Novelty Assessment Report

Paper: OSCAR: Online Soft Compression for RAG
PDF URL: https://openreview.net/pdf?id=ideKAUWvFE
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by integrating external knowledge, leading to improved accuracy and relevance. However, scaling RAG pipelines remains computationally expensive as context length grows. On one hand, hard compression methods have recently proposed to prune the retrieved text on-the-fly with a limited compression ration. On the other hand, soft compression method performs a costly offline compression thanks a dedicated LLM but with a higher compression rate. In this paper, we introduce OSCAR, a novel query-dependent online soft compression method for RAG. OSCAR bridges the gap between online hard and offline soft compression methods, bringing the best of both: OSCAR dynamically compresses retrieved information at inference time, eliminating storage overhead and enabling higher compression rates than existing methods. Our experiments demonstrate state-of-the-art performance with a 2-5x speed-up in inference and minimal, if any, accuracy loss, for LLMs ranging from 1B to 24B parameters.

Core Task Landscape

This paper addresses: **Online Soft Compression for Retrieval-Augmented Generation**

A total of **13 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Query-Dependent Online Soft Compression Methods**
- **Hybrid Compression with Selective Retrieval**
- **Task-Aware Dynamic Compression for Long Contexts**
- **Robustness and Interpretability in Compressed RAG**
- **Specialized Compression for Non-Text Retrieval**
- **Domain-Specific Adaptive Retrieval Systems**

Complete Taxonomy Tree

- Online Soft Compression for Retrieval-Augmented Generation Survey Taxonomy
- Query-Dependent Online Soft Compression Methods
  - Online Compression with Reranking Integration ★ (2 papers)
  - [0] OSCAR: Online Soft Compression for RAG (Anon et al., 2026) View paper
  - [4] OSCAR: Online Soft Compression And Reranking (Louis, 2025) View paper
  - Pretraining-Free Compression Architectures (2 papers)
  - [1] PISCO: Pretty Simple Compression for Retrieval-Augmented Generation (Maxime Louis, 2025) View paper
  - [6] Simple Context Compression: Mean-Pooling and Multi-Ratio Training (Artzi, 2025) View paper
- Hybrid Compression with Selective Retrieval
  - Snippet-Based Compression with Semantic Vectors (1 papers)
  - [2] SARA: Selective and Adaptive Retrieval-augmented Generation with Context Compression (Jin, 2025) View paper
  - Cache-Augmented Adaptive Compression (1 papers)
  - [11] Enhancing Cache-Augmented Generation (CAG) with Adaptive Contextual Compression for Scalable Knowledge Integration (Agrawal, 2025) View paper
- Task-Aware Dynamic Compression for Long Contexts
  - KV Cache Compression for Long-Context LLMs (1 papers)
  - [3] Dynamickv: Task-aware adaptive kv cache compression for long context llms (Xiabin Zhou, 2024) View paper
  - Semantic Proximity-Based Redundancy Reduction (1 papers)
  - [13] Semantic Proximity for Redundancy-Aware Context Compression in Large Language Models (C Boutros, n.d.) View paper
- Robustness and Interpretability in Compressed RAG
  - Robustness Under Noisy Retrieval Conditions (1 papers)
  - [8] Robustness of Fine-Tuned Llms Under Noisy Retrieval Inputs (Yinghao Sang, 2025) View paper
  - Explainable Influence Analysis in RAG (1 papers)
  - [5] Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation (Sang, 2025) View paper
- Specialized Compression for Non-Text Retrieval
  - Multi-Vector Document Retrieval Compression (1 papers)
  - [10] Hierarchical Patch Compression for ColPali: Efficient Multi-Vector Document Retrieval with Dynamic Pruning and Quantization (Bach Duong, 2025) View paper
  - Vision-Language-Action Memory Augmentation (1 papers)
  - [7] Map-vla: Memory-augmented prompting for vision-language-action model in robotic manipulation (Runhao Li, 2025) View paper

