BEYOND VISUAL TOKEN COUNT: TARGET INCLUSION AND READER RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Rendering text as images can fit more source information into a model's input budget, but token savings alone do not establish whether that information remains usable. We distinguish target inclusion—whether a queried fact reaches the reader—from conditional recovery—whether the reader can recover a fact that is present—and study both with three fixed vision-language models. On fresh held-out cases, changing image geometry improves exact identifier recovery across all three readers while preserving source content, native token counts, and input-ID sequences. In a separate finite-budget comparison on 150 fresh long-source cases, GLM answers 126 queries correctly from full-source optical input, compared with 90 from compact query-independent text. Optical input includes every target, whereas text recovers every target it retains, revealing an inclusion advantage despite imperfect optical recovery. This comparison concerns one simple static textual storage policy, not optimized textual memory. Density and dialogue-memory controls further delimit the findings. Together, these measurements show why useful compressed context depends on both inclusion and recovery, and why visual-token counts alone cannot predict task performance.

1 INTRODUCTION

Rendering text as images can place more source content within a multimodal reader's input budget. Recent optical encoders, learned renderers, and visual memory systems exploit this possibility (Wei et al., 2025; Cheng et al., 2026; Liang et al., 2026; Feng et al., 2026; Shi et al., 2026; Li et al., 2026). Evaluating it requires two questions: is the requested source span included, and can the reader recover its answer?

The broad principles are established. Fixed readers can use rendered text (Li et al., 2025); fidelity changes can affect utility without changing the reported compression ratio (Tu et al., 2026); coverage and readable evidence are distinct concerns (Xie et al., 2026); and compressed full context can improve factual recall over truncation (Lee et al., 2026). Our question is narrower: what remains unresolved about a particular target when source inclusion and native input allocation are explicitly controlled?

We test recovery through paired geometry interventions, beginning with an unusually narrow identifier layout. The same source and native allocation yield different recovery after reflow; a separate scale study holds wrapping and page assignments fixed. We then test a different failure stage: pre-query retention can omit a target that text would otherwise read correctly. GLM's advantage persists against a compact, query-independent text control on fresh LONG sources; it remains a reader- and task-specific operating point, not a comparison with the best attainable textual memory.

Our contributions are three controlled measurements: **(i)** a held-out, native-allocation-matched layout intervention and a narrower scale follow-up; **(ii)** per-case target inclusion and answer success under specified original-format and compact hash-retention policies; and **(iii)** same-source, same-evidence, and budgeted-text controls that delimit the resulting claims. RULER-derived density and LoCoMo selection measurements provide supporting operating regimes; separate systems profiles document implementation costs. The experiments characterize existing readers without introducing a trained compressor.

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.

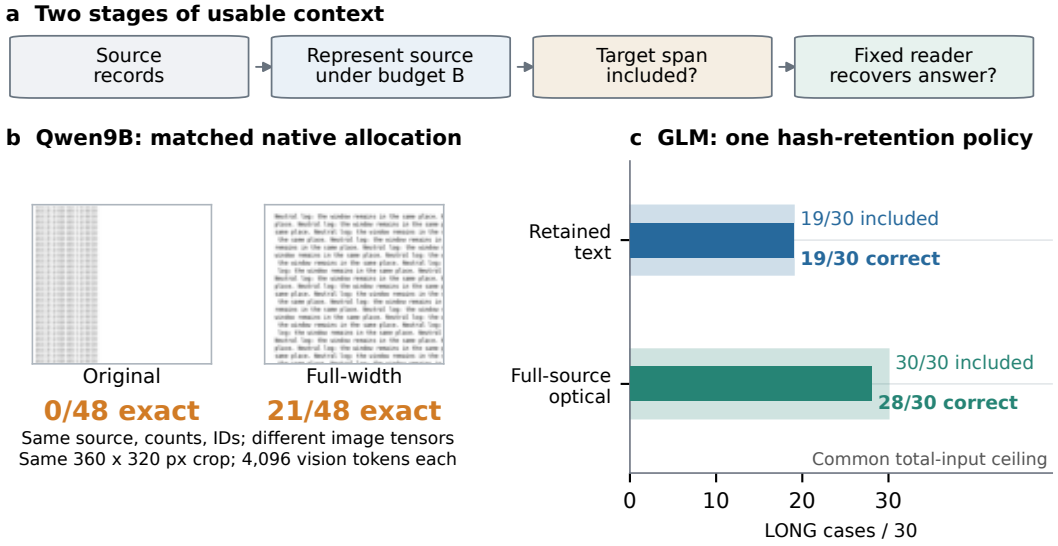


Figure 1: **Matched allocation does not establish target inclusion or recovery.** (a) Inclusion and conditional recovery are separate measurements; a correct guess need not imply inclusion. (b) Qwen9B's Original and Full-width layouts preserve the source, native counts, and input IDs (4,096 vision tokens), yet recover 0/48 versus 21/48 identifiers. Identical top-left 360×320-pixel crops illustrate occupied geometry; full pages determine the outcomes. (c) Under the hash-retention policy, GLM's LONG text includes and answers 19/30 targets; full-source optical input includes 30/30 and answers 28/30 at the same total-input upper bound. Only GLM establishes a corrected LONG advantage. Fresh compact-text control: Section 5. Panels (b) and (c) use different cases and readers.

2 Native budgets, target inclusion, and conditional recovery

Let X be the source, $\tilde{X}$ the source content assigned to the representation, V the measured vision-token count, and N_M reader M's source-text token count. Optical rate and token-level source coverage are

$$r = \frac{V}{N_M(\tilde{X})}, \qquad \rho = \frac{N_M(\tilde{X})}{N_M(X)}. \tag{1}$$

We report rate's reciprocal, $C_{\text{eff}} = 1/r$, as realized visual density so it can be compared with the nominal renderer targets $C \in \{0.8, 2, 4\}$. It counts included-source tokens per vision token, not original-source tokens per total input token. Full-source optical input has $\tilde{X} = X$; with selection, $V/N_M(X) = r\rho$. Retained text has no optical rate and is compared under the native *total-input* ceiling B, which includes prompt overhead and is not a physical storage-byte budget. Vision tokens are processor-specific units, not bits; rate–distortion theory supplies only the distinction between rate and loss, not an optimality claim (Cover & Thomas, 2006, Chapter 10).

For single-target binary correctness Y, let τ indicate inclusion of the requested record. Conditional recovery is $P(Y = 1 \mid \tau = 1)$; overall success obeys the accounting identity

$$\begin{aligned} P(Y = 1) = {} & P(\tau = 1)P(Y = 1 \mid \tau = 1) \\ & + P(\tau = 0)P(Y = 1 \mid \tau = 0). \end{aligned} \tag{2}$$

Token coverage ρ does not determine target inclusion τ. Inclusion means that the audited source span was assigned to the reader's representation, not that perceptually distinct characters survived its visual encoding. Correct answers to omitted targets are possible through guessing or prior knowledge; none occurs in our retained-text synthetic cases. For task loss $D_{\text{task}} = \mathbb{E}[L(A, \hat{A})]$, L is one minus the study's EM, native retrieval score, or normalized F1, with its stated sampling weights. Lower loss means better overall task performance, not necessarily better conditional recovery. LoCoMo may require multiple evidence spans; we do not estimate its evidence-inclusion decomposition.

Pre-query retention $S_{\text{store}}(X; B)$ commits content before the question and cannot retrieve discarded records; query-time selection $S_{\text{query}}(X, Q; B)$ can consult the full source using the question. The

original hash study conditions fit on actual question overhead; the compact-text follow-up uses a fixed future-query reserve and commits both text representations strictly before the query. LoCoMo BM25 instead studies query-time selection. Preparation, prefill, latency, and memory remain separate costs: neither native token rate nor equal total input counts determine them.

3 EXPERIMENTAL SETTING

We evaluate Qwen3.5-2B (Qwen2B), Qwen3.5-9B (Qwen9B), and GLM-4.6V-Flash (GLM), with pinned model and processor snapshots (Qwen Team, 2026; Z.ai Team, 2025). Readers are not fine-tuned for our experiments. Greedy decoding, native prompts, and generation caps are frozen within each study. The original study contains 500 LoCoMo-derived questions from ten conversations, 180 RULER-derived cases across six task configurations, and 96 controlled identifier cases. The geometry follow-ups use distinct held-out sets of 48 and 32 cases; fixed-budget retrieval uses 30 paired base cases at three lengths. A separate GLM compact-text follow-up uses 150 fresh LONG sources and three paired representations. We keep these samples separate rather than pooling interventions across studies.

Metrics reflect the requested information. LoCoMo uses normalized answer F1, averaged equally across conversations. RULER uses its scoped task-native containment scores, including partial credit for multiple targets; these scores are not exact-output accuracy. Identifiers use canonical exact match (EM), which normalizes the frozen answer representation without accepting arbitrary explanatory prose. Both finite-budget studies use native single-answer containment, with EM and character error rate (CER) as diagnostics. An identifier prompt includes a worked example. The synthetic experiments are diagnostic tasks, not evaluations of an interactive agent.

Uncertainty follows the sampling unit: conversations for LoCoMo and paired cases for the controlled studies. We preserve each study's prospective testing family. Reported bootstrap intervals are pointwise 95% intervals, not simultaneous intervals. The appendix gives full tests, sample construction, decoder settings, and provenance.

4 GEOMETRY CHANGES RECOVERY AT MATCHED MEASURED RATE

A visual-token count alone cannot specify exact recovery if different image geometries at the same native allocation yield different results. Exact identifiers offer a demanding test because plausible paraphrases cannot replace the requested string.

