FORESIFT: SPECULATIVE REASONING FOR SMALL-TO-LARGE PROMPT COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt compression reduces the cost of processing long contexts with large language models. However, selecting context before target inference can discard evidence needed for reasoning. These evidence needs vary with the task and its intermediate relations. A useful selection view must expose additional evidence cues while remaining aligned with the request. We therefore combine request-conditioned and hint-conditioned attention for selection. We propose *ForeSift*, a training-free small–large compressor that obtains both views in one draft prefill. Window-level agreement weights the hint view while preserving the request as a Task Anchor. To retain local evidence structure, ForeSift selects complete sentences within an exact target-tokenizer budget. The method requires no draft decoding and invokes the unmodified target once. Across LongBench and RULER-32K, ForeSift achieves the highest compressed-method macro in all 12 combinations of two model families and three budgets. With 32K-token inputs and a 1024-token budget, the evaluated Llama pipeline achieves a 1.71× end-to-end speedup over Dense; beyond the shared 37.1 GiB target residency, its allocation is 8.3 GiB versus 30.7 GiB for Dense, a 73% reduction that includes 6.0 GiB for the draft and 2.3 GiB for inference.

1 INTRODUCTION

Long-context language models process retrieved documents, demonstrations, histories, and tool traces, increasing target prefill and KV-cache costs (Lewis et al., 2020; Liu et al., 2024). Prompt compression reduces these costs by retaining an informative subset without modifying the target (Jiang et al., 2023; Pan et al., 2024; Choi et al., 2024). We study a small–large pipeline in which a lightweight draft reads the source and constructs a compact prompt for one target call.

Existing compressors estimate importance through small-model likelihood, distilled token classification, reader attention, or sentence representations (Jiang et al., 2023; Pan et al., 2024; Choi et al., 2024; Liskavets et al., 2025). LongLLMLingua and SpecPC further condition selection on the question or draft lookahead attention (Jiang et al., 2024b; Galim et al., 2026). Despite these different signals, selection precedes the target reasoning that can expose intermediate entities, relations, or variable bindings; an omitted bridge fact or assignment is then unavailable to the target, creating the *selection-before-reasoning mismatch* illustrated in Figure 1.

Reasoning-conditioned retrieval shows how intermediate computations can guide further evidence searches (Wei et al., 2022; Trivedi et al., 2023), and our diagnostics show that draft reasoning can expose bridge evidence missed by a question-only view (Figure 2). However, generating such traces adds decoding latency, and their selection benefits vary across tasks. These observations motivate obtaining an additional view within prefill while retaining the request as an anchor that controls its influence on selection.

To this end, we propose *ForeSift*, a training-free compressor that obtains a request-conditioned Task Anchor and a fixed-hint Latent Reasoning View in one causal draft prefill. Adaptive Confirmation weights the hint view by its window-level agreement with the anchor, after which Exact-budget Sentence Commitment retains complete sentences within the target-tokenizer budget. The hint is discarded before the target receives the compressed canonical prompt, yielding one target call without draft decoding.

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.

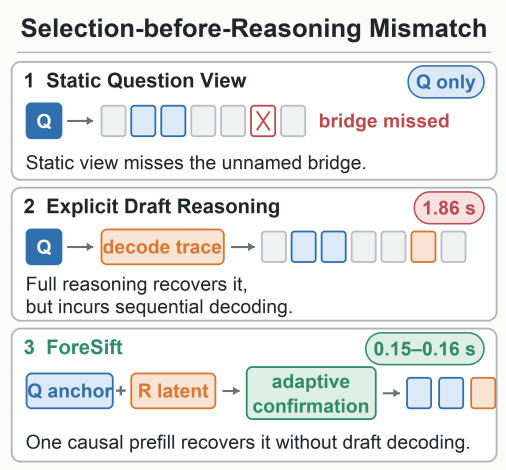

Figure 1: The selection-before-reasoning mismatch: generated draft reasoning can expose bridge evidence but adds decoding latency. ForeSift combines fixed-hint and Question attention in one prefill through adaptive confirmation.