- Domain-Specific Adaptive Retrieval Systems
  - Multi-Interest Retrieval with Soft Probabilistic Compression (1 papers)
  - [9] SPARC: Soft Probabilistic Adaptive multi-interest Retrieval Model via Codebooks for recommender system (Shi, 2025) View paper
  - Adaptive Query Routing for Resource-Constrained LLMs (1 papers)
  - [12] An adaptive query-routing framework for optimizing small languages models in resource-constrained environments (Ribeiro, 2025) View paper

Narrative

Core task: online soft compression for retrieval-augmented generation. The field addresses the challenge of efficiently integrating retrieved documents into language models by compressing context on-the-fly rather than relying solely on hard filtering or static summarization. The taxonomy reveals several complementary directions: Query-Dependent Online Soft Compression Methods focus on tailoring compression to each query's needs, often leveraging learned representations to distill retrieved passages; Hybrid Compression with Selective Retrieval combines compression with intelligent document selection to balance coverage and efficiency; Task-Aware Dynamic Compression for Long Contexts adapts compression strategies based on downstream task requirements and context length constraints; Robustness and Interpretability in Compressed RAG examines how compression affects model reliability and explainability; Specialized Compression for Non-Text Retrieval extends these ideas to multimodal or structured data; and Domain-Specific Adaptive Retrieval Systems tailor compression and retrieval jointly to particular application areas. Representative works like PISCO[1] and SARA[2] illustrate query-dependent approaches, while DynamicKV[3] exemplifies task-aware dynamic methods.

A central tension across branches involves the trade-off between aggressive compression for efficiency and preserving sufficient signal for accurate generation, with many studies exploring learned compression modules that can be fine-tuned for specific retrieval pipelines. OSCAR[0] sits within the Query-Dependent Online Soft Compression Methods branch, specifically in the Online Compression with Reranking Integration cluster alongside OSCAR Reranking[4]. This positioning reflects its emphasis on integrating reranking signals directly into the compression process, allowing the model to prioritize salient content based on both query relevance and retrieval confidence. Compared to simpler compression schemes like Simple Context Compression[6], OSCAR[0] and its neighbor OSCAR Reranking[4] leverage richer reranking feedback to guide soft compression decisions. This contrasts with works in the Robustness and Interpretability branch, such as Explainable RAG[5], which prioritize transparency over compression efficiency, highlighting ongoing questions about how to balance compactness, fidelity, and interpretability in retrieval-augmented systems.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. OSCAR: Online Soft Compression And Reranking

Authors: Louis, Maxime, Formal, Thibault, Maxime Louis, et al. (12 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating external knowledge, leading to improved accuracy and relevance. However, scaling RAG pipelines remains computationally expensive as retrieval sizes grow. To address this, we introduce OSCAR, a novel query-dependent online soft compression method that reduces computational overhead while preserving performance. Unlike traditional hard compression methods, which shorten retrieved texts, or soft compression ap...

⚠ Similarity Notice

These papers share nearly identical titles ('OSCAR: Online Soft Compression for RAG' vs 'OSCAR: Online Soft Compression And Reranking'), describe the same core system architecture (query-dependent online soft compression with optional reranking), report identical performance metrics (2-5x speedup, 1B-24B parameter LLMs), and present the same technical approach. This appears to be the same paper or a very close variant with minor title modification.

Contributions Analysis

Overall novelty summary. The paper introduces OSCAR, a query-dependent online soft compression method for RAG that dynamically compresses retrieved documents at inference time using continuous embeddings. Within the taxonomy, OSCAR resides in the 'Online Compression with Reranking Integration' leaf under 'Query-Dependent Online Soft Compression Methods', sharing this leaf with only one sibling paper. This positioning indicates a relatively sparse research direction focused specifically on combining dynamic soft compression with reranking mechanisms, distinguishing it from broader compression approaches that lack explicit reranking components or operate offline.

