© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716603

Compressing Without Losing Context: A Novel Framework for Text Summarization

ARKESH KUMAR SATAPATHY[1], SURYA LN PRADHAN[2], SNEHA PATNAIK[3], SWAIN KANHEYA BHIMA[4], PROF. SANJIT KUMAR ACHARYA[5]

[1, 2, 3, 4]*Student, Department of Computer Science and Engineering, NIST University, Berhampur, Odisha, India*

[5]*Assistant Professor, Department of Computer Science and Engineering, NIST University, Berhampur, Odisha, India*

Abstract- ***The expansion of Large Language Models (LLMs) into long-context reasoning has introduced critical challenges regarding computational efficiency and information integrity. While increasing context windows provides more space for data, it often leads to information dilution, "lost-in-the-middle" phenomena, and prohibitive key-value (KV) cache costs. This paper presents a comprehensive framework for context-aware text summarization utilizing information theory, discourse analysis, and agentic refinement. We specifically investigate the COMI (COarse-to-fine Context Compression) architecture, which leverages Marginal Information Gain (MIG) to balance relevance and diversity. Furthermore, we explore the shift from passive retention to active, iterative reasoning through paradigms like InftyThink and extreme compression algorithms such as TurboQuant. Experimental results across benchmarks such as NaturalQuestions, GovReport, and LongBench-v2 demonstrate that these techniques maintain high fidelity even at 32x to 40x compression ratios, bridging the gap between computat ional constraints and semantic completeness.***

Index Terms- ***Automatic Text Summarization, Context Compression, Information Bottleneck, Large Language Models, Marginal Information Gain, Key-Value Cache Optimization, Agentic Refinement***

I. INTRODUCTION

The landscape of natural language processing (NLP) has undergone a fundamental transition from constrained, task-specific architectures to the expansive, zero-shot capabilities of Large Language Models (LLMs) [1],.[2] This evolution has redefined Automatic Text Summarization (ATS) not merely as a task of shortening text, but as a complex exercise in semantic distillation and context preservation.[3] While the industry moves toward processing massive corpora—ranging from legal briefs and clinical notes to multi-document news threads—a significant computational and cognitive bottleneck has emerged [2],.[4]

The global document summarization AI market, valued at $3.8 billion in 2025, is projected to reach $22.6 billion by 2034, driven by the accelerating digitization of enterprise workflows and the proliferation of unstructured data . However, the primary obstacle in contemporary long-document understanding remains the quadratic complexity of the self-attention mechanism [2],.[5] As input sequence length increases, the memory requirements for maintaining the key-value (KV) cache grow linearly, creating prohibitive latency and cost barriers [6],.[7]

Consequently, the focus of research has shifted toward context compression as a more sustainable alternative to increasing physical window sizes [2],.[8] This paper presents a comprehensive analysis of a novel, coarse-to-fine framework for context compression, integrating information theory and agentic refinement to ensure that the process of summarization enhances, rather than dilutes, the utility of large-scale document processing.

II. THEORETICAL FOUNDATIONS

The quest to compress context without information loss is theoretically grounded in the Information Bottleneck (IB) principle [8],.[9] The IB method seeks to find a representation $\tilde{X}$ of an input X that is as concise as possible while being maximally predictive of a relevant variable Y [8],.[10]

A. The Information Bottleneck Lagrangian

The optimization objective is the IB Lagrangian, which balances the compression rate against the

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716603

preservation of task-relevant information [8], . The objective function is formulated as:

$$L_{IB} = I(\tilde{X}; X|Q) - \beta I(\tilde{X}; Y|Q)$$

In this formulation, Q represents the query or summarization objective, $I(\tilde{X}; X|Q)$ denotes the mutual information between the compressed and original context (to be minimized to reduce redundancy), and $I(\tilde{X}; Y|Q)$ denotes the mutual information between the compressed context and the target output (to be maximized to ensure performance) [8], . The parameter β acts as a weight that determines the intensity of the bottleneck; as β increases, the model prioritizes information retention over the degree of compression [8], .