Figure 2: View diagnostics with Qwen3-1.7B → Qwen3-32B, SpecPC, and $C_{\max} = 512$. (a) QA F1 (200 examples per dataset). (b) Judged support recall and synchronized draft latency over reasoning prefixes. (c) Per-example view winners by task. (d) Mean JS divergence from Question on 200 paired 2WikiMQA examples, using H1–H3 together or alone. Intervals are 95% bootstrap CIs over existing outputs.

Our contributions can be described as follows:

First, controlled-view, trajectory, and task-level diagnostics characterize the selection-before-reasoning mismatch.

Second, ForeSift combines one-prefill adaptive confirmation with exact-budget sentence selection.

Third, ForeSift leads compressed-method macros in all 12 LongBench and RULER-32K settings across two model pairs and three budgets, with latency and memory savings at the evaluated 32K Llama operating point.

2 Motivation

Observation: Draft reasoning can expose evidence missed by static selection views. Figure 2(a) compares views on 200 HotpotQA and 200 2WikiMQA examples, fixing the model pair, SpecPC selector, budget, target prompt, decoding, and evaluator. Question improves over task-agnostic Context Salience by 16.90 and 17.21 F1 points. Draft Answer does not consistently improve on Question, whereas Draft Reasoning performs best, adding 10.32 and 16.64 points. The comparison

captures joint differences in view content and length. The stronger Draft Reasoning results motivate additional evidence cues beyond the request.

Trajectory: Later reasoning reveals more evidence, but decoding is costly. We next examine support retention as draft reasoning progresses. Figure 2(b) evaluates the 0%, 25%, 50%, 75%, and full prefixes on the same 200 2WikiMQA examples. Qwen3-32B judges whether each 512-token compressed output from 16k input retains gold supporting facts from cited candidate spans. Judged recall rises through 39.25, 58.38, 67.88, 81.13, and 84.25, compared with 57.75 for Question. Prefixes from 50% onward exceed Question, and the full prefix gains 26.50 points (95% CI $[21.88, 31.25]$). Later prefixes thus provide more useful selection views. On 20 fixed examples, processing supplied prefixes autoregressively raises CUDA-synchronized draft latency from 0.30 to 1.86 s, while causal prefilling takes 0.15–0.16 s. This selection- and target-free control motivates obtaining the additional view through prefill, without generating a draft trace.

Constraint: Reasoning-conditioned selection is not uniformly reliable. The preferred view also depends on the task (Figure 2(c)). Draft Reasoning wins on 27.0% of pooled multi-hop QA examples versus 8.3% for Question, but loses on word aggregation (3.1% vs. 60.2%) and variable tracking (12.5% vs. 34.4%). Classification is mostly tied, and summarization is nearly split (48.0% vs. 51.5%). This task dependence motivates retaining the request view alongside additional conditioning.

Design implication: Gate a one-prefill hint view against the Task Anchor. The design must obtain additional cues without draft decoding and control their influence relative to the request. Fixed Reasoning Hints induce distinct attention views in one causal prefill (Figure 2(d)). ForeSift pairs the H1-conditioned profile with Question as the Task Anchor. Window agreement weights the hint before exact-budget sentence selection and one target call.

3 Related Work

Prompt compression. Prompt compressors estimate which source content to retain. LLMLingua uses small-LM token scores, and LongLLMLingua adds question-aware document and token compression (Jiang et al., 2023; 2024b). LLMLingua-2 distills selection into token classification (Pan et al., 2024); Selective Context and CPC use self-information or sentence representations (Li et al., 2023; Liskavets et al., 2025). R2C uses reader cross-attention, while RECOMP trains retrieval-augmented compressors (Choi et al., 2024; Xu et al., 2024). ForeSift combines a task anchor with an additional hint-conditioned view.

Draft-based approximate inference. Speculative decoding uses a draft to propose tokens for target verification (Leviathan et al., 2023). Speculative Prefill and SpecPC use draft attention to estimate context importance (Liu et al., 2025; Galim et al., 2026); SnapKV similarly uses pre-generation attention (Li et al., 2024). SpecPC pools trailing-prompt and lookahead attention before fixed-token selection. ForeSift uses separate request and hint profiles, adaptive fusion, and sentence selection to construct the target input. Its speculative step selects evidence before target inference, rather than accepting tokens during decoding.