The original identifier study recovers 90/96, 93/96, and 93/96 values from text, but none of the 1,152 answers across compressed $C = 2$ and $C = 4$ arms is exact. Source-span auditing confirms that the represented text is present. Geometric inspection instead reveals an unusually narrow occupied region and small effective glyphs. This motivates intervention on representation geometry rather than an information-capacity interpretation of the failures.

Matched-budget layout intervention. We select an answer-blind full-width wrapping layout using a separate development set and evaluate 48 fresh held-out cases. Original and full-width conditions preserve the source string, native vision/nonvision/total counts, input-ID sequences, and recorded image-grid allocation. The IDs fix text and image-placeholder positions; image tensors and their visual embeddings differ. This is matched allocation, not identical effective neural input. Both retain the complete source across four pages. The layout package changes wrapping, whitespace, and pagination jointly.

Exact retrieval rises from zero in the original layout to 6/48, 21/48, and 7/48 (Figure 2a). All three paired exact McNemar tests pass Holm correction. This establishes different exact-recovery rates for different layouts at the same measured native allocation, not a new information-theoretic bound. Absolute optical recovery remains below the full-text controls: even enlarged-image controls reach only 17/48, 35/48, and 29/48, all below the prespecified 36/48 criterion. Figure 2 shows the remaining deficit to full text and the paired uncertainty; the enlarged and full-text references spend more input tokens than the matched layouts.

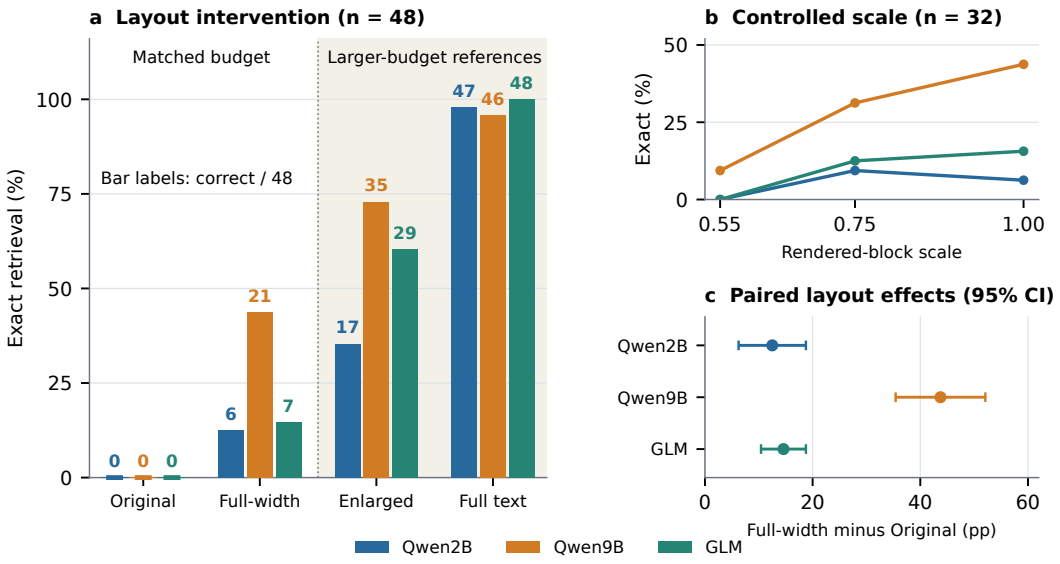


Figure 2: **Geometry changes recovery at matched native allocation.** (a) Original and Full-width are the matched-budget comparison. Shaded Enlarged and Full text are larger-budget diagnostic references; neither is a matched-budget alternative. Labels give exact successes out of 48. (b) Independent scale triplets hold source, wrapping, pagination, canvas, and counts fixed; only Qwen9B passes the corrected primary test. (c) Paired Full-width minus Original effects with pointwise 95% design-stratified bootstrap intervals; each cell contains only two cases. The layout and scale studies use different cases (Appendix D).

Controlled rendered-block scale. To narrow the intervention, we render each of 32 fresh cases once, then uniformly scale the page content to 0.55, 0.75, or 1.00 and center it on the same final canvas. Line breaks, page assignments, source, question, and native input counts remain fixed. This removes reflow and pagination changes while varying occupied area and effective visual scale.

Qwen9B exact recovery increases from 3/32 to 10/32 to 14/32. The exact conditional Cochran Q test is significant after correction across readers ($p = 0.0110$). Its endpoint difference is 34.375 percentage points (95% interval [15.625, 53.125]); the endpoint test also passes the separate nine-comparison family. Qwen2B yields 0, 3, and 2 successes; GLM yields 0, 4, and 5. Neither passes its corrected primary test. All ten generation-cap hits remain in the analysis.

Together, the interventions establish a matched-allocation layout effect across readers and a narrower scale response for Qwen9B. Uniform scaling still changes glyph size, occupied area, resampling, and physical positions together. It neither isolates glyph height alone nor establishes a universal monotonic law. Qwen2B's observed response is nonmonotonic; a difference in significance across readers does not establish a reader-by-scale interaction.

5 Source coverage under a fixed input budget

When full text no longer fits, pre-query storage may omit a future target. We construct 30 fresh single-key cases with SHORT, MEDIUM, and LONG versions that preserve the query and target while adding distractors. Optical input renders the complete source on four pages. Text greedily retains complete hash-ranked records while the native templated input fits the same reader-specific ceiling, then restores source order. Ranking is query-independent, but fit reserves actual question overhead; strictly pre-query storage instead needs a fixed reserve. Omitted records are unavailable. This S_{store} diagnostic bounds native input, not storage bytes or an optimized memory policy. Query-time lookup with access to discarded records is untested; LoCoMo BM25 separately studies selection from the complete history.

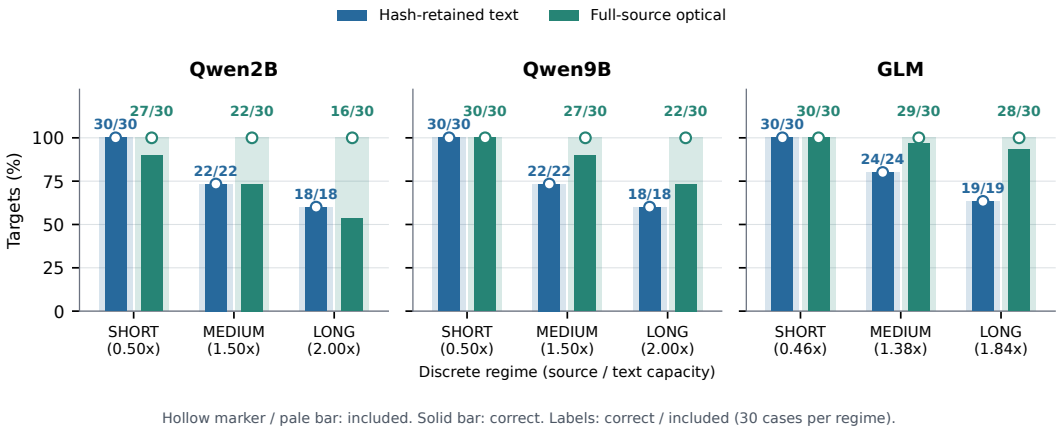


Figure 3: **Inclusion and correctness under one hash-retention policy.** Each regime has 30 paired cases and a common total-input upper bound, not equal consumed tokens. Pale bars and hollow markers show included targets; solid bars show correct answers. Labels are correct/included counts: text recovers every retained target, while optical includes all 30 but can fail to recover them. Only GLM establishes a corrected LONG advantage. Source/capacity ratios are nominal; complete templated-input checks govern fit (Appendix E).

Full text fits at SHORT. Source length is approximately one-half, one-and-a-half, and twice the nominal text-source capacity across the three regimes (Figure 3). The constraint is an equal total-input upper bound, not equal consumed tokens: SHORT text is not padded. Optical renders the complete source in all regimes, with target spans audited as included; its realized visual density therefore increases with length. Appendix E gives native budgets and target-retention diagnostics.

Under this hash-retention policy, GLM recovers 28/30 LONG targets from optical input versus 19/30 from retained text, a 30.00-point difference (95% interval [10.00, 50.00]; adjusted $p = 0.0352$). Precisely 19/30 queried records survive in the GLM text input, and every one is answered correctly; the optical input includes all 30 target spans but has two recovery failures. The primary interaction passes its separate corrected family with the same value. Because GLM ties at SHORT on every case, the LONG and interaction tests are the same case-level contrast, not independent confirmations. Qwen9B's LONG difference is positive but uncertain; Qwen2B's is negative (Figure 3; all interaction tests appear in Appendix E, Table 13).

At every length and for every reader, text answers every surviving target and no omitted target. Optical includes the full source but sometimes fails on an included target. For GLM LONG, greater inclusion outweighs imperfect recovery: conditional text recovery is 19/19, versus optical 28/30, with 10 optical-only and 1 text-only successes. The balance is less favorable for the Qwen readers. Token coverage alone would not establish these case-level facts.

No reader meets the prespecified descriptive crossover rule, which requires text to be strictly better at SHORT and optical strictly better at LONG. GLM and Qwen9B tie at SHORT; Qwen2B does not win optically at LONG. The original primary sign-flip test assumes conditional sign symmetry of case interactions; its uncertainty is conditional on these 30 cases and their construction.

