BCAR: BUDGETED CONTEXT ADAPTIVE RETENTION FOR QUERY-AWARE COMPRESSION IN LONG-RUNNING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-running agents must answer under a fixed prompt budget even after relevant turns have been retrieved. Keeping the full pool is slow and noisy, and compressing it as one flat token stream often drops the evidence that the current query needs. We formalize this decision as query-adaptive budgeted context allocation: each retrieved turn is assigned a retention ratio under a global character budget. BCAR implements this with a train-free query–turn value score, chunked water-filling quotas, and BEAVER hierarchical span selection (HSP) inside each quota. At a 10% keep ratio, BCAR reaches 64.8% on LongMemEval-S versus 57.4% for retrieve-all (+7.4 points), and outperforms the other compressors under the same retrieval and a shared character or $\approx$5800-token budget. On AgentLongBench the mean difference versus retrieve-all is $+2.8$ points at 32k and $+2.5$ at 64k; neither pooled comparison is significant, and the two 64k knowledge-intensive splits are slightly negative. Under the same quotas on LongMemEval-S, HSP with role-aware span ordering is 7.0 points above prefix head truncation. Perceived latency (prep+TTFT) drops by 9.7–18.1$\times$.

1 INTRODUCTION

Long-running agents must answer under a fixed prompt budget while histories span far beyond the native window (Wu et al., 2025; Liu et al., 2024). Even after retrieval returns a candidate pool of turns, the agent still has to decide how much of each turn to keep for the current question (Jiang et al., 2023; Hu et al., 2026). Relevance is skewed across turns, so a uniform slice of the retrieved pool need not keep the turns the current question depends on. At compression ratio $r{=}0.1$ (about 10% of retrieved characters kept), uniform packing, truncation extractive, and span selection over the whole concatenated pool all fall below retrieve-all on LongMemEval-S. BCAR exceeds retrieve-all while cutting prompt tokens by about $8\times$.

Given retrieved blocks with lengths L_i and estimated values v_i, we choose retention ratios $\rho_i \in [0, 1]$ under budget $B{=}r\sum_i L_i$ by chunked water-filling on $\sum_i v_i \log(1 + \rho_i)$ (Boyd & Vandenberghe, 2004).

Budgeted Context Adaptive Retention (BCAR) scores each turn with a train-free value $v_i{=}p_i{\cdot}d_i$ (query–turn relevance times surface density), splits the budget by that allocator, and runs BEAVER hierarchical span selection (HSP) (Hu et al., 2026) inside each quota. The continuous program is characterized by a shadow price λ; BCAR implements the quotas by chunked water-filling (Algorithm 1). Retrieval and the generator are shared across arms (Section 4); arms differ in how they spend the prompt budget on that frozen pool.

Our contributions are as follows. (1) We formulate post-retrieval packing as **Budgeted Context Allocation (BCA)**, a character-quota program under a global B, computed by chunked water-filling on a log utility. (2) We present BCAR: the train-free score $v{=}p{\cdot}d$ and BEAVER HSP inside each quota; under the same quotas, HSP with role-aware span ordering is compared with prefix head truncation (Section 4.5). (3) At $r{=}0.1$, BCAR beats retrieve-all on LongMemEval-S by 7.4 points; the other evaluated compressors fall below retrieve-all. On AgentLongBench the mean difference versus retrieve-all is $+2.8$ points at 32k and $+2.5$ at 64k; neither pooled comparison is significant, and the two 64k knowledge-intensive splits are slightly negative (Table 6).

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.

2 Related Work

Agent memory and long-context benchmarks. Agent workloads accumulate multi-session histories that exceed fixed windows (Wu et al., 2025; Packer et al., 2023; Zhong et al., 2023). MemGPT and MemoryBank decide what to write into a store across sessions. We start after a retriever has returned a candidate turn pool and ask how to spend the current prompt budget across those turns. LongMemEval (Wu et al., 2025) and AgentLongBench (Fang et al., 2026) test whether a model can find the relevant turns in agent-style histories. LongBench and RULER (Bai et al., 2024; Hsieh et al., 2024) evaluate long-input tasks more generally; we use LongMemEval and AgentLongBench as the agent-memory setting.

