SARA: Selective and Adaptive Retrieval-augmented Generation with Context Compression

Yiqiao Jin [1] **Kartik Sharma** [1] **Vineeth Rakesh** [2] **Yingtong Dou** [2] **Menghai Pan** [2] **Mahashweta Das** [2]
Srijan Kumar [1]

Abstract

Retrieval-augmented Generation (RAG) extends large language models (LLMs) with external knowledge but faces key challenges: restricted *effective* context length and redundancy in retrieved documents. Pure compression-based approaches reduce input size but often discard fine-grained details essential for factual accuracy. We propose SARA, a unified RAG framework that balances *local precision* and *global knowledge coverage* under tight context budgets. SARA combines natural-language text snippets with semantic compression vectors to jointly enhance context efficiency and answer correctness. It represents contexts at two complementary levels: 1) *fine-grained* natural-language spans that preserve critical entities and numerical values, and 2) *compact*, *interpretable* vectors that summarize high-level semantics. An iterative evidence-selection module employs the compression vectors for dynamic reranking of contexts. Across 9 datasets and 5 open-source LLMs spanning 3 model families (Mistral, Llama, and Gemma), SARA consistently improves answer relevance (+17.71), answer correctness (+13.72), and semantic similarity (+15.53), demonstrating the importance of integrating textual and compressed representations for robust, context-efficient RAG.

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various natural language understanding and generation tasks (Xiao et al., 2024; Zhao et al., 2024). Meanwhile, as LLMs are parametric in nature, their knowledge is inherently constrained by the scope, domain, and recency of their training data (Liu et al., 2025). Retrieval-augmented generation (RAG) (Lewis et al., 2020) addresses this by retrieving from external non-parametric knowledge sources, essential for knowledge-intensive tasks.

Challenges. Despite its promise, RAG still faces key challenges in effectively retrieving, selecting, and integrating external evidence. *1) Limited Effective Context.* While some LLMs support long inputs, their attention is biased toward earlier tokens (Li et al., 2024b), making them sensitive to input order and prone to overlooking important information near the end of the input (Yu et al., 2024). Extending usable context often requires costly, model-specific architectural changes (Ding et al., 2023). *2) Context Redundancy.* Retrieved documents often include *redundant* or loosely structured content (e.g. transcripts or news articles) (Yu et al., 2024; Ge et al., 2024). Without careful post-processing, duplicate or irrelevant content inflates token usage, distracts the model, degrades answer quality or even leads to hallucinations. *3) Compression-fidelity Trade-off.* Existing context compression techniques reduces input length but often sacrifice fine-grained details (e.g. numeric values, organization names, and geographical locations), leading to hallucinated or incomplete responses. While existing methods achieve high compression rates, aggressive compression process risk discarding critical information essential for factual accuracy.

This Work. We present SARA, a unified RAG framework that improves both *retrieval* and *generation* stages through structured evidence compression and adaptive selection. From the *generation* perspective, SARA represents long contexts using a small number of semantically rich, self-contained *compression vectors*, which act as lightweight abstractive summaries that preserve essential information while significantly reducing input length. Specifically, we leverage state-of-the-art embedding models (Meng et al., 2024; Muennighoff et al., 2023) to encode retrieved documents into multiple, semantically rich compression vectors. These vectors are also *explainable* and can be interpreted through auto-encoding to reveal their underlying semantics. From the *retrieval* perspective, SARA introduces an *iterative evidence selection mechanism* that leverages the com-

*Equal contribution [1]Georgia Institute of Technology, GA, United States [2]Visa Research, CA, United States. Correspondence to: Yiqiao Jin <yjin328@gatech.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

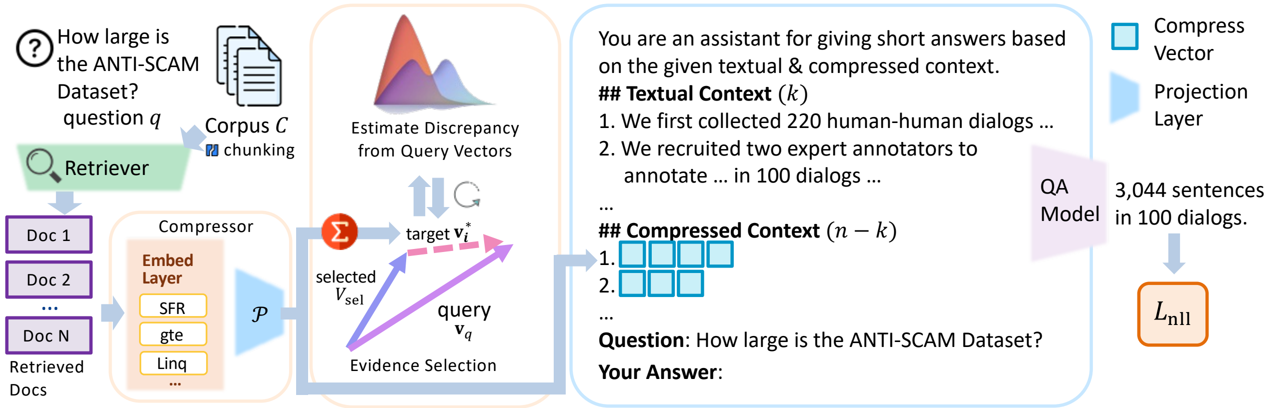

Figure 1: SARA reasons over a mixture of compressed evidence and natural language contexts to balance local precision and global coverage when generating responses. An iterative evidence reranking step selects contexts for relevance and diversity. The retriever, compressor, and QA model uses a variety of embedding models.

pression vectors to dynamically refine the set of top-ranked documents. SARA progressively selects contexts based on the knowledge required to properly address the query and knowledge coverage of existing contexts, minimizing redundancy while maximizing informativeness. SARA is agnostic to the choice of embedding models, open-source LLMs, and retrievers. Our contributions are as follows:

- We propose SARA, a novel RAG framework for long-context tasks. SARA introduces a **hybrid compression strategy**, balancing *local precision* using natural language spans and *global abstraction* via compression vectors, enabling fine-grained reasoning and holistic understanding within strict context budgets.
- We propose an **iterative context refinement** mechanism based on the compression vectors to dynamically optimize the retrieved context by reducing redundancy and prioritizing query-relevant content.
- Comprehensive experiments on 5 LLMs spanning 3 model families, including Mistral-7B, MistralNemo-12B, MistralSmall-24B, Llama-3.1-8B, and Gemma3-4B, demonstrate that SARA consistently improves performance and generalizes well across LLMs (Section B.2), retrievers (Section B.3), and embedding models (Appendix B.7).

2. Method

2.1. Problem Formulation

A retrieval-augmented generation (RAG) pipeline consists of a *retriever* that fetches relevant evidence from a large-scale corpus based on the input query and a *generator* that synthesizes the evidence to answer the query. Given a query q and corpus C, the retriever $\mathcal{R}(\cdot)$ selects the top-n relevant contexts $\mathcal{V}_{\text{sel}} \subseteq C$. To improve effectiveness, RAG may incorporate a *reranking* step to reorder the input documents, prioritizing the most relevant ones for answer generation.

2.2. Overview

LLMs have limited effective context windows, and performance degrades when key information is buried in long inputs (Jiang et al., 2024). SARA mitigates this by compressing long context into compact vectors while selectively retaining essential evidence in natural language, preserving model capacity for the most relevant content.

SARA follows a two-stage training procedure: During **Compression Learning**, SARA learns to reconstruct original context from compression vectors. In **Instruction-tuning**, SARA is adapted to rerank the evidence using the compression vectors and reason over mixed inputs–combining natural language and compressed evidence. Our method is *model-agnostic*, compatible with any retrievers, embedding models, and open-source LLMs. A lightweight *projection layer* aligns the embedding space with the LLM space, requiring no significant changes to internal components like the attention mechanism, enabling seamless integration with future embedding models and LLMs. Sample prompts for all stages are provided in Table 8.

2.3. Compression Learning

An effective compression mechanism should meet three core principles: 1) *Semantic Fidelity*–preserving sufficient information for accurate context reconstruction; 2) *Token Compatibility*–producing compression vectors interpretable by LLMs via prompting; and 3) *Scalability*–requiring minimal adaptation across retrievers and LLMs.

To meet these goals, SARA leverages sentence embeddings (Reimers & Gurevych, 2019) aligned with the LLM's token space, enabling compact and interpretable representations that support reasoning under tight context budgets.

Embedding Alignment. SARA encodes each text chunk into a compression vector that fits within a single token's embedding space (Figure 3). A lightweight compressor–combining a sentence embedding model and an MLP–is trained via an autoencoding task (Liu et al., 2023b; Cheng et al., 2024) to align sentence embeddings with the LLM's token space:

$$\mathcal{L}_{\text{align}}(s_i) = -\log P_\theta\left(s_i \mid \text{Enc}(s_i), x_{\text{ins}}\right). \quad (1)$$

Here, s_i is a text chunk, $\text{Enc}(\cdot)$ is the compressor, θ is the model's parameter, and x_{ins} is the decoding instruction such as "The token <C> can be interpreted as: `[CHUNK]`." As one compression vector has limited representation capacity, we segment each document into chunks, and encode each chunk as a separate compression vector. We adopt a *curriculum learning* strategy (Bengio et al., 2009; Wang et al., 2021) to improve training stability (Appendix A.3).

Context Reconstruction. After learning to decode individual compression vectors, we extend the model to full context reconstruction:

$$\mathcal{L}_{\text{recons}}(c) = -\log P_\theta(c \mid \{\text{Enc}(s_i), \forall s_i \in c\}, x_{\text{ins}}). \quad (2)$$

Here, c is a document composed of multiple chunks $\{s_i\}$, each encoded as a separate vector. Unlike traditional extractive or abstractive summarization methods (Xu et al., 2024) that require multiple passes, these vectors naturally serve as high-ratio, parallelizable summaries.

