# CREAM-RAG: Enhanced Retrieval Augmented Generation to Limit Hallucination through Consistency-based Self-RAG

**Yuliah Louis, Vivek Sekhadia, James Vaisman, Kingston Huynh, Sri Yanamandra, Kevin Zhu, Ryan Lagasse**

Paper ID: [32]

## Abstract

Retrieval-Augmented Generation (RAG) systems enhance large language models (LLMs) by grounding responses in external evidence, mitigating hallucinations, and enabling access to up-to-date, domain-specific knowledge. However, existing RAG frameworks often suffer from unstable self-supervised optimization signals and inconsistent factual grounding. We introduce CREAM-RAG (Consistency-Regularized Enhanced Augmented Model for RAG) (Wang et al. 2025a), a unified framework that integrates retrieval, Direct Preference Optimization (DPO)-based self-reward reinforcement learning, and a consistency regularization objective to stabilize reward dynamics during fine-tuning. By enforcing alignment between multiple retrieved contexts and generated responses, CREAM-RAG improves factual faithfulness and semantic coherence without external supervision. Empirical evaluations on the LLaMA-2-7B model demonstrate that CREAM-RAG achieves a 35.04% average improvement over the base model across reasoning and factuality benchmarks, highlighting its effectiveness in reducing hallucinations and enhancing retrieval-grounded reasoning.

## Introduction

Retrieval-Augmented Generation (RAG) enhances language models by grounding outputs in retrieved, relevant documents, effectively addressing key limitations such as hallucinations, updated knowledge, and restricted domain expertise (Lewis et al. 2021; Kwiatkowski, Palomaki, et al. 2019). By integrating external sources during the inference time, RAG provides accurate and up-to-date facts. Recent advancements, including autonomous retrieval frameworks, further improve efficiency and scalability by reducing dependency on larger context windows and lowering computational costs (Zhou and Chen 2025).

Despite these improvements, existing self-rewarding RAG approaches, which focus on refining retrieval and reward design, overlook a critical weakness: the inherent instability of self-generated reward signals during training. Such systems are prone to failure modes like reward hacking and retrieval-blind collapse, with recent work identifying temporal inconsistency as a fundamental yet unsolved issue (Niu et al. 2024).



We propose stabilizing the self-reward process in RAG systems to reduce hallucinations and reward misalignment. Rather than redesigning reward types, our method enhances their reliability over time by applying consistency regularization to self-reward signals, drawing on recent advances from Consistency Regularized Self-Rewarding Language Models (CREAM) (Wang et al. 2025a).

We introduce CREAM-RAG, which combines a retrieval module with an actor-critic generator where the critic assigns self-rewards. Crucially, we augment standard Direct Preference Optimization (DPO) with a consistency loss that minimizes divergence in reward signals across training steps. By incorporating a frozen reference model to penalize shifts in reward judgments, our approach ensures more stable and trustworthy reinforcement signals throughout the generation process.

## Related Works

Our method integrates advances in RAG, self-rewarding RL, and consistency regularization to reduce hallucinations and improve reliability.

### Other Variants to RAG and Early RAG

Early RAG models showcased how coupling large language models with external retrieval can improve accuracy. For example, REALM introduced an end-to-end retriever-generator framework that significantly boosted open-domain question and answering (QA) performance and interpretation capabilities (Guu et al. 2020). Building on this, newer variants such as PIKE-RAG adapt retrieval and rationale generation to better serve specialized applications (Wang et al. 2025b). At the same time, systematic surveys of RAG methods emphasize both the versatility and persistent challenges of balancing retrieval efficiency and robustness (Oche et al. 2025).

More recent work has focused on addressing two major challenges. Self-RAG and other self-rewarding methods train models to adaptively decide when to retrieve, how to generate, and even how to critique their own outputs, though they remain vulnerable to noisy or limited evidence (Asai et al. 2023). On the other hand, consistency-regularization approaches like CORD encourage models to produce stable outputs under retrieval uneasiness while accounting for passage ranking, but often strike the balance poorly, either

ignoring or over-emphasizing rank information (Lee et al. 2024).