Prompt compression and extractive selection. LLMLingua-family methods compress toward a target budget (Jiang et al., 2023; 2024; Pan et al., 2024), and RECOMP trains extractive compressors over sentence pools (Xu et al., 2024). LongLLMLingua assigns per-context keep rates by rank and then drops tokens by perplexity. BCAR sets a character quota from $v{=}p{\cdot}d$ by chunked water-filling (marginal $v_i/(L_i{+}c_i)$) and fills that quota with HSP pages. Recent query-aware compressors refine selection inside the pool: EXIT (Hwang et al., 2025), LooComp (Do et al., 2026), EHPC (Fei et al., 2025), and FAVICOMP (Jung et al., 2025) use classifiers, leave-one-out deltas, evaluator heads, or target-model familiarity; R2C (Choi et al., 2024) prunes chunks and then sentences using Fusion-in-Decoder attention; Perception Compressor (Tang et al., 2025) allocates ratios over in-context demonstrations; ILRe (Liang et al., 2025) and cross-family speculative prefill (Upasani et al., 2026) select tokens or chunks from intermediate-layer or draft-model attention. Several of these methods add a trained or evaluator-side selector, or they act on a concatenated stream. BCAR does not train a compressor and does not query the generator while packing; BEAVER HSP (Hu et al., 2026) is the in-block engine under the assigned quotas. Retrieve-then-filter pipelines (Li et al., 2023) concatenate or sentence-prune the pool.

Budget allocation. Classical knapsack treats items as indivisible (Kellerer et al., 2004). Budgeted maximum coverage is NP-hard (Khuller et al., 1999). If a turn were an atomic item, the packer would keep or drop it. BCA lets ρ_i sit in $[0, 1]$, so a turn can be kept in part. The objective $\sum v_i \log(1+\rho_i)$ has no overlap term and is a separable water-filling program (Boyd & Vandenberghe, 2004), related to weighted logarithmic utilities (Kelly, 1997). Algorithm 1 implements that rule in 256-character steps.

3 Method: BCAR

After retrieval returns a candidate pool, BCAR spends a character budget $B{=}r\sum_i L_i$ across turns and then fills each turn's quota. The score $v{=}p{\cdot}d$ is train-free; quotas come from the water-filling program below; in-block selection is BEAVER HSP. Figure 1 is the request path.

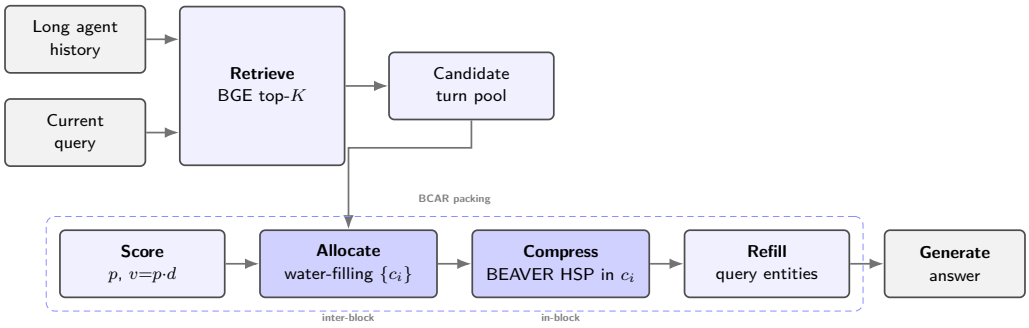

Figure 1: BCAR request path. History and the current query feed BGE retrieval into a candidate turn pool; BCAR then scores turns ($v{=}p{\cdot}d$), allocates quotas by water-filling, compresses each turn with BEAVER HSP under its quota, refills query entities, and generates. Retrieve-all skips packing; whole-pool BEAVER skips water-filling.

3.1 BUDGETED CONTEXT ALLOCATION

Text blocks $\mathcal{T}=\{t_1,\ldots,t_n\}$ have lengths L_i and values $v_i \geq 0$; the budget is $B=r\sum_i L_i$ ($r=0.1$ in the main tables), and block i receives $c_i=\rho_i L_i$ characters. We formulate the budget as a separable water-filling program (Boyd & Vandenberghe, 2004). The log utility makes the marginal $v_i/(L_i+c_i)$ fall as a turn is filled, and leftover budget then moves onto secondary turns. The continuous KKT solution is closed-form in λ; Algorithm 1 is the chunked approximation we run. Budgeted Context Allocation (BCA) is

$$\max_{\boldsymbol{\rho}} \sum_{i:v_i>0} v_i \log(1+\rho_i) \text{ s.t. } \sum_i L_i\rho_i \leq B,\ \rho_i \in [0,1]. \tag{1}$$

In characters, the marginal utility of block i is $m_i(c_i)=v_i/(L_i+c_i)$, which starts at the value per character $u_i=v_i/L_i$ and decreases in c_i. The KKT conditions give $c_i^*(\lambda)=[v_i/\lambda-L_i]_0^{L_i}$, equivalently $\rho_i^*=[u_i/\lambda-1]_0^1$, with λ chosen so the unsaturated positive-value blocks exhaust B. A block is dropped when $u_i \leq \lambda$, truncated when $\lambda < u_i < 2\lambda$, and kept whole when $u_i \geq 2\lambda$.