Training Corpus Selection. Since the goal is to align the embedding spaces, the pretraining corpus is domain-agnostic and can be drawn from any natural language dataset. We use the Wikipedia dataset (Izacard et al., 2023), which provides broad topical diversity and diverse narrative styles, and has proven effective for language model pretraining (Gao et al., 2023). In Section B.6 and Tables 11/9, we demonstrate that these compression vectors are able to encode detailed information, such as exact organization names, academic terms, and numeric values.

2.4. Instruction-tuning and Inference

Simple 'retrieve-and-read' pipelines often implies redundant evidence and overlook interdependencies between previously retrieved and newly needed information (Wang et al., 2024). In long-context understanding, *what* should be retrieved next hinges on *what* has already been inferred from previously retrieved evidence (Sarthi et al., 2024; Li et al., 2024a). To address this, SARA leverages a 2-stage context refinement, which interleaves *retrieval* and *reasoning*: 1) a *coarse* retrieval step eliminating irrelevant documents while maintaining computational efficiency; 2) a *fine-grained* reranking step that iteratively refines contexts for informativeness, relevance, and diversity.

Instruction-tuning. Initially, SARA is instruct-tuned to holistically reason over both formats–the top-k passages are input as natural text, while the remaining are passed as compression vectors (Figure 1). For faster training, we instruct-tune the LLM generator on downstream tasks with LoRA (Hu et al., 2021) using top-n contexts retrieved via BM25 (Robertson et al., 2004).

Dynamic Evidence Reranking. Effective RAG requires balancing *relevance*–which ensures alignment with the user query–and *novelty*–which introduces new information beyond existing evidence. To achieve this, we adopt an iterative evidence selection method (Algorithm 1) that dynamically selects context based on its incremental value to model understanding.

Embedding-based Novelty ranks candidates based on their contribution to the model's discrepancy in knowledge, selecting the vector that minimizes the discrepancy between the selected set $\mathcal{V}_{\text{sel}}$ with the query representation $\mathbf{v}_q$ in the embedding space:

$$\text{SelectEvi}(q, \mathcal{V}_{\text{sel}}, \mathcal{V}) \quad (3)$$

$$= \underset{v_i \in \mathcal{V} \setminus \mathcal{V}_{\text{sel}}}{\operatorname{argmin}} \left\| \mathbf{v}_q - \text{Aggregate}\left(\{\text{Enc}(v) \mid v \in \mathcal{V}_{\text{sel}} \cup \{v_i\}\}\right) \right\|_2 . \quad (4)$$

Since the user query is usually succinct, we supplement the query representation $\mathbf{v}_q$ by aggregating the embeddings of both the question and the top-1 retrieved context: $\mathbf{v}_q = \text{Avg}(\text{Enc}(q), \text{Enc}(v_1))$.

Conditional Self-information (CSI). An alternative is to select evidence based on CSI (Shannon, 1948), which quantifies the surprisal of new evidence given previously selected evidence:

$$\text{SelectEvi}(q, \mathcal{V}_{\text{sel}}, \mathcal{V}) = \underset{v_i \in V \setminus \mathcal{V}_{\text{sel}}}{\operatorname{argmax}} I(v_i | \mathcal{V}_{\text{sel}}) \quad (5)$$

$$I(v_i | \mathcal{V}_{\text{sel}}) = \frac{1}{|v_i|} \sum_{j=1}^{|v_i|} -\log P(w_i^j \mid v_i \in \mathcal{V}_{\text{sel}}, w_i^1, \ldots, w_i^{j-1}) \quad (6)$$

where $I(v_i | \mathcal{V}_{\text{sel}}) = -\log P(v_i | \mathcal{V}_{\text{sel}})$ is the conditional self-information of context v_j given selected contexts $\mathcal{V}_{\text{sel}}$, estimated using a smaller proxy language model. Higher CSI introduces novel information, while lower CSI suggests redundancy with previously selected content. Filtering low-CSI candidates reduces repetition and enhances context *diversity* with minimal impact on overall informativeness.

3. Evaluation

3.1. Overall Performance

Results under Context Constraints. Tables 2 and 1 compare SARA and strong compression-based methods under strict context length constraints (512 and 1024 tokens). SARA consistently outperforms baselines on both lexical (F1, ROUGE-L) and LLM-based evaluation metrics. Under

512 tokens	QASPER		NarrativeQA		TriviaQA		QuALITY		HotpotQA	
	F1	R-L	F1	R-L	F1	R-L	F1	R-L	F1	R-L
ICAE	26.64	23.53	37.58	38.08	53.47	49.16	26.79	28.20	53.18	44.05
LLMLingua	31.29	32.38	50.26	48.58	63.22	58.95	30.53	31.48	57.36	49.30
LongLLMLingua	29.49	28.31	41.90	39.27	66.28	62.76	36.13	38.03	64.34	60.32
SARA	**36.23**	**39.17**	**55.64**	**54.90**	**82.50**	**81.74**	**42.27**	**43.62**	**83.03**	**75.56**
Impr. %	15.8	21.0	10.7	13.0	24.5	30.2	17.0	14.7	29.0	25.3
1024 tokens	QASPER		NarrativeQA		TriviaQA		QuALITY		HotpotQA	
	F1	R-L	F1	R-L	F1	R-L	F1	R-L	F1	R-L
ICAE	31.82	33.32	36.70	38.35	51.07	49.78	28.15	29.88	64.51	55.60
LLMLingua	33.18	32.19	50.09	52.46	71.92	67.01	33.82	34.90	62.80	60.71
LongLLMLingua	34.09	33.47	52.48	51.17	72.64	67.47	36.57	33.18	69.21	67.88
SARA	**40.37**	**42.24**	**55.96**	**56.01**	**83.67**	**82.16**	**42.40**	**44.19**	**83.77**	**76.37**
Impr. %	18.4	26.2	6.6	6.8	15.2	21.8	15.9	26.6	21.0	12.5

Table 1: Performance of compression methods under context length constraints (512/1024 tokens) in terms of F1 scores and ROUGE-L (R-L). Improvements over the best models are shown with *Impr.%*.

512 tokens, SARA improves F1 by 19.4% and ROUGE-L by 20.8% on average. We observe that the gains are particularly significant on knowledge-intensive tasks like TriviaQA (+24.5%) and HotpotQA (+29.0%), which require facts and reasoning. Improvements on narrative-style tasks (e.g. NarrativeQA) are more modest, particularly under 1024 tokens (+6.6% F1 and 6.8% ROUGE-L), likely because chunking and compression can change the narrative flow and obscure subtle discourse-level cues. Unlike factoid questions, narrative questions demand holistic coherence that is harder to retain under chunking and summarization (Ge et al., 2024).

Impact of Context Budgets. Increasing the context budget from 512 to 1024 tokens generally improves performance. Baselines that produce natural language compression (e.g., LongLLMLingua) see substantial gains–up to +10.6 F1 on NarrativeQA–as the additional budget reduces the need to truncate or overly compress passages, allowing inputs to better reflect their original structure. SARA outperforms the strongest baseline by 6-12 F1 on knowledge-intensive tasks (e.g. TriviaQA and HotpotQA). As SARA has already captured key content efficiently under a lower context budget using its hybrid compression strategy, it exhibits relatively modest gains on certain datasets (e.g., +4.1 F1 on QASPER).

Balancing Compression Efficiency and Answer Faithfulness. A central challenge in RAG is balancing compression efficiency with faithfulness. Aggressive approaches like xRAG, which compress entire evidence sets into a single dense vector, optimize for efficiency but often at the cost of factuality and hallucination. As shown in Table 4, baselines like xRAG especially struggle on knowledge-intensive tasks, achieving only 43.4 F1 and 35.5 ROUGE-L on TriviaQA, in contrast to SARA's 85.1 F1 and 83.9 ROUGE-L. Qualitative analysis in Table 6 reveals that baselines can hallucinate content, generating answers with fabricated entities or tasks ('sentiment analysis' and 'machine translation') ungrounded in the original documents. Methods that over-compress inputs (e.g. ICAE) risk discarding critical content. As a result, the model tends to become overly conservative-frequently concluding that the answer is not present. These failures underscore the drawbacks of one-shot compression when multiple facts must be retained. In contrast, SARA can accurately recovers fine-grained content, such as specific task names (e.g. NLI, document and intent classification) prompted in the question) with high fidelity, even under tight context budgets. Thus, SARA's hybrid approach preserves salient content, simplifying key information while mitigating factual distortion under tight context budgets.

Comparison with Summarization-based methods SARA consistently outperforms standard RAG and state-of-the-art summarization-based baselines, including Raptor and GraphRAG, despite their use of stronger base models like GPT-4o (OpenAI, 2025) for question-answering and summarization. On HotpotQA, which requires multi-hop reasoning, SARA achieves +15% F1 and +14.6% ROUGE-L. These results highlight the effectiveness of our compression approach in helping the model accommodate and reason over multiple discrete evidence within constrained context.

4. Conclusion

We present SARA, a unified and efficient RAG framework that enhances both retrieval and generation through structured evidence compression and adaptive document selection without significant architectural changes to the LLM. Experiments across multiple LLM backbones, retrievers, and embedding models demonstrate that SARA significantly improves answer correctness and relevance.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2023.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *ACL*, pp. 3119–3137, 2024.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *ICML*, pp. 41–48, 2009.

Cheng, X., Wang, X., Zhang, X., Ge, T., Chen, S.-Q., Wei, F., Zhang, H., and Zhao, D. xrag: Extreme context compression for retrieval-augmented generation with one token. *arXiv:2405.13792*, 2024.