Compact pre-query text control. On 150 fresh sources, we prospectively test GLM's positive LONG operating point. Both text arms use the same pre-query hash priority and fixed query reserve; compact key/value lines retain 15.85% more records. At a common 4,025-token ceiling, optical answers 126/150, compact text 90/150, and original-format text 85/150. The primary optical–compact difference is 24.0 points (95% stratified paired bootstrap interval [14.67, 33.33]; exact McNemar $p = 4.82 \times 10^{-6}$). Text recovers every included target; optical includes all but misses 24. The ceiling was amended before inference, 2.78% above the original budget (Appendix E.1, including conservative sensitivity and position diagnostics). This supports the specific compact control, not optimized memory or query-time retrieval with source access.

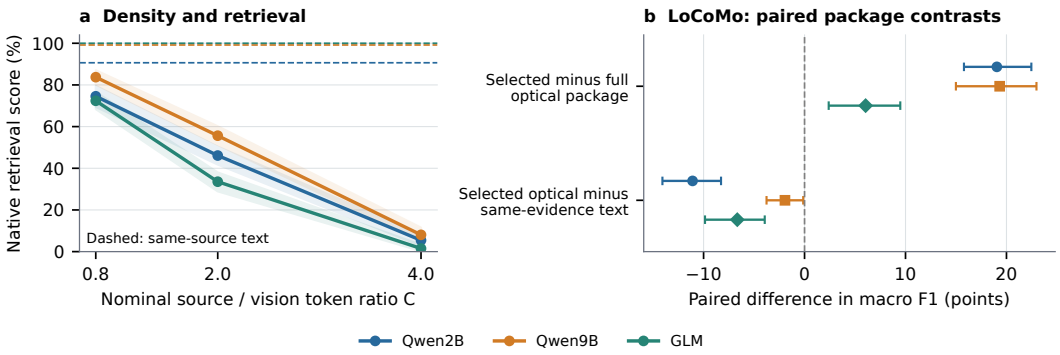


Figure 4: **Density and selection change task performance in different ways.** (a) Six-task RULER aggregate across nominal renderer targets; Table 1 reports realized density. Shading gives pointwise task-stratified 95% intervals; dashed lines show same-source text. (b) Paired LoCoMo contrasts: selected minus full optical measures a selection-plus-rendering package; selected optical minus same-evidence text compares modalities. Intervals resample ten conversations and are pointwise. Condition means appear in Table 2; no continuous frontier or isolated retrieval effect is implied.

Table 1: **Realized RULER visual density.** Median per-case $C_{\text{eff}} = N_M(X)/V$ at each nominal target. Both Qwen readers have the same recorded counts. Appendix C gives native counts and ranges.

Reader	Nominal $C = 0.8$	$C = 2$	$C = 4$
Qwen2B / Qwen9B	0.800	1.970	4.056
GLM	0.797	2.020	4.115

6 Density-dependent retrieval within one rendering family

How does retrieval change as the same source is rendered more densely within one page-construction procedure? We use six RULER-derived configurations, with 30 cases each at an upstream reference length of 8,192 tokens (Hsieh et al., 2024). Each case is evaluated as text and on four optical pages at nominal $C \in \{0.8, 2, 4\}$. Native processors determine the realized token counts; a shared nominal ratio does not imply identical pixels across readers. The same renderer produced the original identifier pages in Section 4, but the realized geometries differ: audited median maximum horizontal text span across four-page inputs is 77.9% of page width here versus 12.3% for the original identifiers (Appendix C). The identifier study's narrow-column failure therefore should not be projected onto this sweep.

Figure 4a shows a steep decline for every reader. At $C = 0.8$, task-native scores are 74.56, 83.72, and 72.39%; at $C = 4$, they are 5.44, 8.06, and 1.50%. Same-source text reaches 90.64, 99.17, and 100.00%. Thus even the least dense optical setting has a lower six-task aggregate in this rendering family. Individual tasks differ: Qwen2B variable tracking scores 84.00% optically at $C = 0.8$ versus 44.67% as text (Appendix C). These are selected-task operating points, not a universal capacity threshold or the official RULER aggregate.

These results do not directly contradict *Text or Pixels?* (Li et al., 2025): its single-needle experiments use different readers, shorter contexts, and a different renderer, and also show degradation beyond successful operating ranges. Our task mixture and protocol do not reproduce that setting; they cannot identify which difference explains a cross-paper gap.

Scoring matters to this conclusion. A frozen diagnostic sample of 20 Qwen text answers that passed native retrieval but failed canonical EM all contained the target with extra prose or Markdown. Such disagreements cannot be read as retrieval failures. We retain native scores as primary and report stricter formatting diagnostics separately. This distinction becomes consequential for identifiers, where every character of the returned value is part of the task.

Table 2: **LoCoMo conversation-macro F1.** All conditions answer the same 500 questions. Retrieved text and optical share evidence; the two budgeted-text controls may retain different turns.

Representation	Qwen2B	Qwen9B	GLM
Full text	48.45	53.04	48.31
Full optical	9.84	15.95	24.84
Retrieved text	39.98	37.22	37.55
Retrieved optical	28.89	35.28	30.89
Text, full-optical budget	48.34	49.77	44.01
Text, selected-optical budget	40.74	39.11	39.35

7 Evidence selection and rendering shift the operating point

How does a query-time selected optical package perform when it represents less irrelevant conversation history? On LoCoMo (Maharana et al., 2024), we compare complete histories with the top 16 turns selected by a frozen BM25 retriever conditioned on the question but never on its answer. This is query-time selection, S_{query}, with the complete history available to the retriever. Retrieved text and retrieved optical consume the same canonical evidence string. Full optical uses approximately $C = 2$; retrieved optical uses a fixed single-page rendering. Two additional textual controls pack whole turns to the respective optical input budgets.

The selected optical package improves F1 over full optical by 19.05, 19.33, and 6.05 points (reader order Qwen2B, Qwen9B, GLM). Each contrast passes the frozen Holm family. The increase is large enough to move retrieved optical well away from the low full-optical scores, but it is an effect of the *retrieved optical package*: selection changes evidence, context burden, layout, and rendering regime together. The experiment does not isolate retrieval as the sole causal mechanism.

With the evidence held equal, optical minus text is -11.09, -1.94, and -6.66 points. None establishes equivalence within the prespecified ± 3-point margin. Qwen9B's smaller gap is sensitive to the interval construction: its conversation-level t interval crosses zero (Appendix B). The point estimates support a fidelity cost, with unequal precision across readers. Both budgeted-text controls also score above their optical counterparts for every reader (Table 2); changing the selected optical package does not establish a practical advantage over those text controls.

Generation length does not account for the large selection gains in the bounded sensitivity we tested. Rerunning only the 126 Qwen2B and 121 Qwen9B full-optical cap hits from 96 to 256 tokens changes reconstructed full-arm F1 by just 0.1511 and 0.1466 points. All reruns preserve their original generated-token prefixes. Some remain capped, so this diagnostic does not establish that generation length is irrelevant; it leaves the original primary results intact.

8 Systems costs in two operating regimes

We profile one predetermined source per length on an A100 40GB with resident BF16 models, SDPA, batch size one, and one generated token. Each of 45 cells has three warmups and ten measured repetitions; these measure execution variation, not accuracy or workload uncertainty. Appendix F reports all timing boundaries and values, including allocated peaks both with and without the resident-model baseline.

For the same LONG source, C2 lowers combined prefill (including vision encoding) and allocated peak for every reader, but lowers end-to-end time only for Qwen9B. C4 lowers LONG end-to-end time for all three, without corresponding C4 quality evaluation on these inputs. In the fixed-budget comparison, optical has lower MEDIUM/LONG end-to-end time but higher prefill and allocated peak than retained text. Text packing repeatedly tokenizes subsets inside timing; optical layout planning and calibration are precomputed and excluded. Caching the text packer could change the ranking. These results distinguish preparation, model execution, and memory costs in the measured implementation; they establish neither quality-preserving nor intrinsic optical acceleration.

Table 3: **Closest prior work and the comparison tested here.** This is a protocol map, not a cross-paper score ranking. The final column states this study's distinct *measurement*, without claiming that earlier work lacked all related controls.

Work	Established result or design	Specific measurement here
Text or Pixels?	Fixed-reader rendered text; density limits, rendering and timing	Held-out exact-string geometry at matched native allocation
Fico	Fidelity perturbations, including unchanged-ratio conditions, across OCR, retrieval and QA	Paired reflow and fixed-canvas scale with audited native counts and IDs
LensVLM	Coverage/fidelity and conditional hit analysis; expansion with trained and untrained readers	Target inclusion and success under original and compact hash retention; actual-query or fixed-reserve packing
Lee et al.	Factual-recall benefit over truncation; direct alternatives outperform optical	Generated answers from fixed readers versus a specified hash policy
Glyph	Trained rendering search; similar-ratio configuration ablation; rare-string difficulty	Exact-string layout interventions preserving native allocation with unmodified readers
AGAR	Evidence magnification; fixed readers; approximately token-matched uniform scaling	Answer-blind static exact-string intervention with identical per-case native counts and IDs

9 Related Work

Closest empirical comparisons. *Text or Pixels?* evaluates fixed readers on rendered-text RULER retrieval and summarization, including rendering, token savings, and timing (Li et al., 2025). Fico studies fidelity across OCR, retrieval, and QA; binarization, JPEG, and upscale perturbations can change utility without changing the reported compression ratio (Tu et al., 2026). AGAR studies evidence localization and recovery through span magnification, with a uniform-scaling control designed to match visual-token use; its reported aggregate counts are close, not identical (Zeng et al., 2026). These works already establish density and geometry sensitivity. We add an answer-blind, held-out exact-string layout intervention with identical per-case native vision, nonvision, and total counts and input IDs, followed by fixed-canvas scaling. These measure task-specific layout and scale effects without isolating a mechanism.

