BEYOND TOKEN PRUNING: PROMPT COMPRESSION VIA MINIMUM DESCRIPTION LENGTH AND SEMANTIC MESSAGE ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt compression accelerates long-context inference by condensing extensive context into concise core messages. Existing approaches follow an extractive score-and-drop paradigm, scoring tokens or text spans via local importance heuristics and discarding low-ranked elements. Hard deletion permanently discards distributed background context, failing on tasks requiring global counts or bounds, while token pruning destroys syntactic integrity and narrative topology. We reframe prompt compression as a coding problem rather than a filtering problem. Under the Minimum Description Length (MDL) principle, a compressed prompt is a two-part code: intact evidence units answer the query directly, while compact sufficient statistics summarize omitted background context. This view yields a select-and-aggregate paradigm beyond pruning, which we instantiate as **S**emantic **M**essage **A**nchors (SMA), an efficient, training-free, and state-agnostic framework requiring no reader internal states or gradient updates. SMA segments text into grammatically self-contained propositional units, greedily packs query-salient anchors while preserving relative document topology, and compiles unselected background spans into compact sufficient statistics. SMA outperforms the strongest evaluated training-free compressor by up to 9.34 points on LongBench, accelerates prefill by up to 7.55×, and reduces peak KV cache by 10.9×, advancing the empirical Quality-Latency Pareto frontier.

1 INTRODUCTION

Processing long textual contexts is central to modern large language models (LLMs), including multi-document question answering, repository-level code synthesis, and long-horizon agent planning (Bai et al., 2024; Shaham et al., 2023). However, ingesting prompts comprising tens of thousands of tokens incurs severe latency and memory overheads from quadratic complexity during prefill attention and substantial Key-Value (KV) cache footprints that constrain serving throughput.

To alleviate these bottlenecks, prompt compression has become an active line of research. Under the *extractive pruning* paradigm, lexical tokens or coarse passage spans are scored via diverse local heuristics, ranging from language model perplexity (Jiang et al., 2023; 2024) and self-information (Li et al., 2023) to distilled token classifiers (Pan et al., 2024) and shallow evaluator heads (Fei et al., 2025); lowest-scoring elements are then excised until a target budget is satisfied.

However, this extractive pruning paradigm exhibits two fundamental limitations. First, downstream reasoning often depends on global statistics, entity inventories, existence markers, or numerical extrema distributed across background context; hard deletion permanently discards unselected spans, losing valuable evidence that might help for downstream tasks. Second, token-level pruning disrupts lexical integrity and syntax. In extreme-length contexts, token-level classification pruning (Pan et al., 2024) disproportionately retains punctuation tokens, which multiply and displace informative words, producing fragmented sequences that downstream readers reject. Both failures stem from a shared design choice: compressors permanently discard unselected background context, whereas downstream readers frequently require global evidence alongside focal anchors.

To address this problem, we reframe prompt compression from a destructive *score-and-drop* filtering process to a *select-and-aggregate* coding problem through the **Minimum Description Length**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(MDL) principle (Grünwald, 2007). While prior information-theoretic studies characterize the theoretical limits of context pruning (Nagle et al., 2024), our MDL framework directly prescribes the structural contents of a compressed prompt. Over discrete *semantic units* (propositional sentences, code blocks, and structured tables), effective compression identifies a compact subset of self-contained message anchors to substantiate the query, while compiling residual background context into a task-relevant summary. Under this paradigm, we present **SMA** (**S**emantic **M**essage **A**nchors), an efficient, training-free, and state-agnostic compression framework that operates without accessing reader internal states, attention weights, or parameter updates. SMA partitions context into propositionally self-contained atomic semantic units, packs query-salient message anchors via a relevance-guided greedy selection that approximates conditional mutual information under stated redundancy conditions, and condenses unselected background text into predicate-aware summary statistics within the residual budget.

Our main contributions are threefold:

- **Information-Theoretic Formulation**. We formulate prompt compression through the Minimum Description Length principle as a two-part code, showing how augmenting discrete factual anchors with sufficient statistics of omitted text achieves lower distortion than pure extractive pruning on aggregable tasks.
- **The SMA Framework**. We introduce Semantic Message Anchors, a training-free and state-agnostic compression framework that pairs greedy relevance packing over intact proposition units with predicate-aware background aggregation, strictly bounded by a tokenizer compliance guard.
- **Empirical Quality-Latency Advantage**. Across two mainstream benchmarks on two reader models, SMA outperforms prior compressors by up to 9.34 points on LongBench, accelerates prefill (TTFT) by up to 7.55×, and reduces peak KV cache by up to 10.9×.

2 Related Work

2.1 Token-Level and Structural Prompt Compression

Prompt compression techniques reduce input prompt length to mitigate the quadratic prefill complexity and substantial KV cache footprints of Transformer architectures (Vaswani et al., 2017). One prominent line of research focuses on explicit token pruning, using statistical likelihood heuristics (Jiang et al., 2023; 2024; Li et al., 2023) or learned token classifiers (Pan et al., 2024) to eliminate uninformative tokens. Beyond standalone language modeling objectives, recent approaches use internal model states to guide compression, such as identifying shallow evaluator attention heads during early prefilling (Fei et al., 2025) or exploiting multi-document cross-attention in Fusion-in-Decoder readers (Choi et al., 2024). Complementary efforts explore sentence-level extractive selection and summarization (Xu et al., 2024), context-aware sentence scoring (Liskavets et al., 2025), recursive observation filtering in multi-turn agent environments (Kang et al., 2026), iterative reasoning-aware compression (Yun & Kim, 2026), and inference-time KV cache eviction for sequence extrapolation (Xiong et al., 2025). These approaches treat context reduction as destructive filtering, discarding the distributed background information downstream tasks require.

2.2 Information-Theoretic Limits and Message Transport

Information theory provides rigorous tools for characterizing lossy context reduction (Cover & Thomas, 2006). Rate-distortion formulations of prompt compression demonstrate that conditioning selection on the target query (such as Adaptive QuerySelect) expands the achievable distortion-rate frontier compared to query-agnostic filtering (Nagle et al., 2024). The Minimum Description Length (MDL) principle complements rate-distortion analysis by formalizing statistical induction as finding compact two-part codes that balance model complexity with data description (Grünwald, 2007). In multimodal compression, GMC (Qian et al., 2026) transports unselected attention messages onto preserved visual representatives, demonstrating that residual background signals can be preserved in continuous representations; however, its carriers are modified hidden activations tied to the multimodal setting, leaving textual prompting without a transparent, theoretically grounded counterpart. These foundations suggest that effective data reduction does not require complete information erasure, where optimal compression balances primary message carriers with compact representations of the residual background.

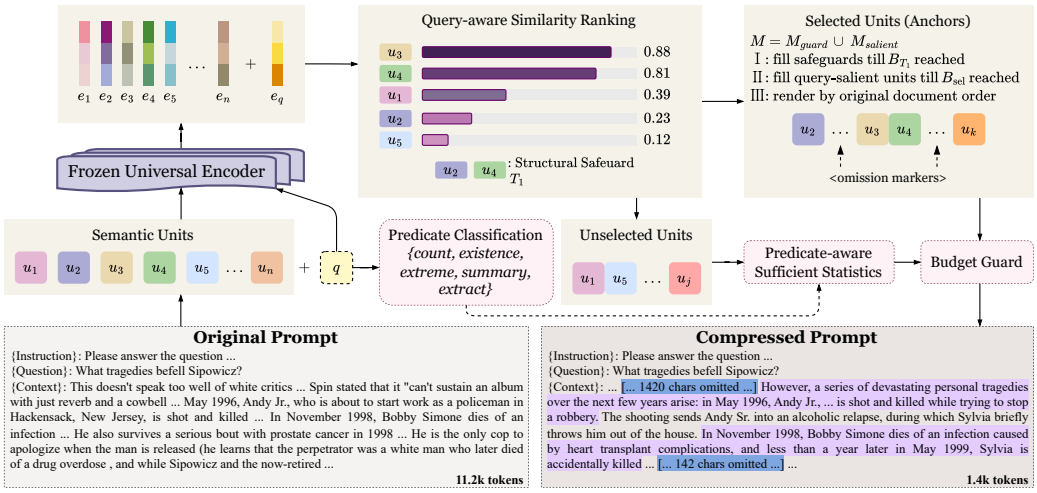


Figure 1: The overall framework of Semantic Message Anchors (SMA). Purple text marks preserved semantic message anchors carrying core factual information; Blue text marks skipped intervals with explicit omission markers preserving original document topology.

3 Methodology: Semantic Message Anchors (SMA)

3.1 Problem Formulation and Operational Boundary

Let $x \in \mathcal{V}^*$ denote an input prompt containing $N = |x|_{\text{tok}}$ tokens, $q \in \mathcal{V}^*$ an accompanying query, and Y the ground-truth downstream response. Prompt compression seeks a mapping $\text{comp} : (x, q) \mapsto C(x, q)$ that condenses x into a compact prompt satisfying a token budget $|C(x, q)|_{\text{tok}} \leq B$ with $B \ll N$.

Compressors in the operational class $\mathcal{C}_{\text{free}}$ contrast with token pruners (Pan et al., 2024) and reader-conditioned scorers (Choi et al., 2024; Fei et al., 2025) by satisfying three properties: (i) zero task-specific fine-tuning, compression-specific dataset distillation, or additional trainable parameters; (ii) state-agnostic execution, accessing no internal activations, attention weights, or KV caches of reader ϕ_{LLM}; and (iii) strict target budget compliance under the reader's tokenizer. Within $\mathcal{C}_{\text{free}}$, representations are extracted via a **frozen universal encoder** $\phi_E : \mathcal{V}^* \to \mathbb{R}^d$, where L_2-normalized embeddings $\mathbf{e}_u = \phi_E(u)/\|\phi_E(u)\|_2$ yield dense cosine similarities $s_i = \cos(\phi_E(u_i), \phi_E(q))$ without task adaptation.

3.2 Two-Part MDL Formulation and Theoretical Limits

We formulate prompt compression through the Minimum Description Length (MDL) principle (Grünwald, 2007). Transmitting long context x to reader ϕ_{LLM} to answer query q corresponds to constructing a two-part structural code $C(x, q) = (M, A)$ over discrete propositional units $U = \{u_1, \ldots, u_n\}$ that partition x: an explicit subset of primary *semantic message anchors* $M \subseteq U$ providing direct evidence, and a compact background code A capturing residual statistics of unselected text $U \setminus M$. The conceptual two-part description length is:

$$L_{\text{MDL}}(M, A) = L(M, A) + \lambda \cdot L_{\text{task}}(Y \mid C(x, q), q) \tag{1}$$

where $L(M, A) = |C(x, q)|_{\text{tok}}$ counts prompt tokens, $L_{\text{task}}(Y \mid C(x, q), q) = -\log P(Y \mid C(x, q), q)$ is the empirical log-loss on response Y, and $\lambda > 0$ calibrates discrete token counts against continuous log-loss. Under the budget constraint $|C(x, q)|_{\text{tok}} \leq B$, the multiplier $\lambda = \lambda(B)$ is the shadow price of budget. SMA does not estimate λ; it imposes the budget constraint directly, viewing Eq. (1) as a rate–distortion framing of the framework design.