Reasoning-conditioned retrieval. Chain-of-thought exposes intermediate computations (Wei et al., 2022); IRCoT uses them for searches, and Self-RAG learns to retrieve and critique evidence (Trivedi et al., 2023; Asai et al., 2024). These approaches reveal evolving evidence needs through sequential generation and retrieval.

Position of this work. ForeSift couples hint-conditioned selection with asymmetric confirmation: the request supplies the Task Anchor, and window-level agreement controls the hint update. It extracts supplied context in one draft prefill, without external retrieval or draft decoding.

4 Method

Overall. ForeSift comprises **Prompt Construction**, **Latent Reasoning View**, **Adaptive Confirmation**, and **Exact-budget Sentence Commitment** (Figure 3). One draft prefill produces two context profiles; confirmation combines them, and sentence selection constructs the target prompt.

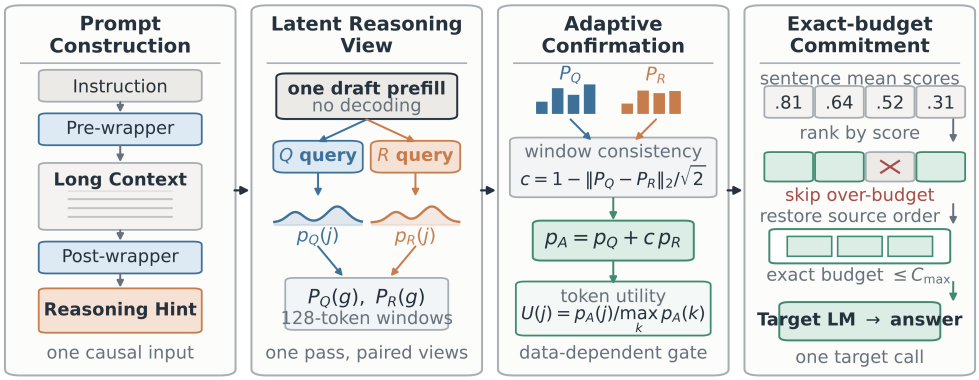

Figure 3: ForeSift obtains request- and hint-conditioned attention in one draft prefill. Window agreement weights the hint before exact-budget sentence selection; the target receives the canonical prompt.

Prompt construction. The draft input orders the frozen instruction d, pre-wrapper w_{pre}, context $\mathbf{x} = (x_1, \ldots, x_L)$, post-wrapper w_{post} containing request q, and the Reasoning Hint, exposing both query spans to the source. Appendix A gives the fixed instruction and templates. H1 asks for sentences important to answering the request; Table 3 compares the templates.

Latent Reasoning View. The Latent Reasoning View supplies an additional selection signal through attention induced by fixed hint tokens. From one small-LM prefill, we derive $\boldsymbol{p_Q}$ from the request span Q and $\boldsymbol{p_R}$ from the complete hint span R, including its two literal <> placeholders. For either span $v \in \{Q, R\}$, we collect causal attention at layers $\ell \geq 8$ and score context position j following SpecPC (Galim et al., 2026):

$$
\begin{aligned}
a_v(j) &= \max_{\ell \geq 8} \max_{m} \max_{r \in v} \alpha_v(r) A^{(\ell,m)}_{r,j}, \\
p_v(j) &= \frac{a_v(j)}{\sum_{k=1}^{L} a_v(k) + \epsilon}, \qquad v \in \{Q, R\},
\end{aligned}
\tag{1}
$$

Here $A^{(\ell,m)}$ is causal attention after the original RoPE transformation, and $\alpha_v(r)$ increases linearly from 0 to 1 across the query span. We retain context-key columns and maximize over query positions, heads, and layers, without neighborhood pooling. Normalization makes the two token-level profiles comparable. Non-overlapping $W = 128$-token windows aggregate each profile as $P_v(g) = \sum_{j:g(j)=g} p_v(j)$ for confirmation.