Under Equation 1, block j enters as soon as λ falls below u_j; by then the leading block holds $c_1=\min\big(L_1,\ L_1(u_1/u_j-1)\big)$ characters (saturating if $u_1 \geq 2u_j$). The next subsection defines $v=p\cdot d$.

Worked example. Take three blocks $(L,v)=(10000,0.8)$, $(8000,0.6)$, $(6000,0.45)$ and $B=5000$. The continuous KKT solution with $\lambda=1.85/29000$ is $(c_1,c_2,c_3)\approx(2541,1405,1054)$. The chunked allocator ($\mathrm{chunk}=256$, Algorithm 1) returns $(2560,1416,1024)$. All three blocks keep a non-zero quota. If $u_1 \geq 2u_j$, block 1 saturates before j enters: with $(L,v)=(1000,0.8)$, $(8000,0.6)$, $(6000,0.45)$ and the same B, $c_1=L_1=1000$ and the leftover 4000 is split as $(c_2,c_3)\approx(2286,1714)$. Once the leading turn is partly covered, the marginal term moves quota onto secondary turns. The LongMemEval pool is larger ($\sim$100 retrieved turns); the three-block quotas above are only a worked example.

3.2 SCORING

(1) Filter. Each retrieved turn carries a query–turn lexical overlap score: 0.8 times unigram recall of the query tokens plus 0.2 times (overlap tokens)/(turn tokens), both in $[0,1]$. Turns scoring below 0.05 are dropped (turns without a score default to 0.5 and are kept). On the main LongMemEval run, a mean of 94.5 of 100 retrieved turns enter the allocator; the rest are removed by this filter or get $v=0$.

(2) Calibrate. BGE similarities (Xiao et al., 2023) of the remaining turns are min–max normalized per query to $p_i \in [0,1]$; rank-only retrieval uses $p_i=1/(1+\mathrm{rank}_i)$. An entity boost adds 2 to p_i when the block contains a query entity of at least three characters (quoted spans, dates, identifiers, multi-digit numbers; bare single digits are skipped), so a boosted p_i may exceed 1. Min–max sets the lowest BGE score in the pool to $p=0$, so that block gets no quota unless an entity boost applies.

(3) Value. The density factor $d_i=1-\#\{\mathrm{digit},\mathrm{quote},\mathrm{comma}\}/L_i$ down-weights punctuation-heavy turns, and the block value is $v_i=p_i\cdot d_i$. Turning d off is reported in Section 4.5.

3.3 ALLOCATION

(4) Allocate across turns. Chunked water-filling implements Equation 1 by repeatedly giving $\mathrm{chunk}=256$ characters to the block with the highest $m_i(c_i)=v_i/(L_i+c_i)$ until B is spent. There is no per-block floor. Algorithm 1 puts only $v_i>0$ blocks in I, so a zero-value block receives $c_i=0$ and is not eligible for the later backfill.

Algorithm 1 Chunked marginal water-filling.

Require: lengths L_i, values v_i, budget B, chunk size
$c_i \leftarrow 0$ for all i; $I \leftarrow \{i : v_i > 0, L_i > 0\}$; $\text{rem} \leftarrow B$
while $\text{rem} > 0$ **and** $\exists\, i \in I : c_i < L_i$ **do**
  $i^\star \leftarrow \arg\max_{i \in I,\, c_i < L_i} v_i/(L_i + c_i)$
  $\Delta \leftarrow \min(\text{chunk},\, L_{i^\star} - c_{i^\star},\, \text{rem})$
  $c_{i^\star} \leftarrow c_{i^\star} + \Delta$; $\text{rem} \leftarrow \text{rem} - \Delta$
end while
return quotas c_i and $\rho_i = c_i/L_i$ for $L_i > 0$ ($\rho_i = 0$ if $L_i = 0$)

The loop spends B in 256-character steps on the current highest marginal rather than re-solving the continuous KKT system.