Chevalier, A., Wettig, A., Ajith, A., and Chen, D. Adapting language models to compress contexts. In *EMNLP*, pp. 3829–3846, 2023.

Chirkova, N., Formal, T., Nikoulina, V., and Clinchant, S. Provence: efficient and robust context pruning for retrieval-augmented generation. *arXiv:2501.16214*, 2025.

Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M. A dataset of information-seeking questions and answers anchored in research papers. In *NAACL*, pp. 4599–4610, 2021.

Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., Zheng, N., and Wei, F. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv:2307.02486*, 2023.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., and Larson, J. From local to global: A graph rag approach to query-focused summarization. *arXiv:2404.16130*, 2024.

Es, S., James, J., Anke, L. E., and Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In *EACL*, pp. 150–158, 2024.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. In *EMNLP*, pp. 6465–6488. ACL, 2023.

Ge, T., Jing, H., Wang, L., Wang, X., Chen, S.-Q., and Wei, F. In-context autoencoder for context compression in a large language model. In *ICLR*, 2024.

Ho, X., Nguyen, A.-K. D., Sugawara, S., and Aizawa, A. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*, pp. 6609–6625, 2020.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

Hwang, T., Cho, S., Jeong, S., Song, H., Han, S., and Park, J. C. Exit: Context-aware extractive compression for enhancing retrieval-augmented generation. *arXiv:2412.12559*, 2024.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Atlas: Few-shot learning with retrieval augmented language models. *JMLR*, 24(251):1–43, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Jiang, H., Wu, Q., Lin, C.-Y., Yang, Y., and Qiu, L. Llmlingua: Compressing prompts for accelerated inference of large language models. In *EMNLP*, pp. 13358–13376, 2023b.

Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *ACL*, 2024.

Jin, H., Han, X., Yang, J., Jiang, Z., Liu, Z., Chang, C.-Y., Chen, H., and Hu, X. Llm maybe longlm: Selfextend llm context window without tuning. In *ICML*, pp. 22099–22114, 2024.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pp. 1601–1611, 2017.

Kim, J., Lee, S., Kwon, J., Gu, S., Kim, Y., Cho, M., Sohn, J.-y., and Choi, C. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog, 2024a. URL https://getlinq.com/blog/linq-embed-mistral/.

Kim, J.-H., Yeom, J., Yun, S., and Song, H. O. Compressed context memory for online language model interaction. In *ICLR*, 2024b.

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*, 33:9459–9474, 2020.

Li, C., Liu, Z., Xiao, S., and Shao, Y. Making large language models a better foundation for dense retrieval. *arXiv:2312.15503*, 2023a.

Li, S., He, Y., Guo, H., Bu, X., Bai, G., Liu, J., Liu, J., Qu, X., Li, Y., Ouyang, W., et al. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *EMNLP*, pp. 12758–12786, 2024a.

Li, W., Li, J., Ramos, R., Tang, R., and Elliott, D. Understanding retrieval robustness for retrieval-augmented image captioning. In *ACL*, pp. 9285–9299, 2024b.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv:2308.03281*, 2023b.

Li, Z., Chen, X., and Jeon, Y. Grappi: A retrieve-divide-solve graphrag framework for large-scale protein-protein interaction exploration. In *NAACL*, 2025.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023b.

Liu, J. LlamaIndex. https://github.com/jerryjliu/llama_index, 11 2022. DOI: 10.5281/zenodo.1234.

Liu, S., Jin, Y., Li, C., Wong, D. F., Wen, Q., Sun, L., Chen, H., Xie, X., and Wang, J. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv:2501.01282*, 2025.

Louis, M., Déjean, H., and Clinchant, S. Pisco: Pretty simple compression for retrieval-augmented generation. *arXiv:2501.16075*, 2025a.

Louis, M., Formal, T., Dejean, H., and Clinchant, S. Oscar: Online soft compression and reranking. *arXiv:2504.07109*, 2025b.

Luo, J., Luo, X., Chen, X., Xiao, Z., Ju, W., and Zhang, M. Semi-supervised fine-tuning for large language models. In *NAACL*, pp. 2795–2808, 2025.

Meng, R., Liu, Y., Joty, S. R., Xiong, C., Zhou, Y., and Yavuz, S. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

Mu, J., Li, X., and Goodman, N. Learning to compress prompts with gist tokens. *NeurIPS*, 36:19327–19352, 2023.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. Mteb: Massive text embedding benchmark. In *ACL*, pp. 2014–2037, 2023.

Munkhdalai, T., Faruqui, M., and Gopal, S. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv:2404.07143*, 2024.

OpenAI. Gpt-4o, 2025. URL https://chat.openai.com/.

Pan, Z., Wu, Q., Jiang, H., Xia, M., Luo, X., Zhang, J., Lin, Q., Rühle, V., Yang, Y., Lin, C.-Y., et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *ACL*, pp. 963–981, 2024.

Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., et al. Quality: Question answering with long input texts, yes! In *NAACL*, pp. 5336–5358, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Rajpurkar, P., Jia, R., and Liang, P. Know what you don't know: Unanswerable questions for squad. In *ACL*, pp. 784–789, 2018.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Risch, J., Möller, T., Gutsch, J., and Pietsch, M. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pp. 149–157, 2021.

Robertson, S., Zaragoza, H., and Taylor, M. Simple bm25 extension to multiple weighted fields. In *CIKM*, pp. 42–49, 2004.

Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., and Manning, C. D. Raptor: Recursive abstractive processing for tree-organized retrieval. In *ICLR*, 2024.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Sharma, K., Kumar, P., and Li, Y. Og-rag: Ontology-grounded retrieval-augmented generation for large language models. *arXiv:2412.15235*, 2024.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *NeurIPS*, 33:16857–16867, 2020.

Wang, X., Chen, Y., and Zhu, W. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., et al. Searching for best practices in retrieval-augmented generation. In *EMNLP*, 2024.

Wei, Z., Chen, W.-L., and Meng, Y. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *ICLR*, 2025.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *EMNLP*, pp. 38–45, 2020.

Xiao, Y., Jin, Y., Bai, Y., Wu, Y., Yang, X., Luo, X., Yu, W., Zhao, X., Liu, Y., Chen, H., et al. Large language models can be good privacy protection learners. In *EMNLP*, 2024.

Xu, F., Shi, W., and Choi, E. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. In *ICLR*, 2024.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pp. 2369–2380, 2018.

Yoon, C., Lee, T., Hwang, H., Jeong, M., and Kang, J. Compact: Compressing retrieved documents actively for question answering. In *EMNLP*, 2024.

Yu, Y., Ping, W., Liu, Z., Wang, B., You, J., Zhang, C., Shoeybi, M., and Catanzaro, B. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *NeurIPS*, 2024.

Zhang, D., Li, J., Zeng, Z., and Wang, F. Jasper and stella: distillation of sota embedding models. *arXiv:2412.19048*, 2024a.

Zhang, Q., Zhang, H., Pang, L., Zheng, H., and Zheng, Z. Adacomp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv:2409.01579*, 2024b.

Zhang, Z., Chen, R., Liu, S., Yao, Z., Ruwase, O., Chen, B., Wu, X., and Wang, Z. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. In *NeurIPS*, 2024c.

Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., and Xie, X. Competeai: Understanding the competition behaviors in large language model-based agents. In *ICML*, 2024.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Hassan, A., Celikyilmaz, A., Liu, Y., Qiu, X., et al. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *NAACL*, pp. 5905–5921, 2021.

Zhu, W., Chen, P., Hu, H., Huang, S., Yuan, F., Chen, J., and Birch, A. Generalizing from short to long: Effective data synthesis for long-context instruction tuning. *arXiv:2502.15592*, 2025.

A. Experimental Details

Model	QASPER				QuALITY			
	Rele.	Correct.	Sim.	Faith.	Rele.	Correct.	Sim.	Faith.
ICAE	75.45	24.03	59.48	21.72	63.33	22.18	59.84	31.05
LLMLingua	79.83	23.97	61.08	25.31	85.58	36.06	79.61	41.19
LongLLMLingua	82.77	22.86	62.17	29.77	86.87	38.90	83.09	40.86
SARA	85.35	25.74	63.99	31.95	89.23	49.71	83.51	43.57

Table 2: Evaluation results across QASPER and QuALITY with a context length budget of 512 tokens. We report Response Relevance (Rele.), Answer Correctness (Correct.), Semantic Similarity (Sim.), and Faithfulness (Faith.) in percentages.

Dataset	SQuAD-v2	
Metrics	F-1	R-L
RAG	63.65	51.26
Raptor	70.69	65.28
GraphRAG	74.82	67.36
xRAG	60.19	49.56
InstructRAG	67.21	57.94
ICAE	50.31	40.82
LLMLingua	70.24	65.12
LongLLMLingua	72.57	67.03
SARA	76.55	69.22

Table 3: Performance comparison on SQuAD-v2.

Algorithm 1 Query Expansion and Novelty-Based Evidence Selection.