Formulation 1 (Two-Part MDL and Rate-Distortion Framing)**.** Normalizing description length by input length $N = |x|_{\text{tok}}$, the two-part Minimum Description Length objective connects context

reduction to Rate-Distortion trade-offs:

$$\frac{1}{N} L_{\text{MDL}}(M, A) = R(M, A) + \lambda \cdot D(M, A) \tag{2}$$

where $R(M, A) = \frac{|C(x,q)|_{\text{tok}}}{N}$ is the normalized transmission rate, and $D(M, A) = \frac{1}{N} L_{\text{task}}$ denotes the normalized downstream task distortion.

While rate-distortion analysis in Nagle et al. (2024) provides diagnostic bounds on token pruning limits, our MDL formulation is constructive: it prescribes two-part codes composed of primary factual anchors plus background statistics. In deployment, instead of solving unconstrained Lagrangian objectives, compressors enforce $|C(x, q)|_{\text{tok}} \leq B$, which SMA addresses via structured proposition packing and a dedicated background headroom reserve τ_{agg}.

Because training-free, state-agnostic compressors in $\mathcal{C}_{\text{free}}$ operate without access to decoder gradients or task supervision, downstream task information is constrained by the upstream representation mapping:

Proposition 2 (Information Ceiling of Encoder-Mediated Compressors)**.** For any compressor in $\mathcal{C}_{\text{free}}$ where candidate anchor selection $M \subseteq U$ is an encoder-mediated decision $M = g(\phi_E(U), q)$, the selection mutual information satisfies $I(Y; M \mid q) \leq I(Y; \phi_E(U) \mid q)$. Emitted verbatim prompt $C(x, q)$ satisfies the data-processing ceiling:

$$I(Y; \hat{Y} \mid q) \leq I(Y; C(x, q) \mid q) \leq I(Y; U \mid q) \tag{3}$$

While encoder representations bound candidate anchor selection via $I(Y; M \mid q) \leq I(Y; \phi_E(U) \mid q)$, emitted prompt $C(x, q)$ extracts verbatim units from U; applying the Data Processing Inequality to the reader generation pipeline yields Eq. (2).

3.3 The SMA Compression Pipeline

The SMA framework operationalizes this theoretical foundation through a bypass check and four tightly coupled stages, illustrated in Figure 1 and formalized in Algorithm 1. SMA segments the raw input prompt x into a sequence of discrete, grammatically self-contained propositional units $U = \{u_1, \ldots, u_n\}$ while safeguarding atomic code blocks and markdown tables from fragmentation. Candidate units are scored via dense cosine similarities $\mathbf{s} = \cos(\phi_E(U), \phi_E(q))$ against the query representation.

Beyond token-level pruning, SMA constructs the primary anchor set M through an orthogonal dual-component partition $M = M_{\text{guard}} \cup M_{\text{salient}}$ ($M_{\text{guard}} \cap M_{\text{salient}} = \emptyset$). Stage 2A greedily allocates up to a structural floor $B_{T_1} = \lfloor \tau_{\text{tl}} B \rfloor$ tokens ($\tau_{\text{tl}} = 0.552$) to safeguard candidates $T_1 \subset U$ (negations, epistemic hedges capped at 0.15, and enumerations) to preserve discourse logic. Stage 2B packs remaining units $U \setminus M_{\text{guard}}$ by descending similarity $\mathbf{s}$ up to the anchor selection budget $B_{\text{sel}} = \lfloor (1 - \tau_{\text{agg}}) B \rfloor$ ($\tau_{\text{agg}} = 0.08$, reserving 8% for aggregation). Anchors are rendered via a topological projection operator $C_{\text{anchor}} = \mathcal{R}_{\text{topo}}(M)$, preserving relative document order. Budget reparameterization and decoupled formulations are detailed in Appendix B.

Conditional on safeguard floor $M_{\text{guard}} \subseteq T_1$, Stage 2B packs the remaining candidates $U_{\text{cand}} = U \setminus M_{\text{guard}}$ within residual budget $B_{\text{rem}} = B_{\text{sel}} - \text{used}$, and only candidates of at most B_{rem} tokens can ever be admitted; write $\tilde{U} = \{u \in U_{\text{cand}} : |u|_{\text{tok}} \leq B_{\text{rem}}\}$ for this admissible set. Exact conditional mutual information maximization being intractable, Stage 2B uses TopSim-Greedy; Proposition 3 bounds its approximation ratio.

Proposition 3 (Approximation Bound for TopSim-Greedy Selection)**.** Let S^* be the candidate subset maximizing conditional task information $I(Y; X_S \mid X_{M_{\text{guard}}}, q)$ under budget B_{rem}, where X_S denotes the joint random variable of text units in S. Under ϵ-conditional additivity of candidate propositions, monotone alignment of encoder similarity with marginal utility, and bounded proposition length ratio $\kappa_l = l_{\max}/l_{\min}$, the Stage-2B TopSim-Greedy selection M_{salient} satisfies:

$$I(Y; X_{M_{\text{salient}}} \mid X_{M_{\text{guard}}}, q) \geq \frac{1 - \epsilon}{\kappa_l (1 + \epsilon)} \frac{B_{\text{rem}} - l_{\max}}{B_{\text{rem}}} I(Y; X_{S^*} \mid X_{M_{\text{guard}}}, q) \tag{4}$$

Here $l_{\max}$ is the longest admissible candidate; the two factors price the gap between value ordering and density ordering. The denominator B_{rem} stays positive under the reported Stage-2A/2B split and units longer than B_{rem} never enter Stage 2B, so no degenerate case arises (Appendix A.3); the bound approaches $(1 - \epsilon)/(\kappa_l(1 + \epsilon))$ as the residual budget grows.

Algorithm 1 SMA: Semantic Message Anchors Compression

Input: Context x, Query q, Target token budget B, Frozen encoder ϕ_E, Hyperparameters $\tau_{\text{agg}}, \tau_{\text{tl}}$
Output: Compressed prompt $C(x, q)$ with $|C(x, q)|_{\text{tok}} \leq B$

if $|x|_{\text{tok}} \leq B$ **then return** x ▷ Bypass check
end if
$\kappa \leftarrow \text{ClassifyPredicate}(q)$
$U = \{u_1, \ldots, u_n\} \leftarrow \text{PropositionalSegment}(x)$ ▷ Stage 1: Segmentation & scoring
$\mathbf{s} \leftarrow \cos(\phi_E(U), \phi_E(q)), \quad T_1 \leftarrow \text{ConstructSafeguards}(U, \mathbf{s}, \text{cap}_{\text{hedge}} = 0.15)$
$B_{\text{sel}} \leftarrow \lfloor (1 - \tau_{\text{agg}})B \rfloor, \quad B_{T_1} \leftarrow \lfloor \tau_{\text{tl}} B \rfloor, \quad M \leftarrow \emptyset, \quad \text{used} \leftarrow 0$ ▷ Stage 2: Relevance packing
$M_{\text{guard}} \leftarrow \emptyset, \quad M_{\text{salient}} \leftarrow \emptyset$
for all $u \in T_1$ in descending order of $\mathbf{s}$ **do**
  if $\text{used} + |u|_{\text{tok}} \leq B_{T_1}$ **then** $M_{\text{guard}} \leftarrow M_{\text{guard}} \cup \{u\}, \quad \text{used} \leftarrow \text{used} + |u|_{\text{tok}}$
  end if
end for
for all $u \in U \setminus M_{\text{guard}}$ in descending order of $\mathbf{s}$ **do**
  if $\text{used} + |u|_{\text{tok}} \leq B_{\text{sel}}$ **then** $M_{\text{salient}} \leftarrow M_{\text{salient}} \cup \{u\}, \quad \text{used} \leftarrow \text{used} + |u|_{\text{tok}}$
  end if
end for
$M \leftarrow M_{\text{guard}} \cup M_{\text{salient}}$
$C_{\text{anchor}} \leftarrow \mathcal{R}_{\text{topo}}(M)$ ▷ Topological order-preserving projection
$B_{\text{agg}} \leftarrow \max(0, B - |C_{\text{anchor}}|_{\text{tok}}), \quad A \leftarrow \emptyset$ ▷ Stage 3: Controlled aggregation
if $B_{\text{agg}} > 0$ **and** $\kappa \neq \text{extract}$ **then**
  $A \leftarrow \text{AggregateStatistics}(U \setminus M, \kappa, \text{limit} = B_{\text{agg}})$
end if
$C_{\text{raw}} \leftarrow C_{\text{anchor}} \circ A$
return $\text{TokenTruncate}(C_{\text{raw}}, B)$ ▷ Stage 4: Budget guard

This bound provides a structural guarantee on worst-case retention; at empirical median $\kappa_l = 13.6$ (Appendix A.3), the asymptotic prefactor is approximately $0.07(1-\epsilon)/(1+\epsilon)$, characterizing worst-case retention rather than a tight numerical certificate.

3.4 Controlled Semantic Aggregation and Budget Guards

While extractive pruning discards unselected context $U \setminus M$, setting its downstream mutual information to zero, SMA recovers background information. The query predicate $\kappa \in \Phi$ identifies the task category. Formally, $\Phi = \{extract, count, existence, extreme, summary, query_free\}$ is a finite, mutually exclusive partition where each predicate class determines the sufficient-statistic mapping between $U \setminus M$ and Y, such as entity frequencies for counting, proper nouns for existence checks, numerical extrema for comparison, and structural deduplication for open summarization; extractive queries admit no such statistics and skip aggregation (Algorithm 1, Stage 3). Within residual headroom $B_{\text{agg}} = \max(0, B - |C_{\text{anchor}}|_{\text{tok}})$, the aggregation operator $\mathcal{A}(U \setminus M, \kappa)$ compiles a compact statistical summary A appended to the anchors, retaining the predicate's sufficient statistics up to $I(Y; U \setminus M \mid M, A, q) \leq \delta_{\text{loss}}$.

For non-extractive queries such as counting and aggregation, correct response generation relies on global properties across the entire document collection. Whereas hard deletion discards this background mass, allocating a dedicated fraction of the budget to sufficient statistics can lower the rate-distortion objective relative to pure extraction for aggregable tasks.

Proposition 4 (Conditional Aggregation Dominance Criterion)**.** Let $D_{\text{Extractive}}(B)$ denote the task distortion of pure extractive pruning under budget B. For tasks admitting task-relevant summary statistics, whenever the per-token net distortion reduction of background statistics exceeds that of marginal evicted anchors, allocating budget to summary statistics lowers downstream distortion: $D_{\text{SMA}}(B) \leq D_{\text{Extractive}}(B)$.

We set $\tau_{\text{agg}} = 0.08$ to establish a compact statistical buffer, ensuring sufficient capacity to capture global background mass without compromising the narrative density of primary anchors.

Budget Compliance via Final Truncation Guard. To strictly enforce recipient model context limits, Stage 4 applies a deterministic truncation guard, $C(x, q) = \text{TokenTruncate}(C_{\text{raw}}, B)$, ensuring that the emitted prompt satisfies $|C(x, q)|_{\text{tok}} \leq B$ under the reader's tokenizer.

3.5 Computational Complexity

Consider an input prompt x of $N = |x|_{\text{tok}}$ tokens and query q, partitioned into $n = |U|$ semantic units. Let d denote the embedding dimension of dense encoder ϕ_E. Stage 1 identifies structural markers and proposition boundaries in $O(N)$ linear time. Because each proposition unit is bounded by a maximum chunk length $l_{\text{cap}} = O(1)$ enforced during segmentation, dense encoding across all n units scales as $O(n \cdot l_{\text{cap}} \cdot d) = O(nd)$, linear in unit count n. Stage 2 computes inner products $\mathbf{s} = \cos(\phi_E(U), \phi_E(q))$ in $O(nd)$ operations, sorts candidate units in $O(n \log n)$ time, and performs linear greedy packing in $O(n)$ steps. Stage 3 extracts sufficient statistics over unselected text $U \setminus M$ via a single linear scan $O(N)$.