Our work integrates core principles from RAG, self-rewarding reinforcement learning, and consistency regularization. Existing RAG frameworks have enhanced factual accuracy by leveraging external knowledge, yet they frequently exhibit unstable reward dynamic during reinforcement learning. For instance, while methods like Self-RAG (Asai et al. 2023) incorporate self-assessment mechanisms, they do not enforce consistency across training steps, leaving them susceptible to reward hacking and preference drift. Conversely, consistency-focused approaches such as CORD (Lee et al. 2024) promote generation stability but fall short in merging retrieval awareness with self-rewarding objectives.

CREAM-RAG bridges these lines of research by embedding consistency-regularized self-reward mechanisms within the RAG process. Central to our approach is the use of a frozen reference model to stabilize self-reinforcing RAG systems. By integrating this consistency mechanism into an actor-critic RAG architecture and evaluating it across diverse tasks — including long-form QA, comprehension, and hallucination mitigation — we show that maintaining internal consistency is essential for reducing hallucinations and improving factual reliability. Our framework thus unifies previously separate advances in retrieval quality, self-supervision, and reward stability.

### RAG Systems

RAG enhances factuality by retrieving external documents during inference time (Lewis et al. 2021). Early systems used loosely coupled retrieval and generation, often yielding ungrounded outputs. Recent training approaches optimize the retriever and generator, although noisy retrieval can still cause hallucinations (Zhou and Chen 2025). Self-rewarding reinforcement learning methods aim to improve outputs, but suffer from unstable rewards and preference drift (Wang et al. 2025c). Consistency regularization helps stabilize training by penalizing shifts in reward preferences (Zhang et al. 2024).

### Long-form QA and Challenges

Long-form QA introduces additional challenges for RAG, requiring synthesis across lengthy and conflicting documents. Early solutions either overloaded context or oversimplified outputs (Kwiatkowski, Palomaki, et al. 2019). Recent methods use aspect-based summarization (Hayashi et al. 2020) and modular document processing (Izacard and Grave 2021), with RL further improving factual accuracy (Wang et al. 2025c).

Reliable RAG requires: (1) tightly coupled retrieval generation, (2) rigorous self-evaluation, and (3) stable reward optimization. Prior approaches addressing these in isolation remain prone to inconsistencies.

CREAM-RAG combines three core ideas: self-reflection-based adaptive retrieval, rank awareness with training, and alignment between the retriever and generator. Together, these reduce extra retrieval steps, make citations more accurate than using self-reflection alone, and give more reliable grounding than recent consistency-based approaches.

## Methodology

### Preprocessing and System Initialization

Before extending the RAG framework, we perform key initialization steps: an instruction-tuned LLM (LLaMA-2-7B) is used as both the generator and the self-reward model; a frozen copy of this model is retained; and a vector database of external documents is incorporated for contextual retrieval.

### Consistency-Regularized Self-Rewarding Reinforcement Learning Within Training Procedure

We utilize CREAM to enhance the stability of RAG: unifying candidate generation, self-evaluation, and regularization into a five-stage loop.

1. Retrieval: The RAG module retrieves relevant documents for the input prompt.
2. Candidate Generation: For the given query and retrieved context, the model produces multiple responses conditioned on both prompt and documents.
3. Self-Scoring: The model ranks its own outputs, assigning preference-based rewards for factual grounding, coherence, and relevance.
4. Consistency Regularization: A KL divergence loss is computed between the current model's preferences and those of a frozen reference model, penalizing inconsistent reward shifts to stabilize learning.
5. Model Optimization: DPO updates the model parameters. The reference model is periodically updated to mirror the current model, maintaining training stability.

This end-to-end framework directly addresses key challenges of unstable rewards, retrieval noise, and hallucinations by ensuring a stable training process aligned with high-quality output.

### Formalization of CREAM-RAG Objective Functions

We now describe the mathematical underpinnings of CREAM-RAG's training objectives, including the reward function, ranking stability via Kendall Tau, and the final per-pair DPO loss.