3.4 IN-BLOCK SELECTION

(5) Select within turns. BEAVER HSP (Hu et al., 2026) fills each quota c_i. It splits the block into 64-token pages, scores each page by a blend of semantic similarity (weight 0.7) and lexical overlap (0.3), and keeps three kinds of pages: opening anchor pages, a sliding flow window, and the highest-scoring remaining pages until c_i is reached. If HSP returns an empty span, the block falls back to its first c_i characters. If HSP returns fewer characters than c_i, that block is topped up with its own prefix up to c_i; blocks with $c_i{=}0$ are not added back. If the assembled context is still shorter than 90% of B, leftover budget is spent the same way on blocks that already have a quota. On LongMemEval and AgentLongBench 32k, span ordering inside a quota prefers user-tagged lines, and assistant and tool lines take leftover quota; AgentLongBench 64k uses plain span ordering (Appendix B). HSP scores pages inside one turn; characters move across turns only through the quotas and the later backfill.

(6) Refill entities. Query entities that HSP dropped are restored by prepending the source lines that contain them and trimming the rest of the block to stay within c_i. Refill does not enlarge B.

3.5 BASELINES

Retrieve-all uses the same retrieval with no compression. **Uniform packing** gives every positive-value block the same ρ_i and fills each quota with the same HSP, role-aware ordering, and entity refill as BCAR, without backfill. **Truncation extractive** keeps at least the first whole block, then stops before adding a block that would exceed B. On AgentLongBench 32k kf-c the first turn is $\approx$9k tokens. **Whole-pool BEAVER** runs HSP once over the concatenated pool at the same r, without turn quotas, backfill, or entity refill. **Head truncation under fixed quotas** uses the same filter, entity boost, backfill, and entity refill as BCAR; only in-block selection is a prefix.

3.6 COMPLEXITY

Water-filling costs $O(n \cdot B/\text{chunk})$ per query for pools of $n{\leq}100$ blocks and budgets of roughly 10–25k characters on LongMemEval. This arithmetic takes milliseconds, far less than the HSP forward passes, and never queries the generator. BGE retrieval is shared with retrieve-all and is not part of the packer cost; the generator sees only the packed prompt. The budget B is in characters, so packing arms share r without depending on the generator tokenizer; token-targeted compressors are aimed at about 5800 tokens (Section 4). The 256-character chunk is about one 64-token HSP page, so a one-chunk quota and a one-page selector see the same span. Packer constants are listed in Appendix B.

4 EXPERIMENTS

4.1 SETUP

Datasets. We use the official LongMemEval-S release (LME; $n{=}500$, all question types) and the non-thinking AgentLongBench (ALB) splits knowledge-free concise (kf-c), knowledge-free verbose (kf-v), knowledge-intensive concise (ki-c), and knowledge-intensive verbose (ki-v) at 32k ($n{=}200$

each, 800 pooled) and 64k (n=175 each, 700 pooled), with the official ALB scorer. ALB histories contain about four long turns per question, so top-50 retrieval keeps the full history. LongMemEval-S is the primary accuracy test (n=500, 70B judge). AgentLongBench is the transfer setting: four official splits at two history lengths, scored with the official metric that may assign half credit.

Models. Retrieval uses BGE-large-en-v1.5 on CPU (top-100 on LongMemEval, top-50 on ALB), and BEAVER runs Qwen3-4B on GPU (a CPU path is in Appendix A). The generator is Qwen3-32B-AWQ served by vLLM with T=0, seed 42, YaRN (f=2), and a maximum model length of 81,920 tokens. Every arm, including retrieve-all, reads the same frozen retrieval and uses this serving stack on the same 2×80GB host (Table 1). BGE runs on CPU so it does not share a GPU with BEAVER or the generator. LongMemEval answers are judged by Llama-3.3-70B.

Baselines and budgets. Arms differ only in the packing and compression stage. Packing arms (BCAR, uniform, truncation extractive, whole-pool BEAVER) share the character budget $B = r \sum_i L_i$. Token-targeted arms (LongLLMLingua, LLMLingua-2, RECOMP) target about 5800 tokens and land within ±10%.

LongLLMLingua uses the official compressor with a Qwen3-4B scorer, question-conditioned ranking, context reordering, coarse context-level filtering, dynamic ratio 0.3, and keep rate 0.1. LLMLingua-2 uses the official MeetingBank XLM-RoBERTa at the token level, with the context-level filter off. RECOMP uses the official Natural Questions extractive compressor zero-shot. We keep BEAVER HSP fixed as the in-block engine.

Table 1: Frozen protocol summary.