Combining all pipeline stages, total computational complexity is $O(Nd + N \log N)$. Because proposition units partition text with bounded average length ($n = \Theta(N)$), feature encoding and background aggregation scale linearly with document length as $O(Nd)$, while candidate ranking introduces an $O(n \log n) = O(N \log N)$ sorting step. In practice, sorting candidate scalar similarities takes less than 0.2 ms on CPU even for extended sequences ($N \approx 175$k, $n \approx 5{,}800$), leaving wall-clock overhead dominated in practice by the linear encoding pass. By encoding bounded local chunks without sequence-wide cross-attention, SMA keeps compression cost linear in N; reader-side prefill then scales with the target token budget $B \ll N$. Detailed mathematical derivations and formal proofs are compiled in Appendix A.

4 Experimental Setup

4.1 Evaluation Benchmarks

We evaluate across two widely adopted long-context benchmark suites: LongBench (Bai et al., 2024) and ZeroSCROLLS (Shaham et al., 2023). LongBench covers 16 tasks spanning six categories: single-document QA, multi-document QA, summarization, few-shot learning, synthetic retrieval probes, and code completion. For ZeroSCROLLS, following prior compression studies (Pan et al., 2024; Fei et al., 2025), we evaluate on the official validation split across 10 long-form reasoning tasks spanning summarization, question answering, multi-hop reasoning, aggregation, and chapter reordering.

4.2 Baselines and Evaluation Protocol

We compare SMA against six prominent prompt compression baselines: Selective Context (Li et al., 2023), LLMLingua (Jiang et al., 2023), LLMLingua-2 (Pan et al., 2024), LongLLMLingua (Jiang et al., 2024), EHPC (Fei et al., 2025), and R2C (Choi et al., 2024), alongside uncompressed full-context inference (NoCompression). Following standard protocols (Fei et al., 2025; Choi et al., 2024), all compressors are evaluated under unified target budgets of 2,000 and 3,000 tokens, measured with each reader's official tokenizer. For semantic encoding, SMA uses frozen BGE-M3 embeddings (Chen et al., 2025). We run downstream evaluation on two open-weight architectures, Qwen3.5-9B (Qwen Team, 2026) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), with deterministic greedy decoding for reproducibility. Complete hardware specifications, software dependencies, and runtime configurations are detailed in Appendix C.

5 Experimental Results and Analysis

5.1 Main Benchmark Comparison

Table 1 presents downstream performance across LongBench and ZeroSCROLLS under 2,000- and 3,000-token budgets. SMA achieves the highest macro-average score across all four reader-budget configurations on LongBench, outperforming the strongest baseline by up to 9.34 points, and secures the top performance on ZeroSCROLLS with Llama-3.1-8B-Instruct under both budgets. Two observations stand out. First, while expanding the token budget from 2,000 to 3,000 tokens improves performance across all methods, SMA's margin is largest under the tighter 2,000-token budget (+9.34 on Qwen3.5-9B and +7.89 on Llama-3.1-8B-Instruct), where severe compression causes destructive token fragmentation in baseline pruners while SMA preserves intact propositional anchors. Second,

Table 1: Main results on LongBench and ZeroSCROLLS validation for both readers, under the 2,000- and 3,000-token budgets. **Bolded** and underlined mark the best and second-best compression method in each column. # Tokens is the mean reader-tokenized prompt length; NoCompression evaluates uncompressed inputs across independent evaluation batches.

Method	LongBench								ZeroSCROLLS	
	SingleDoc	MultiDoc	Summ.	FewShot	Synth.	Code	Avg.	# Tokens	Avg.	# Tokens
2,000-token budget on Qwen3.5-9B										
NoCompression	47.14	57.34	25.88	68.09	85.38	14.12	49.65	11,313.7	36.90	11,209.8
Selective Context	30.70	30.88	22.34	56.15	33.46	10.46	31.75	2,047.9	13.93	2,017.6
LLMLingua	23.95	20.78	22.40	**65.13**	29.93	10.75	29.88	1,945.4	10.96	1,811.8
LLMLingua-2	36.11	35.89	23.75	48.59	20.17	10.95	30.95	2,172.4	19.58	2,382.4
LongLLMLingua	25.98	24.97	21.55	57.47	22.77	10.23	28.49	1,857.1	16.94	1,680.1
EHPC	30.27	36.59	23.04	61.33	25.00	11.41	32.91	2,199.8	17.50	2,061.6
R2C	29.12	26.73	22.75	64.36	23.63	11.66	31.22	2,206.9	**22.22**	1,972.7
SMA (Ours)	**40.74**	**49.69**	**24.49**	52.94	**73.90**	**12.35**	**42.25**	2,103.9	20.34	2,037.2
2,000-token budget on Llama-3.1-8B-Instruct										
NoCompression	43.80	43.44	29.14	39.66	79.86	12.85	40.85	10,884.5	27.99	10,930.2
Selective Context	29.12	26.40	24.44	**50.72**	27.60	11.59	29.40	2,013.3	18.26	2,015.8
LLMLingua	23.25	20.38	24.90	44.72	31.61	11.68	26.64	1,910.4	16.32	1,787.3
LLMLingua-2	34.12	32.00	25.42	43.37	33.50	**12.53**	31.05	2,267.2	19.72	2,119.3
LongLLMLingua	24.70	25.93	23.41	43.99	32.14	10.80	27.50	1,818.8	13.91	1,671.5
EHPC	31.24	30.38	25.02	37.60	41.06	11.91	29.91	2,207.2	18.49	2,082.5
R2C	28.12	27.03	25.63	45.25	31.58	11.43	29.01	2,174.0	19.33	2,052.1
SMA (Ours)	**39.87**	**44.81**	**27.12**	41.40	**69.55**	12.20	**38.94**	2,115.1	**22.96**	2,051.8
3,000-token budget on Qwen3.5-9B										
NoCompression	46.97	57.06	25.80	67.86	85.41	14.09	49.50	11,313.7	36.91	11,209.8
Selective Context	35.10	36.86	23.26	59.25	44.55	10.98	35.90	3,008.1	17.38	2,978.9
LLMLingua	30.28	25.16	23.03	64.20	31.79	11.73	32.19	2,761.9	15.87	2,658.3
LLMLingua-2	38.56	41.99	24.17	52.38	49.50	12.26	37.18	3,466.8	**28.28**	3,275.4
LongLLMLingua	30.18	28.94	22.26	59.60	28.03	10.30	31.23	2,493.3	21.05	2,391.8
EHPC	35.61	38.91	23.80	61.30	29.98	12.12	35.19	3,098.8	26.93	2,995.3
R2C	36.10	34.41	23.64	**64.81**	43.97	12.24	36.83	3,190.4	22.25	3,048.3
SMA (Ours)	**41.28**	**53.07**	**24.61**	53.12	**79.23**	**12.85**	**43.77**	3,007.9	26.16	2,974.5
3,000-token budget on Llama-3.1-8B-Instruct										
NoCompression	43.80	43.44	29.14	39.66	79.86	12.85	40.85	10,884.5	27.99	10,930.2
Selective Context	33.83	30.46	25.81	**53.69**	34.11	12.03	32.73	2,953.5	20.67	2,975.1
LLMLingua	28.40	24.49	25.45	48.61	33.16	11.97	29.44	2,700.5	18.76	2,614.5
LLMLingua-2	36.46	38.01	26.54	45.29	41.18	12.77	34.18	3,156.8	22.03	3,003.8
LongLLMLingua	27.92	29.10	24.24	48.43	32.00	10.82	29.67	2,433.3	14.52	2,369.4
EHPC	33.42	32.99	26.33	39.17	52.21	12.60	32.83	3,096.6	21.76	3,007.1
R2C	33.17	30.14	26.63	44.39	36.03	11.65	31.15	3,076.6	21.30	2,988.1
SMA (Ours)	**40.44**	**44.01**	**27.54**	41.08	**70.27**	**12.98**	**39.11**	3,008.0	**22.90**	2,980.6

the single exception on ZeroSCROLLS with Qwen3.5-9B is instructive: R2C slightly overtakes SMA (22.22 vs. 20.34) by leveraging supervised cross-attention from the reader itself. This gap marks the limit of what unsupervised query representations can extract here, consistent with the data-processing ceiling formalized in Proposition 2. Finally, empirically satisfying the aggregation dominance criterion (Proposition 4), SMA achieves its most pronounced gains on tasks requiring global counts and bounds, where hard-deletion baselines permanently lose the relevant evidence. All compressed contexts strictly satisfy $|C(x, q)|_{\text{tok}} \leq B$ via the Stage 4 tokenizer compliance guard, confirming that structured message anchors preserve task-critical information across diverse reader architectures without budget drift.

5.2 In-Depth Analysis

The empirical results in Table 1 exhibit task trade-offs rooted in compression granularity. SMA achieves its largest margins on synthetic retrieval probes (exceeding Selective Context by 40.44 points on Qwen3.5-9B) and multi-document reasoning (outperforming EHPC by 13.10 points), where tracking interconnected factual chains across passages is critical. As illustrated in Figure 2, token pruners retain isolated entities while stripping functional syntax and modifying clauses, producing an ungrounded entity inventory that misleads reader reasoning. In comparison, coarse chunk pruners rank by passage attention, where query bias can cause paragraph-level omissions. By preserving propositional units in original document order, SMA retains both modifying clauses and narrative topology. In contrast, on few-shot classification, token pruners such as LLMLingua score

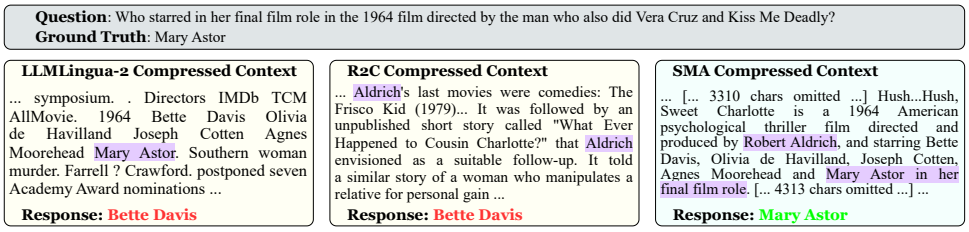


Figure 2: Qualitative comparison of compressed contexts on multi-hop QA (HotpotQA (Yang et al., 2018)). Token pruning (LLMLingua-2) excises functional syntax and disconnects attribution modifiers; coarse chunk scoring (R2C) misses the evidence paragraph; SMA preserves intact proposition units in original document order, enabling correct reader grounding.

higher than SMA, reaching 65.13 against 52.94 on Qwen3.5-9B, because stripping non-essential words packs more demonstrations into the budget, whereas complete propositions accommodate fewer exemplars. On NarrativeQA (Kočiský et al., 2018), where answer evidence has minimal lexical overlap with the query, dense query representations struggle to index semantically disparate evidence, consistent with the selection bottleneck formalized in Proposition 2. On ZeroSCROLLS with Qwen3.5-9B, R2C gains a modest advantage of 22.22 over 20.34 by exploiting supervised reader cross-attention; SMA maintains competitive accuracy while operating strictly without task supervision or internal state access. Full case transcripts are in Appendix E.