LensVLM distinguishes coverage from fidelity, analyzes conditional evidence hits, and tests selective expansion with trained and untrained readers (Xie et al., 2026). Neither that distinction nor fixed readers alone is new here. We audit target inclusion and answer success under frozen hash ranking, including a compact-text follow-up with a fixed pre-query reserve and no adaptive expansion. Lee et al. (2026) find optical language modeling comparable to truncation; factual recall, measured by answer log-likelihood, can improve over truncation but favors direct alternatives. Our GLM result concerns generated answers on one synthetic LONG task under specified original-format and compact hash-retention policies, not superiority to learned or query-aware alternatives.

Measure-transport analyses distinguish token savings from task-relevant visual loss; their coverage cost concerns cross-patch fragmentation rather than omitted source records (Tang et al., 2026). ZeroSense probes full-text recovery using low-predictability contexts to reduce semantic compensation (Gao et al., 2026). Our measurements instead audit source-span inclusion and reader-specific answer recovery.

Systems and learned memory. DeepSeek-OCR trains an optical encoder–decoder and measures reconstruction across compression levels (Wei et al., 2025). Glyph combines training with rendering search, ablates similar-ratio configurations, and identifies rare order-sensitive strings as difficult (Cheng et al., 2026); our contribution is stricter native-allocation control, not discovering that fragility. VIST learns a resampler and cross-attention interface for distant visual context (Xing et al., 2025). AgentOCR learns optical history compression through agentic reinforcement learning (Feng et al., 2026); MemOCR learns layouts under budget-aware rewards (Shi et al., 2026). OCR-Memory fetches original text through optical anchors (Li et al., 2026); our comparator cannot access discarded records. VizoMem trains a visual-note retriever for conversational memory (Liang et al., 2026). These systems use training, adaptive selection, or retrieval beyond our static representations.

Cross-paper score rankings would conflate readers, training, native accounting, rendering, tasks, and selection timing. Our internal controls compare the same source or evidence across modalities, match native allocation across layouts, and diagnose omission under original-format and compact hash-retention policies. The last comparison does not establish the best attainable textual memory. Table 3 summarizes these boundaries.

10 Discussion and limitations

The experiments separate three quantities: measured representation rate, whether the requested source record or span is included in the reader input, and whether a fixed reader recovers an included target. The six-task RULER aggregate and LoCoMo condition means favor available text over the corresponding optical arms, with task-level exceptions in RULER. Under a finite input budget, hash-retained text can instead fail because it omits the requested record. GLM's LONG results, including the compact-text follow-up, show one regime where full-source optical inclusion outweighs imperfect recovery. They do not establish the same balance for another reader, untested retention policy, or memory workload.

Geometry and selection alter these stages through different routes. Matched layout changes recovery while leaving the represented source and native token allocation fixed. LoCoMo selection changes the evidence, context burden, and rendering package together, so its gains cannot be assigned to retrieval alone. Reporting a visual-token ratio without target inclusion and task recovery misses both a coverage benefit and a fidelity cost. Systems measurements add a separate distinction: lower end-to-end time need not mean lower prefill or peak allocation.

The scope is limited to three readers and selected benchmark subsets. Ten dialogue clusters and the small independent case sets constrain precision. Public benchmark exposure during training is unknown. Exact identifiers probe brittle state recovery rather than all memory use, and native containment does not establish verbatim reconstruction. Generation caps remain binding for some conditions. The unusually narrow original layout limits generalization to typical documents. The fresh geometry studies are separate samples; covarying visual dimensions prevent a glyph-only mechanism claim. Nonsignificance in other readers is not equivalence.

The finite-budget results depend on deterministic hash retention; the follow-up tests only one simple compact serializer on the previously positive GLM LONG operating point. Optimized textual memory and query-time retrieval with source access remain untested. Budgets count native input, not storage bytes. The original packer uses actual question overhead; the follow-up commits storage with a fixed reserve at a 2.78% larger ceiling. Similarly, local freezes establish prospective specification within this project, not independent preregistration. Systems results use one predetermined source per length, a single hardware setup, one-token output, and implementation-dependent preprocessing; they cannot establish workload-wide latency or compute savings. The appendix reports these boundaries alongside the complete secondary results.

Evaluation implications. A budgeted context evaluation can diagnose the two failure stages directly: record whether each requested source span reaches the reader, then report answer accuracy for included and omitted targets as well as overall accuracy. State whether selection occurs before the query or can still consult the complete source. For optical inputs, report processor-native vision, nonvision, and total counts together with representative rendered pages; for a geometry claim, pair cases and check input IDs. Measure representation preparation, prefill, and peak allocation separately when making systems claims. These checks make the tested operating point interpretable without assuming that token count predicts utility.

11 Conclusion

A visual-token count does not specify whether a future target is present or recoverable. At identical native allocation, geometry changes exact recovery. GLM answers more LONG targets optically than under original-format hash retention and a fresh compact-text control; the latter tests GLM only. Density, query-time selection, and systems profiles delimit these operating points. Compressed-context evaluations should report inclusion, conditional recovery, overall performance, native input counts, and execution costs separately.

REPRODUCIBILITY STATEMENT

The accompanying evidence index maps claims and generated numbers to frozen analyses. The repository preserves model and processor pins, input manifests, rendering plans, raw outputs, generated-token journals, independent scoring and processor-input audits, hashes, and prospective follow-up freezes. Appendix G summarizes verification; the supplementary reproducibility manifest provides snapshot identities, artifact paths, hashes, and replay boundaries. The manuscript's values and plots are generated from audited machine-readable results. No new inference or outcome-driven case selection was performed during manuscript preparation.

AI-USE DISCLOSURE

Generative AI tools assisted with experimental implementation, debugging, testing, audit and analysis code, data formatting, and synthetic diagnostic data generation. For the final compact-text follow-up, assistance included proposing experimental alternatives, hypothesis refinement, protocol and parameter feedback, and interpretation of audited results. Additional uses included literature discovery and analysis; manuscript drafting, editing and polishing; reference formatting; figures and reproducibility artifacts. No formal mathematical proof was produced or assisted. Translation and qualitative/thematic coding were not part of this work.

AI-assisted experimental code was covered by deterministic tests and audited against frozen inputs and outputs. Reported numerical results were independently recomputed from preserved artifacts; citations and literature claims were checked against primary sources; and manuscript claims were reconciled with the claim–evidence ledger and authoritative analyses. The authors reviewed all AI-assisted work and take responsibility for the final submission.

REFERENCES

Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Weihan Wang, Zhe Su, Xiaotao Gu, Xiao Liu, Yushi Bai, Jie Tang, Hongning Wang, and Minlie Huang. Glyph: Scaling Context Windows via Visual-Text Compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37145–37158, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1722. URL https://aclanthology.org/2026.acl-long.1722/.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X. URL https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X.

Lang Feng, Fuchao Yang, Feng Chen, Xin Cheng, Haiyang Xu, Zhenglin Wan, Ming Yan, and Bo An. AgentOCR: Reimagining Agent History via Optical Self-Compression. arXiv:2601.04786, 2026. URL https://arxiv.org/abs/2601.04786.

Yonghan Gao, Zehong Chen, Lijian Xu, Jingzhi Chen, Jingwei Guan, and Xingyu Zeng. Decoupling Semantics from Vision: A Framework for Faithful Visual-Text Compression Evaluation. arXiv:2608.01848v1, 2026. URL https://arxiv.org/abs/2608.01848v1.

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? arXiv:2404.06654, 2024. URL https://arxiv.org/abs/2404.06654.

Ivan Yee Lee, Cheng Yang, and Taylor Berg-Kirkpatrick. Optical Context Compression Is Just (Bad) Autoencoding. arXiv:2512.03643v2, 2026. URL https://arxiv.org/abs/2512.03643v2. Revised April 2026; first posted December 2025.

Jinze Li, Yang Zhang, Xin Yang, Jiayi QU, Jinfeng Xu, Shuo Yang, Junhua Ding, and Edith Cheuk-Han Ngai. OCR-Memory: Optical Context Retrieval for Long-Horizon Agent Memory. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of*

the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10409–10420, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.474. URL https://aclanthology.org/2026.acl-long.474/.

Yanhong Li, Zixuan Lan, and Jiawei Zhou. Text or pixels? evaluating efficiency and understanding of LLMs with visual text inputs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10564–10578, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.558. URL https://aclanthology.org/2025.findings-emnlp.558/.

Weijie Liang, Yuanfeng Song, Xing Chen, Caleb Chen Cao, Sirui Han, and Yike Guo. VizoMem: A Visual-Textual Memory Framework for Efficient Long-Horizon Reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7399–7422, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.365. URL https://aclanthology.org/2026.findings-acl.365/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Qwen Team. Qwen3.5 model cards, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Yaorui Shi, Shugui Liu, Yu Yang, Wenyu Mao, Yuxin Chen, Qi GU, Hui Su, Xunliang Cai, Xiang Wang, and An Zhang. MemOCR: Layout-Aware Visual Memory for Efficient Long-Horizon Reasoning. arXiv:2601.21468, 2026. URL https://arxiv.org/abs/2601.21468.

Lv Tang, Tianyi Zheng, Yang Liu, Bo Li, and Xingyu Li. Visual Text Compression as Measure Transport. arXiv:2605.06708v1, 2026. URL https://arxiv.org/abs/2605.06708v1.

Jianhong Tu, Nicholas Crispino, Kyle Montgomery, Chenguang Wang, and Dawn Song. Fico: Evaluating Vision-Language Models under Visual Fidelity and Compression at Scale. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35261–35277, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1758. URL https://aclanthology.org/2026.findings-acl.1758/.

Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts Optical Compression. arXiv:2510.18234, 2025. URL https://arxiv.org/abs/2510.18234.

Roy Xie, Dan Friedman, Donghan Yu, Bowen Pan, Christopher Fifty, Jang-Hyun Kim, Xianzhi Du, Zhe Gan, Vivek Rathod, and Bhuwan Dhingra. LensVLM: Selective Context Expansion for Compressed Visual Representation of Text. arXiv:2605.07019, 2026. URL https://arxiv.org/abs/2605.07019.

Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. Vision-centric Token Compression in Large Language Model. arXiv:2502.00791, 2025. URL https://arxiv.org/abs/2502.00791.

Z.ai Team. GLM-4.6V-Flash model card, 2025. URL https://huggingface.co/zai-org/GLM-4.6V-Flash.

Shenglai Zeng, Qirui Wang, Kai Guo, Xinnan Dai, Xianxuan Long, and Hui Liu. Magnifying What Matters: Attention-Guided Adaptive Rendering for Visual Text Comprehension. arXiv:2606.12898v1, 2026. URL https://arxiv.org/abs/2606.12898v1.

A PROTOCOLS, SAMPLES, AND STATISTICAL INTERPRETATION

The original experiments, the layout intervention, the bounded cap diagnostic, and the final studies retain their separate frozen designs. Local timestamped freezes precede each follow-up's evaluation outputs; they are not independently registered preregistrations. Development, smoke, historical legacy, and systems-timing outputs do not enter held-out task estimates.

Table 4: Evaluation inventory. Outputs are executions of conditions, not independent samples. The cap diagnostic is conditional on the original capped subset.

Study	Cases	Outputs	Independent unit
LoCoMo	500 questions	9,000	10 conversations
RULER-derived	180	2,160	Case, stratified by task
Original identifiers	96	2,016	Case, balanced factor design
Layout intervention	48	576	Paired held-out case
Controlled scale	32	288	Paired triplet
Fixed-budget retrieval	30 base cases	540	Base case across lengths
Compact static storage	150	450	Paired source triplet
LoCoMo cap sensitivity	247 capped questions	247	Original conversation
Systems profiling	1 source per length	585	Repeated invocation

The original study contributes 13,176 outputs. The layout intervention additionally uses 36 development calls on 12 cases, excluded from held-out estimates. The final round contributes 828 quality outputs and 585 profiling invocations, of which 135 are warmups and 450 are measured. The profiling instrumentation observes 585 top-level model forwards. This count does not measure internal forward calls of the quality runs or count their outputs as additional accuracy samples. The later compact-storage follow-up contributes 450 additional quality outputs on 150 independent sources; its three answer-blind development sources and separate smoke cases do not enter estimates.

Decoding and scoring. The original LoCoMo, RULER-derived, and identifier caps are 96, 128, and 64 generated tokens, respectively. Both earlier final quality studies and the compact-storage follow-up use 64. Generation is greedy, with thinking disabled where the native interface supports that setting. The final studies use BF16, SDPA, batch one, and the inherited model generation configuration; the complete configuration accompanies each output. Token IDs are preserved as well as decoded text. Canonical identifier EM uses the frozen scorer's whitespace and answer normalization; it does not strip arbitrary prose to rescue an otherwise incorrect output. CER can exceed one when extra predicted characters outnumber reference characters. Literal-raw EM is a separate diagnostic because model-specific output wrappers can differ from the canonical answer.

Families and intervals. The original confirmatory family consists of three directional LoCoMo package tests and three original identifier contrasts. Its six-test Holm correction is retained. For LoCoMo H1, the statistic is the mean of ten paired conversation-level F1 differences, selected optical minus full optical. The one-sided upper-tail test enumerates all sign assignments, conditions on the observed magnitudes, and assumes sign symmetry under the null; condition assignment was not randomized. The LoCoMo same-evidence comparison estimates a difference and assesses a prespecified equivalence margin; it is not silently added as a new directional discovery. The RULER sweep is descriptive. Layout uses three two-sided exact McNemar tests in its own Holm family. Scale uses three exact conditional omnibus tests, with nine pairwise secondary tests in a separate family. Fixed-budget retrieval uses three primary interaction tests and a separate family of three secondary LONG tests. The compact-storage follow-up has one confirmatory optical–compact McNemar contrast, with other arm and position comparisons descriptive. There is no single global correction across these prospectively separated studies.

Paired percentile bootstrap intervals use 50,000 resamples. LoCoMo resamples whole conversations with equal weight, not individual questions. RULER resampling is stratified by task and keeps a case's representations together. Layout resampling is stratified within its 24 two-instance cells. With only two cases per cell, these intervals warrant caution: cells with identical observed differences contribute no bootstrap variation. The exact paired tests provide the confirmatory layout evidence. Scale and the original fixed-budget study resample paired cases; the compact-storage follow-up

resamples paired triplets within its three position strata. These intervals are pointwise. Boundary intervals from zero successes or zero observed differences do not imply population certainty. Cluster sign tests and t intervals below are sensitivity analyses, not additional confirmatory tests.

B DIALOGUE MEMORY: COMPLETE COMPARISONS AND CAP SENSITIVITY

The LoCoMo subset comprises the recorded 500 questions in ten previously studied conversation clusters, with 105, 108, and 287 questions in categories 1, 2, and 4. Expansion of question coverage does not create new independent conversations. We preserve dated chronological memories, speaker and turn identifiers, text, and shared-image captions. The BM25 retriever uses $k_1 = 1.5$, $b = 0.75$, lowercase ASCII alphanumeric terms, and stable source-index tie-breaking. It selects the top 16 turns within the question's conversation and restores source order. It never uses reference answers, evidence annotations, or reader scores.

The same-evidence text and optical conditions receive identical canonical strings and turn metadata. Retrieved optical uses a single 1,372-pixel square RGB page, an 8-pixel source font, and 24-pixel margins. Full optical is calibrated to approximately $C = 2$. The full-optical-budget text control greedily packs whole turns using the same ranking and skips oversized turns. The selected-optical-budget text control uses the frozen ranked-prefix rule without splitting turns or skipping an earlier ranked turn. These controls share a budget with their optical counterpart, not necessarily the same evidence.

Table 5: All principal LoCoMo paired differences, in F1 points. H1: retrieved optical minus full optical. H2: retrieved optical minus same-evidence text. Selected/full budget: the relevant optical arm minus its budgeted-text control.

Reader	Contrast	Difference (F1 points)	95% interval
Qwen2B	H1	19.05	[15.79, 22.46]
Qwen2B	H2	-11.09	[-14.08, -8.27]
Qwen2B	selected_budget	-11.85	[-15.70, -8.58]
Qwen2B	full_budget	-38.51	[-41.61, -35.18]
Qwen9B	H1	19.33	[14.99, 22.97]
Qwen9B	H2	-1.94	[-3.76, -0.13]
Qwen9B	selected_budget	-3.83	[-4.69, -3.10]
Qwen9B	full_budget	-33.81	[-37.73, -29.94]
GLM	H1	6.05	[2.39, 9.48]
GLM	H2	-6.66	[-9.86, -3.94]
GLM	selected_budget	-8.46	[-10.52, -6.38]
GLM	full_budget	-19.17	[-22.61, -15.50]

H1 adjusted p values are 0.00292969, 0.00292969, and 0.00878906. The Qwen readers improve in all ten conversations; GLM improves in eight. GLM's directional sign test is less decisive than the original magnitude-sensitive sign-flip test. All leave-one-conversation-out H1 point estimates remain positive. For H2, the Qwen9B t interval crosses zero even though its percentile bootstrap interval is below zero. Neither that sensitivity nor failure of the equivalence criterion licenses an equivalence claim.

Table 6: Conversation-level sensitivity. $+/-$ count positive/negative differences. Sign p_+ tests a positive direction and should not be interpreted as a test of disadvantage for H2. Intervals are descriptive t intervals over ten cluster differences.

Reader	Contrast	$+$	$-$	Ties	Sign p_+	t interval (pp)
Qwen2B	H1	10	0	0	0.0010	[14.99, 23.11]
Qwen2B	H2	0	10	0	1.0000	[-14.63, -7.55]
Qwen9B	H1	10	0	0	0.0010	[14.41, 24.24]
Qwen9B	H2	1	9	0	0.9990	[-4.17, 0.29]
GLM	H1	8	2	0	0.0547	[1.73, 10.37]
GLM	H2	0	10	0	1.0000	[-10.29, -3.04]

Bounded cap follow-up. Precisely the original 126 Qwen2B and 121 Qwen9B full-optical cap hits are rerun once at 256 tokens. All original first-96-token prefixes match, permitting a cap-extension interpretation. Original uncapped answers remain in the reconstructed full arm. Qwen2B's macro F1 changes from 9.8378 to 9.9889, a 0.1511-point change [-0.0010, 0.3315]. Qwen9B changes from 15.9514 to 16.0980, a 0.1466-point change [-0.0365, 0.3340]. Neither reaches the prespecified three-point descriptive materiality threshold. There are still 22 and 8 cap hits, respectively. The diagnostic preserves the original primary tests and does not justify extrapolation to unlimited decoding.

C RULER-DERIVED RETRIEVAL AND ORIGINAL EXACT IDENTIFIERS

The RULER-derived subset uses `niah_single_1`, `niah_single_2`, `niah_single_3`, `niah_multikey_3`, `niah_multivalue`, and `vt`. A pinned reference tokenizer generates one upstream sequence-length setting, 8,192 tokens. This setting does not promise precisely that many source tokens for every reader. The generator's insertion distributions, templates, and source backgrounds are preserved; native prompts are shared across modalities. Native multiple-answer metrics admit partial credit. The aggregate equally weights the six configurations.