B. Variational Information Bottleneck (VIB)

To make the IB objective tractable for deep neural networks, researchers utilize the Variational Information Bottleneck (VIB), which provides a lower bound through variational approximation.[5] The VIB objective function is derived as:

$$L_{VIB} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{p_\theta(Z|x_i)}[\log q_\phi(y_i|Z)] - \beta D_{KL}[p_\theta(Z|x_i)\|r(Z)]$$

Here, the first term represents the expected log-likelihood (the prediction term), while the second term uses Kullback-Leibler (KL) divergence to match the posterior distribution to a fixed prior $r(Z)$, typically $N(0, I)$, enforcing the information bottleneck [5], .

III. THE COMI FRAMEWORK

To operationalize these theoretical insights, researchers have proposed the COMI (COarse-to-fine Context Compression via Marginal Information Gain) framework [2], . COMI addresses the non-uniform distribution of information in long documents via a two-stage adaptive process [8], .

A. Marginal Information Gain (MIG) Formulation

The innovation at the heart of COMI is the Marginal Information Gain (MIG) metric [2], . Unlike standard attention scores, which only reflect relevance, MIG incorporates a penalty for semantic redundancy [2],.[8]

For a given token x_i, query vector q, and context X, MIG is defined as:

$$G(x_i, q, X) = \frac{x_i^\top q}{\|x_i\|\|q\|} - \max_{x_j \in X, j \neq i}\left(\frac{x_i^\top x_j}{\|x_i\|\|x_j\|}\right)$$

The first term measures the cosine similarity between the token and the query (relevance), while the second term captures the maximum cosine similarity between x_i and any other token in X (redundancy) [11], . This ensures that each retained token provides "novel" information, maximizing diversity within the compressed window [2],.[8]

B. Two-Stage Implementation

1) Coarse-Grained Group Reallocation:
The input context is partitioned into several segments or compression groups [8],.[12] The model evaluates the inter-group MIG to determine each segment's contribution. Rather than applying a uniform ratio, the framework reallocates the token budget toward high-MIG regions, effectively mitigating the "lost-in-the-middle" bias where models ignore the central sections of a prompt [2], .

2) Fine-Grained Token Merging:
Within each group, tokens are weighted by their intra-group MIG and merged through semantic fusion [8], . This "soft" approach allows the model to preserve the "gist" of a sequence by combining embedding representations into a compact set of vectors that still carry the primary semantic signal [2],.[6] Synergistic effects between coarse reallocation and fine fusion allow COMI to achieve a 25-point Exact Match (EM) improvement over traditional baselines .

IV. STRUCTURAL PRESERVATION AND DISCOURSE ANALYSIS

Traditional compression often loses rhetorical structure—the logical relationship between sentences—resulting in summaries that are factually correct but incoherent [13],.[14]

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716603

A. LingoEDU and Structural Relation Trees

The EDU-based Context Compressor reformulates compression as a "structure-then-select" process [15], . It transforms linear text D into a Structural Relation Tree $T = (V, E)$,.

1) Nodes (V):

Represent Elementary Discourse Units (EDUs), which are the minimal variable-length units capable of conveying coherent semantics. Each unit e_i is represented as a triplet: $e_i = (t_i, pos_i, id_i)$, where t_i is text, pos_i is character offset, and id_i is a unique index.

2) Edges (E):

Represent discourse linkages and dependency relations (e.g., elaboration, contrast), capturing the logical flow often lost in standard retrieval ,. A lightweight ranking module then selects query-relevant sub-trees for linearization, ensuring high fidelity even under extreme reduction , .