The taxonomy reveals that OSCAR's immediate neighbors include 'Pretraining-Free Compression Architectures' within the same parent branch, which emphasizes lightweight compression without extensive pretraining. Adjacent branches address 'Hybrid Compression with Selective Retrieval' (combining compression with adaptive document selection) and 'Task-Aware Dynamic Compression for Long Contexts' (optimizing compression based on downstream task requirements). OSCAR's focus on query-dependent online soft compression with reranking integration positions it at the intersection of efficiency and relevance optimization, diverging from purely efficiency-driven methods or those requiring offline preprocessing.

Among 30 candidates examined, the contribution-level analysis shows mixed novelty signals. The core OSCAR method (Contribution 1) examined 10 candidates with zero refutations, suggesting relative novelty in its specific approach. However, the two architectural contributions—efficient compressor architectures (Contribution 2) and simultaneous compression with reranking (Contribution 3)—each found one refutable candidate among 10 examined. This indicates that while the overall OSCAR framework appears novel within the limited search scope, specific architectural choices and the compression-reranking integration concept have some overlap with existing work in the examined literature.

Based on the top-30 semantic matches and taxonomy structure, OSCAR appears to occupy a moderately novel position, particularly in its query-dependent online soft compression approach. The limited search scope and sparse taxonomy leaf suggest the work addresses an emerging research direction, though certain architectural and integration aspects show partial overlap with prior methods. A more exhaustive literature review would be needed to definitively assess novelty across the broader RAG compression landscape.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: OSCAR: Online Soft Compression Method for RAG

Description: The authors propose OSCAR, which dynamically compresses retrieved documents into query-optimized representations for efficient answer generation. OSCAR bridges the gap between online hard compression and offline soft compression methods, achieving 2-5× inference speed-up with minimal accuracy loss.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Query-Aware Graph Neural Networks for Enhanced Retrieval-Augmented Generation

URL: View paper

Brief Assessment

Query-Aware GNN[28] focuses on graph-based retrieval architectures with query-aware attention for multi-hop reasoning, not on document compression methods for RAG efficiency.

---

2. Efficient Dynamic Clustering-Based Document Compression for Retrieval-Augmented-Generation

URL: View paper

Brief Assessment

Dynamic Clustering Compression[30] focuses on clustering-based document organization and compression to handle redundancy and noise, while OSCAR addresses query-dependent soft compression with dynamic representation optimization. The candidate does not demonstrate prior work on online soft compression methods that bridge hard and offline soft compression approaches.

---

3. PISCO: Pretty Simple Compression for Retrieval-Augmented Generation

URL: View paper

Brief Assessment

PISCO[1] is an offline soft compression method that requires no pretraining but operates independently of queries. OSCAR is specifically designed for online, query-dependent compression, which is a fundamentally different approach.

---

4. ACoRN: Noise-Robust Abstractive Compression in Retrieval-Augmented Language Models

URL: View paper

Brief Assessment

ACoRN[31] focuses on noise-robust abstractive compression using smaller language models (T5-large) to handle factual errors and irrelevant documents in retrieval. OSCAR addresses a different problem: query-dependent online soft compression that bridges hard and offline soft compression methods, achieving 2-5× speed-up with minimal accuracy loss. The technical approaches differ fundamentally—ACoRN uses offline data augmentation for noise robustness training, while OSCAR introduces novel online compression architectures (oscar-n-layers, oscar-llama) with query-dependent embeddings.

---

5. Oreo: A plug-in context reconstructor to enhance retrieval-augmented generation

URL: View paper

Brief Assessment

Oreo[25] focuses on refining and reorganizing retrieved chunks through a three-stage training paradigm, while OSCAR specifically addresses query-dependent online soft compression with dynamic document-to-embedding mapping for efficiency gains. The candidate does not demonstrate prior work on online soft compression methods.

---

6. AttentionRAG: Attention-Guided Context Pruning in Retrieval-Augmented Generation

URL: View paper

Brief Assessment

AttentionRAG[26] focuses on hard compression through attention-guided context pruning at the text level, while OSCAR performs soft compression by mapping documents to continuous embeddings. These are fundamentally different compression paradigms.