Table 7: Complete task-level RULER-derived native scores (percent). MK3: multikey 3; MV: multivalue; S1–S3: single 1–3; VT: variable tracking. The optical conditions all use four pages.

Reader	C	MK3	MV	S1	S2	S3	VT
Qwen2B	Text	100.00	99.17	100.00	100.00	100.00	44.67
Qwen2B	0.8	56.67	60.00	96.67	96.67	53.33	84.00
Qwen2B	2	3.33	40.00	90.00	70.00	10.00	63.33
Qwen2B	4	0.00	0.00	30.00	0.00	0.00	2.67
Qwen9B	Text	100.00	98.33	100.00	100.00	100.00	96.67
Qwen9B	0.8	93.33	58.33	100.00	100.00	76.67	74.00
Qwen9B	2	23.33	55.83	96.67	93.33	20.00	44.67
Qwen9B	4	0.00	8.33	26.67	10.00	0.00	3.33
GLM	Text	100.00	100.00	100.00	100.00	100.00	100.00
GLM	0.8	23.33	95.00	100.00	100.00	33.33	82.67
GLM	2	0.00	50.00	63.33	66.67	0.00	21.33
GLM	4	0.00	1.67	3.33	3.33	0.00	0.67

Table 8: Native RULER accounting from preserved per-case records. Entries are marginal medians across 180 cases per condition; they need not add or divide to the median of another column. $S = N_M(X)$, V is vision tokens, $H = T - V$ is nonvision input, and T is total input. Density is median per-case S/V, with its observed range.

Reader	C	Source	Vision	Nonvision	Total	C_{eff} median [range]
Qwen2B	0.8	7853	10000	81	10080	0.800 [0.784, 0.816]
Qwen2B	2	7853	3844	81	4175	1.970 [1.952, 2.064]
Qwen2B	4	7853	1936	81	2017	4.056 [3.828, 4.134]
Qwen9B	0.8	7853	10000	81	10080	0.800 [0.784, 0.816]
Qwen9B	2	7853	3844	81	4175	1.970 [1.952, 2.064]
Qwen9B	4	7853	1936	81	2017	4.056 [3.828, 4.134]
GLM	0.8	7387	9216	78	9682	0.797 [0.787, 0.802]
GLM	2	7387	3600	78	3922	2.020 [1.944, 2.070]
GLM	4	7387	1936	78	2012	4.115 [3.815, 4.186]

The score-disagreement audit samples 20 raw-text Qwen outputs conditional on native correctness and canonical-EM failure. Every sampled answer includes the correct target plus additional prose or Markdown. This is evidence about the selected disagreement stratum, not an estimate of formatting-error prevalence. Native scoring, canonical scoring, source partition checks, and token decoding reproduce the original records. Full task and shared-background diagnostics remain in the machine-readable supplement.

RULER versus original identifier geometry. The original RULER-derived and identifier studies call the same renderer, but their realized four-page layouts differ. At $C = 2$, the audited median maximum horizontal text span across four-page inputs is 77.9% of page width for RULER and 12.3% for the original identifiers. Effective font-em ranges are 8.4–9.0 versus 4.1–4.5 processor pixels for Qwen and 6.9–7.8 versus 3.5–3.8 for GLM, respectively. These measurements explain why a shared rendering code path does not make the identifier pages representative of RULER geometry. They are descriptive cross-dataset diagnostics, not a causal estimate of line width or font size alone.

Original identifier design. Identifiers cross payload lengths 8 and 24, decimal and lowercase-alphanumeric alphabets, repeated four-character motifs and independently sampled characters, 4 and 32 distractors, and early/middle/late target positions, with two instances per joint cell. The resulting 96 cases have text and six optical conditions (two or four pages at nominal $C = 0.8, 2, 4$). Nominal string length is not a measure of source entropy. The four compressed conditions are the $C = 2$ and $C = 4$ cells; the $C = 0.8$ cells are reported separately below.

Table 9: Complete original identifier EM counts, each out of 96. The `p2/p4` suffix identifies the number of pages. Nonzero $C = 0.8$ performance is retained and is not included in the claim of zero recovery over compressed arms.

Arm	Qwen2B	Qwen9B	GLM
full_raw	90	93	93
optical_c0p8_p2	0	3	0
optical_c0p8_p4	13	22	0
optical_c2_p2	0	0	0
optical_c2_p4	0	0	0
optical_c4_p2	0	0	0
optical_c4_p4	0	0	0

Source coverage is complete despite exact failure in the compressed arms. The source-width and effective-font diagnostics identify unfavorable geometry but cannot isolate its causal role from cross-dataset comparisons. GLM produces 382 instances of the instructed abstention "No information available" and two incorrect uppercase key-shaped strings among its 384 compressed outputs. This is not classified as a safety refusal or proof of absent source evidence. A separate prior ExactStrip development comparison (14/18 versus 12/18, five gains and three losses, exact McNemar $p = 0.72656$) fails its gate; it is preserved as a negative development result, not a successful method.

D GEOMETRY INTERVENTION DETAILS AND ALL SECONDARY TESTS

The layout intervention freezes 12 development and 48 held-out cases from new seeds, with approximately 8,192 reference source tokens and four distractors. Development evaluates three answer-blind layouts on Qwen2B. The selection hierarchy is EM, effective font size, width utilization, then layout identifier. Development scores of 0/12, 1/12, and 1/12 select full-width word wrapping through the frozen font-size tie-break. Development and evaluation templates are distinct; their absolute scores are not treatment contrasts.

Table 10: All layout-intervention conditions, exact correct out of 48.

Representation	Qwen2B	Qwen9B	GLM
Full text	47	46	48
Original layout	0	0	0
Full-width layout	6	21	7
Enlarged-image control	17	35	29

The original and efficient C2 conditions exactly match native source, vision, nonvision and total token counts, as well as input IDs. Recorded image-grid shapes, patch sizes, and merge sizes also match in all 144 source-paired comparisons. Input IDs contain fixed image placeholders; visual tensors and embeddings vary. Qwen uses 4,096 vision tokens; GLM uses 3,844. Each condition represents all

source characters on four pages. The enlarged control uses 10,000 Qwen vision tokens and 13,456 GLM vision tokens. It requires at least 16 processor pixels of font-em height and ten pixels for the smallest bounding-box height in a fixed lowercase/digit alphabet. These geometry criteria are answer-blind and do not establish recognizability. Raw controls must reach 42/48; enlarged controls must reach 36/48. All raw controls pass, and all enlarged controls fail their respective descriptive criterion.

Full-width minus original effects are 12.50 [6.25, 18.75], 43.75 [35.42, 52.08], and 14.58 [10.42, 18.75] percentage points. Gains/losses are 6/0, 21/0, and 7/0. The corresponding adjusted p values are 0.03125, 2.86102e-06, and 0.03125. Post-hoc observations of key echoing or explanatory transcription do not alter frozen exact-match scores.

Controlled scale construction. The 32 fresh five-record cases cross two lengths, two alphabets, and two entropy patterns with four instances per cell. Target positions rotate early/middle/late (11/11/10). We compute a single full-width source plan, rasterize it once, and resize the complete rendered page with LANCZOS to the rounded side length implied by each scale. The scaled block is centered on a white fixed canvas: 1,008 pixels for Qwen and 868 for GLM. Source content, line breaks, pagination, font choice before scaling, and canvas are identical within a triplet. Native vision, nonvision, and total counts are equal. Source-span and ink-box checks reject omissions or clipping before evaluation, but cannot guarantee preservation of perceptually distinct characters after resampling.

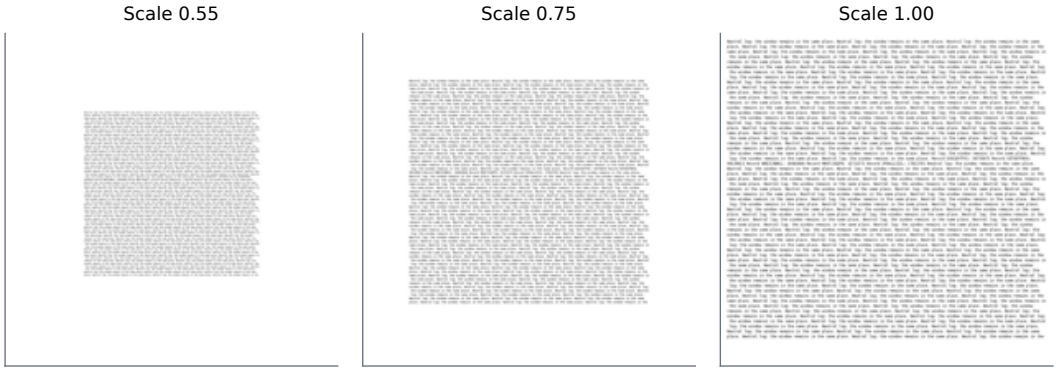


Figure 5: First native page of the predeclared example `geometry-000`, Qwen9B, at the three scales. These are full-canvas images, displayed at equal size. Source and page assignment are fixed.

Table 11: All controlled-scale conditions. Scale columns show correct / mean CER / cap hits, each with 32 outputs. Q is the exact conditional omnibus statistic; adjusted p uses three readers.