B. Page-Specific Alignment (PTSPI)

The PTSPI (Page-specific Target-text alignment Summarization with Page Importance) model extends sequence-to-sequence methods by dividing source documents into pages.[16] An additional layer provides dynamic page weightage, focusing the model's energy on the most informative segments. PTSPI has demonstrated a 6.32% improvement in ROUGE-1 and an 8.08% improvement in ROUGE-2 scores over standard state-of-the-art models [17],.[18]

V. EXTREME MEMORY ARCHITECTURES

The technical implementation of context compression has expanded into extreme memory architectures designed to optimize the KV cache without performance degradation .

A. TurboQuant: High-Fidelity Quantization

Introduced in 2026, the TurboQuant algorithm addresses the memory overhead of vector quantization through a two-step process .

1) PolarQuant Method:

The data vectors are randomly rotated to simplify their geometry. This enables the application of a high-quality quantizer that captures the majority of the vector's semantic strength using the primary bits .

2) QJL Algorithm:

TurboQuant uses a single residual bit to apply the Quantized Johnson-Lindenstrauss (QJL) algorithm. This serves as a mathematical error-checker that eliminates bias in attention scores, ensuring that the compressed KV cache maintains optimal recall .

B. Titans and MIRAS Variants

The Titans architecture introduces long-term memory (MAC) modules that compress past data into a summary before incorporating it into the current context . Titans transcends the mean squared error (MSE) paradigm by using robust MIRAS variants like YAAD and MONETA, which utilize Huber loss and generalized norms to handle outliers and messy data, scaling effectively to context windows larger than 2 million tokens .

VI. AGENTIC REFINEMENT AND SELF-REGULATION

The latest generation of frameworks has moved toward "agentic" models where the LLM acts as an active manager of its own context window [19],.[20]

A. InftyThink and the Sawtooth Pattern

The InftyThink paradigm, published in early 2026, transforms monolithic reasoning into an iterative process . By interleaving short reasoning segments with intermediate summarization, it enables unbounded reasoning depth while maintaining bounded computational costs . This creates a characteristic "Sawtooth" memory pattern, where context grows during active exploration and collapses during consolidation [20], . Findngs indicate that enforcing this behavior can result in a 22.7% token saving with zero loss in accuracy .

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716603

B. Multi-Stage Inference via AgenticSum

AgenticSum addresses hallucinations in clinical summarization by separating the task into coordinated stages: context selection, draft generation, verification using internal attention grounding signals, and targeted correction [6], . This multi-stage process ensures that every statement in the output is traceable to a specific path in the source material, providing interpretable signals of content grounding .

VII. EVALUATION RIGOR AND BENCHMARKING

As techniques become more sophisticated, metrics must move beyond n-gram overlap to evaluate factual consistency and semantic relevance [21], .

A. QA-Based Evaluation (QAFactEval)

QA-based metrics like QAGS and QAFactEval have emerged as the most reliable way to measure consistency [22], . These metrics decompose a summary into "atomic facts" and generate questions for each fact. A question-answering model then answers these questions based on both the summary and the source document [21], . Discrepancies between the answers reveal factual inconsistencies, leading to a 14% improvement in identifying hallucinations compared to traditional metrics , .

B. LLM-as-a-Judge (G-Eval)

The G-Eval framework leverages GPT-4 to score summaries on coherence, consistency, fluency, and relevance using chain-of-thought (CoT) prompts [23], . By using probability-weighted summation of output scores, G-Eval achieves a Spearman correlation of 0.514 with human judgment, significantly outperforming ROUGE and BLEU in open-ended generation tasks [10], .

VIII. EMPIRICAL PERFORMANCE ANALYSIS

The efficacy of these novel frameworks is validated across standardized long-document benchmarks such as LongBench-v2 and RULER , .