Input: Corpus $\mathcal{C} = \{v_i\}_{i=1}^{|\mathcal{C}|}$, query q, number of top contexts n, k
Output: Ranked evidence set $\mathcal{V}_{\text{sel}}$

$\mathcal{V} = \text{Retriever}(q, \mathcal{C})$ ▷ Retrieve top n contexts.
$\mathbf{v}_q = \text{Avg}(\text{Enc}(q), \text{Enc}(v_1))$ ▷ Initialize query embedding with top-1 retrieval v_1.
$\mathcal{V}_{\text{sel}} \leftarrow \{v_1\}$ ▷ Initialize the set of selected contexts.
for $j = 2$ to k **do**
  $\hat{\mathbf{v}} = \text{Aggregate}(\text{Enc}(v), v \in \mathcal{V}_{\text{sel}})$ ▷ Aggregate embeddings of $\mathcal{V}_{\text{sel}}$.
  $v_i^* = \text{SelectEvi}(q, \mathcal{V}_{\text{sel}}, \mathcal{V})$ ▷ Evaluate and select context via Eq. 4 or 5.
  $\mathcal{V}_{\text{sel}} \leftarrow \mathcal{V}_{\text{sel}} \cup \{v_i^*\}$ ▷ Update the selected context set.
end for
return $\mathcal{V}_{\text{sel}}$

A.1. Baselines

We compare our methods with 8 baselines spanning 3 categories: 1) *Standard RAG* (Lewis et al., 2020), which directly feed retrieved documents to the input prompt; 2) *Compression-based methods*, which condense input passages before feeding them into the LLM, including LLMLingua (Jiang et al., 2023b), LongLLMLingua (Jiang et al., 2024), ICAE (Ge et al., 2024), and xRAG (Cheng et al., 2024); 3) *Summarization-based methods*, which generate intermediate summaries over retrieved documents to support more focused reasoning, including Raptor (Sarthi et al., 2024), GraphRAG (Edge et al., 2024), and InstructRAG (Wei et al., 2025). For summarization-based approaches such as Raptor and GraphRAG, which rely on community-level summarization and long-context reasoning, we adopt the more powerful GPT-4o (OpenAI, 2025) as the base model, following prior work (Luo et al., 2025; Li et al., 2025), as open-source models struggle with reasoning over long complex inputs.

A.2. Generalizability Experiments.

To demonstrate the modularity and robustness of our approach, we evaluate its generalizability across different retrieval, embedding, and generation components. For the *retrieval* module, we experiment with both sparse and dense retrievers, including BM25 (Robertson et al., 2004), `bge-reranker-v2-m3` (Li et al., 2023a) and `SFR-Embedding` (Meng et al., 2024).

Dataset	QASPER		NarrativeQA		TriviaQA		QuALITY		HotpotQA	
Metrics	F1	R-L	F1	R-L	F1	R-L	F1	R-L	F1	R-L
RAG	22.73	16.71	40.23	40.16	58.43	49.07	31.79	31.63	48.56	40.06
Raptor	31.77	25.26	56.60	56.91	70.51	65.46	34.27	34.49	68.26	63.14
GraphRAG	37.05	36.66	64.93	63.55	77.52	72.35	37.21	38.15	73.23	68.21
xRAG	32.36	33.72	33.43	32.15	43.36	35.52	32.65	33.84	60.19	49.56
InstructRAG	32.83	33.92	41.79	39.85	76.47	72.19	37.98	38.30	66.77	60.18
SARA-CSI	38.83	41.52	**69.46**	**68.02**	**85.08**	83.85	**42.78**	44.18	**84.21**	**78.16**
SARA-EMB	**40.55**	**41.71**	69.15	66.55	84.74	**84.17**	42.59	**44.31**	83.77	76.37
Impr. %	9.4%	13.8%	7.0%	7.0%	9.8%	16.3%	12.6%	15.7%	15.0%	14.6%

Table 4: Performance of SARA, vanilla RAG, and state-of-the-art summarization-based methods.

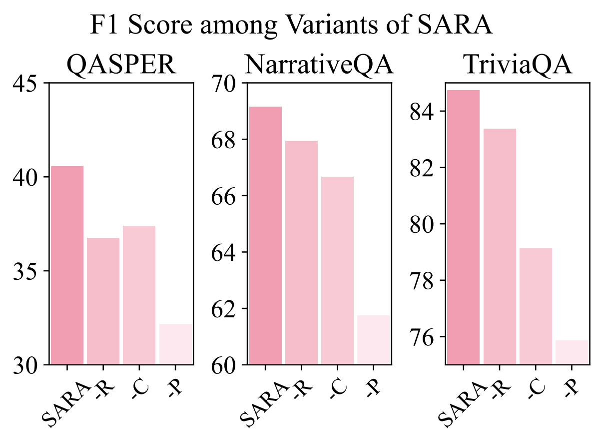

Figure 2: Performance of SARA's variants.

A.3. Implementation Details

Our implementation is based on `PyTorch` (Paszke et al., 2019), `transformers` (Wolf et al., 2020), and `llama-index` (Liu, 2022). All models and data use the `bfloat16` data type. For LoRA setup, we adopt a rank attention dimension of 16, scaling factor $\alpha = 32$, and dropout of 0.1. For chunking, we set the chunk size to 256. The model processes at most $n = 10$ chunks. Our method further selects the top $k = 5$ as natural language evidence, and encode the rest as compression vectors. To reduce the effects of stochasticity, we fix the sampling temperature at 0. Experiments were performed on a Linux server with 6 NVIDIA A100 GPUs.

For embedding alignment (Section 2.3), we adopt a *curriculum learning* strategy, starting with shorter sentences and gradually transition into complex examples. Specifically, we use spaCy[1] for NER and rank sentences by token count and the number of named entities in categories such as `PER`, `ORG`, `LOC`, `GPE`, `Date`, `Time`, and `Event`. The embedding models we experimented with are in Table 3.

[1] `https://spacy.io/`

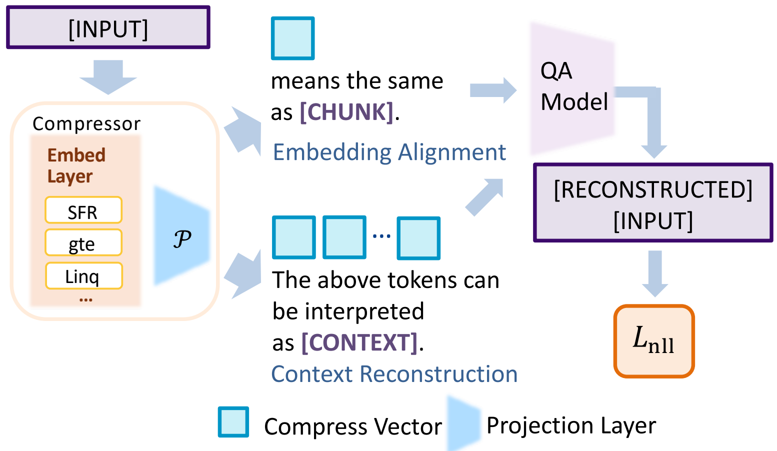

Figure 3: During *Compression Learning*, SARA learns to reconstruct text from compression vectors.

Model	Full Name	Base LLM	Size
SFR (Meng et al., 2024)	`Salesforce/SFR-Embedding-Mistral`	Mistral-7B	4096
Linq (Kim et al., 2024a)	`Linq-AI-Research/Linq-Embed-Mistral`	Mistral-7B	4096
GTE (Li et al., 2023b)	`Alibaba-NLP/gte-Qwen2-7B-instruct`	Qwen2-7B	3584
Stella (Zhang et al., 2024a)	`NovaSearch/stella_en_1.5B_v5`	Qwen2-1.5B	8960

Table 5: Embedding models used in the compressor and their embedding sizes.

A.4. Dataset Descriptions

We evaluate our approach across diverse datasets spanning different domains, input length, and task types: 1) *Short-context question answering*, including SQuAD-v2.0 (Rajpurkar et al., 2018) 2) *Long-context question answering*, which requires responses based on a single long document, including NarrativeQA (Kočiskỳ et al., 2018), QASPER (Dasigi et al., 2021), QuALITY (Pang et al., 2022), and MultifieldQA-en (Bai et al., 2024); 3) *Multi-hop reasoning*, which requires multi-hop inference across documents, including HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), 2WikiMultihopQA (Ho et al., 2020); 4) *Summarization*, including QMSum (Zhong et al., 2021). We use SQuAD-v2, NarrativeQA, QASPER, QuALITY, HotpotQA, and TriviaQA for both training and evaluation. MultifieldQA-en, 2WikiMultihopQA, and QMSum are held out for out-of-domain evaluation only.

- NarrativeQA (Kočiskỳ et al., 2018): question-answering based on books and movie transcripts.
- QASPER (Dasigi et al., 2021): information seeking over scientific research papers with supporting evidence spans.
- QuALITY (Pang et al., 2022): reading-comprehension benchmark with $\sim$ 5000-token passages and unambiguous questions that require consolidating information from multiple text segments.
- TriviaQA (Joshi et al., 2017): trivia questions paired with web evidence (news, encyclopedia, and blogs).
- HotpotQA (Yang et al., 2018): natural questions that require multi-hop reasoning. The questions are annotated with supporting facts.
- SQuAD-v2.0 (Rajpurkar et al., 2018): questions are based on Wikipedia articles, and the answers are text segments from the corresponding reading passage. We select questions that are marked as "answerable"
- QMSum (Zhong et al., 2021): query-focused meeting summarization from dialogue transcripts.
- MultifieldQA-en (Bai et al., 2024) single-doc QA from diverse sources (arXiv, C4, Wikipedia, WuDaoCorpora, etc.)
- 2WikiMultihopQA (Ho et al., 2020): multi-hop QA combining structured and unstructured evidence with reasoning paths.

All corpora are split into 256-token chunks aware of the sentence structures. The token-count distribution is in Figure 4, and the overall statistics is in Figure 10. To improve fine-tuning, we use GPT-4o (OpenAI, 2025) to convert the fine-tuning dataset into instruction-following format, following previous works (Liu et al., 2023a;b).