---

7. Searching for best practices in retrieval-augmented generation

URL: View paper

Brief Assessment

RAG Best Practices[24] focuses on optimizing the entire RAG pipeline through systematic evaluation of different module combinations (retrieval, reranking, summarization), rather than proposing novel compression methods. The paper does not present query-dependent soft compression techniques.

---

8. Neural-Symbolic Dual-Indexing Architectures for Scalable Retrieval-Augmented Generation

URL: View paper

Brief Assessment

Neural-Symbolic Dual-Indexing[27] focuses on graph-based retrieval architectures with vector compression for search efficiency, not on query-dependent soft compression for RAG answer generation.

---

9. Familiarity-Aware Evidence Compression for Retrieval-Augmented Generation

URL: View paper

Brief Assessment

Familiarity-Aware Compression[32] focuses on making compressed evidence familiar to the target model through ensemble decoding at inference time, while OSCAR addresses the challenge of creating efficient query-dependent compression models that can operate online. The technical approaches differ fundamentally: Familiarity-Aware Compression[32] uses ensemble decoding between compression and target models to lower perplexity, whereas OSCAR trains specialized compressor architectures (oscar-n-layers and oscar-llama) with distillation objectives to achieve fast online compression.

---

10. Autorag-hp: Automatic online hyper-parameter tuning for retrieval-augmented generation

URL: View paper

Brief Assessment

AutoRAG-HP[29] focuses on hyper-parameter tuning for RAG systems, not on developing novel compression methods. The minimal context provided does not demonstrate prior work on query-dependent online soft compression techniques.

Contribution 2: Two Efficient Compressor Architectures

Description: The authors design two compressor variants: OSCAR-N-Layers uses the first N layers of the pretrained generator backbone, while OSCAR-llama employs a smaller 1B parameter LLM with alignment layers. These architectures enable fast online compression while maintaining generation quality.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A survey on model compression for large language models

URL: View paper

Brief Assessment

Model Compression Survey[14] is a general survey covering quantization, pruning, and knowledge distillation techniques for LLMs. It does not describe specific compressor architectures using pretrained language model layers for document compression in RAG systems.

---

2. Prompt compression for large language models: A survey

URL: View paper

Brief Assessment

Prompt Compression Survey[23] discusses various soft prompt compression architectures but does not specifically propose using the first N layers of a pretrained generator or a smaller 1B parameter LLM with alignment layers as compressor architectures. The survey categorizes existing methods rather than introducing these specific architectural designs.

---

3. Pretraining context compressor for large language models with embedding-based memory

URL: View paper

Brief Assessment

Pretraining Context Compressor[21] uses a decoupled compressor-LLM framework with two components (encoder and converter), not the specific architectures described in OSCAR (OSCAR-N-Layers using first N layers of generator backbone, or OSCAR-llama with alignment layers). The candidate focuses on a different architectural design principle.

---

4. In-context autoencoder for context compression in a large language model

URL: View paper

Prior Art Analysis

In-context Autoencoder[20] demonstrates prior work on efficient compressor architectures for context compression in LLMs. The candidate paper proposes ICAE, which introduces lightweight compression by adding approximately 1% additional parameters to compress contexts. This directly challenges the novelty of OSCAR's claim to be the first to design efficient compressor architectures using pretrained LLM components. Both papers address the same fundamental problem of creating computationally efficient compressors that can operate with minimal additional parameters while maintaining generation quality. The candidate's approach of using a lightweight architecture with minimal parameter overhead (1%) predates OSCAR's variants (OSCAR-N-Layers using first N layers, and OSCAR-llama using 1B parameters).

Evidence

Evidence 1 - **Rationale**: Both papers propose efficient compressor architectures based on LLM components. The candidate's ICAE introduces only 1% additional parameters for compression, while OSCAR proposes using either the first N layers of the pretrained backbone or a 1B parameter LLM. This shows prior work existed on designing lightweight, efficient compressor architectures for LLM-based compression. - **Original**: we propose two different architectures for the compressor backbone: • oscar-n-layers: we construct headless transformers using the first n layers of the pretrained backbone (same architecture as the generator). • oscar-llama: we use a smaller llm, primarily llama-1b2, as our compressor. - **Candidate**: we propose the in-context autoencoder (icae), leveraging the power of a large language model (llm) to compress a long context into short compact memory slots that can be directly conditioned on by the llm for various purposes. experiments demonstrate that our lightweight icae, introducing about 1% a...