Method	Compression Ratio	Performance Metric	Primary Advantage
COMI	32x	+25 EM Score	Balance of relevance and diversity
C3	40x	93% Accuracy	Pure-text latent-space upper bound
TurboQuant	Extreme	Optimal Recall	Minimal KV memory footprint
LingoEDU	Explicit	-51.11% HLE Error	Structural integrity and detail
DeepSeek-OCR	10x	97% Accuracy	Vision-language synergy

Experimental data demonstrate that context-aware predictors consistently achieve lower performance prediction error than context-agnostic ones, allowing for "Performance-oriented Context Compression" (PoC) that meets specific user-defined accuracy bounds.[24]

IX. CONCLUSION

The "Compressing Without Losing Context" framework represents a synthesis of information

theory, discourse-aware architectures, and agentic self-regulation. By utilizing Marginal Information Gain and structural relation trees, the framework addresses the fundamental trade-off between computational efficiency and semantic fidelity. As we scale toward millions of tokens, the transition from "reading every token" to "understanding every unit of value" via iterative reasoning and extreme quantization will define the next era of high-performance automated intelligence.

REFERENCES

[1] T. Tishby et al., "The information bottleneck method," *Proc. 37th Allerton Conf. Commun. Control Comput.*, pp. 368-377, 1999.

[2] S. Ko et al., "Evidence-focused fact summarization for knowledge-augmented zero-shot question answering," *Proc. EMNLP 2024*, pp. 10636–10651, 2024.

[3] H. Pan et al., "COMI: Coarse-to-fine context compression via marginal information gain," *arXiv preprint*, arXiv:2602.01719v3, 2026.

[4] Y. Gao et al., "PTSPI: Page-specific target-text alignment summarization with page importance," *arXiv preprint*, arXiv:2509.16539v1, 2025.

[5] H. Yang et al., "Structured information bottleneck (SIB) framework for IB Lagrangian methods," *AAAI*, 2025.

[6] J. Ren et al., "AgenticSum: A multi-stage agentic framework for clinical text summarization," *arXiv preprint*, arXiv:2602.20040v1, 2026.

[7] H. Pan et al., "COMI: Coarse-to-fine adaptive context compression framework," *arXiv preprint*, arXiv:2602.01719, 2026.

[8] X. Pu and V. Demberg, "RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization," *Proc. NAACL 2024*, 2024.

[9] J. Ren et al., "AgenticSum for clinical text summarization," *arXiv*, 2602.20040, 2026.

[10] S. Fang et al., "Active compression: Autonomous context regulation for cost-aware agents," *arXiv preprint*, arXiv:2601.07190v1, 2026.

[11] W. Scialom et al., "QuestEval: Summarization asks for fact-based evaluation," *Proc. EMNLP 2021*, 2021.

[12] B. Fabbri et al., "QAFactEval: Improved QA-based factual consistency evaluation," *Proc. NAACL 2022*, 2022.

[13] Dataintelo Analysis, "Global Document Summarization AI Market Report 2025-2034," *March 2026*.

[14] X. Sun et al., "InftyThink: Iterative reasoning with intermediate summarization," *ICLR 2026*, 2026.

[15] Bulatov et al., "Titans: Learning to compress at the memory-augmented context," *Google Research*, 2025.

[16] Microsoft Foundry, "G-Eval metric for abstractive summarization evaluation," *Microsoft Build 2026*, 2026.

[17] S. Si et al., "EDU-based context compressor: A novel explicit compression framework," *arXiv preprint*, arXiv:2512.14244, 2025.

[18] A. Zandieh and V. Mirrokni, "TurboQuant: Redefining AI efficiency with extreme compression," *Google Research / ICLR 2026*, 2026.

[19] Y. Wei et al., "C3: Context cascade compression for efficient long context processing," *arXiv preprint*, arXiv:2511.15244v1, 2025.

[20] S. Si et al., "LingoEDU: Transforming linear context into structural relation trees," *arXiv preprint*, arXiv:2512.14244v2, 2025.