Adaptive Confirmation. Adaptive Confirmation controls how strongly the hint-conditioned view updates the request-conditioned Task Anchor. We measure agreement between the paired window distributions as

$$
c = 1 - \frac{\|P_Q - P_R\|_2}{\sqrt{2}}. \tag{2}
$$

For normalized profiles, $c \in [0, 1]$: identical window distributions give $c = 1$, while concentration on different windows gives $c = 0$; ϵ makes normalization approximate in implementation. The request contribution is retained at full weight, and greater agreement admits a larger hint contribution:

$$
p_A = p_Q + c\,p_R. \tag{3}
$$

Raw-MaxNorm converts this combined profile to token utility:

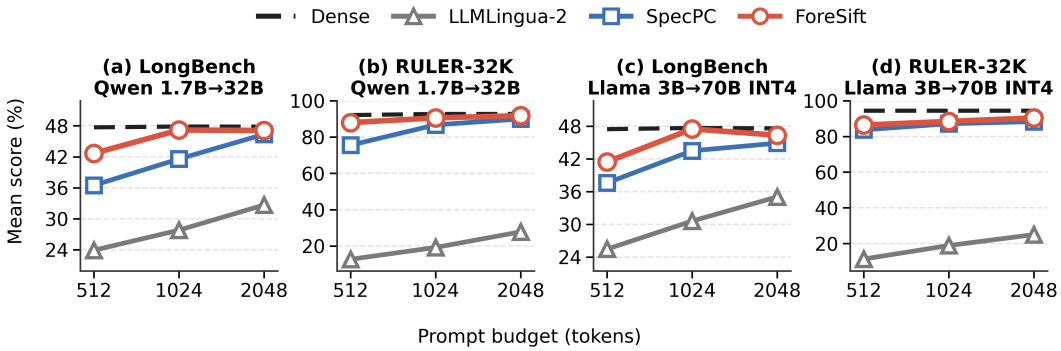

Figure 4: Task macro versus target-prompt budget for Qwen3-1.7B → Qwen3-32B and Llama-3.2-3B-Instruct → Llama-3.1-70B-Instruct-AWQ-INT4 on LongBench and RULER-32K. Qwen RULER uses 12/13/13 tasks at $C_{\max} = 512/1024/2048$; Dense is re-scored on each budget-specific UID intersection.

$$U(j) = \frac{p_A(j)}{\max_k p_A(k) + \epsilon}. \tag{4}$$

Dividing Equation (3) by $1 + c$ preserves selection rankings; fixed ϵ changes only the common scale. The bounded update requires no tuned fusion coefficient, iterative memory, or target feedback. Agreement serves as a control heuristic rather than calibrated confidence: both views can favor irrelevant evidence, and useful complementary evidence can differ from the anchor.

Exact-budget sentence commitment. We select complete sentences to retain the local structure around high-scoring evidence. Sentence units include passage markers, titles, and attached headings (Appendix A). Units receive mean utility $\bar{U}(s) = |s|^{-1} \sum_{j \in s} U(j)$ and are visited by decreasing score, with ties broken by source position. A unit is admitted only when the fully rendered payload satisfies

$$|w_{\text{pre}} \oplus \mathbf{x}_{\mathcal{S}} \oplus w_{\text{post}}|_{\mathcal{T}_t} \leq C_{\max}, \tag{5}$$

Here wrappers remain unchanged, and $\mathcal{T}_t$ is the target tokenizer. An over-budget unit is skipped so that a shorter later unit can use the remaining capacity. Retained sentences return to source order between the canonical wrappers; d and the Reasoning Hint are discarded before the single target call. The exact budget bounds the pre-chat payload, and indivisible sentences can leave unused capacity.

5 Experiments

5.1 Experimental Setup

We evaluate whether ForeSift preserves task performance under tight budgets and whether target savings outweigh compression overhead. The model pairs are Qwen3-1.7B → Qwen3-32B (Qwen Team, 2025) and Llama-3.2-3B-Instruct → Llama-3.1-70B-Instruct-AWQ-INT4 (Dubey et al., 2024). LongBench supplies 14 English non-code tasks; RULER supplies 13 tasks at 32K (Bai et al., 2024; Hsieh et al., 2024). The full pre-chat payload must fit $C_{\max} \in \{512, 1024, 2048\}$ target tokens; target decoding is greedy with task-specific output limits.

