5 Experimental Details

Our experimental results for Mistral-7B-Instruct v0.1, Alpaca-7B & Llama-2-7B differ from those reported by other works such as Self-RAG[1] & CRAG[23], and Speculative RAG due to methodological variations:

1. **Evaluation Function:** We employ a different evaluation criteria for assessing accuracy between Large Language Model (LLM) generations and gold labels in tasks such as ARC-

Table 7: Examples of Step-back questions created from original questions in the three datasets.

Dataset	Original Question and Step-back Question
ARC-Challenge	**Original Question:** An astronomer observes that a planet rotates faster after a meteorite impact. Which is the most likely effect of this increase in rotation? [SEP] **Stepback Question:** What effects do meteorite impacts on planets have?
ARC-Challenge	**Original Question:** A group of engineers wanted to know how different building designs would respond during an earthquake. They made several models of buildings and tested each for its ability to withstand earthquake conditions. Which will most likely result from testing different building designs? [SEP] **Stepback Question:** What are the testing methods used by the engineers to determine the earthquake resilience of the different building models?
PopQA	**Original Question:** What is Henry Feilden's occupation? [SEP] **Stepback Question:** What are the important aspects of Henry Feilden's academic work?
PubHealth	**Original Question:** A mother revealed to her child in a letter after her death that she had just one eye because she had donated the other to him. [SEP] **Stepback Question:** What are the circumstances surrounding the donation of the mother's second eye to her child after her death?

Challenge, PopQA, and PubQA. Our approach considers an LLM generation correct based on the principle of "inclusion," i.e., if the generation includes the correct answer as a substring, post-normalization.

2. **Number of DPR-retrieved passages (K):** We set $K = 11$ for all models, where 10 passages are from the Wikipedia KB mixed with a web search result from CRAG.
3. **Prompt Engineering:** Our prompts differ slightly from those used in Self-RAG and C-RAG. We have engineered our prompts to adhere more closely to the recommended Instruction Tuning format, particularly for Alpaca-7B [18] and Llama-2-7B-chat [19].
4. **Stepback-LLM:** In all experiments, we use Mistral-7B-Instruct v0.1 as the step-back LLM.

A.6 Example Prompts

Examples of the prompts utilized in our study are as follows:

- **ARC-Challenge**
  - Mistral-7B-Instruct:

```
Refer to the following documents, follow the instruction and answer the question.\n\n
Documents:{highlighted_passages}\n\n
Question: {question}\n\n
Instruction: Given four answer candidates, A, B, C and D, choose the best answer choice.
Please answer with the capitalized alphabet only, without adding any extra phrase or period.\n
Choices: {choices_str}
```

  - Alpaca-7B:

```
Below is an instruction that describes a task. Write a response that appropriately completes
the request.\n\n
### Instruction: Given four answer candidates, A, B, C and D, choose the best answer choice.
Please answer with the capitalized alphabet only, without adding any extra phrase or period.\n\n
### Input\n
Documents: {highlighted_passages}\n
Question: {question}\n
Choices: {choices_str}\n\n
### Response:
```

  - Llama-2-7B-chat:

```
Below is an instruction that describes a task. Write a response that appropriately completes
the request.\n\n
### Instruction: Given four answer candidates, A, B, C and D, choose the best answer choice.
Please answer with the capitalized alphabet only, without adding any extra phrase or period.\n\n
### Input\n
```

```
Documents: {highlighted_passages}\n
Question: {question}\n
Choices: {choices_str}\n\n
### Response:
```

- **PopQA**
  - Mistral-7B-Instruct:

```
Refer to the following documents, follow the instruction and answer the question.\n\n
### Input:\n
Documents: {highlighted_passages}\n\n
### Instruction: Answer the question: {question}
### Response:
```

  - Alpaca-7B:

```
Below is an instruction that describes a task. Write a response that appropriately completes
the request.\n\n
### Instruction: Refer to the following documents and answer the question.\n
### Input:\n
Documents: {highlighted_passages}\n\n
Question: {question}\n
### Response:
```

  - Llama-2-7B:

```
<s>[INST] <<SYS>>
   You are a helpful, respectful and honest assistant. Always answer as helpfully as possible,
   while being safe. Your answers should not include any harmful, unethical, racist, sexist,
   toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased
     and positive in nature.

   If a question does not make any sense, or is not factually coherent, explain why instead of
   answering something not correct. If you don't know the answer to a question, please don't
     share false information.
<</SYS>>

Below is an instruction that describes a task. Write a response that appropriately completes
the request.

Instruction: Refer to the following documents and answer the question.

Documents: {highlighted_passages}

Question: {question}
### Response: [/INST]
```

- **PubHealth**
  - Mistral-7B-Instruct:

```
Read the documents and answer the question: Is the following statement correct or not?
Only say true if the statement is true; otherwise say false. Don’t capitalize or add periods,
just say "true" or "false".\n\n
Documents: {highlighted_passages}\n\n
Statement: {question}\n
### Response:
```

  - Alpaca-7B:

```
Below is an instruction that describes a task. Write a response that appropriately completes
the request.\n\n
### Instruction: Read the documents and answer the question: Is the following statement correct
or not? Only say true if the statement is true; otherwise say false. Don’t capitalize or add
periods, just say "true" or "false".\n\n
### Input:\n
Documents: {highlighted_passages}\n\n
```

```
Statement: {question}\n
### Response:
```

- Llama-2-7B:

```
<s>[INST] <<SYS>>
   You are a helpful, respectful and honest assistant. Always answer as helpfully as possible,
   while being safe. Your answers should not include any harmful, unethical, racist, sexist,
   toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased
     and positive in nature.

   If a question does not make any sense, or is not factually coherent, explain why instead of
   answering something not correct. If you don't know the answer to a question, please don't
     share false information.
<</SYS>>

Below is an instruction that describes a task. Write a response that appropriately completes
the request.

### Instruction: Read the documents and answer the question: Is the following statement correct
or not? Only say true if the statement is true; otherwise say false. Don’t capitalize or add
periods, just say ”true” or ”false”.

### Input:
Documents: {highlighted_passages}

Statement: {question}
### Response: [/INST]
```

These methodological distinctions should be considered when comparing our results with those of previous studies.

A.7 Compute Resources

All experiments were conducted on a hardware instance consisting of 2 Nvidia H100 GPUs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim that highlighting sentences based on pragmatics principles in retrieved context can help improve accuracy of standard RAG systems such as DPR. Table 2 indicates that this claim mostly holds for 2 different LLMs across 2 datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of this work are discussed in Appendix A.1

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is an empirical work that does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the approach taken to inject pragmatic principles in RAG by describing the retrieval process, the LLMs used, stepback prompting and evidence highlighting with illustrated examples, including the datasets evaluated on and including the prompts used (see Appendix A.6). The accompanying source code to reproduce all experiments will also be released upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All of the datasets used in this work are publically available as benchmark test sets. The code will be provided upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The method presented in the paper does not include a supervised or training component. The test details (which test sets were used) are mentioned in Section 4. Other hyperparameters are specified in footnotes on pages 2, 3 and 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not include statistical significance information for the results. To be as deterministic as possible, the accuracy metrics reported are obtained by setting the model temperature to be close to 0.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify the compute resource (GPUs) used for this work in Appendix A.7

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not anticipate any ethical concerns related to this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We leave as future work, the discussion on the societal impact of RAG systems such as ours.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used in this code are fully open source and have been properly cited along with their versions used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.