5.3 Compression Efficiency and Latency Analysis

Table 2 details empirical execution profiles on LongBench under the 2,000-token budget (Qwen3.5-9B), with the Quality-Latency trade-off of both evaluated models plotted in Figure 3.

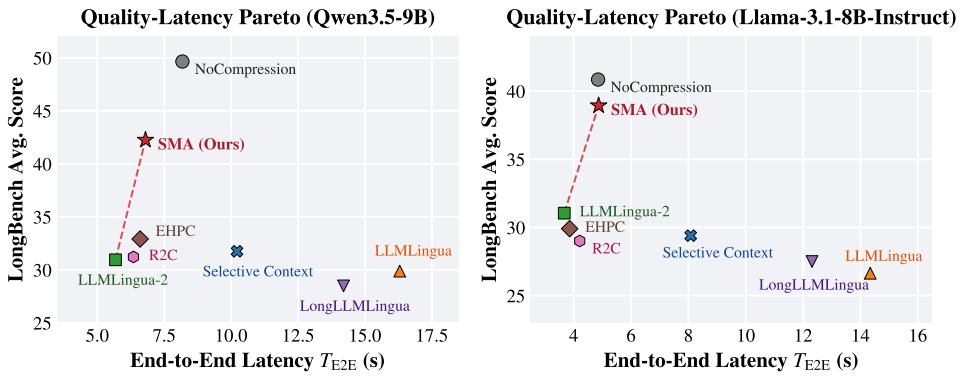
Table 2: Inference efficiency on LongBench under the 2,000-token budget (Qwen3.5-9B). T_{comp} excludes the compressor's one-time model load. Best value per column is bolded.

Method	Pre-proc.	Prefill (TTFT)		End-to-End Latency		Peak Memory	
	T_{comp} (ms)	TTFT (ms)	Speedup	T_{E2E} (ms)	Speedup	Peak KV (MB)	Reduction
NoCompression	0.0	3,075.6	1.00×	8,178.2	1.00×	2,219.6	1.0×
Selective Context	5,498.5	436.5	7.05×	10,222.1	0.80×	**192.0**	**11.6**×
LLMLingua	11,565.2	438.7	7.01×	16,299.6	0.50×	200.1	11.1×
LLMLingua-2	**721.8**	469.1	6.56×	**5,682.0**	**1.44**×	200.7	11.1×
LongLLMLingua	9,554.9	451.8	6.81×	14,198.2	0.58×	300.5	7.4×
EHPC	1,834.0	429.0	7.17×	6,600.3	1.24×	205.7	10.8×
R2C	1,430.9	488.0	6.30×	6,346.3	1.29×	278.2	8.0×
SMA (Ours)	1,963.4	**407.2**	**7.55**×	6,797.0	1.20×	202.9	10.9×

As detailed in Table 2, SMA eases the latency–quality trade-off in long-context inference. Prior iterative compressors (LLMLingua, LongLLMLingua) incur substantial scoring overhead ($T_{\text{comp}} \geq 9.5$ s), running slower than raw context end-to-end (0.50–0.58× the uncompressed speed). In contrast, SMA's single-pass propositional pipeline operates with lightweight pre-processing ($T_{\text{comp}} \approx 1.9$ s). By condensing input context, SMA unlocks the fastest prefill speed across evaluated compressors (407.2 ms, a 7.55× TTFT acceleration), converting prompt reduction into positive end-to-end speedup (1.20×) even under single-batch sequential decoding. SMA also collapses peak Key-Value cache demand by 10.9× (from 2,219.6 MB to 202.9 MB), relieving the primary memory bottleneck in concurrent serving. On the Quality-Latency Pareto frontier (Figure 3), while token classifier LLMLingua-2 compresses faster, it loses 11.30 points on LongBench; SMA stays on the high-accuracy end of the frontier across both reader families (Appendix D.2).

5.4 Ablation Studies and Mechanistic Analysis

Table 3 reports knock-out ablations across representative task archetypes, including multi-hop reasoning (HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022)) and counting (PassageCount), under a strictly enforced 2,000-token budget on Qwen3.5-9B.



(a) (b)

Figure 3: Quality-latency trade-offs on LongBench under a 2,000-token budget across (a) Qwen3.5-9B and (b) Llama-3.1-8B-Instruct. The dashed red line marks the empirical Pareto frontier across evaluated methods, with SMA occupying the highest-accuracy operating regime.

Table 3: Ablation study across representative long-context tasks (Qwen3.5-9B, 2,000-token budget). Each variant removes one core mechanism while remaining stages and the budget are held fixed. Parentheses show performance degradation relative to full SMA.

Variant	HotpotQA	MuSiQue	PassageCount	Avg.	# Tokens
SMA (ours)	55.88	40.65	53.29	49.94	1,996
w/o query-guided selection	38.45 (-17.43)	19.03 (-21.62)	49.98 (-3.31)	35.82 (-14.12)	1,997
w/o structural safeguards	55.25 (-0.63)	40.65 (-0.00)	38.22 (-15.07)	44.71 (-5.23)	1,997
w/o background aggregation	52.76 (-3.12)	37.18 (-3.47)	50.73 (-2.56)	46.89 (-3.05)	1,834

w/o query-guided selection. Replacing dense similarity ranking with random selection drops average score by 14.12 points (17.43 on HotpotQA, 21.62 on MuSiQue), verifying that dense semantic alignment drives approximation to the Oracle CMI subset (Proposition 3).

w/o structural safeguards. Omitting T_1 degrades PassageCount by 15.07 points (53.29 to 38.22). With comparable label retention (28.7% vs. 29.8%), gains stem from budget restructuring: T_1 preallocates up to $\tau_{t1}B$ to structurally load-bearing units (bounds, enumerations), shielding them from displacement by superficial high-similarity text.

w/o background aggregation. Omitting Stage 3 aggregation drops average score by 3.05 points across all three task archetypes (2.56 on PassageCount, 3.12 on HotpotQA, 3.47 on MuSiQue). On counting queries, this empirical result is consistent with Proposition 4: compact statistics preserve distributed evidence that pure extraction discards. On multi-hop reasoning, reserving headroom prevents omitted context from complete informational loss, consistent with the two-part coding framing.

We present extended stress diagnostics and limitations in Appendices D and F.

6 Conclusion

In this work, we reframed prompt compression from destructive score-and-drop filtering to a structural coding problem under the Minimum Description Length principle, and introduced a novel compression framework, namely Semantic Message Anchors (SMA). By coupling logic-preserving structural safeguards and query-salient propositions with compact background sufficient statistics, SMA mitigates token-level syntactic fragmentation and irreversible context erasure. Operating in a training-free, state-agnostic manner, SMA shows that external semantic representations can preserve essential context without accessing reader internal states. Experimental results demonstrate the effectiveness of SMA across mainstream benchmarks, highlighting the potential of select-and-aggregate prompt compression as an effective paradigm for long-context LLM inference.

AI USE STATEMENT

In this work, we used generative AI tools to assist in identifying literature gaps, formulating theoretical models, writing proofs and editing the paper for clarity and readability. We have reviewed all AI-assisted work, manually verifying all mathematical arguments against formal information-theoretic derivations (Appendices A and B) and evaluating all empirical pipelines under deterministic execution. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

For theoretical results, complete mathematical proofs, derivations, and formal explanations of all stated assumptions and formal properties are provided in Appendix A and Appendix B. For empirical results, all evaluations are conducted on publicly available benchmarks (LongBench and ZeroSCROLLS) using open-weight readers with deterministic greedy decoding. Source code, evaluation configurations, and prompt traces are included in the supplementary material to ensure full reproducibility.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2025. URL https://arxiv.org/abs/2402.03216.

Eunseong Choi, Sunkyung Lee, Minjin Choi, Jun Park, and Jongwuk Lee. From reading to compressing: Exploring the multi-document reader for prompt compression. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14734–14754. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.864. URL https://aclanthology.org/2024.findings-emnlp.864/.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. ISBN 978-0-471-24195-9. doi: 10.1002/047174882X.

Weizhi Fei, Xueyan Niu, Guoqing Xie, Yingqing Liu, Bo Bai, and Wei Han. Efficient prompt compression with evaluator heads for long-context transformer inference. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 154566–154593. Curran Associates, Inc., 2025. doi: 10.52202/085713-5171. URL https://papers.nips.cc/paper_files/paper/2025/hash/e356ed5f27885c79c7cb597bb1107c94-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Peter D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, 2007. ISBN 9780262256292. doi: 10.7551/mitpress/4643.001.0001. URL https://doi.org/10.7551/mitpress/4643.001.0001.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. ACON: Optimizing context compression for long-horizon LLM agents. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5EmOOLtH5P.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6342–6353. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.391. URL https://aclanthology.org/2023.emnlp-main.391/.

Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane K. Luke. Prompt compression with context-aware sentence encoding for fast and improved llm inference. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, volume 39, pp. 24595–24604, 2025. doi: 10.1609/aaai.v39i23.34639. URL https://ojs.aaai.org/index.php/AAAI/article/view/34639.

Alliot Nagle, Adway Girish, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. Fundamental limits of prompt compression: A rate-distortion framework for black-box language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94934–94970. Curran Associates, Inc., 2024. doi: 10.52202/079017-3009. URL https://papers.nips.cc/paper_files/paper/2024/hash/ac8fbba029dadca99d6b8c3f913d3ed6-Abstract-Conference.html.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Long Qian, Jiaqi Wei, Bingke Zhu, Yingying Chen, and Jinqiao Wang. Messages, not tokens: Grounded coresets for faithful vlm compression, 2026. URL https://arxiv.org/abs/2608.02134.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.536. URL https://aclanthology.org/2023.findings-emnlp.536/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Jing Xiong, Jianghan Shen, Chuanyang Zheng, Zhongwei Wan, Chenyang Zhao, Chiwun Yang, Fanghua Ye, Hongxia Yang, Lingpeng Kong, and Ngai Wong. ParallelComp: Parallel long-context compressor for length extrapolation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68998–69016. PMLR, 2025. URL https://proceedings.mlr.press/v267/xiong25b.html.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/bda88ed2892f5e61c9a9bf215c566913-Abstract-Conference.html.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

JungMin Yun and YoungBin Kim. IterCOMP: Reasoning-aware adaptive prompt compression for multi-hop question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33827–33840. Association for Computational Linguistics, 2026. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1559. URL https://aclanthology.org/2026.acl-long.1559/.

A Complete Mathematical Proofs and Theoretical Analysis

A.1 Two-Part MDL and Rate-Distortion Framing (Formulation 1)

In classical Minimum Description Length (MDL) theory (Grünwald, 2007), learning from data corresponds to discovering regularities that yield the most compact encoding. A two-part code splits the total description length into hypothesis complexity $L(H)$ and data given hypothesis $L(D \mid H)$.

In prompt compression for large language models, transmitting long context x to reader ϕ_{LLM} to answer query q can be cast as constructing a two-part structural code $C(x, q) = (M, A)$ over the discrete proposition units $U = \{u_1, \ldots, u_n\}$ that partition x. Here, the compressed prompt $C(x, q)$ acts as the structural hypothesis presented to the reader, while the downstream task response Y represents the target data to be predicted. The total description length under this two-part coding formulation is:

$$L_{\text{MDL}}(M, A) = L(M, A) + \lambda \cdot L_{\text{task}}(Y \mid C(x, q), q) \tag{5}$$

with terms as defined in §3. Under the target budget constraint $|C(x, q)|_{\text{tok}} \leq B$, $\lambda = \lambda(B)$ is the Lagrange multiplier (shadow price) of the budget constraint, governing the trade-off between prompt length and prediction fidelity.