Question	How do they define rumors?
xRAG	Rumors are false, unverified information that spreads quickly.
LLMLingua	Rumors are not explicitly defined in the context provided.
LongLLMLingua	They define rumors as unverified information spread on social media.
SARA	They define rumors as unverified information that spreads rapidly through a population.
Ground-truth	They define rumors as the presence of rapidly spreading information that is unconfirmed by official media.
Question	What are the five downstream tasks?
xRAG	1. Named Entity Recognition (NER) 2. Dependency Parsing (DP) 3. Sentiment Analysis (SA) 4. Machine Translation (MT) 5. Question Answering (QA)
LLMLingua	The five downstream tasks mentioned in the context are: machine translation, question answering, text summarization, sentiment analysis, and named entity recognition.
LongLLMLingua	The five downstream tasks related to grammatical error correction are text summarization, sentiment analysis, named entity recognition, question answering, and text classification.
SARA	The five downstream tasks are text / intent classification, NLI, named entity recognition, and POS tagging.
Ground-truth	The five downstream tasks we perform using MMTE are three classification tasks: NLI (XNLI dataset), document classification (MLDoc dataset), and intent classification, as well as two sequence tagging tasks: POS tagging and NER.

Table 6: Comparison of answers generated by different compression methods.

A.5. Evaluation Metrics

We adopt standard evaluation protocols consistent with prior work (Asai et al., 2023; Cheng et al., 2024; Sarthi et al., 2024; Edge et al., 2024). For holistic evaluation, we report both traditional *lexical metrics*–including ROUGE (R-L) (Lin, 2004), F1 match scores–and *LLM-based metrics* (Es et al., 2024), including response relevance, answer correctness, semantic similarity, and faithfulness.

Automatic Evaluation. For free-form answer generation, we report ROUGE-L (R-L) (Lin, 2004) and F1 match scores to measure lexical overlap between predicted and ground-truth answers.

LLM-based Evaluation. To complement traditional lexical scores, we adopt four LLM-based metrics that capture orthogonal dimensions essential for reliable RAG deployment (Es et al., 2024; Risch et al., 2021). Each metric returns a value in $[0, 1]$, with higher values indicating better performance.

- **Faithfulness** measures whether the generated answer is grounded in the retrieved context. The answer is decomposed into atomic claims with GPT-4o. Each claim is then tested for entailment against the retrieved context. Answers fully supported by the evidence are favored, and hallucinations are penalized.
- **Answer Relevance** (Response Relevance) judges how directly the answer addresses the user's question. Redundant, off-topic, or missing information lowers the score. It does not take factual accuracy into consideration.
- **Factual Correctness** uses claim decomposition and natural language inference to verify the model's claims against reference texts.
- **Semantic Similarity** uses a cross-encoder to compute the semantic overlap between the generated answer and the ground-truth reference.

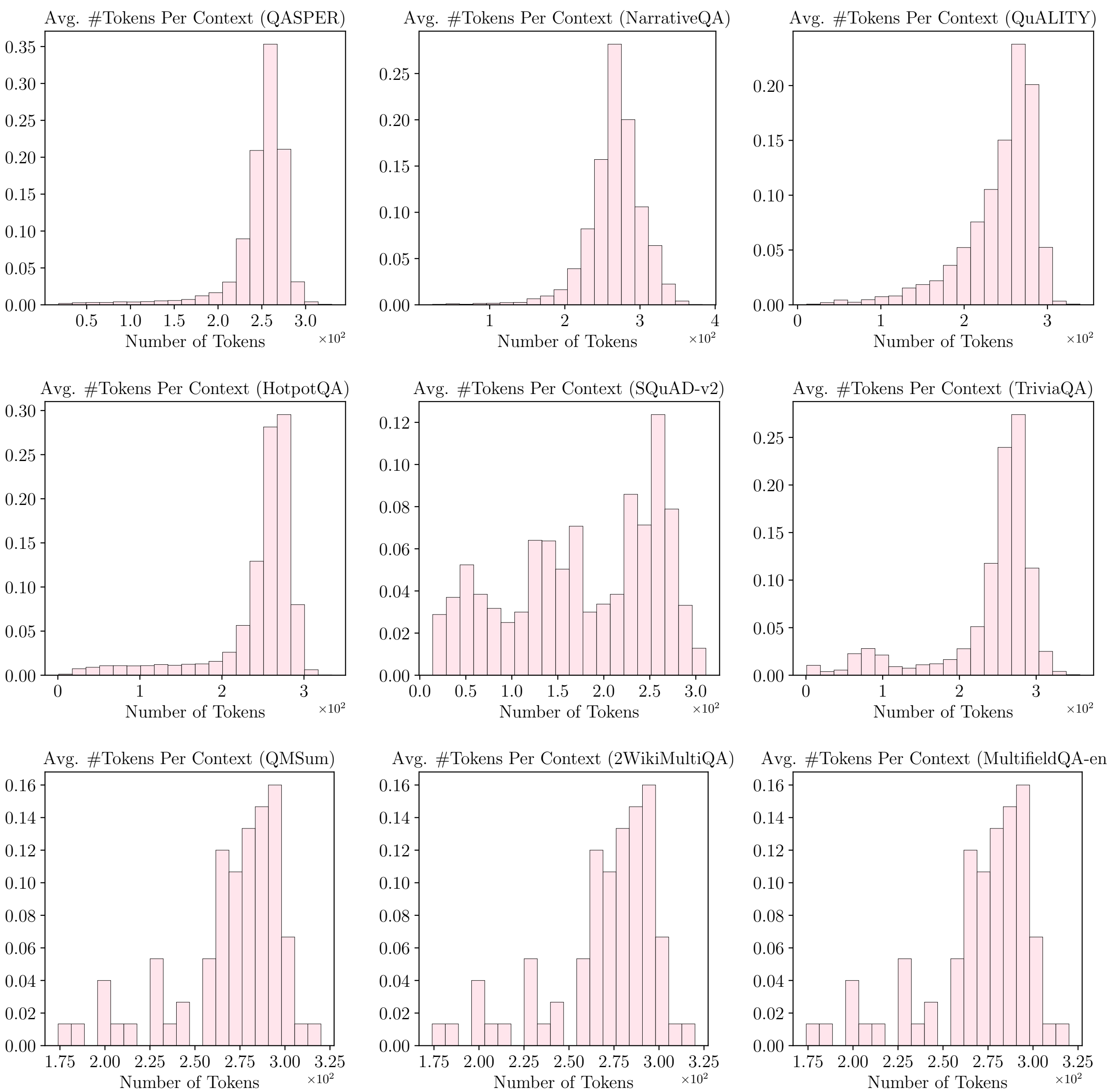

Figure 4: Distribution of number of tokens per chunk in each dataset.

Question	Which NER dataset do they use?
Evidence	• CoNLL2003 is one of the most evaluated English NER datasets, which contains four different named entities: PERSON, LOCATION, ORGANIZATION, and MISC ... • OntoNotes 5.0 is an English NER dataset whose corpus comes from different domains, such as telephone conversation, newswire. We exclude ... • ... OntoNotes 4.0 ... we use the Chinese part. We adopted the same pre-process ... • The corpus of the Chinese NER dataset MSRA came from news domain ... • Weibo NER was built based on text in Chinese social media Sina Weibo, and it contained 4 kinds of entities ... • Resume NER was annotated by ...
Ground-truth	The datasets include CoNLL2003, OntoNotes 5.0, OntoNotes 4.0, the Chinese NER dataset MSRA, Weibo NER, and Resume NER.
Predictions	
0/10	They use the **CoNLL-2003 NER dataset**.
2/8	The NER dataset they use is **CoNLL-2003, OntoNotes-5.0** and data based on Chinese social media.
5/5	The NER datasets used are **CoNLL-2003, OntoNotes-5.0, MSRA, Weibo, and Resume.**

Table 7: Sample responses when using `Llama-3.1-8B-Instruct` as the base model with varying numbers of natural language and compressed contexts. '2/8' means using 2 natural language and 8 compressed context. Exact matches with the ground-truth answer is in **bold** and semantic similar parts are in gray. As the number of natural language contexts increase, the model answers are more detailed.

[Embedding Alignment]
<C> means the same as: <Sentence>

[Context Reconstruction]
Interpret the following tokens as a single document: <C> <C> . . . <C>: <Paragraph>

[Instruction-tuning / Inference]
Using the context and additional context, answer the following question: *<question>*
Context: *<context>*
Additional Context:
1. <C>, <C>, . . . , <C>;
2. <C>, <C>, . . . , <C>;
Question: *<Question>*
Your Answer: *<Answer>*
Judgment:

Table 8: Prompt for pretraining, instruction-tuning, and inference. <C> indicate positions for the compression vectors

B. Additional Experiments

B.1. Performance on Short-context QA

SQuAD-v2 presents minimal challenges in context length, as each query is paired with a single passage that fits within the model's input window in most cases. Accordingly, the performance gap across models narrows. SARA achieves the highest results (76.55 F1, 69.22 ROUGE-L; Table 3), outperforming the strongest baseline by a modest margin (+3.98 F1, +2.19 ROUGE-L). In contrast, aggressively compressed systems such as xRAG and ICAE perform significantly worse ($\leq$ 60.19 F1), likely due to summaries that obscures critical details–such as entity names, numeric values, and specific events–reducing accuracy even when full text fits into the model.

B.2. Generalization across LLM Architectures & Sizes.

Beyond Mistral-7B, we evaluate SARA on 4 additional models from 3 families–Mistral, Llama, and Gemma–spanning various sizes and architectures: MistralNemo-12B, MistralSmall-24B, Llama3.1-8B, and Gemma3-4B. As shown in Figures 6 and 5, SARA consistently outperforms the baseline, with up to +40 in Answer Relevance, +14 in Answer Correctness, and +21 in Semantic Similarity. Improvements are particularly pronounced on smaller models. On Mistral-7B, SARA boosts answer relevance by 17.71, answer correctness by 13.72, and semantic similarity by 15.53. These results highlight the method's ability to optimize context usage under tighter context budgets, making it especially effective for smaller models. In some cases, SARA enables a 7B model to match or surpass much larger ones (e.g., MistralSmall-24B), highlighting that reasoning over mixed-format contexts can close the performance gap without increasing model sizes.