Evidence 2 - **Rationale**: Both papers describe training strategies for their compressor architectures. The candidate's ICAE uses pretraining with autoencoding and language modeling objectives, while OSCAR-N-Layers avoids pretraining by using pretrained backbone layers. This demonstrates that efficient compressor architectures with various training approaches were already explored before OSCAR. - **Original**: oscar-n-layers models require no pre-training to align hidden representations with the generator llm. efficiency is controlled by the choice of n. we typically set n to 1/4-1/3 the total number of layers. - **Candidate**: icae is first pretrained using both autoencoding and language modeling objectives on massive text data, enabling it to generate memory slots that accurately and comprehensively represent the original context. then, it is fine-tuned on instruction data for producing desirable responses to various pro...

---

5. Integrating context compression and structural representation in large language models for financial text generation

URL: View paper

Brief Assessment

Financial Text Generation[17] focuses on financial document summarization using context compression mechanisms but does not describe compressor architectures using pretrained language model layers. The candidate employs attention-guided sentence-level scoring and graph-based structural modeling, which differs fundamentally from OSCAR's approach of using the first N layers of a pretrained generator backbone or a smaller LLM with alignment layers for fast online compression in RAG pipelines.

---

6. Adapting language models to compress contexts

URL: View paper

Brief Assessment

Adapting Context Compression[16] focuses on autocompressors that compress long contexts into summary vectors using unsupervised training on document segments. The original paper's OSCAR architectures (N-Layers using first N layers of generator, OSCAR-llama with 1B LLM plus alignment layers) are designed specifically for query-dependent online compression in RAG pipelines with supervised distillation training, representing a different architectural approach and use case.

---

7. Language modeling is compression

URL: View paper

Brief Assessment

Language Modeling Compression[15] focuses on using large language models as general-purpose compressors for various data types (text, images, audio), not on designing efficient compressor architectures specifically for RAG pipelines with query-dependent compression.

---

8. Lossless data compression by large models

URL: View paper

Brief Assessment

Lossless Compression[22] focuses on general data compression using large models for text, images, video and audio. It does not describe compressor architectures using pretrained language model layers specifically for RAG document compression as OSCAR does.

---

9. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding

URL: View paper

Brief Assessment

mplug-docowl2[19] focuses on document image compression using vision encoders and cross-attention mechanisms for OCR-free understanding, not on general RL framework compression using pretrained LLM layers as in OSCAR.

---

10. Extending context window of large language models via semantic compression

URL: View paper

Brief Assessment

Semantic Compression[18] uses a pre-trained summarization model (BART-large-cnn) for compression, not compressor architectures built from pretrained generator layers. The candidate's approach is fundamentally different from OSCAR's layer-based compression variants.

Contribution 3: Simultaneous Compression and Reranking

Description: The authors extend OSCAR to perform both document compression and reranking in a single forward pass by adding a reranking token and training objective. This makes compression essentially free in standard RAG pipelines that already include reranking.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Efficient Dynamic Clustering-Based Document Compression for Retrieval-Augmented-Generation

URL: View paper

Brief Assessment

Dynamic Clustering Compression[30] does not address simultaneous compression and reranking in a single forward pass. The candidate's clustering-based approach is distinct from OSCAR's reranking token mechanism.

---

2. Exit: Context-aware extractive compression for enhancing retrieval-augmented generation

URL: View paper

Brief Assessment

Exit[33] focuses on extractive sentence-level compression for RAG, not simultaneous compression and reranking in a single forward pass. The candidate paper does not discuss reranking capabilities or joint training objectives for both tasks.

---

3. PISCO: Pretty Simple Compression for Retrieval-Augmented Generation

URL: View paper

Brief Assessment

PISCO[1] does not mention or demonstrate any capability for simultaneous document compression and reranking in a single forward pass. This contribution remains unchallenged by the candidate paper.