Normalizing by input context length $N = |x|_{\text{tok}}$ establishes the connection to Rate-Distortion theory:

$$\frac{1}{N} L_{\text{MDL}}(M, A) = \frac{|C(x, q)|_{\text{tok}}}{N} + \lambda \cdot \frac{1}{N} L_{\text{task}}(Y \mid C(x, q), q) = R(M, A) + \lambda \cdot D(M, A) \tag{6}$$

where $R(M, A)$ is the normalized transmission rate and $D(M, A)$ is the normalized empirical task distortion.

This formulation extends rate-distortion perspectives on prompt compression (Nagle et al., 2024) by providing a constructive structural partition: allocating budget between primary verbatim anchors $M \subseteq U$ for narrative evidence and compact summary statistics $A = \mathcal{A}(U \setminus M, \kappa)$ for omitted background context (where "sufficient" is understood in the MDL sense of capturing task-relevant regularities up to residual loss δ_{loss}). In deployment under target budget cap B, SMA directly enforces $|C(x, q)|_{\text{tok}} \leq B$ by capping anchor selection at $B_{\text{sel}} = \lfloor (1 - \tau_{\text{agg}})B \rfloor$ while leaving the remaining headroom for background statistics.

A.2 Proof of Proposition 2 (Information Ceiling of Encoder-Mediated Compressors)

Proof. For any compressor in $\mathcal{C}_{\text{free}}$, candidate anchor selection $M \subseteq U$ is an encoder-mediated decision $M = g(\phi_E(U), q)$. Because the universal encoder ϕ_E has no access to downstream target labels Y, the anchor selection dependency forms the Markov chain:

$$Y \longleftrightarrow (U, q) \longleftrightarrow (\phi_E(U), q) \longleftrightarrow M \tag{7}$$

Applying the Data Processing Inequality (Cover & Thomas, 2006, Theorem 2.8.1) yields:

$$I(Y; M \mid q) \leq I(Y; \phi_E(U) \mid q) \tag{8}$$

The emitted prompt $C(x, q) = \text{Render}(U, M)$ is a deterministic extraction of verbatim textual units from raw context U. The overall end-to-end pipeline from context to reader output $\hat{Y} \sim P_{\text{LLM}}(\cdot \mid C(x, q), q)$ satisfies:

$$Y \longleftrightarrow (U, q) \longleftrightarrow C(x, q) \longleftrightarrow \hat{Y} \tag{9}$$

Applying the Data Processing Inequality to this Markov chain establishes the global operational ceiling:

$$I(Y; \hat{Y} \mid q) \leq I(Y; C(x, q) \mid q) \leq I(Y; U \mid q) \tag{10}$$

Thus, while encoder representations constrain the anchor selection mutual information $I(Y; M \mid q) \leq I(Y; \phi_E(U) \mid q)$, verbatim prompt emission is bounded globally by the raw input context $I(Y; U \mid q)$. □

A.3 Proof of Proposition 3 (Approximation Bound for TopSim-Greedy Selection)

In Algorithm 1, Stage 2A first commits up to $B_{T_1} = \lfloor \tau_{\text{tl}} B \rfloor$ tokens to safeguard set T_1, fixing anchor floor M_{guard}. Proposition 3 characterizes the subsequent Stage 2B relevance packing over candidate units $U_{\text{cand}} = U \setminus M_{\text{guard}}$ under residual budget $B_{\text{rem}} = B_{\text{sel}} - |M_{\text{guard}}|_{\text{tok}}$.

Assumption 5 (Conditional Information Additivity)**.** Conditional on query q and structural safeguard anchors M_{guard}, candidate proposition units in natural discourse provide approximately modular information about target Y, with bounded cross-proposition interaction:

$$\left| I(Y; X_S \mid X_{M_{\text{guard}}}, q) - \sum_{u \in S} I(Y; u \mid X_{M_{\text{guard}}}, q) \right| \leq \epsilon \cdot I(Y; X_S \mid X_{M_{\text{guard}}}, q) \tag{11}$$

for any candidate subset $S \subseteq U_{\text{cand}}$ (with $X_{\{u\}} \equiv u$ for singleton units), where $\epsilon \in [0, 1)$ parameterizes the degree of multi-proposition synergistic coupling.

Assumption 6 (Monotone Relevance Alignment)**.** The dense similarity score $s(u, q) = \cos(\phi_E(u), \phi_E(q))$ is monotonically aligned with marginal task information:

$$s(u_i, q) \geq s(u_j, q) \iff I(Y; u_i \mid X_{M_{\text{guard}}}, q) \geq I(Y; u_j \mid X_{M_{\text{guard}}}, q) \tag{12}$$

Property 7 (Bounded Proposition Length Variation)**.** By Propositional Segmentation (§3), atomic structures are isolated as indivisible units, while the candidate prose units entering Stage 2B are bounded by short-sentence merging (mean 24.5 tokens, median 19.0 tokens under the BGE-M3 tokenizer, with 95% of units containing ≤ 59 tokens). Within each Stage-2B candidate set, the length ratio $\kappa_l = \max_{u \in \tilde{U}} |u|_{\text{tok}} / \min_{u \in \tilde{U}} |u|_{\text{tok}}$ over the admissible set $\tilde{U}$ is bounded ($\kappa_l = 13.6$ median, 49.2 at the 95th percentile over 150 LongBench documents). The continuous relaxation of the knapsack problem sorts by density $\rho(u) = I(Y; u \mid \cdot)/|u|_{\text{tok}}$, whereas TopSim-Greedy sorts by value $s(u, q)$. For any two candidates u_i, u_j with $s(u_i, q) \geq s(u_j, q)$, bounded unit lengths constrain their densities one-sidedly:

$$\frac{I(Y; u_j \mid X_{M_{\text{guard}}}, q)}{|u_j|_{\text{tok}}} \leq \kappa_l \frac{I(Y; u_i \mid X_{M_{\text{guard}}}, q)}{|u_i|_{\text{tok}}}, \qquad s(u_i, q) \geq s(u_j, q) \tag{13}$$

Proof. Let $S^* = \arg\max_{S \subseteq \tilde{U}, \sum |u|_{\text{tok}} \leq B_{\text{rem}}} I(Y; X_S \mid X_{M_{\text{guard}}}, q)$ denote the optimal admissible subset. Under Assumption 5, the true combinatorial information objective is bounded by its additive approximation:

$$(1 - \epsilon) I(Y; X_{S^*} \mid X_{M_{\text{guard}}}, q) \leq \sum_{u \in S^*} I(Y; u \mid X_{M_{\text{guard}}}, q) \leq (1 + \epsilon) I(Y; X_{S^*} \mid X_{M_{\text{guard}}}, q) \tag{14}$$

Consider the modular knapsack problem: $\max_S \sum_{u \in S} I(Y; u \mid X_{M_{\text{guard}}}, q)$ subject to $\sum_{u \in S} |u|_{\text{tok}} \leq B_{\text{rem}}$. The continuous relaxation is solved by packing in descending order of information density $\rho(u) = I(Y; u \mid \cdot)/|u|_{\text{tok}}$.

By Assumption 6, TopSim-Greedy considers candidate units in descending value $I(Y; u \mid \cdot)$, fixed by $s(\cdot, q)$ alone. Let $m = |S^*|$ be the cardinality of the additively optimal subset; since each prose candidate carries at least $l_{\min}$ tokens, $m \leq B_{\text{rem}}/l_{\min}$. For the largest prefix length k fitting within B_{rem}, the subsequent candidate exceeds headroom ($p_k + l_{k+1} > B_{\text{rem}}$), giving $k \geq (B_{\text{rem}} - l_{\max})/l_{\max}$; with $l_{\max} = \kappa_l \, l_{\min}$, the cardinality ratio satisfies $m/k \leq \kappa_l B_{\text{rem}}/(B_{\text{rem}} - l_{\max})$. The value-sorted prefix maximizes total value among all cardinality-k subsets. Because subsequent units each carry marginal value at most $v_{(k)} \leq G_{\text{prefix}}/k$, extending the value ordering bounds the optimum. Furthermore, because Algorithm 1 continues probing subsequent smaller admissible candidates (skip-and-continue), the final packed subset M_{salient} contains this prefix, so $\sum_{u \in M_{\text{salient}}} I(u) \geq \sum_{u \in \text{prefix}} I(u)$, ensuring the bound holds:

$$\sum_{u \in M_{\text{salient}}} I(Y; u \mid X_{M_{\text{guard}}}, q) \;\geq\; \frac{1 - \epsilon}{\kappa_l} \frac{B_{\text{rem}} - l_{\max}}{B_{\text{rem}}} I(Y; X_{S^*} \mid X_{M_{\text{guard}}}, q) \tag{15}$$

Since Assumption 5 also gives $I(Y; X_{M_{\text{salient}}} \mid \cdot) \geq \sum_{u \in M_{\text{salient}}} I(Y; u \mid \cdot)/(1 + \epsilon)$, dividing the modular bound by $(1 + \epsilon)$ yields $I(Y; X_{M_{\text{salient}}} \mid \cdot) \geq \frac{1 - \epsilon}{\kappa_l (1 + \epsilon)} \frac{B_{\text{rem}} - l_{\max}}{B_{\text{rem}}} I(Y; X_{S^*} \mid \cdot)$ as stated.

The budget factor makes explicit that a tight residual budget inflates the rank gap m/k, while the bound degrades gracefully and approaches $(1-\epsilon)/(\kappa_l(1+\epsilon))$ as the residual budget grows. The denominator B_{rem} is positive by construction, since Stage 2A admits at most $\lfloor \tau_{\text{tl}} B \rfloor$ tokens against a Stage-2B budget of $\lfloor (1-\tau_{\text{agg}})B \rfloor$ with $\tau_{\text{tl}} < 1-\tau_{\text{agg}}$; candidates longer than B_{rem} simply never enter Stage 2B, so $l_{\max} \le B_{\text{rem}}$ holds automatically and no degenerate denominator arises. TopSim-Greedy runs in $O(n \log n)$ time. □

A.4 Proof of Proposition 4 (Aggregation Dominance for Aggregable Tasks)

Proof. Let B be a fixed token budget. For a fixed reader model ϕ_{LLM}, the expected downstream prediction loss on target Y given compressed prompt $C(x,q)$ decomposes as:

$$\begin{aligned}\mathbb{E}[-\log P_{\text{LLM}}(Y \mid C(x,q),q)] = H(Y \mid C(x,q),q) \\ + \mathbb{E}\big[D_{\text{KL}}\big(P^*(\cdot \mid C(x,q),q) \,\|\, P_{\text{LLM}}(\cdot \mid C(x,q),q)\big)\big]\end{aligned} \tag{16}$$

where $H(Y \mid C(x,q),q)$ is the intrinsic conditional entropy of Y given the compressed text, and D_{KL} is the reader's representation decoding mismatch.