Component	Setting
LME retrieval	BGE-large-en-v1.5, top-100, CPU
ALB retrieval	BGE-large-en-v1.5, top-50, CPU
Generator	Qwen3-32B-AWQ, vLLM, YaRN f=2, 81,920 tok (all arms)
Compressor	BEAVER Qwen3-4B, GPU (CPU path: App. A)
LME judge	Llama-3.3-70B
ALB score	Official deterministic metric
Budget	r=0.1 char.; ±10% token closure (Lingua/RECOMP)

4.2 Statistical testing

Every comparison is paired on the same questions and the same frozen retrieval. On LongMemEval the 70B judge returns a binary label, so we compare each arm with retrieve-all by a continuity-corrected McNemar test on those n=500 labels. A win is a flip from retrieve-all wrong to the other arm right, and a loss is the reverse. Table 6 reports that McNemar p and a paired Wald interval on the mean per-question score difference.

The official ALB scorer may assign half credit, so Acc need not equal k/n. The ALB rows in Table 6 are a sign test on any official-score difference, with a paired Wald interval on the mean official-score difference. The 32k pool is n=800 and the 64k pool is n=700. Table 6 is the paired summary for the main LongMemEval run and those two pools.

4.3 Main results

At r=0.1, BCAR is the only evaluated compressor above retrieve-all (+7.4 points; McNemar p=0.002, wins/losses 88/51; Figure 2, Table 2). Packing arms share the r=0.1 character budget, and the token-targeted compressors land near 5800 tokens. Uniform packing is 12.8 points below retrieve-all, so equal per-turn quotas are not enough; BCAR is 20.2 points above uniform, but that gap mixes adaptive quotas and backfill and this comparison does not separate them. RECOMP, LongLLMLingua, LLMLingua-2, truncation extractive, and whole-pool BEAVER all fall below retrieve-all. Retrieve-all reads the uncompressed pool; BCAR keeps about an 8× smaller prompt (5886 vs. 48277 tokens).

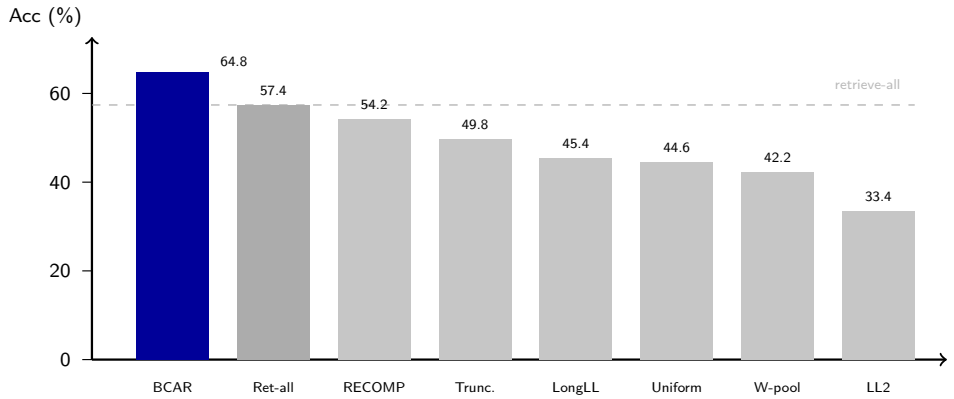

Figure 2: LongMemEval-S accuracy at r=0.1 (same retrieval). Only BCAR exceeds retrieve-all (dashed). Ret-all = retrieve-all, LongLL = LongLLMLingua, W-pool = whole-pool BEAVER, LL2 = LLMLingua-2.

Table 2: LongMemEval-S (n=500, Llama-3.3-70B). Δ: points vs. retrieve-all. Tokens are generator-tokenizer means; packing arms share the r=0.1 character budget.

Method	Acc (%)	Tokens	Δ
BCAR @r=0.1	64.8	5886	+7.4
Retrieve-all	57.4	48277	—
RECOMP-extractive	54.2	5278	−3.2
Truncation extractive	49.8	5345	−7.6
LongLLMLingua	45.4	5644	−12.0
Uniform packing	44.6	5174	−12.8
Whole-pool BEAVER	42.2	∼5290	−15.2
LLMLingua-2	33.4	5776	−24.0

Whole-pool BEAVER is the weakest packing arm. The $+7.4$ points are not spread evenly by question type (Table 3). The net is over all six official types: temporal-reasoning ($+22.6$, n=133) and knowledge-update ($+17.9$, n=78) raise it, and single-session-assistant (-16.1, n=56) lowers it.