We compare with released SpecPC (Galim et al., 2026), exact-budget LLMLingua-2 (Pan et al., 2024), and uncompressed Dense. ForeSift and SpecPC share the draft checkpoint and do not query the target during compression. We report official metrics and unweighted task macros: group columns average tasks within a group, while All averages individual tasks. Scores use four-method UID intersections without imputing missing outputs; the table captions specify their variation by budget. Each main-table configuration has one frozen evaluation with greedy decoding and seed

Table 1: LongBench task-group macros (%) on the four-method UID intersection at each target-prompt budget. The intersections contain 2,524/2,634/2,707 Qwen examples and 2,543/2,638/2,709 Llama examples at $C_{\max} = 512/1024/2048$. Dense is budget-exempt but re-scored on each intersection; bold marks the best compressed method.

		Qwen3-1.7B → Qwen3-32B					
$C_{\max}$	Method	SingleQA↑	MultiQA↑	Summ.↑	Few-shot↑	Synthetic↑	All↑
512	Dense	42.30	51.33	27.34	62.79	58.25	47.70
	LLMLingua-2	21.23	22.05	22.45	43.84	3.01	23.91
	SpecPC	37.20	39.68	24.11	**64.93**	6.46	36.48
	ForeSift	**40.06**	**50.01**	**25.05**	61.93	**33.12**	**42.67**
1024	Dense	42.30	51.33	27.34	63.53	58.25	47.86
	LLMLingua-2	28.79	27.99	23.98	45.63	5.00	27.80
	SpecPC	39.55	48.62	25.96	**65.89**	21.00	41.58
	ForeSift	**41.51**	**53.87**	**26.84**	63.11	**52.25**	**47.18**
2048	Dense	42.30	51.33	27.34	63.43	58.25	47.83
	LLMLingua-2	35.26	37.04	25.49	49.49	7.75	32.67
	SpecPC	40.50	53.64	26.87	**65.98**	44.00	46.35
	ForeSift	**40.59**	**56.38**	**27.39**	59.64	**53.75**	**47.11**
		Llama-3.2-3B → Llama-3.1-70B (AWQ-INT4)					
512	Dense	46.10	59.27	28.61	48.16	58.97	47.45
	LLMLingua-2	22.46	30.50	22.70	41.31	2.79	25.46
	SpecPC	41.48	51.65	24.01	50.78	11.33	37.60
	ForeSift	**42.31**	**53.43**	**25.77**	**53.91**	**26.94**	**41.44**
1024	Dense	46.10	59.27	28.61	49.13	59.00	47.67
	LLMLingua-2	31.29	36.28	24.52	43.19	11.25	30.60
	SpecPC	**45.13**	58.69	26.30	50.35	33.75	43.49
	ForeSift	44.11	**60.46**	**27.44**	**53.26**	**54.50**	**47.49**
2048	Dense	46.10	59.27	28.61	48.76	59.00	47.59
	LLMLingua-2	34.80	43.83	25.81	**48.40**	16.00	35.04
	SpecPC	44.46	59.62	27.69	44.51	49.75	44.88
	ForeSift	**45.49**	**60.35**	**28.19**	46.04	**54.25**	**46.34**

42. The main tables report these scores without uncertainty intervals; bootstrap intervals elsewhere resample existing examples or task means and quantify sample uncertainty, not variation across independent runs.

5.2 Overall Effectiveness

Main finding. ForeSift achieves the highest compressed-method macro in all 12 benchmark–model–budget combinations (Tables 1 and 2). The comparison covers complete pipelines, including scoring, selection granularity, and budget enforcement. The ablations hold sentence selection fixed; isolating scoring gains over SpecPC would require a selector-matched comparison.

Budget sensitivity on LongBench. At $C_{\max} = 512$, ForeSift leads SpecPC by 6.19 points with Qwen and 3.84 with Llama (Figure 4(a,c)). At 1024 tokens, its overall macro approaches Dense: 47.18 versus 47.86 for Qwen and 47.49 versus 47.67 for Llama. The ForeSift–SpecPC margin narrows to 0.76 and 1.46 points at 2048 tokens. The larger tight-budget margins suggest that evidence selection matters more when capacity is limited.