A pure extractive pruning method selects an anchor subset M_{ext} ($|M_{\text{ext}}|_{\text{tok}} \le B$) and completely discards the residual background context $U \setminus M_{\text{ext}}$. Decomposing the conditional entropy gives:

$$H(Y \mid X_{M_{\text{ext}}}, q) = H(Y \mid U, q) + I(Y; U \setminus M_{\text{ext}} \mid X_{M_{\text{ext}}}, q) \tag{17}$$

Because hard deletion irreversibly discards $U \setminus M_{\text{ext}}$, the background mutual information $I(Y; U \setminus M_{\text{ext}} \mid X_{M_{\text{ext}}}, q)$ is permanently lost. For aggregable query classes $\kappa \in \{count, existence, extreme, summary, query_free\}$ admitting task-relevant summary statistics, this background mutual information can remain non-zero even after salient anchors have been selected.

In contrast, SMA partitions the budget into focal anchor allocation $B_{\text{sel}} = \lfloor (1-\tau_{\text{agg}})B \rfloor$ and background aggregation headroom $B_{\text{agg}} = \max(0, B - |C_{\text{anchor}}|_{\text{tok}})$. The predicate-specific aggregation operator $A = \mathcal{A}(U \setminus M, \kappa)$ compiles compact summary statistics of the unselected text, bounding residual background uncertainty:

$$H(Y \mid X_M, A, q) = H(Y \mid U, q) + \delta_{\text{loss}} \tag{18}$$

where $\delta_{\text{loss}} = I(Y; U \setminus M \mid X_M, A, q)$ represents the statistical summary loss.

Provided the instruction-tuned reader model decodes the natural-language structured summary cards effectively (keeping representation mismatch D_{KL} controlled), allocating a budget slice to A instead of packing marginal verbatim units lowers task distortion whenever the net distortion reduction per token of summary statistics exceeds that of marginal evicted anchor propositions. Specifically, let $M_{\text{ext}} \setminus M$ denote the marginal anchor units that could have been packed into the $|A|_{\text{tok}}$ budget slice under pure extractive selection. The mutual information preserved by these marginal anchors is $\Delta H_{\text{anchor}} = I(Y; X_{M_{\text{ext}} \setminus M} \mid X_M, q)$, with reader decoding discrepancy $\Delta D_{\text{KL}}^{\text{anchor}} \approx 0$ under verbatim text. In contrast, the summary-statistic block A captures global background evidence with mutual information gain $\Delta H_{\text{agg}} = I(Y; U \setminus M \mid X_M, q) - \delta_{\text{loss}}$, while inducing reader interpretation shift $\Delta D_{\text{KL}}^{\text{agg}} = \mathbb{E}[D_{\text{KL}}(P^* \,\|\, P_{\text{LLM}}(\cdot \mid X_M, A, q))] - \mathbb{E}[D_{\text{KL}}(P^* \,\|\, P_{\text{LLM}}(\cdot \mid X_{M_{\text{ext}}}, q))]$. Aggregation dominance ($D_{\text{SMA}}(B) \le D_{\text{Extractive}}(B)$) holds whenever both sides are non-degenerate ($|A|_{\text{tok}} > 0$ and $|M_{\text{ext}} \setminus M|_{\text{tok}} > 0$) and the per-token densities satisfy:

$$\frac{\Delta H_{\text{agg}} - \Delta D_{\text{KL}}^{\text{agg}}}{|A|_{\text{tok}}} > \frac{\Delta H_{\text{anchor}} - \Delta D_{\text{KL}}^{\text{anchor}}}{|M_{\text{ext}} \setminus M|_{\text{tok}}} \tag{19}$$

If either side is empty, the per-token density criterion is undefined and outside the scope of the dominance inequality; in particular, $B_{\text{agg}} = 0$ skips Stage 3 ($A = \emptyset$), but anchor selection remains capped by B_{sel}. For aggregable query classes (such as counting, existence, or numerical extrema), ΔH_{agg} scales with the aggregate mass of omitted context $U \setminus M$, whereas ΔH_{anchor} suffers from diminishing marginal relevance of low-ranked units. The criterion is therefore a testable condition for aggregation dominance across diverse task distributions. In practice, setting $\tau_{\text{agg}} = 0.08$ caps anchor selection at 92% of B, reserving the remaining headroom for predicate-specific summary statistics. □

A.5 Budget Compliance and Final Truncation Guard

In Algorithm 1, Stage 4 enforces token budget compliance via a deterministic truncation guard: $C(x,q) = \text{TokenTruncate}(C_{\text{raw}}, B) \triangleq C_{\text{raw}}[: \min(B, L_{\text{raw}})]$, where $C_{\text{raw}} = C_{\text{anchor}} \circ A$ and $L_{\text{raw}} = |C_{\text{raw}}|_{\text{tok}}$. By construction under the recipient reader's tokenizer, $|C(x,q)|_{\text{tok}} \leq B$ is satisfied identically.

By the chain rule for conditional mutual information, the information difference between C_{raw} and $C(x,q)$ is:

$$I(Y; C_{\text{raw}} \mid q) - I(Y; C(x,q) \mid q) = I(Y; C_{\text{raw}}[B:] \mid C(x,q), q) \tag{20}$$

When $L_{\text{raw}} \leq B$, the overflow slice $C_{\text{raw}}[B:] = \emptyset$ is empty, and information loss is zero. When discrete unit concatenation yields $L_{\text{raw}} > B$, the information discrepancy corresponds to the conditional mutual information of the residual tail span $C_{\text{raw}}[B:]$.

In SMA, selected proposition units are rendered in original document order $\mathcal{R}_{\text{topo}}(M)$ to protect syntactic discourse dependencies and narrative flow. Delimiters (`[... N chars omitted ...]`) mark omitted spans to maintain structural clarity. Anchor packing bounds the sum of unit lengths, whereas rendered anchors additionally carry these delimiters; Stage 3 therefore evaluates residual headroom $B_{\text{agg}} = \max(0, B - |C_{\text{anchor}}|_{\text{tok}})$ against the rendered anchor sequence itself, reserving summary cards only when sufficient capacity remains. The truncation guard accordingly operates on the rendered sequence C_{raw}, whose length is validated under the reader's tokenizer before emission.

B Hyperparameter Reparameterization and Budget Partitioning

Lemma 8 (Decoupled Hyperparameter Reparameterization)**.** Let $\alpha \in (0,1]$ be the anchor budget ratio and $\gamma \in (0,1]$ the safeguard reserve ratio, implying domain $\tau_{\text{agg}} \in [0,1)$ and $\tau_{\text{t1}} \in (0, 1-\tau_{\text{agg}}]$. Defining:

$$\tau_{\text{agg}} = 1 - \alpha, \quad \tau_{\text{t1}} = \gamma\alpha \tag{21}$$

yields a unique inverse:

$$\alpha = 1 - \tau_{\text{agg}}, \quad \gamma = \frac{\tau_{\text{t1}}}{1 - \tau_{\text{agg}}} \tag{22}$$

Proof. Since $\alpha \in (0,1]$, $1 - \tau_{\text{agg}} = \alpha \neq 0$. Substituting $\alpha = 1 - \tau_{\text{agg}}$ into $\tau_{\text{t1}} = \gamma\alpha$ directly gives $\gamma = \frac{\tau_{\text{t1}}}{1-\tau_{\text{agg}}}$. Because the forward mapping $(\alpha, \gamma) \mapsto (\tau_{\text{agg}}, \tau_{\text{t1}})$ is continuous and strictly monotonic, it is a bijection onto its image (not onto the whole square), with the inverse given above. □

Under budget $B = 2{,}000$, setting $\tau_{\text{agg}} = 0.08$ reserves exactly 160 tokens for background aggregation. We set a majority reserve ($\gamma = 0.60$) within the anchor selection budget to prioritize logic-preserving structural units, yielding $\tau_{\text{t1}} = \gamma\alpha = 0.60 \times 0.92 = 0.552$ as a fixed heuristic default without fine-tuning. Empirical unit lengths across LongBench following propositional segmentation average 24.5 tokens (median 19.0 tokens under BGE-M3, 95% containing ≤ 59 tokens), verifying that prose units are well-behaved.

C Implementation Details and System Environment

All experiments were conducted on a dedicated GPU server equipped with eight NVIDIA RTX A6000 GPUs (48 GB GDDR6), running CentOS Linux 7.9.2009. Our implementation is built on Python 3.10 with PyTorch 2.6.0 (CUDA 12.4), HuggingFace Transformers 5.14.1, and Accelerate 1.13.0, and we adopt the official evaluation harnesses of LongBench (Bai et al., 2024) and ZeroSCROLLS (Shaham et al., 2023) without modification. Semantic representations for SMA are computed with the frozen BGE-M3 dense text encoder (Chen et al., 2025) at default FP32 precision, with unit chunk length capped at 384 tokens. Downstream reader models Qwen3.5-9B (Qwen Team, 2026) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) are loaded in official configurations, and all reader evaluations employ deterministic greedy decoding (`do_sample=False`, `temperature=0.0`) to eliminate sampling variance. Token budgets ($B \in \{2{,}000, 3{,}000\}$) are enforced against each reader's official tokenizer, so that reported compression rates are directly comparable across methods.

Deterministic Rules for Algorithmic Primitives. To ensure deterministic reproducibility without auxiliary neural predictors:

- **Propositional Segmentation**: Context x is parsed via regex in linear time $O(N)$. Atomic structures (code fences ```` ```...``` ```` and markdown tables) are never split at sentence boundaries and are admitted as indivisible units, so they survive segmentation and packing whole; the budget guard of Stage 4 is the only mechanism that can cut trailing content. Natural discourse is partitioned at standard sentence boundaries (`.!?;` followed by newline) and merged for short fragments below 8 words (24.5 mean tokens per unit under BGE-M3).
- **Structural Safeguards** (T_1): Units matching structural regexes (numbered lists `^(\d+\.|\([a-z\d]\)|[A-Z]\.)`), explicit negations (`not`, `never`, `neither`, `nor`, `cannot`, `without`), or epistemic hedges (`likely`, `probably`, `possibly`, `maybe`, `might`, `could`, `suggest`, `indicate`, `potential`, `perhaps`) are routed to candidate set T_1, with hedge allocation restricted by a heuristic safeguard cap $\text{cap}_{\text{hedge}} = 0.15 \cdot B_{\text{sel}}$ to prevent speculative expressions from dominating the structural budget.
- **Predicate Classification** ($\kappa \in \Phi$): Query q is classified into six canonical archetypes via case-insensitive regex keyword matching: `count` (`how many`, `number of`, `count`), `existence` (`is there`, `does`, `did`, `whether`), `extreme` (`highest`, `lowest`, `earliest`, `most`, `least`), `summary` (`summarize`, `summary`, `overview`), and `extract` (default factoid inquiries).
- **Controlled Background Aggregation**: Stage 3 extracts frequency distributions, extreme bounds, or existence indicators across unselected context $U \setminus M$ via a deterministic linear scan, formatted into compact string cards within residual token budget B_{agg}.

D Extended Experimental Results and Diagnostics

D.1 Ultra-Long Context Stress Testing on ZeroSCROLLS Test Suite

While standard evaluation on ZeroSCROLLS is conducted on the official validation split (Pan et al., 2024; Fei et al., 2025), we additionally benchmark SMA and all six baselines across the extended ZeroSCROLLS test benchmark (Shaham et al., 2023) under the unified 2,000-token budget (Qwen3.5-9B) as an extreme-length stress test. The test suite comprises documents spanning 5k to 1.59M characters (median 31k); the BookSumSort subset, evaluated for sequence-length scaling and chapter ordering (Figure 4), ranges up to 18.2k tokens, while NarrativeQA reader inputs reach 175,000 tokens, subjecting prompt compression methods to severe scaling pressure.