In general, performance gains are more significant when the compressor and LLM share the same architecture (e.g. Mistral). Among the Mistral family, we observe an average boost in Answer Relevance of 20.12 and Answer Correctness of 7.07. MistralNemo and MistralSmall achieve improvements in response relevance of +19.65 and +23.01, and semantic similarity of +20.44 and +14.38, respectively. This suggests that architectural alignment between the compressors and LLMs enhances semantic compatibility between compressed inputs and answer generation. In contrast, Gemma-3 shows modest gains (e.g. +6.83 in answer relevance and +5.82 in answer correctness), likely due to its architectural mismatch.

Note that SARA does not aim to directly enhance the QA model's intrinsic generation capability. Instead, its strength lies in refining and reorganizing retrieved contexts to support finer-grained reasoning. Since both SARA and RAG leverage the same initial retriever, they operate over comparable evidence. As a result, faithfulness–the factual consistency with the retrieved context–shows modest improvements.

B.3. Generalization Across Retrievers

We evaluate SARA with dense retrievers like multi-qa-mpnet-base-cos-v1 (Song et al., 2020) and SFR (Meng et al., 2024) in addition to BM25 (Robertson et al., 2004). As shown in Table 12, SARA performs consistently across retrievers, confirming its model-agnostic design. Dense retrievers, especially SFR, yield stronger results–achieving +19 F1

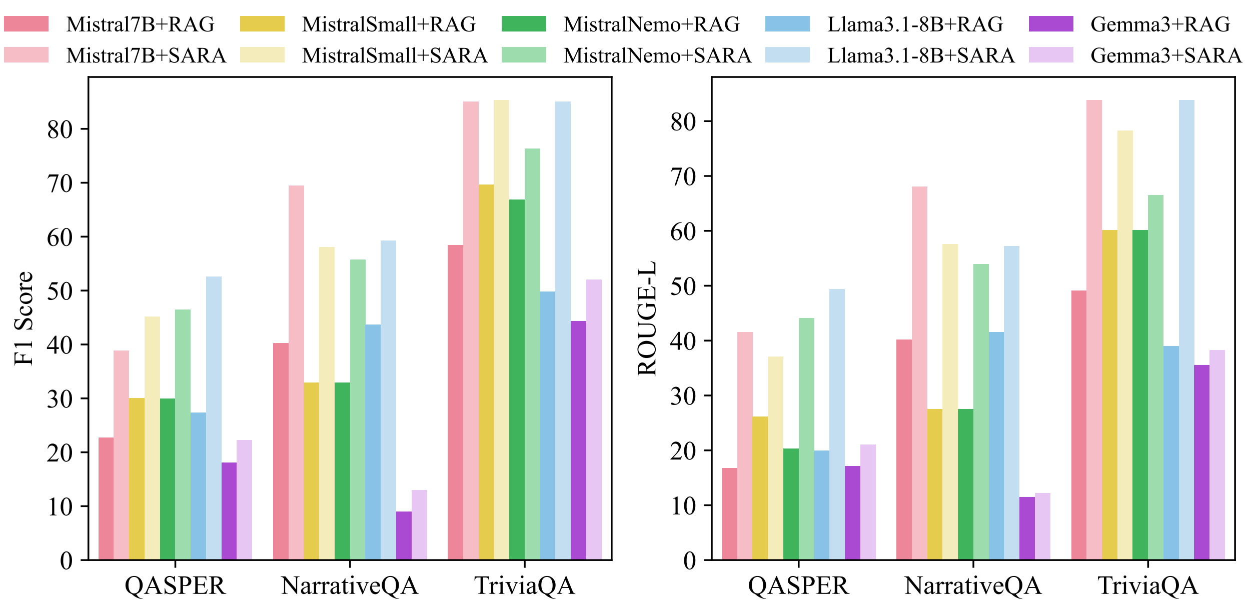

Figure 5: Generalizability across models. We report lexical metrics (F1 score and ROUGE-L) on QASPER (Dasigi et al., 2021) before and after applying SARA.

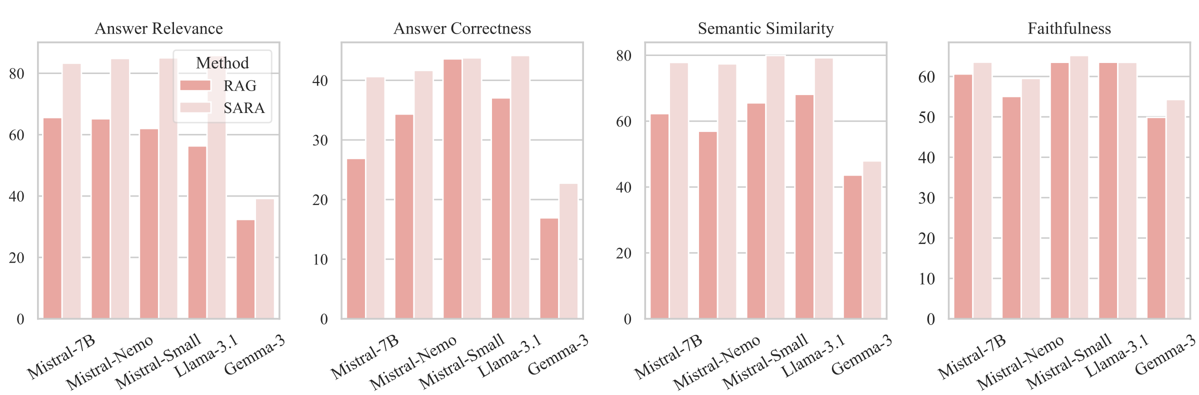

Figure 6: Performance of RAG and SARA across LLMs in terms of LLM-based metrics on QASPER (Dasigi et al., 2021).

over BM25 on QASPER–highlighting the value of semantically richer base retrievers for complex, multi-hop QA. Overall, SARA remains robust to retriever choice while benefiting from higher-quality evidence.

B.4. Ablation Studies

To quantify the contribution of each major component–compression, reconstruction, and reranking–we evaluate 3 variants of SARA. **SARA-C** removes the **C**ompression vectors and only process contexts in natural language formats. **SARA-P** removes the context reconstruction objective during training (Section 2.3). **SARA-R** skips the adaptive reranking stage, relying solely on initial BM25 retrieval (Section 2.4).

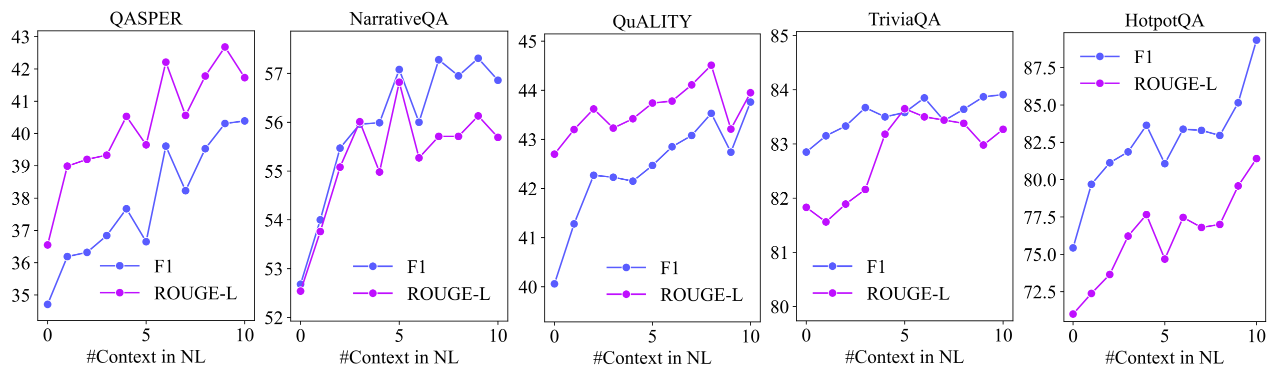

Figure 7: Sensitivity analysis with total contexts fixed at $N = 10$, varying the number of natural language contexts k. Performance improves as k increases, peaking around k = 7-8, and slightly declines beyond 8. SARA achieves strong performance by optimally balancing natural language and compressed contexts, effectively minimizing token overhead without sacrificing accuracy.

Context Reconstruction is Critical. Removing the reconstruction objective (SARA-P) results in the most substantial performance drop (Figure 2)–7-9 F1 across all datasets. This confirms that learning to reconstruct full contexts from compressed vectors is essential for preserving semantic and leveraging these vectors for accurate answer generation.

Compression Enhances Robustness. Disabling compression (SARA-C) also leads to consistent performance declines, especially on TriviaQA (-5.6 F1) where the long-form contexts are potentially noisy or irrelevant. Compression helps filter salient content and suppress redundancy, enhancing answer correctness.

Reranking offers Measurable Gains. Removing reranking (SARA-R) yields modest but consistent drops, confirming that compression-aware reranking improves evidence selection beyond lexical similarity–especially when initial retrieval are suboptimal–at minimal computational cost.

B.5. Sensitivity Analysis

We evaluate SARA's ability to leverage compressed context by fixing the total number of retrieved contexts ($N = 10$) and varying k, the number of top-ranked passages retained in natural language. As shown in Figure 7, performance remains strong even with minimal natural language input (e.g., $k = 1$, F1$= 38.54$, ROUGE-L$= 39.89$), indicating that compression vectors retain essential information. Performance improves with larger k but plateaus around $k = 8$ (F1$= 41.6$, ROUGE-L$= 43.12$), and slightly drops at $k = 9$, suggesting diminishing returns or noise from excessive natural language content. These results highlight the effectiveness of our hybrid strategy in balancing context utility, informativeness, and efficiency.