First, the reward function $r_{ij}$ reflects the improvement in logarithmic likelihood of the current model $P_\theta$ over a reference model $P_{\text{ref}}$, with an optional normalization term.

$$r_{ij} = \beta\Big[\log P_\theta(y_{ij} \mid x_j) - \log P_{\text{ref}}(y_{ij} \mid x_j)\Big] + \beta \log Z(x_j) \tag{1}$$

To stabilize learning, we introduce a consistency measure between model and reference preferences using Kendall Tau:

$$\tau_j = \frac{2}{N(N-1)} \sum_{1 \le i < i' \le N} \begin{bmatrix} \mathbf{1}((J_{ij} - J_{i'j})(K_{ij} - K_{i'j}) > 0) \\ -\mathbf{1}((J_{ij} - J_{i'j})(K_{ij} - K_{i'j}) < 0) \end{bmatrix} \tag{2}$$

Finally, we train the model using a per-pair DPO loss, which aligns generation likelihood with self-assessed preferences while preserving consistency with the reference model:

$$\mathcal{L}_{\text{DPO}}(\theta; y, y', x, z) = -z(y, y', x) \log \left( \frac{P_\theta(y \mid x)}{P_{\text{ref}}(y \mid x)} \right) - \big(1 - z(y, y', x)\big) \log \left( \frac{P_\theta(y' \mid x)}{P_{\text{ref}}(y' \mid x)} \right) \quad (3)$$

# Experiments

## Training and Evaluating

Our RAG contains 3 main parts for training and evaluation: long-form parsing, comprehension, and hallucination training. In total, we used the full training and evaluation methods of 5 datasets (Hotpot-QA, Natural Questions, Trivia-QA, RAGTruth, and SQuADv2) (Yang et al. 2018; Kwiatkowski, Palomaki, et al. 2019; Joshi et al. 2017; Niu et al. 2024; Rajpurkar, Jia, and Liang 2018) and compared all scores against a baseline. 10,000 samples were pulled from each dataset when training our baseline and model. Additionally, when running tests on both our baseline and model, we utilized BERT scoring to ensure effective semantic evaluation of both models (Zhang et al. 2020). This allowed us to test our model with strong, reliable metrics that resonate more with human judgment.

Long-form parsing tasks include two sets: a question-answer dataset that contains real user questions from Google search where the answers are found in Wikipedia (Natural Questions; Kwiatkowski, Palomaki, et al. 2019), and a multi-hop question-answer dataset with 113,000 Wikipedia-based question-answer pairs (HotpotQA; Yang et al. 2018). Both datasets used F1 and exact match as evaluation metrics; we tested each dataset against a baseline.

Comprehension tasks include two question-answering datasets: TriviaQA (Joshi et al. 2017) and SQuAD v2 (Rajpurkar, Jia, and Liang 2018). Trivia QA consists of over 650,000 trivia-style question-answer-evidence triples, primarily to train the ability to answer factual knowledge-based questions. SQuAD v2 contains over 50,000 unanswerable questions that look like answerable ones, challenging models to both answer and filter between possible and impossible for questions. These two datasets gave the same scoring metrics as HotpotQA and Natural Questions, Exact Match, and F1 score.

Hallucination testing was run on RAGTruth, which consists of more than 18,000 naturally generated responses, annotated for evaluation of hallucinations (Niu et al. 2024). RAGTruth uses similar data evaluation metrics to the aforementioned datasets.

## Baselines

We employed one main baseline to test against our CREAM-RAG model. Ablation testing of our model occurred without the retrieval and CREAM.

Self-RAG (Asai et al. 2023), is a self-rewarding language model that furthered Retrieval Augmented Generation by adding a self-rewarding process. At first, RAG lacked versatility and struggled to complete tasks without human-based reinforcement learning. Self-RAG was the first to break this mold and depend upon Actor-Critic in RAG. We ran the same datasets for evaluation that we ran on our model and compared the scoring.

# Discussion

Our work introduces CREAM-RAG, a framework aimed at stabilizing reward signals and reducing hallucinations. Our experiments across various long-form QA, comprehension, and hallucination-specific tasks demonstrate how stabilizing the self-reward process leads to more reliable outputs, even in noisy applications.