WORKS CITED

[1] A Comprehensive Survey on Automatic Text Summarization with Exploration of LLM-Based Methods - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2403.02901v3

[2] COMI: Coarse-to-fine Context Compression via Marginal Information Gain - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2602.01719v3

[3] A Hierarchical Representation Model Based on Longformer and Transformer for Extractive Summarization - MDPI, accessed on April 13, 2026, https://www.mdpi.com/2079-

9292/11/11/1706

[4] A comparative study of pretrained language models for long clinical text - PMC, accessed on April 13, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC9846675/

[5] Long Document Summarization with Top-Down and Bottom-Up Representation Inference, accessed on April 13, 2026, https://openreview.net/forum?id=xiXOrugVHs

[6] Simple Context Compression: Mean-Pooling and Multi-Ratio Training - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2510.20797v1

[7] Simple Context Compression: Mean-Pooling and Multi-Ratio Training - arXiv, accessed on April 13, 2026, https://arxiv.org/pdf/2510.20797

[8] QUITO-X: A New Perspective on Context ... - ACL Anthology, accessed on April 13, 2026, https://aclanthology.org/2025.findings-emnlp.362.pdf

[9] Theory and Application of the Information Bottleneck Method - PMC, accessed on April 13, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC10968930/

[10] QuestEval: Summarization Asks for Fact-based Evaluation | Request PDF - ResearchGate, accessed on April 13, 2026, https://www.researchgate.net/publication/357120095_QuestEval_Summarization_Asks_for_Fact-based_Evaluation

[11] Long Document Classification Benchmark 2025 - Procycons, accessed on April 13, 2026, https://procycons.com/en/blogs/long-document-classification-benchmark-2025/

[12] DeepSeek-OCR: How Optical Compression Redefines Long Context | IntuitionLabs, accessed on April 13, 2026, https://intuitionlabs.ai/articles/deepseek-ocr-optical-compression

[13] Long-Text Abstractive Summarization using Transformer Models: A ..., accessed on April 13, 2026, https://journals-sol.sbc.org.br/index.php/jbcs/article/view/5786

[14] DISRetrieval: Harnessing Discourse Structure for Long Document Retrieval - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2506.06313v1

[15] Evidence-Focused Fact Summarization for Knowledge-Augmented Zero-Shot Question Answering - ACL Anthology, accessed on April 13, 2026, https://aclanthology.org/2024.emnlp-main.594/

[16] Long document summarization using page specific target text alignment and distilling page importance - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2509.16539v1

[17] COMI: Coarse-to-fine Context Compression via Marginal Information Gain - ResearchGate, accessed on April 13, 2026, https://www.researchgate.net/publication/400369308_COMI_Coarse-to-fine_Context_Compression_via_Marginal_Information_Gain

[18] BERTScore and ROUGE: Two Metrics for Evaluating Text Summarization Systems, accessed on April 13, 2026, https://haticeozbolat17.medium.com/bertscore-and-rouge-two-metrics-for-evaluating-text-summarization-systems-6337b1d98917

[19] AgenticSum: An Agentic Inference-Time Framework for Faithful Clinical Text Summarization, accessed on April 13, 2026, https://arxiv.org/html/2602.20040v1

[20] Active Context Compression: Autonomous Memory Management in LLM Agents - arXiv, accessed on April 13, 2026, https://arxiv.org/html/2601.07190v1

[21] Evaluate the text summarization capabilities of LLMs for enhanced decision-making on AWS, accessed on April 13, 2026, https://aws.amazon.com/blogs/machine-learning/evaluate-the-text-summarization-capabilities-of-llms-for-enhanced-decision-making-on-aws/

[22] QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization - ACL Anthology, accessed on April 13, 2026, https://aclanthology.org/2022.naacl-main.187.pdf

[23] How to evaluate a summarization task - OpenAI Developers, accessed on April 13, 2026,

https://developers.openai.com/cookbook/examples/evaluation/how_to_eval_abstractive_summarization

[24] arXiv:2112.08542v2 [cs.CL] 29 Apr 2022, accessed on April 13, 2026, https://arxiv.org/pdf/2112.08542