To further illustrate such effects, Table 7 shows how increasing k within a specific range improves factual specificity. With only compressed context (0/10), the model is able to identify a single entity name (CoNLL-2003), whereas increasing $k = 5$ enables the model to produce answers with high fidelity. Our hybrid approach allows for such precision without overwhelming the context budget.

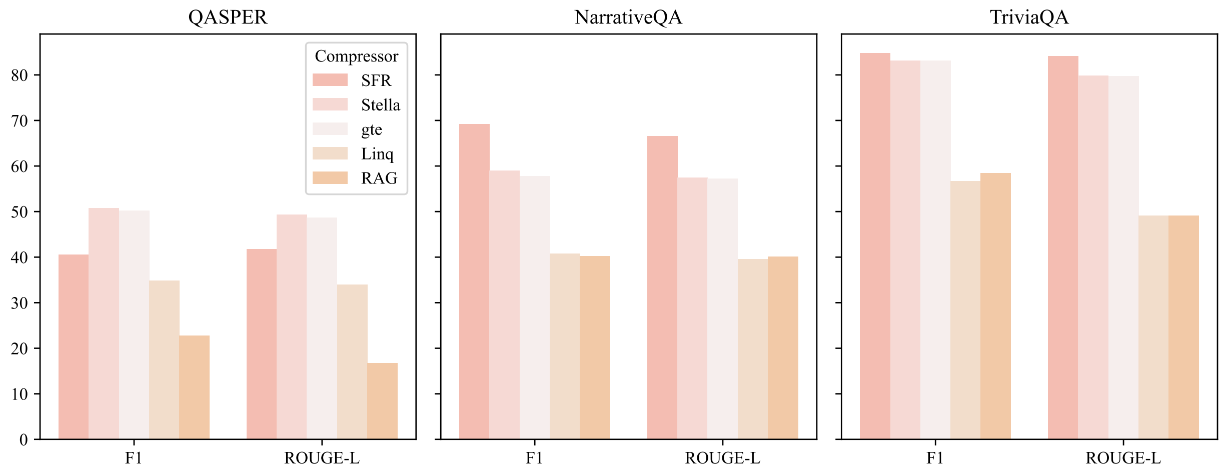

Figure 8: Results on different compressors.

B.6. Intrinsic Analysis of Compression Vectors

B.7. Generalization on Additional Embedding Models

Aside from `Salesforce/SFR-Embedding-Mistral` (SFR), we experimented with additional embeddings, including `Linq-AI-Research/Linq-Embed-Mistral` (Linq) embedding (Kim et al., 2024a), `Alibaba-NLP/gte-Qwen2-7B-instruct` (GTE) (Li et al., 2023b), and `NovaSearch/stella_en_1.5B_v5` (Stella). The profiles of base sentence embedding models are shown in Table 5. Results are shown in Figure 8.

B.8. Generalization on Unseen Datasets

We evaluate generalization by testing the fine-tuned models on three out-of-domain (OOD) datasets from LongBench (Bai et al., 2024): MultiFieldQA-en, 2WikiMultihopQA, and QMSum, which differ substantially in domain and task format from the training data (See Appendix A.4 for details). As shown in Table 13, SARA consistently improves performance across all benchmarks. It boosts RESPONSE RELEVANCE by wide margins—+18.5 on QMSum, +47.7 on MultifieldQA-en, and +55.0 on 2WikiMultiHopQA. These gains highlight the strength of combining natural language spans with compression vectors, which helps leverage more relevant evidence despite domain shifts. The improvements are especially pronounced on QA-style tasks, suggesting that the QA data in the fine-tuning dataset contributes to SARA's performance on other QA datasets. Improved relevance also leads to cleaner answers, hallucinations and off-topic content, leading to cleaner answers.

In contrast, Answer Correctness rises more modestly (+0.3 to +2.2), suggesting that while retrieval quality generalizes well, reasoning over the retrieved content might be partially domain-dependent. For example, TriviaQA and QASPER (used in training) are based on Wikipedia and academic literature, respectively. MultiFieldQA-en involves answering questions based on articles from multiple domains. In this case, in-domain adaptation or instruction tuning could help further improve this performance.

C. Related Work

C.1. Retrieval-augmented Generation (RAG)

Retrieval-augmented Generation has become a standard practice for knowledge-intensive tasks. Instead of treating LLMs as knowledge repositories, RAG generates answers using an external knowledge base (Lewis et al., 2020; Sharma et al., 2024). This approach helps them address model knowledge cutoffs and insufficient training coverage. As a common challenge for RAG models, LLMs struggle to process long, chunked retrieved contexts effectively, even with extended context windows (Yu et al., 2024). Recent work such as Raptor (Sarthi et al., 2024), GraphRAG (Edge et al., 2024) and GraphReader (Li et al., 2024a) focus on improving the *retrieval* and *augmentation* stages by structuring retrieved content, enhancing RAG through semantic or graph-based organization of knowledge, leading to more relevant and compact inputs for generation.

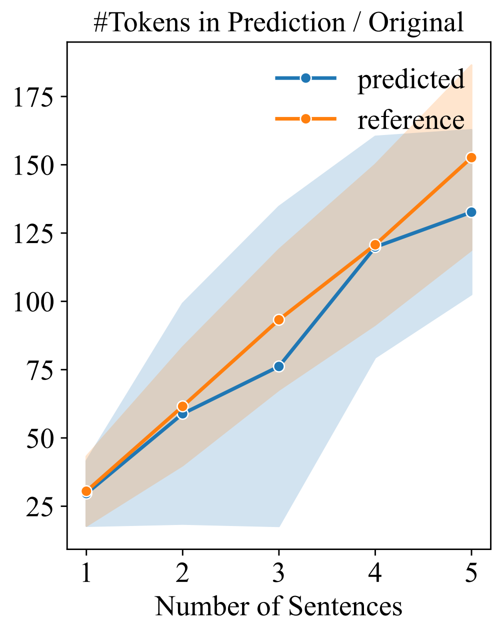

Figure 9: Number of words generated from compression vectors when we vary from 1 to 3 sentences.

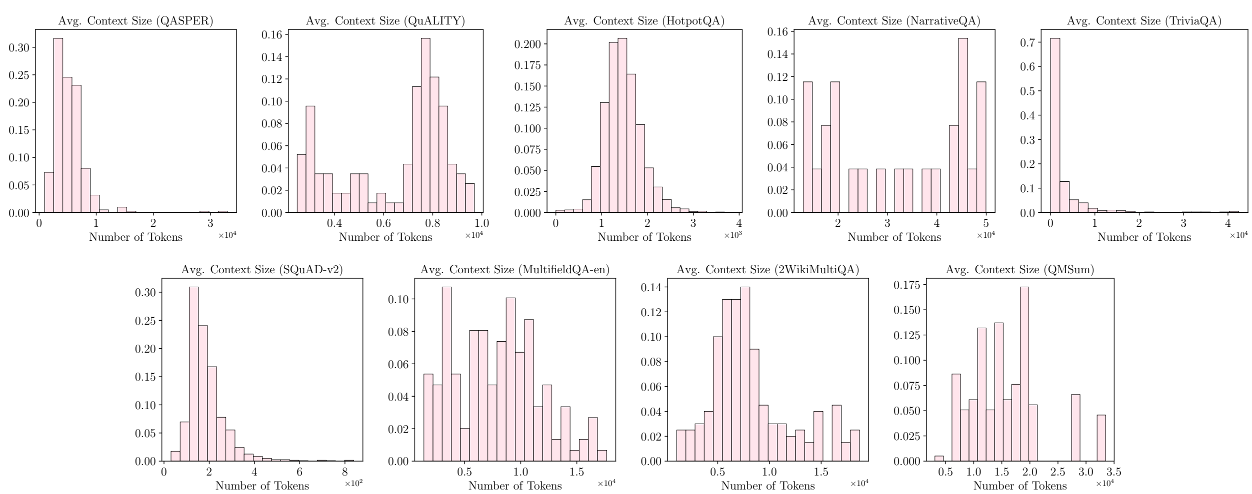

Figure 10: Context Size in terms of number of tokens according to Mistral-7B's tokenizer. All datasets except SQuAD-v2 focus on long context.

C.2. Context Compression

Context compression is essential for reducing inference costs and maintaining language understanding capabilities in long-context (Pan et al., 2024) or multi-turn scenarios (Kim et al., 2024b). Prior work approach this in two main directions: natural-language (NL)-based compression and representation-level compression. *NL-based compression* (Zhang et al., 2024b; Chirkova et al., 2025) like ADACOMP (Zhang et al., 2024b), COMPACT (Yoon et al., 2024), and EXIT (Hwang et al., 2024) condense prompts or histories into concise natural language summaries, typically using extractive or abstractive summarization. These methods are generally model-agnostic and applicable across both open-source and proprietary LLMs (Zhu et al., 2025). Representation-based methods (Chevalier et al., 2023; Munkhdalai et al., 2024; Louis et al., 2025b;a), on the other hand, treat the LLM as a white box and modify attention calculation (Munkhdalai et al., 2024), positional encodings (Jin et al., 2024; Zhang et al., 2024c), or embeddings (Cheng et al., 2024). Methods such as xRAG (Cheng et al., 2024), GIST (Mu et al., 2023), and ICAE (Ge et al., 2024) project instructions demonstrations, or the context into the language models' space. While compression improves efficiency, it often introduces a performance trade-off. Our work focuses on leveraging compression to improve retrieval and generation quality in RAG settings.

D. Discussion

Extension to New Decoders SARA is designed to be *model-agnostic*. All components–retriever, compressor, and the QA model–can be replaced with minimal effort. Note that the same decoder must be used across both *Compression Learning* (Section 2.3), *Instruction-tuning*, and *Generation* (Section 2.4). This is because the model learns to *interpret* compression vectors through its own decoder weights.