## Key Findings and Interpretation

Our results across multiple benchmark datasets show that CREAM-RAG improves answer quality in both short-answer and long-form question-answering tasks:

The model achieved high scores on TriviaQA, demonstrating accurate answer selection. On SQuAD v2, its performance shows a reduced tendency to hallucinate. For long-form questions on HotPotQA and Natural Questions, it maintained strong results, proving its ability to synthesize complex answers from multiple documents.

Our experimental results across a diverse set of data suggest that the key to improving factuality and reliability lies not only in better retrieval but in more stable reward modeling during training. When the reward signal is erratic, models struggle to learn consistent patterns of accuracy, leading to frequent hallucinations or degraded performance in noisy circumstances. Contrastingly, CREAM-RAG's stabilized self-reward mechanism enables the model to better distinguish between accurate and inaccurate generations, even when retrieval results are partially irrelevant or convoluted.

These findings have several important implications. First, they underscore the critical role of reward signal quality in the success of RAG systems, specifically for tasks that require high factual accuracy. Second, they demonstrate that improving internal dynamics (e.g., stabilizing the self-reward process) can be as influential as external improvements such as improved retrieval or model scaling.

## Broader Context and Significance

CREAM-RAG advances beyond traditional RAG by mitigating reward hacking and enabling reliable self-evaluation, which is vital for creating autonomous RAGs in high-stakes domains.

Beyond its technical significance, CREAM-RAG paves the way for more autonomous and adaptive RAG systems. Its use of consistency regularization for self-evaluation allows deployment in fields with scarce human feedback, such as healthcare diagnostics, legal analysis, and scientific research. However, the self-rewarding functionality introduces serious ethical risks, including potential reward hacking, bias, and obscured accountability. To ensure safe and fair use, future applications require careful monitoring, clear reward logic, and domain-specific safeguards.

## Unexpected Observations

Consistency regularization increased recall without degrading precision. The model retrieved more relevant facts (e.g.,

Table 1: Overall experimental results on four tasks. Balanced scores were calculated by allowing small token differences and numeric overlap. Token-based F1 with pre-processing took place, where the output text becomes normalized, allowing for minor capitalization or grammar differences. Normalized ground truth tokens are compared to normalized prediction tokens.

| **Benchmark** | *Balanced F1* | *Balanced EM* | *BERT Precision* | *BERT Recall* | *BERT F1* |
|---|---|---|---|---|---|
| **Llama-2-7B (No RAG)** | | | | | |
| SQuADv2 | 34.8 | 39.4 | 24.2 | 24.6 | 24.4 |
| NQ | 32.9 | 32.6 | 34.2 | 35.1 | 27.6 |
| TriviaQA | 24.5 | 28.5 | 19.8 | 27.3 | 10.5 |
| HotpotQA | 21.1 | 30.5 | 24.0 | 16.3 | 10.9 |
| RAGTruth | 18.7 | 38.2 | 26.5 | 16.8 | 11.3 |
| **Self-RAG (RAG)** | | | | | |
| SQuADv2 | 46.9 | 41.5 | 37.4 | 37.8 | 37.5 |
| NQ | 35.6 | 34.0 | 28.3 | 29.0 | 28.6 |
| TriviaQA | 26.3 | 31.4 | 44.9 | 41.9 | 43.3 |
| HotpotQA | 40.8 | 29.8 | 26.5 | 27.3 | 36.8 |
| RAGTruth | 56.3 | 36.5 | 11.5 | 28.5 | 20.1 |
| **CREAMRAG (RAG)** | | | | | |
| SQuADv2 | 43.1 | 52.0 | 84.3 | 86.6 | 85.3 |
| NQ | 46.5 | 42.3 | 82.0 | 84.1 | 82.9 |
| TriviaQA | 45.2 | 41.7 | 83.5 | 82.9 | 83.1 |
| HotpotQA | 45.9 | 44.2 | 82.6 | 85.2 | 83.8 |
| RAGTruth | 82.7 | 43.0 | 84.5 | 86.2 | 85.3 |

0.5980 recall on Natural Questions) without introducing noise (0.4147 precision), indicating a new ability to take informed risks. Additionally, significantly higher balanced exact match scores versus specific exact match confirm that answers are semantically, if not stringently, correct.