Reader	0.55	0.75	1.00	Q	Adj. p
Qwen2B	0 / 1.962 / 4	3 / 1.132 / 1	2 / 1.260 / 2	4.6667	0.2588
Qwen9B	3 / 1.007 / 2	10 / 0.355 / 1	14 / 0.135 / 0	10.9412	0.0110
GLM	0 / 1.638 / 0	4 / 0.945 / 0	5 / 0.521 / 0	5.2500	0.2588

The exact conditional test permutes scale labels within each case while holding its number of successes fixed. Cases with one or two successes have three equally likely assignments; cases with zero or three successes are uninformative for the permutation. Dynamic programming evaluates the finite null distribution, independently checked by multinomial enumeration. This is not an asymptotic chi-squared test.

Only Qwen9B's endpoint comparison passes this family. Its intermediate contrasts do not establish a monotonic population response. Qwen2B's observed response is nonmonotonic. The ten cap hits (seven Qwen2B and three Qwen9B) remain unchanged, and there are none for GLM. No case or output was replaced after inspection.

Table 12: All nine secondary paired scale contrasts. Differences and intervals are percentage points. Holm correction covers the entire nine-comparison secondary family.

Reader	Scales	Δ (pp)	95% interval	Gain/loss	Adj. p
Qwen2B	1.00–0.55	6.250	[0.000, 15.625]	2/0	1.000000
Qwen2B	1.00–0.75	-3.125	[-9.375, 0.000]	0/1	1.000000
Qwen2B	0.75–0.55	9.375	[0.000, 21.875]	3/0	1.000000
Qwen9B	1.00–0.55	34.375	[15.625, 53.125]	12/1	0.030762
Qwen9B	1.00–0.75	12.500	[-9.375, 34.375]	8/4	1.000000
Qwen9B	0.75–0.55	21.875	[6.250, 37.500]	8/1	0.312500
GLM	1.00–0.55	15.625	[3.125, 28.125]	5/0	0.437500
GLM	1.00–0.75	3.125	[-12.500, 18.750]	4/3	1.000000
GLM	0.75–0.55	12.500	[3.125, 25.000]	4/0	0.750000

Table 13: **Fixed-budget retrieval under hash retention.** Entries are retained text / full-source optical correct out of 30 at the same total-input upper bound. Interaction is the LONG optical-minus-text difference minus its SHORT value. Adjusted p is from the prespecified three-reader primary family.

Reader	SHORT	MEDIUM	LONG	Interaction (pp)	Adj. p
Qwen2B	30 / 27	22 / 22	18 / 16	3.33	1.0000
Qwen9B	30 / 30	22 / 27	18 / 22	13.33	0.7754
GLM	30 / 30	24 / 29	19 / 28	30.00	0.0352

E FINITE-BUDGET SOURCE COVERAGE: CONSTRUCTION AND DIAGNOSTICS

Each of 30 base cases has a unique ten-uppercase-letter key and a seven-digit numeric answer. Distractors use the same format and do not visually distinguish the target. The query and target persist across lengths as a prefix of the distractor stream grows. Target positions are balanced across 10%, 50%, and 90% of the record sequence, ten base cases per position. Complete-record sources are selected near 2,048, 6,144, and 8,192 reference-Qwen tokens before inference. Their native lengths differ by reader.

The input budgets are 4,173 total tokens for both Qwen readers and 3,916 for GLM. Optical uses four pages and respectively 4,096 and 3,844 vision tokens. The remaining allowance is the maximum native nonvision overhead across fresh queries with four images. Nonvision counts include the question, template, and image boundaries. The text arm uses the same total-input upper bound without padding. Source tokens and retained-source tokens are counted separately for coverage; neither should be added to a total that already contains them. Cross-boundary tokenization also prevents interpreting standalone source counts as an exact additive source/scaffolding decomposition. Figure 3 preserves the frozen coordinate $N_M(X)/(B-h_M(Q))$, where $h_M(Q)$ is the native templated-input count with an empty source and the actual question. We call $B - h_M(Q)$ the nominal text-source capacity. It is the source allowance after prompt overhead, not an exact, content-independent maximum: complete templated-input checks govern packing. All tested SHORT sources fit and all MEDIUM/LONG sources require retention, so this distinction does not change regime membership. Ranking remains query-independent even though the fit predicate accounts for question length. These inputs instantiate commitment conditional on prompt overhead; a store committed strictly before the query would need a predetermined reserve. No comparison of physical storage bytes is intended.

When full text does not fit, complete records are ranked by SHA256 of the frozen retention seed, base-case identifier, and record bytes. The packer greedily tests the full templated input and then restores retained records to original order. It receives no query, answer, or target-position information except the budget predicate's required templating context. Hash ranking is answer-blind and exchangeable in expectation, not a guarantee of identical realized target survival by position in a small sample.

The primary interaction for case i is

$$d_i = (Y^{\text{opt}}_{i,\text{LONG}} - Y^{\text{text}}_{i,\text{LONG}}) - (Y^{\text{opt}}_{i,\text{SHORT}} - Y^{\text{text}}_{i,\text{SHORT}}).$$

The exact two-sided sign-flip test enumerates the signed-magnitude distribution of these integer case interactions. Its null requires conditional sign symmetry; it is not an assumption-free test for any

Table 14: Fixed-budget diagnostics. Source/capacity is the mean native-source/nominal-text-source-capacity ratio. Differences are optical minus text. Target retained counts refer to the text condition out of 30; optical source coverage is complete.

Reader	Length	Source/capacity	Δ (pp)	95% interval	Target retained
Qwen2B	SHORT	0.4990	-10.00	[-23.33, 0.00]	30
Qwen2B	MEDIUM	1.4960	0.00	[-23.33, 23.33]	22
Qwen2B	LONG	1.9952	-6.67	[-30.00, 16.67]	18
Qwen9B	SHORT	0.4990	0.00	[0.00, 0.00]	30
Qwen9B	MEDIUM	1.4960	16.67	[-3.33, 36.67]	22
Qwen9B	LONG	1.9952	13.33	[-10.00, 36.67]	18
GLM	SHORT	0.4592	0.00	[0.00, 0.00]	30
GLM	MEDIUM	1.3783	16.67	[3.33, 30.00]	24
GLM	LONG	1.8384	30.00	[10.00, 50.00]	19

distribution with mean zero. The LONG secondary test is exact two-sided McNemar. For Qwen2B, the interaction is 3.33 [-20.00, 26.67] points (adjusted $p = 1.0000$). Qwen9B's is 13.33 [-10.00, 36.67] ($p = 0.7754$). GLM's is 30.00 [10.00, 50.00] ($p = 0.0352$).

Secondary LONG adjusted p values are 0.7754, 0.7754, and 0.0352. The Qwen9B and GLM primary/secondary contrasts coincide because SHORT is tied case by case. All primary native-containment decisions equal canonical EM in this particular experiment; this observed agreement does not change the prespecified primary metric. All quality outputs finish below the cap. Exact per-case input counts, x-coordinate ranges, target-survival records, gains/losses, literal EM, and CER are preserved in the accompanying analysis files.

E.1 Fresh compact query-independent storage control

This follow-up addresses whether GLM's finite-budget advantage survives a materially denser, still query-independent text representation. GLM LONG was selected because it was the historical positive operating point; the source cases, not the reader/regime choice, are fresh. The protocol, capacity gate, primary contrast, scorer, and claim matrix were frozen before held-out outputs. All 450 planned calls completed and passed the result audit, with no exclusions or selective reruns.

Cases and representations. The 150 independent sources contain approximately 8,192 reference-Qwen tokens (tolerance 40), with 50 targets each at early, middle, and late source positions. Both text packers rank complete records by SHA256 of the retention seed, case identifier, and original record bytes, greedily admit records, and restore source order. Original-format text preserves `Record KEY: VALUE`; compact text uses complete `KEY=VALUE` lines beneath the fixed `Records (key=number):` header. Neither storage policy receives the future query, target key, answer, evidence annotations, or model outcomes. The optical arm renders the entire source using full-width word wrapping on four 868×868 RGB pages, without query-conditioned highlighting or magnification. The fixed renderer uses a 12-pixel source font, 24-pixel margins, and complete source-span coverage checks. Model, tokenizer, and processor use the same pinned GLM snapshot as the preceding studies. Greedy BF16/SDPA batch-one inference disables thinking and permits 64 output tokens.

Prospective budget amendment. The common upper bound is $B = 4,025 = 3,844 + 53 + 128$: native vision allocation plus empty-query optical scaffold plus a fixed future-query reserve. Text packing checks empty-question native input plus the same 128-token reserve, then validates the actual query input without repacking. The original 3,916-token draft failed answer-blind native validation by one token on a fresh optical input; it was stopped and preserved before any inference. The amended ceiling is 2.78% larger, so the follow-up is not an exact replication of the old budget or old question-conditioned packing. Equal ceilings do not mean equal consumed tokens (Table 16); neither arm is padded. Native counts come from the pinned processor, including image-grid/placeholder checks, rather than token estimates.

Compact serialization increases mean per-source record capacity by 15.85%, with every source increasing. Its additional 5 included/correct targets over original-format text are descriptive; no

Table 15: **GLM LONG: fresh compact-storage control.** Each arm has 150 cases under $B = 4,025$. Inclusion is audited record presence, not visual readability. Conditional recovery is correct among included targets. Optical omits no target; neither text arm answers an omitted target. Mean records denotes the number of complete retained records, not token coverage.

Representation	Included	Correct	Conditional	Mean records
Original-format text	85/150	85/150	85/85	261.05
Compact text	90/150	90/150	90/90	302.44
Full-source optical	150/150	126/150	126/150	480.81

secondary significance test is added. Optical includes all 150 targets but fails on 24; text's observed failures are all omissions. These observations support the inclusion/recovery decomposition rather than better optical conditional reading.