E. Impact Statement

This work advances Retrieval-Augmented Generation (RAG) by improving efficiency and performance under strict context budgets. SARA's hybrid compression techniques reduces computational overhead and improve a variety of LLMs, especially smaller models. While the proposed method raises no immediate ethical concerns, care should be taken to ensure compressed contexts do not inadvertently exclude critical or biased information, especially in high-stakes or sensitive domains like healthcare.

F. Expressivity of compression vectors

Faithful representation of semantics is pivotal for our compression vectors to serve as reliable contexts. To evaluate this, we decode the compression vectors into natural language and compare the reconstructed evidence with their sources. Representative successes for both chunk-level and paragraph-level reconstructions are shown in Table 11 and 9. We observed that the decoded text are usually shorter and serve as higher level summarizations for the input. In most cases, the decoded text preserves core propositions, causal links, and sentiment. SARA is able to recover key information, such as exact entities (e.g. 'Amazon customer service') and numeric values (e.g. '220'). Losses are mostly fine-grained–exact dates ('1903' $\rightarrow$ '1900s') or numeric magnitudes ('3400 years' $\rightarrow$ over 3,000 years) may be paraphrased or omitted. When contexts are longer, the risk of recovery failure is higher. This necessitates reasoning over mixed evidence formats.

Crucially, the decoder rarely invents new facts: missing detail is typically dropped rather than hallucinated. This behavior implies that the vectors encode stable, high-level meaning while suppressing fewer specifics–a valuable feature for knowledge-intensive tasks that demand both factual precision and robust hallucination control.

G. Limitations & Future Work

While SARA demonstrates strong performance, several open questions suggest promising directions for future work.

First, although the compression vectors serve as effective summaries and preserve salient factual details (e.g. numbers, organization names), the granularity of retained information varies. This uncertainty poses challenges on faithfully interpreting the original context. A promising direction is to train a lightweight probing model that estimates the likelihood of accurately recovering key contents from compressed representations, guiding more adaptive compression strategies. Another approach is to evaluate fidelity through an *evidence reconstruction step*, where compression vectors are decoded back into text and compared to the original using metrics such as conditional self-information (CSI), token overlap, or embedding similarity. One way to ensure compressed representations retain sufficient semantic content is through an

evidence reconstruction step, which evaluates whether a set of compression vectors can faithfully recover the original evidence. High reconstruction fidelity suggests that the compression vectors are sufficient. Second, determining the appropriate number of compression vectors per context remains an open challenge. Our current method–assigning one vector per sentence–may not be optimal across all inputs. Future research could explore adaptive mechanisms that tailor the number of vectors to the complexity of the contexts. Lastly, developing mechanisms that predict the number of contexts required for a given query would lead to more context-aware and resource-efficient RAG. Future work can also extend the use of compression vectors beyond generation, exploring their utility in tasks such as document reranking and citation resolution. We also envision applying our framework to specialized domains (e.g., legal or scientific retrieval) where concise yet faithful evidence representation is critical.

Prediction	Ground-truth
# Anti-scam dataset Collecting **human-human conversational data** to create a dataset for training and evaluating **anti-scam** models. We **collect conversations** between **users** and **attackers who aim to gather customer information** from **Amazon customer service scam scenarios**. We collected 220 anti-scam conversational data from **Amazon customers** through a Turkers' platform, which are **human-human dialogues**. The average length of a conversation is 11.5 turns and the average length is 11 words. **172 out of 220 users successfully identified attackers**, indicating that **the attackers are well-trained** in their scam attack strategy. We recruited **two experienced annotators** to evaluate the quality of the annotated data.	**## AntiScam Dataset** To enrich available non-collaborative task datasets, we created a corpus of **human-human anti-scam dialogs** in order to learn human elicitation strategies. We chose a popular **Amazon customer service scam scenario** to collect dialogs between users and attackers who aim to collect users information. We posted a role-playing task on the Amazon Mechanical Turk platform and collected a typing conversation dataset named AntiScam. We collected 220 human-human dialogs. The average conversation length is 12.45 turns and the average utterance length is 11.13 words. **Only 172 out of 220 users successfully identified their partner as an attacker**, suggesting that **the attackers are well trained** and not too easily identifiable. We recruited **two expert annotators** who have linguistic training to annotate 3,044 sentences in 100 dialogs, achieving a 0.874 averaged weighted kappa value.
Exploration of oil in Nigeria began around 1900, when oil was discovered in commercial quantities in the Niger Delta region. However, **large-quantities** was only **discovered later** in **1956** in **Oloibiri**.	Although the history of **oil exploration in Nigeria dates back to 1903**, non-commercial quantities of oil were not discovered there until 1953. **Commercial amounts of crude oil** were **later discovered** in **Oloibiri, Nigeria** in **1956**.
The Great Trek was a series of **migrations** of **Dutch-speaking settlers** from **Cape Colony in South Africa**, which **began in 1836** and lasted for several years.	**The Great Trek** was an eastward **migration** of **Dutch-speaking settlers** who travelled by wagon trains from **the Cape Colony** into the interior of modern **South Africa from 1836 onwards**. The exploratory treks, however, arrived at the bay of Port Natal in February 1835.
The history of music is **the study of music and its development over time**, from **prehistoric times** to **the present day**. The oldest known written music is the song "Hymn to the Sun" from the Sumerian civilization, which is believed to be **over 3,000 years** old.	**The history of music** covers **the historical development and evolution of music** from **prehistoric times** to **present day**. The "**oldest known song**" was written in cuneiform, dating to 3400 years ago from **Ugarit in Syria**. The first piece of unwritten music was made prior to the Paleolithic age 3.3 million years ago.

Table 9: Reconstruction quality of compression tokens in SARA. Source-aligned spans are shown in **bold** and errors are underlined. SARA faithfully reproduces most original semantics with only minor hallucinations.

Model	NarrativeQA				SQuAD-v2			
	Rele.	Correct.	Sim.	Faith.	Rele.	Correct.	Sim.	Faith.
ICAE_Mistral7B	52.08	16.75	51.27	21.19	67.17	51.93	75.25	69.64
LLMLingua	84.42	37.03	79.95	39.66	86.63	70.66	89.70	75.76
LongLLMLingua	84.17	34.38	76.67	30.86	83.73	67.90	87.72	73.98
SARA	87.87	44.09	82.26	43.83	90.66	77.21	92.16	80.12
Model	TriviaQA				HotpotQA			
	Rele.	Correct.	Sim.	Faith.	Rele.	Correct.	Sim.	Faith.
ICAE_Mistral7B	54.70	36.48	58.21	58.05	47.81	21.59	53.19	39.37
LLMLingua	71.95	68.95	82.26	61.58	61.43	41.72	73.63	75.94
LongLLMLingua	70.44	70.52	82.67	72.53	61.56	41.97	74.02	77.49
SARA	88.92	70.63	88.14	76.47	83.09	55.55	86.94	80.03

Table 10: LLM-based evaluation results across four datasets under context constraint of 512 tokens. We report Response Relevance (Rele.), Answer Correctness (Correct.), Semantic Similarity (Sim.), and Faithfulness (Faith.) in percentages.

Decoded Text	**Original Text**
We release the code and the data.	We release the code and data.
Also, we build a persuasive dialogue system to persuade people to donate to charity.	Furthermore, we also build a persuasion dialog system to persuade people to donate to charities.
Rigid templates limit <u>creativity</u> and diversity, resulting in loss of user engagement.	However, rigid templates lead to limited diversity, causing the user losing engagement.
The generation model is good at producing diverse responses but lacks coherence.	On the other hand, language generation models can generate diverse responses but are bad at being coherent.
<u>Collaborative</u> end-to-end systems have been developed to a great extent for the goal to build a user-friendly system that enables participants to work together with the system to achieve a common goal.	Considerable progress has been made building end-to-end dialog systems for collaborative tasks in which users cooperate with the system to achieve a common goal.
We use a hierarchical annotation scheme. This generic annotation method can be applied to different tasks.	To handle social content, we introduce a hierarchical <u>intent</u> annotation scheme, which can be generalized to different <u>non-collaborative dialog</u> tasks.

Table 11: Decoded text from compression vectors using `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023a) as the base model. Information omitted from one text but present in the other is <u>underlined</u>. Compared to the original, SARA retains concise semantics and excels at capturing high-level concepts. In some cases, it may lose fine-grained details such as specific entities and numerical values.

Retriever	QASPER		NarrativeQA		TriviaQA	
	F-1	ROUGE-L	F-1	ROUGE-L	F-1	ROUGE-L
SFR	55.44	52.93	58.03	56.39	84.13	83.61
BGE	44.47	45.24	54.05	53.98	85.41	84.58
BM25	36.15	39.54	56.79	55.76	83.58	83.65

Table 12: Generalizability across different retrievers.

QMSum	**Relevance**	**Correctness**	**Similarity**	**Faithfulness**
Mistral7B	51.82	8.97	52.90	69.39
+SARA	70.37	11.17	53.51	70.68
MultifieldQA-en	**Relevance**	**Correctness**	**Similarity**	**Faithfulness**
Mistral7B	42.32	21.97	42.09	31.61
+SARA	90.04	22.24	45.13	32.56
2WikiMultiHopQA	**Relevance**	**Correctness**	**Similarity**	**Faithfulness**
Mistral7B	31.50	35.69	29.91	42.82
+SARA	86.53	37.87	31.58	44.13

Table 13: Results on out-of-domain datasets. We report Response Relevance (Relevance), Answer Correctness (Correctness), Semantic Similarity (Similarity), and Faithfulness (Faithfulness).