## Metrics

For scoring we used Balanced F1, Balanced EM, BERT Precision, BERT Recall, and BERT F1. We utilized BERT score to give metrics on the models ability to perform in practical settings by computing semantic similarity between the embeddings of predicted and reference sentences using BERT models (utilized Roberta-Large) (Zhang et al. 2020).

## Ablation

Table 2 reports benchmark performance of the full-scale model in comparison to two ablation studies. The removal of a consistency loss resulted in a notable drop of semantic recall and precision. The elimination of retrieval (Zero-Context) led to steep degradation across all benchmarks, thus underlining the importance of external evidence. The large performance gaps, RAGTruth especially, demonstrate the importance of consistency regularization for factual faithfulness in RAG.

Table 1's report of the full-scale CREAM-RAG Model showcases substantially better results in comparison to the ablation studies. This demonstrates an effective effect on factuality due to consistency regularization, which in turn allows for better semantic precision and recall for RAG Systems.

Table 2: Ablation Results for CREAM-RAG with Llama-2-7B using DPO. Balanced scores allow small token differences and numeric overlap. Outputs are normalized for token-based F1 evaluation.

| | | | | | |
|---|---|---|---|---|---|
| **Remove Consistency Loss (No CREAM)** | | | | | |
| SQuADv2 | 37.8 | 36.3 | 64.9 | 72.9 | 68.6 |
| NQ | 34.4 | 39.2 | 69.2 | 74.3 | 70.9 |
| TriviaQA | 31.9 | 29.1 | 67.7 | 71.4 | 72.4 |
| HotpotQA | 35.8 | 33.7 | 59.1 | 63.5 | 64.4 |
| RAGTruth | 51.3 | 37.6 | 50.2 | 54.7 | 58.3 |
| **Remove Retrieval Entirely (Zero-Context)** | | | | | |
| SQuADv2 | 24.3 | 18.2 | 22.0 | 28.6 | 25.2 |
| NQ | 21.1 | 18.3 | 21.3 | 27.7 | 25.9 |
| TriviaQA | 20.8 | 19.5 | 20.1 | 24.5 | 22.6 |
| HotpotQA | 18.9 | 15.6 | 17.4 | 21.5 | 21.1 |
| RAGTruth | 17.2 | 13.4 | 15.5 | 20.4 | 19.7 |

## Future Directions and Implications

This paves the way for several applications: multimodal retrieval with unified self-evaluation across text, images, and tables; reliable multi-step reasoning via chain-wide consistency checks; and hybrid reward models that effectively blend automated scoring with limited human guidance. On top of this, CREAM-RAG can be specialized for fields like medicine and law, increasing credibility by mitigating hallucinations while also maintaining a strong comprehension of texts. This could allow for AI trust and integration in important situations in a multitude of fields.

## Conclusion

In this paper, we present CREAM-RAG, a consistency-regularized framework for Retrieval-Augmented Generation that mitigates hallucinations and enhances factual reliability. By integrating document retrieval with self-rewarding reinforcement learning and consistency loss, our approach stabilizes training signals. Extensive experiments on long-form QA, comprehension, and hallucination benchmarks demonstrate that CREAM-RAG delivers significant gains in output quality, particularly when retrieval is imperfect. These results validate the role of consistency-based optimization in the development of more reliable, autonomous models for high-stakes applications.

## Limitations

Although CREAM-RAG sparks improvements in various benchmarks, our dependence on BERT Score as a primary metric has limitations. Although BERT Score effectively captures semantic similarity, it can overvalue the factuality of fluent or incorrect answers. This is especially problematic in long-form or multi-hop QA (such as HotpotQA), where it may fail to identify hallucinations or slight inaccuracies. Future evaluations would benefit from incorporating human judgment or task-specific factual accuracy metrics to better gauge real-world applications.