Table 4: Full evaluation on the ZeroSCROLLS extended test suite (Qwen3.5-9B, 2,000-token budget). **Bold** and underline denote the best and second-best compressed results.

Method	GovReport	SummScreenFD	QMSum	SQuALITY	Qasper	NarrativeQA	MuSiQue	SpaceDigest	BookSumSort	Avg.	# Tokens
NoCompression	23.23	14.09	16.99	16.99	54.10	31.87	43.79	32.05	49.52	31.40	15,082.1
Selective Context	19.20	9.71	9.96	15.69	25.67	3.11	43.51	46.28	10.03	20.35	1,987.9
LLMLingua	18.24	9.30	10.94	15.55	12.20	3.74	25.10	0.00	4.26	11.04	1,812.9
LLMLingua-2	19.21	10.89	12.24	**16.13**	**38.17**	3.74	**44.66**	**64.09**	19.12	**25.36**	2,682.6
LongLLMLingua	16.90	8.14	11.02	13.53	33.71	5.46	28.91	55.37	3.37	19.60	1,638.2
EHPC	20.51	**12.01**	12.51	15.57	28.57	8.21	37.97	42.28	16.84	21.61	2,055.2
R2C	19.93	11.15	11.50	16.00	14.31	**13.67**	35.88	46.22	6.43	19.45	1,966.4
SMA (Ours)	**20.65**	11.11	**13.66**	15.21	22.85	5.29	29.34	37.23	**29.51**	20.54	2,030.4

Table 4 presents evaluation results across all nine ZeroSCROLLS test tasks. Evaluating under these extreme context lengths highlights distinct failure modes in existing heuristic compression methods under sequence scaling.

As plotted in Figure 4(a) on BookSumSort across scaling input lengths (3.8k–18.2k tokens), existing heuristic compressors exhibit severe budget instability. LongLLMLingua experiences aggressive budget shrinkage (collapsing from 1,857 tokens down to 1,235 tokens), while Selective Context drifts downward to approximately 1,780 tokens. In contrast, LLMLingua-2 overshoots the budget up to 2,192 tokens. SMA maintains exact budget fidelity (1,988–2,000 tokens) along the target 2,000-token threshold across all sequence lengths.

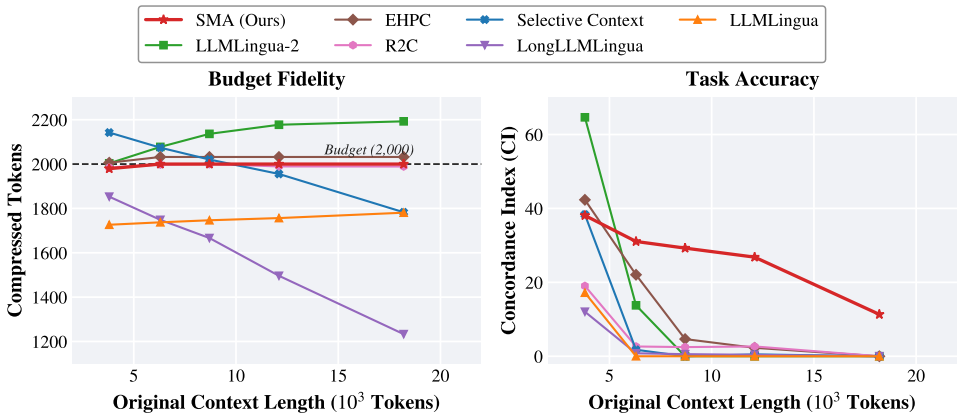


Figure 4: Stress testing under scaling input length on the ZeroSCROLLS test benchmark (BookSumSort, Qwen3.5-9B, 2,000-token budget): (a) Budget fidelity; (b) Task accuracy (Concordance Index).

Figure 4(b) shows that scaling context length triggers catastrophic degradation in baseline sequence ordering accuracy. While LLMLingua-2 achieves an initial concordance index of 64.9 on short inputs (3.8k tokens), its performance plummets to near zero ($\leq$ 0.5 CI) once context length exceeds 8.7k tokens. All six baselines suffer complete collapse across extended sequences (dropping below 5.0 CI). In contrast, SMA reliably preserves the relative topological ordering of narrative chapters, sustaining 26.8–31.0 CI across 6.3k–12.1k tokens and remaining the only method that retains meaningful ordering capability (11.5 CI) at 18.2k tokens. Across the full benchmark suite in Table 4, token classifiers like LLMLingua-2 exhibit inflated macro-averages (25.36) primarily due to unconstrained token overruns on tasks like NarrativeQA (where punctuation accumulation expands compressed length to 2,682 tokens on average, exceeding budget by $> 34\%$) and SpaceDigest, whereas chat readers drift into lengthy conversational prefaces. In comparison, SMA enforces strict budget compliance across all tasks while achieving 29.51 on BookSumSort.

D.2 Architectural Profiles and Serving Diagnostics on Llama-3.1-8B-Instruct

Figure 3(b) plots the empirical Quality-Latency Pareto frontier on LongBench under the 2,000-token budget for Llama-3.1-8B-Instruct. Llama-3.1-8B employs Grouped-Query Attention (GQA with 8 KV heads) with optimized FlashAttention kernels, reducing uncompressed prefill time to 2,186.5 ms from 3,075.6 ms on Qwen3.5-9B. Prompt compression further accelerates TTFT from 2,186.5 ms to 356.1 ms (a 6.14$\times$ prefill speedup) and reduces peak KV cache by 11.6$\times$ (from 8,572.8 MB to 738.1 MB; the larger uncompressed footprint reflects Llama-3.1's layer and hidden-width configuration under the same measurement protocol). However, because GQA already makes uncompressed prefill inexpensive, the auxiliary encoding pass ($T_{\text{comp}} \approx 1{,}947$ ms) offsets the prefill saving in single-batch evaluation, leaving end-to-end latency essentially unchanged ($T_{\text{E2E}} = 4{,}871.5$ ms vs. 4,847.5 ms, a 0.5% overhead). On such architectures, the operational value of prompt compression therefore lies in KV cache reduction and the resulting gains in concurrent serving capacity, instead of single-stream acceleration.

E Qualitative Case Studies and Transcripts

Figures 5–9 present five case studies spanning HotpotQA, NarrativeQA, PassageRetrieval-en, PassageCount, and GovReport, covering multi-hop reasoning, long narrative retrieval, needle retrieval, counting, and summarization. Each figure contrasts the original context with the SMA-compressed prompt and the reader's response.

F LIMITATIONS AND FUTURE WORK

We outline four avenues for future investigation and system refinement.

Scope of Theoretical Assumptions. Proposition 3 relies on conditional additivity (Assumption 5) and monotone alignment (Assumption 6). Strong semantic coupling from repetitive text or multi-hop dependencies inflates ϵ, weakening the bound, while abstract queries with minimal lexical overlap can degrade dense similarity alignment. Quantifying ϵ and rank correlation across diverse real-world corpora remains an open direction for future work.

Omission Framing and Delimiter Design. SMA currently inserts explicit textual delimiters (`[... N chars omitted ...]`) between non-adjacent retained units to indicate discontiguous spans. In heavily fragmented documents, delimiter tokens can encroach on residual headroom B_{agg}, occasionally suppressing Stage 3 aggregation. An open question is whether explicit omission markers are strictly necessary for modern instruction-tuned readers, or whether compact single-token markers or learned delimiter embeddings could convey discourse boundaries without consuming prompt budget.

Anchor Ordering Dynamics. SMA preserves original chronological document order $\mathcal{R}_{\text{topo}}$ to respect narrative causality and discourse structure. Exploring task-adaptive ordering strategies remains open, for instance placing query-relevant anchors near context boundaries to counteract "lost in the middle" attention biases.

Continuous Task Conditioning for Aggregation. The query classifier κ currently maps queries to discrete canonical archetypes (counting, summarization, extreme queries). Extending κ to continuous task conditioning or soft mixture-of-predicates could enable finer-grained statistical summaries for complex multi-faceted instructions.

Case Study of Multi-Hop Task in LongBench

Query: Who starred in her final film role in the 1964 film directed by the man who also did Vera Cruz and Kiss Me Deadly?
Original length: 16,041 tokens

SMA Compressed Prompt with $B = 2{,}000$

[Passage 1: Robert Aldrich Filmography & Directorial Breakthrough]
His most notable credits include Vera Cruz (1954), Kiss Me Deadly (1955), The Big Knife (1955), Autumn Leaves (1956), Attack (1956), What Ever Happened to Baby Jane? (1962), Hush...Hush, Sweet Charlotte (1964), The Flight of the Phoenix (1965), The Dirty Dozen (1967), and The Longest Yard (1974).
[... 6357 chars omitted ...]
Wellman's The Story of G.I. Joe, 1946), the "importance of communication with actors" (Joseph Losey's The Prowler, 1951), and "establishing visual empathy between camera and audience" (Charlie Chaplin's Limelight, 1952).He also worked on Pardon My Past (1945, directed by Leslie Fenton) and The Private Affairs of Bel Ami (1947, directed by Albert Lewin).
[... 2197 chars omitted ...]
Force of Evil's cinematically excessive visuals and striking sound would later appear in Aldrich's films Kiss Me Deadly and Twilight's Last Gleaming.A number of Aldrich's associates at Enterprise came under scrutiny by the HUAC in the late 1940s after Enterprise had closed its doors.

[Passage 2: Production of Hush...Hush, Sweet Charlotte and Full Cast Attribution]
Final films: The Frisco Kid and All the Marbles
Aldrich's last movies were comedies: The Frisco Kid (1979), set in the West with Gene Wilder and Harrison Ford, and ...All the Marbles (1981), set in the world of female wrestling with Peter Falk.
[... 3310 chars omitted ...]
Hush...Hush, Sweet Charlotte is a 1964 American psychological thriller film directed and produced by Robert Aldrich, and starring Bette Davis, Olivia de Havilland, Joseph Cotten, Agnes Moorehead and Mary Astor in her final film role.
[... 4313 chars omitted ...]
(1962), Aldrich wanted to make a film with similar themes for Joan Crawford and Bette Davis.
[... 770 chars omitted ...]
were cast in Hush...Hush, Sweet Charlotte: Wesley Addy, Dave Willock and Victor Buono.
[... 81 chars omitted ...]
Astor retired from acting and died in 1987.
[... 1330 chars omitted ...]
By August 4, 1964, the production had been suspended indefinitely, and the studio's insurance company insisted that Crawford be replaced, or else the film would have to be cancelled entirely.Aldrich sought several actresses to replace Crawford, including Loretta Young, and Vivien Leigh, but they were each either unable or unwilling to take the role.
[... 3771 chars omitted ...]
List of American films of 1964
[... 18 chars omitted ...]
A Kiss for Corliss (retitled Almost a Bride before its release) is a 1949 American comedy film directed by Richard Wallace, written by Howard Dimsdale, and starring David Niven and Shirley Temple.
The film, which was the last for both Wallace and Temple, was released on November 25, 1949, by United Artists.
It is a sequel to the 1945 film Kiss and Tell, also directed by Wallace and starring Temple.
⋮

Prediction (Qwen3.5-9B): Mary Astor
Ground Truth: Mary Astor

Figure 5: Case study of HotpotQA in LongBench. The purple colorbox indicates the core information to generate answers.