Table 3: LongMemEval-S Acc (%) by official question type (same run as Table 2).

Type	n	BCAR	Retrieve-all	Δ
temporal-reasoning	133	56.4	33.8	+22.6
knowledge-update	78	76.9	59.0	+17.9
multi-session	133	54.9	53.4	+1.5
single-session-user	70	94.3	92.9	+1.4
single-session-preference	30	23.3	26.7	−3.3
single-session-assistant	56	76.8	92.9	−16.1

The three single-session types together (n=156) trail retrieve-all by 5.8 points. The same run split by whether the entity extractor fired: the 113 questions with a non-empty entity set are $+2.7$ points versus retrieve-all, and the 387 with an empty set are $+8.8$.

AgentLongBench is the transfer setting (Table 4). Histories have about four long turns, so top-50 retrieval keeps the full history; role-aware span ordering is on at 32k and off at 64k. On the four 32k splits BCAR is above retrieve-all (mean $+2.8$ points). At 64k the mean is $+2.5$, and the two knowledge-intensive splits sit slightly below retrieve-all. Neither pooled comparison is significant (Table 6). On the 32k kf-c multi-arm split (Table 5), the other packing arms sit at or below retrieve-all.

Table 4: AgentLongBench transfer (official Acc as a fraction; Δ in points). The official scorer may assign half credit, so Acc need not equal k/n. Role-aware span ordering is on at 32k and off at 64k (Appendix B).

Split	BCAR	Retrieve-all	Δ (pts)
32k			
kf-c	0.350	0.295	+5.5
kf-v	0.295	0.287	+0.8
ki-c	0.320	0.280	+4.0
ki-v	0.275	0.265	+1.0
Mean	**0.310**	**0.282**	**+2.8**
64k			
kf-c	0.343	0.320	+2.3
kf-v	0.397	0.303	+9.4
ki-c	0.303	0.309	-0.6
ki-v	0.309	0.320	-1.1
Mean	**0.338**	**0.313**	**+2.5**

Table 5: ALB 32k kf-c multi-arm (n=200; frozen retrieval).

Method	Acc	Tokens
BCAR	**0.350**	2943
Retrieve-all	0.295	33801
Truncation extractive	0.280	8928
Uniform packing	0.280	2553
Whole-pool BEAVER	0.275	3364

Table 6: Paired tests vs. retrieve-all. Δ and the interval are the mean score difference with a paired Wald 95% interval. LongMemEval p is McNemar on binary labels; AgentLongBench p is a sign test on official scores.

Setting	n	Acc	Δ (pts)	95% CI	p
LME: BCAR	500	.648/.574	+7.40	[+2.82,+11.98]	0.002
ALB 32k pooled	800	.310/.282	+2.81	[-0.26,+5.89]	0.094
ALB 64k pooled	700	.338/.313	+2.50	[-0.92,+5.92]	0.165

4.4 Latency

Perceived latency is compression prep plus time-to-first-token on a serial runner (Table 7); reported BCAR totals include compression. The 9.7–$18.1\times$ speedups sit against token cuts of about $8\times$ on LongMemEval and $11.5\times$ on AgentLongBench 32k kf-c; retrieval cost is unchanged. Prep is a small share of the retrieve-all totals, so the speedup is mostly a shorter prefill. A CPU compressor path is in Appendix A.

Table 7: Perceived latency (prep + TTFT p50, serial runner, same host).

Setting	Prep	Total	Retrieve-all	Speedup
ALB 32k kf-c	175ms	1.21s	18.1s	$15.0\times$
ALB 64k kf-c	389ms	3.01s	54.5s	$18.1\times$
LME n=500	867ms	3.11s	30.2s	$9.7\times$

4.5 Analysis

Leave-one-out values. On AgentLongBench kf-c 32k we use leave-one-out (LOO) values only as a pool diagnostic: v^* is the official-score drop when that block is removed (774 blocks). A zero LOO value means removing that block did not change the official score (a redundant pair can both be zero). Pools hold about four blocks per question; 84% of blocks have zero LOO value ($653/774$), and when a necessary block exists $p{\cdot}d$ ranks it first in $29/36$ questions. We did not re-run allocation with v^* in place of $v{=}p{\cdot}d$. LOO values were not computed on LongMemEval.