Furthermore, while tested on five diverse datasets to demonstrate validity, CREAM-RAG is not specialized for each, leading to performance variations. For instance, it achieves high precision on TriviaQA but struggles with exact match scores on the multi-hop reasoning required by HotpotQA. This diversity in datasets introduces variables like retrieval noise and answer styles, complicating direct comparisons and contributing to inconsistent output. These results reinforce that, while CREAM-RAG has broad applicability, achieving optimal performance on specialized tasks may require specific tuning.

## References


Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv preprint arXiv:2310.11511.

Bensal, S.; Jamil, U.; Bryant, C.; et al. 2025. Reflect, Retry, Reward: Self-Improving LLMs via Reinforcement Learning. arXiv preprint arXiv:2505.24726.

Cheng, M.; Luo, Y.; Ouyang, J.; Liu, Q.; Liu, H.; Li, L.; Yu, S.; Zhang, B.; Cao, J.; Ma, J.; Wang, D.; and Chen, E. 2025. A Survey on Knowledge-Oriented Retrieval-Augmented Generation. arXiv preprint arXiv:2503.10677.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv preprint arXiv:2002.08909.

Hasling, D. W.; Clancey, W. J.; and Rennels, G. 1984. Strategic Explanations for a Diagnostic Consultation System. *International Journal of Man-Machine Studies* 20(1):3–19.

Hayashi, H.; Budania, P.; Wang, P.; Ackerson, C.; Neervannan, R.; and Neubig, G. 2020. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. arXiv preprint arXiv:2011.07832.

Hayashi, H.; Budania, P.; Wang, P.; et al. 2021. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*.

Izacard, G., and Grave, É. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of EACL (Main Volume)*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv preprint arXiv:1705.03551.

Kumar, A.; Zhuang, V.; Agarwal, R.; et al. 2024. Training Language Models to Self-Correct via Reinforcement Learning. arXiv preprint arXiv:2409.12917.

Kwiatkowski, T.; Palomaki, J.; et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*.

Lee, Y.; Hwang, S.-w.; Campos, D.; Graliński, F.; Yao, Z.; and He, Y. 2024. CORD: Balancing COnsistency and Rank Distillation for Robust Retrieval-Augmented Generation. arXiv preprint arXiv:2412.14581.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

Niu, C.; Wu, Y.; Zhu, J.; et al. 2024. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. arXiv preprint arXiv:2401.00396.

Oche, A. J.; Folashade, A. G.; Ghosal, T.; and Biswas, A. 2025. A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions. arXiv preprint arXiv:2507.18910.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. arXiv preprint arXiv:1806.03822.

Su, D.; Li, X.; Zhang, J.; Shang, L.; Jiang, X.; Liu, Q.; and Fung, P. 2022. Read before Generate! Faithful Long Form Question Answering with Machine Reading. arXiv preprint arXiv:2203.00343.

Wang, Z.; He, W.; Liang, Z.; et al. 2025a. CREAM: Consistency Regularized Self-Rewarding Language Models. arXiv preprint arXiv:2410.12735.

Wang, J.; Fu, J.; Wang, R.; Song, L.; and Bian, J. 2025b. PIKE-RAG: sPecIalized KnowledgE and Rationale Augmented Generation. arXiv preprint arXiv:2501.11551.

Wang, Y.; Ren, R.; Wang, Y.; Zhao, W. X.; Liu, J.; Wu, H.; and Wang, H. 2025c. Reinforced Informativeness Optimization for Long-Form Retrieval-Augmented Generation. Corpus record (Semantic Scholar).

Yang, Z.; Qi, P.; Zhang, S.; et al. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv preprint arXiv:1809.09600.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2025. Self-Rewarding Language Models. arXiv preprint arXiv:2401.10020.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675.

Zhang, J.; Yu, Y.; Zhang, Y.; et al. 2024. CREAM: Coarse-to-Fine Retrieval and Multi-modal Efficient Tuning for Document VQA. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024. Calibrated Self-Rewarding Vision Language Models. arXiv preprint arXiv:2405.14622.

Zhou, J., and Chen, L. 2025. OpenRAG: Optimizing RAG End-to-End via In-Context Retrieval Learning. arXiv preprint arXiv:2503.08398.