---

4. Contextual compression in retrieval-augmented generation for large language models: A survey

URL: View paper

Brief Assessment

Contextual Compression Survey[38] is a survey paper that provides an overview of compression paradigms in RAG but does not present a specific technical method for simultaneous compression and reranking. The provided context contains only abstract-level descriptions without technical details about any specific compression-reranking approach.

---

5. Context embeddings for efficient answer generation in retrieval-augmented generation

URL: View paper

Brief Assessment

Context Embeddings[35] focuses solely on context compression for RAG without any reranking capability. The paper does not mention or implement simultaneous reranking alongside compression.

---

6. FACTS About Building Retrieval Augmented Generation-based Chatbots

URL: View paper

Brief Assessment

FACTS[37] focuses on building enterprise RAG chatbots with emphasis on pipeline engineering, security, and cost considerations. It does not address simultaneous document compression and reranking in a single forward pass as a technical contribution.

---

7. Searching for best practices in retrieval-augmented generation

URL: View paper

Brief Assessment

RAG Best Practices[24] evaluates reranking as a separate module using existing methods (MonoT5, MonoBERT, RankLLaMA, TildeV2) but does not propose simultaneous compression and reranking in a single forward pass. The paper treats these as distinct sequential steps in the RAG workflow.

---

8. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation

URL: View paper

Brief Assessment

RECOMP[36] focuses on document compression for retrieval-augmented generation but does not perform simultaneous reranking. The paper describes extractive and abstractive compression methods trained to optimize end-task performance, but reranking is not mentioned as a joint capability in a single forward pass.

---

9. OSCAR: Online Soft Compression And Reranking

URL: View paper

Prior Art Analysis

OSCAR Reranking[4] demonstrates that the concept of simultaneous compression and reranking was already implemented in their work. Both papers describe adding a reranking token to enable joint compression and reranking in a single forward pass, using the same architectural approach and training methodology. The candidate paper explicitly states they 'extend oscar to simultaneously perform reranking' and use 'a reranking token [rr]' with 'an additional dense layer' for relevance scoring, which directly matches the original

paper's description of adding 'a reranking token and training objective' to perform 'both document compression and reranking in a single forward pass.'

Evidence

Evidence 1 - **Rationale**: Both papers acknowledge that reranking makes compression essentially free in RAG pipelines, and both cite the same prior work (Chirkova et al.) as inspiration for this approach. - **Original**: lastly, we notice, as discussed by chirkova et al. (2025), that the compression operation can be exploited to simultaneously re-rank the initial pool of retrieved documents. since re-ranking is an integral part of efficient rag pipelines (rau et al., 2024a), this enables us to obtain the compression... - **Candidate**: building on recent observations by chirkova et al.[4], we equip oscar with document reranking capabilities. since reranking is an integral part of standard rag pipelines, it makes the compression almost free.

Evidence 2 - **Rationale**: These passages are nearly identical, describing the exact same architectural approach: adding a reranking token [rr] and using a dense layer to map hidden states to relevance scores. The candidate paper uses the same methodology and even similar wording. - **Original**: simultaneous reranking building on insights from chirkova et al. (2025), query-dependent online context compression closely resembles document reranking. rerankers, such as cross-encoders (nogueira & cho, 2019), refine the ranking from the initial retrieval step. unlike retrieval models, which encod... - **Candidate**: simultaneous reranking building on the insights from chirkova et al.[4], query-dependent online context compression shares significant similarities with the document reranking task. rerankers, such as crossencoders[32],refinetherankingproducedbytheinitial retrieval step. unlike retrieval models, whi...

Evidence 3 - **Rationale**: Both papers use pointwise distillation from a reference reranker to train the reranking component, demonstrating the same training methodology. - **Original**: we train this added layer with a point-wise distillation objective from a reference reranker: we add λ pk i=1(ri - r′ i)2 to equation 1, where λ balances generation and reranking and r′ i are scores from a reference reranker. - **Candidate**: to train the reranking component, we simply rely on a pointwise distillation of reranking scores of an effective cross-encoder. distillation is now a standard technique to train ranking models [8,15,25,39], and was already shown to work in the context of prompt compression in provence [4].