Case Study of Movie Script QA Task in LongBench

Query: How did Shizuko Yamamura die?
Original length: 33,866 tokens

SMA Compressed Prompt with $B = 2{,}000$

⋮
Ryuji tears the package open. He squats down on the tatami in
[... 1374 chars omitted ...]
Ryuji turns off the TV,
[... 4792 chars omitted ...]
Asakawa turns a page and then stops, frowning.
[... 2009 chars omitted ...]
Yoichi is passed out asleep on the tatami mats.
[... 713 chars omitted ...]
Her name is Yamamura Shizuko.
committed suicide forty years ago
[... 674 chars omitted ...]
Tomoko calls out to her.
[... 1192 chars omitted ...]
Asakawa freezes, her eyes wide.
[... 372 chars omitted ...]
Ryuji nods his head, lowers it sadly.
[... 23 chars omitted ...]
died, then those three others.
[... 698 chars omitted ...]
Back in the old days, the Yamamuras
[... 65 chars omitted ...]
You know, one of Shizukos
[... 303 chars omitted ...]
Why did Yamamura Shizuko commit
[... 92 chars omitted ...]
After a while she just lost it.
[... 60 chars omitted ...]
Shizuko was getting a lot of
[... 313 chars omitted ...]
Shizuko to go to Tokyo with him,
[... 294 chars omitted ...]
died at one of those demonstrations.
[... 95 chars omitted ...]
university, Ikuma just vanished,
[... 224 chars omitted ...]
OUTSIDE YAMAMURA VILLA - DAY
[... 1146 chars omitted ...]
Yamamura regards them silently, balefully.
[... 148 chars omitted ...]
a few questions, about Shizuko...
[... 26 chars omitted ...]
Its about Shizukos daughter.
⋮

Prediction (Qwen3.5-9B): She committed suicide.
Ground Truth: She committed suicide.

Figure 6: Case study of NarrativeQA in LongBench. The purple colorbox indicates the core information to generate answers.

Case Study of Needle Retrieval Task in LongBench

Query: The following is an abstract.
The text summarizes the life and struggles of a character named Sipowicz. He works to become a better person with the help of his partner Bobby Simone. He confronts his bigotry and homophobia, and deals with personal tragedies such as the death of his son and wife. He also survives prostate cancer and realizes his mistake in putting an innocent man in prison.
Please enter the number of the paragraph that the abstract is from.
Original length: 11,399 tokens

SMA Compressed Prompt with B = 2,000

⋮
"These are very important documents." I was put in a tiny cell and released on bail in the morning.
The charge was that I was living here illegally.
[... 2161 chars omitted ...]
Stay with me and I will protect you." Another story holds that Diệm said "We have always been together during these last years.
[... 130 chars omitted ...]
Paragraph 15: On his route to becoming a better man, Sipowicz struggles to overcome his bigotry with the help of his partner Bobby Simone, he and Fancy continue to have beefs but always resolve them.
He also eventually comes to terms with his homophobia, mainly due to his initially grudging friendship with precinct Police Administrative Assistant John Irvin.
With the birth of his second son, Sipowicz's life seems to be going well.
However, a series of devastating personal tragedies over the next few years arise: in May 1996, Andy Jr., who is about to start work as a policeman in Hackensack, New Jersey, is shot and killed while trying to stop a robbery.
The shooting sends Andy Sr. into an alcoholic relapse, during which Sylvia briefly throws him out of the house.
In November 1998, Bobby Simone dies of an infection caused by heart transplant complications, and less than a year later in May 1999, Sylvia is accidentally killed by the distraught father of PAA Dolores Mayo (whose killer she had been prosecuting) outside the courtroom.
[... 142 chars omitted ...]
He also survives a serious bout with prostate cancer in 1998.
With the exception of Andy Jr.'s death, however, Sipowicz remains sober in the face of all of these tragedies.
He also has to deal with the fact that he had been instrumental in putting an innocent black man in prison for 18 years for the murder of a teenager, remembering that he had no experience as a detective and deferred to a lazy veteran cop.
He is the only cop to apologize when the man is released (he learns that the perpetrator was a white man who later died of a drug overdose, and while Sipowicz and the now-retired veteran cop basically knew he'd been murdered, the end result of the case was left unresolved).
[... 2751 chars omitted ...]
Commenting on the season's exploration of issues facing working women, Paskin wrote, "what is so satisfying about its treatment of sexism...
is not the extent to which the sexism conforms to our expectations, but that the women involved do not." Emily Nussbaum of The New Yorker called season two "such a startling upgrade of the first that it begs for technological metaphors".
She said that the chemistry between Donna and Cameron "is looser, releasing the show from the burdens of its gloomy forerunners", and that the marriage between Gordon and Donna felt nuanced.
[... 224 chars omitted ...]
⋮

Prediction (Qwen3.5-9B): Paragraph 15
Ground Truth: Paragraph 15

Figure 7: Case study of PassageRetrieval-en in LongBench. The purple colorbox indicates the core information to generate answers.

Case Study of Synthetic Deduplication & Counting Task in LongBench

Query: There are some paragraphs below sourced from Wikipedia. Some of them may be duplicates. Please carefully read these paragraphs and determine how many unique paragraphs there are after removing duplicates. In other words, how many non-repeating paragraphs are there in total?
Original length: 18,175 tokens

SMA Compressed Prompt with B = 2,000

Paragraph 1: Schwarz began his professional career as a forward for Ferencvárosi TC when he was seventeen.
[... 1951 chars omitted ...]
While the Americans defeated the St. Louis Shamrocks in the 1937 National Challenge Cup, Schwarz did not play in the final game as he had broken his leg in February 1937.
[... 114 chars omitted ...]
Paragraph 2: Santin's main title was in the Italian Formula Three Championship in 1984, in which he drove a Ralt-Alfa to four wins.
From 1985 to 1988, Santin drove in the European Formula 3000 championship for many teams (Sanremo in 1985; Coloni, Sanremo, Lola Sport, Eddie Jordan Racing and Formula Team in 1986; Genoa in 1987 and Eddie Jordan Racing in 1988), without great results.
[... 219 chars omitted ...]
Paragraph 3: His composition Scaldis and Antverpia (also referred to as Allegory of the Scheldt) of 1609 (Royal Museum of Fine Arts Antwerp) is a key work of Janssens' Caravaggesque period.
[... 1228 chars omitted ...]
Paragraph 4: Schwarz began his professional career as a forward for Ferencvárosi TC when he was seventeen.
[... 1951 chars omitted ...]
While the Americans defeated the St. Louis Shamrocks in the 1937 National Challenge Cup, Schwarz did not play in the final game as he had broken his leg in February 1937.
[... 114 chars omitted ...]
Paragraph 5: His composition Scaldis and Antverpia (also referred to as Allegory of the Scheldt) of 1609 (Royal Museum of Fine Arts Antwerp) is a key work of Janssens' Caravaggesque period.
[... 1228 chars omitted ...]
⋮
[... 219 chars omitted ...]
Paragraph 7: On September 1, 2020, New York Yankees pitcher Aroldis Chapman threw a pitch that narrowly missed the head of Rays batter Mike Brosseau.
[... 589 chars omitted ...]
⋮
Paragraph 8: Sherman died on August 18, 2020 at the age of 64 of a heart attack at his home in Savannah, Georgia.
[... 839 chars omitted ...]
Paragraph 9: Sherman died on August 18, 2020 at the age of 64 of a heart attack at his home in Savannah, Georgia.
[... 839 chars omitted ...]
Paragraph 10: The Puerto Rican professional wrestling style has been influenced by several countries, beginning with the settlement of local wrestlers in New York during the 1950s Great Migration.
⋮
His better results were two sixth places in Silverstone and Vallelunga in 1986, driving a Lola T86/50.
His best start position was in Mugello in 1986, when he started from the second position, also driving a Lola T86/50.
It was commissioned by the Antwerp city magistrate to decorate the chimney in the city hall's Assembly Room where the Tw Rubens also received a commission for the same occasion.

Prediction (Qwen3.5-9B): 7
Ground Truth: 7

Figure 8: Case study of PassageCount in LongBench. The purple colorbox indicates the core information to generate answers.

Case Study of Long Context Summarization Task in ZeroSCROLLS

Query: You are given a report from the government. Write a detailed and accurate executive summary of the report.
Original length: 10,723 tokens

SMA Compressed Prompt with B = 2,000

⋮
The IDFC is authorized by statute to be a "wholly owned Government corporation ...
under the foreign policy guidance of the Secretary of State" in the executive branch.
[... 194 chars omitted ...]
development assistance objectives and foreign policy interests (§1412).
[... 871 chars omitted ...]
The BUILD Act consolidates functions currently carried out primarily by OPIC and certain elements of USAID (see Appendix).
[... 271 chars omitted ...]
They would also include the authority to take minority equity positions in investments, subject to limitations.
[... 82 chars omitted ...]
The extent to which USAID functions will be transferred to the new IDFC is less clear.
The act specifies that the Development Credit Authority (DCA), the existing legacy credit portfolio under the Urban and Environment Credit Program, and any other direct loan programs and non-DCA guarantee programs shall be transferred to the IDFC.
It also provides the authority for, but does not require, the transfer of USAID's Office of Private Capital and Microenterprise, the existing USAID-managed enterprise funds (it gives the IDFC authority to establish new funds), and the sovereign loan guarantee portfolio.
⋮
[... 1121 chars omitted ...]
exports in the case of TDA.
[... 2026 chars omitted ...]
values on governance, transparency, and environmental and social safeguards.
[... 1104 chars omitted ...]
Government Development and/or Finance Functions

[SUMMARY] What is the U Its purpose is to "mobilize and facilitate the participation of private sector c In other words, the IDFC's mission is to promote private investment in support o... [The rest of the report is omitted]

Prediction (Qwen3.5-9B): The International Development Finance Corporation (IDFC), established by the BUILD Act, is a wholly owned U.S. government corporation designed to mobilize private sector investment to support inclusive economic growth and development in less developed countries. Authorized by statute under the foreign policy guidance of the Secretary of State, the IDFC serves as a strategic alternative to state-directed investments by authoritarian regimes and strategic competitors.
⋮
The Board oversees two key officers: the Chief Risk Officer and the Chief Development Officer, both of whom report directly to the Board and are removable only by a majority vote. This structure ensures a balance between financial prudence and developmental impact. The IDFC is explicitly required to mitigate risks to U.S. taxpayers by sharing risks with the private sector and qualifying sovereign entities through cofinancing.
Ground Truth: Members of Congress and Administrations have periodically considered reorganizing the federal government's trade and development functions to advance various U.S. policy objectives. The Better Utilization of Investments Leading to Development Act of 2018 (BUILD Act), which was signed into law on October 5, 2018 (P.L. 115-254), represents a potentially major overhaul of U.S. development finance efforts.
⋮
As the IDFC is operationalized, Members of Congress may examine whether the current statutory framework allows the IDFC to balance both its mandates to support U.S. businesses in competing for overseas investment opportunities and to support development, as well as whether it enables the IDFC to respond effectively to strategic concerns especially vis-à-vis China. Congress also may consider whether to press the Administration to pursue international rules on development finance comparable to those that govern export credit financing. More broadly, the IDFC's establishment could renew legislative debate over the economic and policy benefits and costs of U.S. government activity to support private investment, and whether such activity is an effective way to promote broad U.S. foreign policy objectives.

Figure 9: Case study of GovReport in ZeroSCROLLS.