HSP with role-aware span ordering versus prefix head truncation. With water-filling quotas held fixed, HSP with role-aware span ordering is 64.8% on LongMemEval against 57.8% for prefix head truncation ($+7.0$ points; McNemar wins/losses $72/37$, $p{=}0.001$; Table 8). The two selectors keep the same page when a block's quota is one 64-token page (about one 256-character chunk). On AgentLongBench kf-c 32k the same role-aware HSP versus prefix contrast is $+0.5$ points (0.350 vs. 0.345). Prefix is 57.8% against retrieve-all 57.4%, using the same quotas, backfill, and refill.

Table 8: In-block selection under fixed water-filling quotas (Acc as a fraction; Δ in points). On both rows, HSP uses role-aware span ordering; head truncation is a prefix.

Setting	HSP	Head trunc.	Δ (pts)
LME-S ($n{=}500$, 70B)	0.648	0.578	$+7.0$
ALB kf-c 32k ($n{=}200$)	0.350	0.345	$+0.5$

Keep ratio. On the AgentLongBench 32k kf-c sweep, accuracy is 0.295, 0.350, 0.350, 0.320, 0.290 at $r{=}0.05, 0.1, 0.2, 0.3, 0.5$. The peak is 0.350 at both $r{=}0.1$ and $r{=}0.2$. At $r{=}0.05$ accuracy matches retrieve-all on that split (0.295); at $r{=}0.3$ and $r{=}0.5$ it falls to 0.320 and 0.290. Accuracy is therefore not monotone in r on that split. The main tables use $r{=}0.1$, which is one of the two peak settings.

Ranking robustness. Table 9 permutes the value ranking and then recomputes quotas: a tail shuffle changes which secondary turns receive a chunk. The identity column is the main-table Acc. Swapping the top two blocks costs 7.4–10.0 points. Shuffling the tail costs 1.5–4.6 points on AgentLongBench and 27.8 points on LongMemEval; a full reverse is worse on every row. AgentLongBench pools hold about four blocks, so a tail shuffle moves quota among a few secondary turns. LongMemEval starts from 100 retrieved turns (94.5 enter the allocator), so the same shuffle moves quota among many secondary turns. A tail shuffle therefore costs little on AgentLongBench and a lot on LongMemEval.

Table 9: Acc under value-order perturbations (quotas recomputed).

Benchmark	Identity	Top-2 swap	Tail shuffle	Full reverse
ALB 32k kf-c	0.350	0.250	0.335	0.245
ALB 64k kf-c	0.343	0.269	0.297	0.220
LME $n{=}500$	0.648	0.560	0.370	0.130

Density factor. Table 10 turns off d ($v{=}p$): only the value vector changes, and quotas follow. On LongMemEval this costs 2.0 points (0.628 vs. 0.648). On AgentLongBench 32k kf-c, $v{=}p$ is 0.360 and $v{=}p{\cdot}d$ is 0.350.

Table 10: Density factor on vs. off (quotas recomputed).

Value	LME $n{=}500$	ALB 32k kf-c
$v{=}p{\cdot}d$ (main tables)	0.648	0.350
$v{=}p$ ($d{\equiv}1$)	0.628	0.360

5 Conclusion

BCAR scores retrieved turns with $v{=}p{\cdot}d$, assigns a character quota by chunked water-filling, and fills each quota with HSP. On LongMemEval-S at a 10% keep ratio it is the only compressor in the comparison above retrieve-all ($+7.4$ points). Under the same quotas, HSP with role-aware span ordering is 7.0 points above prefix head truncation. On AgentLongBench the mean difference is $+2.8$ points at 32k and $+2.5$ at 64k; neither pooled comparison is significant, and the two 64k knowledge-intensive splits are slightly negative. Perceived latency (prep+TTFT) drops by 9.7–$18.1\times$, against token cuts of about $8\times$ on LongMemEval and $11.5\times$ on AgentLongBench 32k kf-c.

Limitations. The main tables use one keep ratio ($r{=}0.1$), one generator (Qwen3-32B-AWQ), and one seed, on static pre-recorded histories; other generators and in-block selectors inside the same

quotas were not run, and neither AgentLongBench pooled comparison reaches $p<0.05$. The overlap filter, backfill, entity refill, $+2$ entity boost, and role-aware span ordering were not ablated. Role-aware ordering is on for LongMemEval and AgentLongBench 32k and off for AgentLongBench 64k; on LongMemEval, user-tagged lines are kept first. Single-session-assistant trails retrieve-all by 16.1 points (n=56). On AgentLongBench 32k kf-c, v=p is 0.360 against the main-table 0.350; we did not compare the log allocator with a concentrating rule under the same quota interface on LongMemEval.