---

10. xrag: Extreme context compression for retrieval-augmented generation with one token

URL: View paper

Brief Assessment

xRAG[34] focuses on modality fusion by projecting document embeddings into LLM representation space, achieving extreme compression to a single token. It does not perform simultaneous compression and reranking in a single forward pass as described in the original contribution.

Appendix: Text Similarity Detection

Textual similarity detection checked 27 papers and found 5 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. OSCAR: Online Soft Compression And Reranking

Detected in: Core Task (sibling), Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] OSCAR: Online Soft Compression for RAG View paper
- [1] PISCO: Pretty Simple Compression for Retrieval-Augmented Generation View paper
- [2] SARA: Selective and Adaptive Retrieval-augmented Generation with Context Compression View paper
- [3] Dynamickv: Task-aware adaptive kv cache compression for long context llms View paper
- [4] OSCAR: Online Soft Compression And Reranking View paper
- [5] Towards Explainable RAG: Interpreting the Influence of Retrieved Passages on Generation View paper
- [6] Simple Context Compression: Mean-Pooling and Multi-Ratio Training View paper
- [7] Map-vla: Memory-augmented prompting for vision-language-action model in robotic manipulation View paper
- [8] Robustness of Fine-Tuned Llms Under Noisy Retrieval Inputs View paper
- [9] SPARC: Soft Probabilistic Adaptive multi-interest Retrieval Model via Codebooks for recommender system View paper
- [10] Hierarchical Patch Compression for ColPali: Efficient Multi-Vector Document Retrieval with Dynamic Pruning and Quantization View paper
- [11] Enhancing Cache-Augmented Generation (CAG) with Adaptive Contextual Compression for Scalable Knowledge Integration View paper
- [12] An adaptive query-routing framework for optimizing small languages models in resource-constrained environments View paper
- [13] Semantic Proximity for Redundancy-Aware Context Compression in Large Language Models View paper
- [14] A survey on model compression for large language models View paper
- [15] Language modeling is compression View paper
- [16] Adapting language models to compress contexts View paper
- [17] Integrating context compression and structural representation in large language models for financial text generation View paper
- [18] Extending context window of large language models via semantic compression View paper
- [19] mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding View paper
- [20] In-context autoencoder for context compression in a large language model View paper
- [21] Pretraining context compressor for large language models with embedding-based memory View paper
- [22] Lossless data compression by large models View paper
- [23] Prompt compression for large language models: A survey View paper
- [24] Searching for best practices in retrieval-augmented generation View paper

- [25] Oreo: A plug-in context reconstructor to enhance retrieval-augmented generation View paper
- [26] AttentionRAG: Attention-Guided Context Pruning in Retrieval-Augmented Generation View paper
- [27] Neural-Symbolic Dual-Indexing Architectures for Scalable Retrieval-Augmented Generation View paper
- [28] Query-Aware Graph Neural Networks for Enhanced Retrieval-Augmented Generation View paper
- [29] Autorag-hp: Automatic online hyper-parameter tuning for retrieval-augmented generation View paper
- [30] Efficient Dynamic Clustering-Based Document Compression for Retrieval-Augmented-Generation View paper
- [31] ACoRN: Noise-Robust Abstractive Compression in Retrieval-Augmented Language Models View paper
- [32] Familiarity-Aware Evidence Compression for Retrieval-Augmented Generation View paper
- [33] Exit: Context-aware extractive compression for enhancing retrieval-augmented generation View paper
- [34] xrag: Extreme context compression for retrieval-augmented generation with one token View paper
- [35] Context embeddings for efficient answer generation in retrieval-augmented generation View paper
- [36] RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation View paper
- [37] FACTS About Building Retrieval Augmented Generation-based Chatbots View paper
- [38] Contextual compression in retrieval-augmented generation for large language models: A survey View paper