The gain is structured, not uniform. For Qwen at 512 tokens, ForeSift improves over SpecPC by 26.66 points on synthetic tasks and 10.33 on multi-document QA, but trails it by 3.00 on few-shot learning. Dispersed entities, relations, and state drive these tasks; demonstrations can instead favor a literal prompt view. Llama shows the same broad pattern, including a 15.61-point gain on synthetic tasks at 512 tokens.

Table 2: RULER-32K task-group macros (%) on the four-method UID intersection at each target-prompt budget; VT denotes variable tracing. Qwen at $C_{\max} = 512$ uses 12 tasks and 1,191 UIDs because the LLMLingua-2 NIAH Multikey-2 cell is incomplete; all other blocks use 13 tasks and 1,300 UIDs. Dense is budget-exempt but re-scored on each intersection; bold marks the best compressed method.

		Qwen3-1.7B → Qwen3-32B				
$C_{\max}$	Method	Needle↑	VT↑	Aggreg.↑	QA↑	All↑
512	Dense	99.86	100.00	81.47	71.53	92.08
	LLMLingua-2	6.36	0.60	36.78	16.68	12.67
	SpecPC	91.39	28.60	46.77	72.15	75.51
	ForeSift	**100.00**	**99.60**	**52.55**	**75.15**	**87.92**
1024	Dense	99.75	100.00	81.47	72.00	92.69
	LLMLingua-2	16.06	0.60	40.60	19.50	19.18
	SpecPC	99.75	71.20	54.35	74.50	86.68
	ForeSift	**99.88**	**99.80**	**62.20**	**77.00**	**90.55**
2048	Dense	99.75	100.00	81.47	72.00	92.69
	LLMLingua-2	27.44	0.80	47.52	23.00	27.79
	SpecPC	**100.00**	92.40	61.52	**77.00**	89.96
	ForeSift	**100.00**	**99.80**	**71.53**	75.00	**91.76**
		Llama-3.2-3B → Llama-3.1-70B (AWQ-INT4)				
512	Dense	99.41	99.20	95.83	70.50	94.39
	LLMLingua-2	3.41	0.00	35.38	24.50	11.31
	SpecPC	99.53	50.20	46.25	74.50	83.69
	ForeSift	**100.00**	**75.80**	**46.45**	**77.00**	**86.36**
1024	Dense	99.41	99.20	95.83	70.50	94.39
	LLMLingua-2	14.13	1.00	41.72	24.00	18.88
	SpecPC	**100.00**	68.80	54.73	**77.00**	87.10
	ForeSift	**100.00**	**80.00**	**58.73**	76.00	**88.42**
2048	Dense	99.41	99.20	95.83	70.50	94.39
	LLMLingua-2	21.03	1.80	47.43	30.00	24.99
	SpecPC	**100.00**	73.80	62.47	74.50	88.29
	ForeSift	**100.00**	**80.40**	**69.22**	**78.00**	**90.37**

RULER localizes the gain by task type. On RULER-32K, ForeSift leads SpecPC by 12.41/3.87/1.80 macro points for Qwen and 2.67/1.32/2.08 for Llama as the budget increases (Figure 4(b,d)). Needle retrieval is high and quickly reaches ceiling, whereas variable tracing gains are 71.00/28.60/7.40 points for Qwen and 25.60/11.20/6.60 for Llama. ForeSift also leads aggregation at every budget. These improvements are consistent with retaining the state-bearing context needed to trace assignments and combine dispersed evidence.

Cross-model consistency. The macro advantage holds for both evaluated dense-architecture model pairs. An MoE target could use the same compressed-prompt interface with its tokenizer and wrappers, while an MoE draft would also require accessible attention and validation of the scoring settings. Expert routing, residency, and communication (Jiang et al., 2024a) introduce additional quality and cost considerations for that extension.