AI USE STATEMENT

Generative AI tools were used to draft sections of the manuscript, to aid and polish writing, and to retrieve related work. The authors reviewed and edited all generated text and take responsibility for the final wording, claims, and experimental results.

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)*, 2024. arXiv:2308.14508.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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*, 2024. arXiv:2410.04139.

Thao Do, Dinh Phu Tran, Khuong Vo, Seon Kwon Kim, and Daeyoung Kim. LooComp: Leverage leave-one-out strategy to encoder-only transformer for efficient query-aware context compression. *arXiv preprint arXiv:2603.09222*, 2026.

Shicheng Fang, Yuxin Wang, Xiaoran Liu, Jiahao Lu, Chuanyuan Tan, Xinchi Chen, Yining Zheng, Xuanjing Huang, and Xipeng Qiu. AgentLongBench: A controllable long benchmark for long-context agents via environment rollouts. *arXiv preprint arXiv:2601.20730*, 2026.

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*, 2025. arXiv:2501.12959.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.06654.

Zhengpei Hu, Kai Li, Dapeng Fu, Chang Zeng, Yue Li, Yuanhao Tang, and Jianqiang Huang. BEAVER: A training-free hierarchical prompt compression method via structure-aware page selection. *arXiv preprint arXiv:2603.19635*, 2026.

Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C. Park. EXIT: Context-aware extractive compression for enhancing retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2412.12559.

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, Singapore, 2023. Association for Computational Linguistics. 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, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.91.

Dongwon Jung, Qin Liu, Tenghao Huang, and Muhao Chen. Familiarity-aware evidence compression for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2409.12468.

Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Springer, Berlin, Heidelberg, 2004.

Frank Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8(1):33–37, 1997.

Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. arXiv:2310.06201.

Manlai Liang, Mandi Liu, Jiangzhou Ji, Huaijun Li, Haobo Yang, Yaohan He, and Jinlong Li. ILRe: Intermediate layer retrieval for context compression in causal language models. *arXiv preprint arXiv:2508.17892*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. arXiv:2307.03172.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

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*, 2024. arXiv:2403.12968.

Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, Hai Lin, and Hai-Tao Zheng. Perception compressor: A training-free prompt compression framework in long context scenarios. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. arXiv:2409.19272.

Shubhangi Upasani, Ravi Shanker Raju, Bo Li, Mengmeing Ji, John Long, Chen Wu, Urmish Thakker, and Guangtao Wang. Cross-family speculative prefill: Training-free long-context compression with small draft models. *arXiv preprint arXiv:2603.02631*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10813.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2023.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.04408.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

A PURE-CPU COMPRESSOR LATENCY

ALB 32k kf-c prep+TTFT p50 is 1.79s; LongMemEval is 3.52s. The compressor is PyTorch FP32 with intra-question batch 32 on the same host as the GPU runs. On that paired CPU/GPU run, $199/200$ official scores match the GPU compressor path.

B FROZEN PACKING HYPERPARAMETERS

Table 11 lists the packer constants used for all reported numbers. There is no per-block floor (floor_share=0).

Table 11: Frozen BCAR packer defaults.

Symbol	Default	Role
r	0.1	Character keep ratio on retrieved pool
chunk	256 chars	Chunked water-filling step
floor_share	0	No per-block budget floor
overlap filter τ	0.05	Drop turns with lexical overlap $<\tau$
entity boost	$+2$	If a query entity occurs in the block
entity refill	on, within c_i	Restore dropped entity lines
p calibration	min–max on pool	Maps retrieval scores to $[0, 1]$
d	surface density	Heuristic factor in $v{=}p{\cdot}d$
role-aware ordering	LME, ALB 32k	Off on ALB 64k
generator T / seed	0 / 42	vLLM decode

Entity boost. Let $E(q)$ be query entities from quoted spans, dates, identifiers, and multi-digit numbers, keeping strings of length $\geq$3 and dropping bare single digits. If block text contains any $e \in E(q)$, the calibrated score is replaced by $p_i \leftarrow p_i + 2$.

Entity refill. After HSP, any $e \in E(q)$ present in the source but missing from the snippet is restored by inserting the original source line, then trimming the body so the block length stays $\leq c_i$. Global B is not increased.

Role-aware span ordering. When a turn contains dialogue role tags, in-block assembly keeps user-tagged lines first within the same c_i, with assistant and tool lines filling leftover quota. The rule is on for LongMemEval and AgentLongBench 32k and off for AgentLongBench 64k: the only packer setting that differs by benchmark.