Failure modes. ForeSift trails SpecPC by 3.00 points on Qwen few-shot learning at 512 tokens, 2.00 on Qwen RULER QA at 2048, and 1.02 on Llama SingleQA at 1024. Shifts away from literal cues or demonstrations may contribute to these reversals. At 2048 tokens, ForeSift still trails Dense on Qwen RULER aggregation (71.53 vs. 81.47), Llama aggregation (69.22 vs. 95.83), and Llama variable tracing (80.40 vs. 99.20). An omitted assignment can break a tracing chain, while aggregation may require evidence beyond the retained sentences. Sentence and wrapper overhead, together with unused budget capacity, further limit how much source content fits. These factors may explain residual gaps; changing UID intersections also affect cross-budget comparisons.

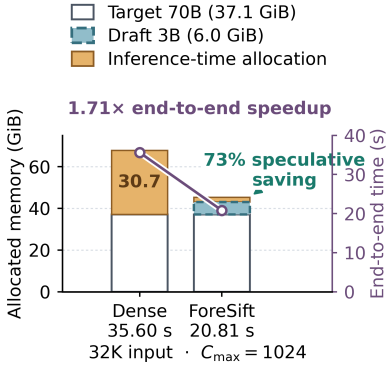

Figure 5: The Llama pair at 32K, $C_{\max} = 1024$, over 32 frozen RULER prompts in pure Transformers. Bars include shared target residency; the 73% saving excludes it.

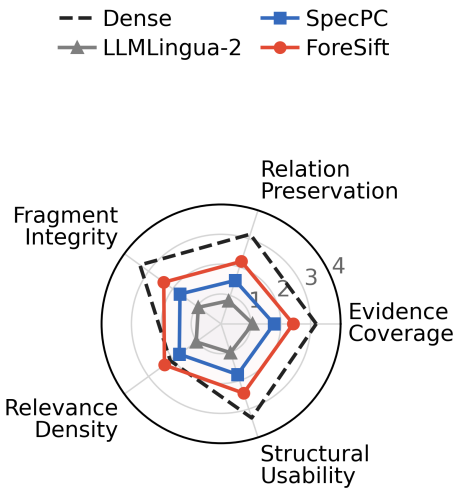

Figure 6: Retained-prompt audit on 70 LongBench prompts for the Qwen pair at $C_{\max} = 512$. A Llama-3.1-70B-Instruct-AWQ-INT4 judge scores each axis from 0 to 4.

5.3 Efficiency and Memory Trade-off

We measure the Llama pair on 32 frozen 32K RULER prompts at $C_{\max} = 1024$ in pure Transformers (Figure 5). ForeSift reduces paired-median end-to-end time from 35.60 to 20.81 s, giving a 1.71× speedup over Dense. The shared target residency is 37.1 GiB for both methods. Excluding it, Dense allocates 30.7 GiB, whereas ForeSift allocates 8.3 GiB, comprising 6.0 GiB for draft residency and 2.3 GiB for inference; this is a 73% reduction relative to Dense. Including target residency gives approximately 67.8 versus 45.4 GiB, a 33% reduction.

Timing includes compression and target generation, with batch size 1, three warmups, and 64 output tokens (Appendix A). Target savings exceed draft overhead here; shorter contexts or different serving regimes can change that balance.

5.4 Why Does ForeSift Help?

We audit 70 LongBench prompts for the Qwen pair at $C_{\max} = 512$ to examine retained evidence. We select five score-blind matched examples per task (Appendix A). A zero-temperature Llama-3.1-70B-Instruct-AWQ-INT4 judge scores evidence coverage, relation preservation, fragment integrity, relevance density, and structural usability. Compressed prompts are blinded and deterministically permuted; 10,000 task-mean bootstrap resamples provide intervals.

ForeSift exceeds SpecPC by 0.61–0.67 points on every LongBench dimension and LLMLingua-2 by 1.31–1.43 (Figure 6). Although below Dense on preservation, it has higher relevance density (2.34 versus 2.10). These scores suggest better retention of useful local evidence. On RULER, audit gaps of only 0.05–0.08 points per axis accompany large variable-tracing gains, suggesting limited sensitivity to exact state transitions.

LLMLingua-2 performs differently across evidence requirements. For Qwen at 512 tokens, it scores 22.45 versus Dense's 27.34 on summarization, but 22.05 versus 51.33 on MultiQA and 3.01 versus 58.25 on Synthetic. Repeated themes can support summaries despite omissions, whereas losing a bridge fact, identifier, or assignment can prevent an exact answer. Task-agnostic token retention may preserve summary content while losing such decisive evidence; metrics also differ in sensitivity to omissions. ForeSift combines task-conditioned scoring with complete sentences, consistent with the evidence and relation-preservation gains in the audit.

Table 3: Hint-template ablation ($C_{\max} = 512$, $N = 497$). Scores $\times 100$; Δ is relative to Question.

Task	Question	+H1	+H2	+H3
Qasper	42.72	**44.90**	39.47	42.72
HotpotQA	66.55	**68.80**	64.97	63.44
GovReport	28.32	28.92	28.77	**29.10**
TREC	40.00	58.75	59.50	**62.50**
PassageRet.	47.00	**54.00**	46.00	43.00
Macro-5	44.92	**51.07**	47.74	48.15
Δ	–	**+6.16**	+2.83	+3.23

Table 4: Confirmation ablation ($C_{\max} = 512$, $N = 497$). Scores $\times 100$; Δ is relative to Question.

Task	Question	Reasoning	Fixed	ForeSift
Qasper	43.12	**46.70**	44.35	44.77
HotpotQA	65.52	63.78	68.13	**68.80**
GovReport	28.34	28.84	28.87	**28.92**
TREC	**60.50**	41.00	54.70	**60.50**
PassageRet.	47.00	**61.00**	54.00	54.00
Macro-5	48.90	48.26	50.01	**51.40**
Δ	–	-0.63	+1.11	**+2.50**

5.5 Ablation and Interpretation

Setup. Both ablations use the Qwen pair at $C_{\max} = 512$ on seed-42 subsets of Qasper, HotpotQA, GovReport, TREC, and Passage Retrieval-en. Each task has 100 examples except GovReport with 97 ($N = 497$); we report official metrics and unweighted Macro-5. The tables use separate frozen references and outputs; comparisons are within-table (Appendix A).

Reasoning hint ablation. We compare Question-only with three separately inserted templates sharing two literal placeholders: **H1** locates sentences important for answering q; **H2** identifies essential supporting tokens; **H3** identifies missing relations or information. H2/H3 reference preceding steps without their outputs, so the comparison tests isolated templates (Appendix A).

Confirmation ablation. One Question+H1 prefill supplies four scoring variants: **Question-only** (p_Q), **Reasoning-only** (p_R), **Fixed fusion** ($p_Q + p_R$), and **ForeSift** ($p_Q + c\,p_R$). Sentence selection, the retained question, and thinking-disabled greedy decoding are identical across variants. We form 95% CIs from 10,000 paired within-task bootstrap resamples (Appendix A).

Results. H1 improves all five tasks and raises Macro-5 by 6.16 points; H2 and H3 gain 2.83 and 3.23 mainly through TREC (Table 3). These results favor the default sentence-localization template.

In the confirmation ablation, Reasoning-only loses 19.50 points on TREC and 0.63 on Macro-5 relative to Question-only. Fixed fusion gains 1.11 points, while ForeSift reaches 51.40 ($+2.50$; Table 4). ForeSift improves over Reasoning-only by 3.14 points (95% CI $[0.66, 5.62]$) and over Fixed fusion by 1.39 ($[0.52, 2.41]$). These comparisons support controlling the hint contribution while retaining the request anchor. The Question-only interval crosses zero ($[-0.10, 5.13]$), leaving that comparison inconclusive. Isolating reasoning-specific wording would further require a non-reasoning control matched for length, position, placeholders, and selection objective.

6 Conclusion

ForeSift addresses the selection-before-reasoning mismatch through request-anchored hint conditioning and exact-budget sentence selection, requiring one draft prefill and one target call. ForeSift achieves the highest compressed-method macro in all 12 LongBench and RULER-32K settings. At the 32K Llama operating point, it provides a $1.71\times$ end-to-end speedup over Dense. Beyond the shared 37.1 GiB target residency, ForeSift allocates 8.3 GiB versus Dense's 30.7 GiB, a 73% reduction that includes 6.0 GiB for the draft and 2.3 GiB for inference.