Primary inference and sensitivity. The only confirmatory contrast is paired optical minus compact-text native single-answer containment: 24.0 percentage points, with 49 optical-only successes, 13 compact-only successes, and 88 ties. The exact two-sided McNemar test gives $p = 4.82 \times 10^{-6}$ at $\alpha = 0.05$; its family contains one test. Conditional discordant signs must be exchangeable under the null; equality within each position stratum suffices, but exactness is not claimed for arbitrary heterogeneous pooled mean-zero effects. The pointwise 95% paired percentile interval is [14.67, 33.33] points, from 50,000 triplet resamples within the three 50-case position strata. The prespecified conservative sensitivity interval, combining Bonferroni Clopper–Pearson bounds for the six stratum-specific gain/loss probabilities, is [-6.25, 49.13] points and includes zero. This wider interval is neither an inverted McNemar interval nor a second hypothesis test. The frozen primary claim rule uses the test and effect direction; neither interval establishes equivalence.

Table 16: Compact-storage native input accounting. All values are processor-native counts; source-native tokens average 7,084.33 (range 7,052–7,122). Text has zero vision tokens, so its nonvision count equals total. Source counts are diagnostics, not an additive part of the already templated total. Unused reserve is retained.

Representation	Vision	Nonvision range	Mean total	Total range
Original-format text	0	3,902–3,917	3,909.92	3,902–3,917
Compact text	0	3,904–3,917	3,910.93	3,904–3,917
Full-source optical	3,844	69–73	3,915.15	3,913–3,917

Descriptive diagnostics and scope. Early/middle/late successes are 35/50, 28/50, 22/50 for original-format text; 35/50, 32/50, 23/50 for compact text; and 48/50, 50/50, 28/50 for optical. Text inclusion equals those counts; optical includes all 50 in each stratum. Thus 22 of 24 optical misses occur late, without a subgroup inference or position-mechanism claim. Native containment, canonical EM, and literal EM agree on every output; all arms have zero cap hits. Abstentions and other incorrect outputs remain in their denominators. The result supports robustness to this specific compact source-only serializer at the selected GLM LONG operating point, not optimized text memory, learned compression, query-time retrieval with discarded-source access, other readers or lengths, or conversational memory. Existing systems profiles are not extended to the compact arm.

F Complete systems measurements

The predetermined source is `cross-000` at each of the three lengths. View A preserves the full source, with text and nearest-native-grid optical $C2/C4$ encodings. View B uses byte-identical budgeted inputs from the finite-budget quality study. A and B are distinct comparisons even when some conditions coincide. We do not pool their repeated measurements or infer cross-reader efficiency.

Before each repetition, the pipeline synchronizes CUDA, releases prior request/output tensors, collects garbage, empties unused allocator cache, and resets peak statistics. Models remain resident. Warmup completes before measured runs, with deterministic condition rotation and seeded length order. GPU

hooks measure the first top-level forward, including visual encoding and instrumentation. The token streamer distinguishes the prompt callback from the first generated token. No extra forward is introduced to measure prefill.

End-to-end measurement starts before representation work and includes rendering or text packing, native template/tokenizer/image processing, transfer, model execution, one output token, and synchronization. It excludes model loading, source generation, optical layout planning, native budget calibration, audit/export, logging, and network transfer. Text retention is recomputed inside timing. CPU filesystem caching and thermal state are not controlled. These boundaries limit deployment extrapolation even when the cell-wise comparisons are reproducible.

Tables report all cells as median [Q1,Q3] across ten invocations, not uncertainty over tasks. Full repetitions and additional summaries are in the supplement. Prefill includes vision encoding; total GiB values include resident weights, while Table 17 subtracts the recorded resident baseline. FLOPs, energy, and monetary costs were not measured.

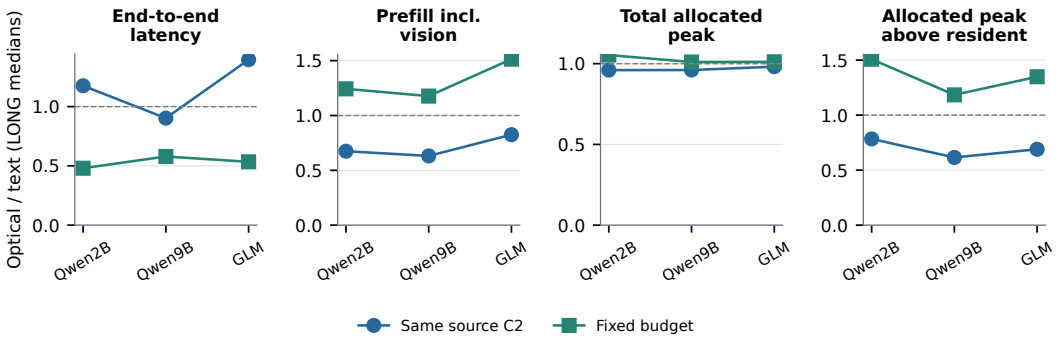


Figure 6: **Token budgets do not determine resource costs.** LONG ratios of optical to text medians; lower means less of the named resource. The final panel removes resident-model allocation using per-invocation differences. Same-source C2 reduces combined prefill and allocated peak for every reader. Fixed-budget optical is faster end to end here, yet has higher prefill and allocated peak than retained text. The end-to-end ranking includes asymmetric preprocessing implementations. Appendix F gives absolute values, all lengths, and dispersion.

Table 17: LONG allocated peak above the resident-model baseline, in GiB. Each cell is median [Q1, Q3] of per-invocation peak minus baseline over ten measured repetitions, not uncertainty over source documents. Total peaks remain in the complete memory table.

LONG arm	Qwen2B	Qwen9B	GLM
A: text	0.935 [0.935, 0.935]	1.990 [1.990, 1.990]	1.229 [1.229, 1.229]
A: C2	0.733 [0.733, 0.733]	1.225 [1.225, 1.225]	0.847 [0.847, 0.847]
A: C4	0.385 [0.385, 0.385]	0.674 [0.674, 0.674]	0.420 [0.420, 0.420]
B: text	0.486 [0.486, 0.486]	1.034 [1.034, 1.034]	0.672 [0.672, 0.672]
B: optical	0.733 [0.733, 0.733]	1.225 [1.225, 1.225]	0.906 [0.906, 0.906]

Table 18: Complete native token accounting. A: same source; B: fixed budget. S/M/L: SHORT/MEDIUM/LONG. Nonvision includes scaffolding; it is not added again to the total.

Reader	Len.	Arm	Source	Retained	Vision	Nonvision	Total	C
Qwen2B	S	A: text	2046	2046	0	2113	2113	–
Qwen2B	S	A: C2	2046	2046	1024	75	1099	1.998
Qwen2B	S	A: C4	2046	2046	484	75	559	4.227
Qwen2B	S	B: text	2046	2046	0	2113	2113	–
Qwen2B	S	B: optical	2046	2046	4096	75	4171	0.500
Qwen2B	M	A: text	6148	6148	0	6215	6215	–
Qwen2B	M	A: C2	6148	6148	3136	75	3211	1.960
Qwen2B	M	A: C4	6148	6148	1600	75	1675	3.842
Qwen2B	M	B: text	6148	4094	0	4161	4161	–
Qwen2B	M	B: optical	6148	6148	4096	75	4171	1.501
Qwen2B	L	A: text	8194	8194	0	8261	8261	–
Qwen2B	L	A: C2	8194	8194	4096	75	4171	2.000
Qwen2B	L	A: C4	8194	8194	2116	75	2191	3.872
Qwen2B	L	B: text	8194	4106	0	4173	4173	–
Qwen2B	L	B: optical	8194	8194	4096	75	4171	2.000
Qwen9B	S	A: text	2046	2046	0	2113	2113	–
Qwen9B	S	A: C2	2046	2046	1024	75	1099	1.998
Qwen9B	S	A: C4	2046	2046	484	75	559	4.227
Qwen9B	S	B: text	2046	2046	0	2113	2113	–
Qwen9B	S	B: optical	2046	2046	4096	75	4171	0.500
Qwen9B	M	A: text	6148	6148	0	6215	6215	–
Qwen9B	M	A: C2	6148	6148	3136	75	3211	1.960
Qwen9B	M	A: C4	6148	6148	1600	75	1675	3.842
Qwen9B	M	B: text	6148	4094	0	4161	4161	–
Qwen9B	M	B: optical	6148	6148	4096	75	4171	1.501
Qwen9B	L	A: text	8194	8194	0	8261	8261	–
Qwen9B	L	A: C2	8194	8194	4096	75	4171	2.000
Qwen9B	L	A: C4	8194	8194	2116	75	2191	3.872
Qwen9B	L	B: text	8194	4106	0	4173	4173	–
Qwen9B	L	B: optical	8194	8194	4096	75	4171	2.000
GLM	S	A: text	1769	1769	0	1831	1831	–
GLM	S	A: C2	1769	1769	900	70	970	1.966
GLM	S	A: C4	1769	1769	484	70	554	3.655
GLM	S	B: text	1769	1769	0	1831	1831	–
GLM	S	B: optical	1769	1769	3844	70	3914	0.460
GLM	M	A: text	5335	5335	0	5397	5397	–
GLM	M	A: C2	5335	5335	2704	70	2774	1.973
GLM	M	A: C4	5335	5335	1296	70	1366	4.117
GLM	M	B: text	5335	3849	0	3911	3911	–
GLM	M	B: optical	5335	5335	3844	70	3914	1.388
GLM	L	A: text	7108	7108	0	7170	7170	–
GLM	L	A: C2	7108	7108	3600	70	3670	1.974
GLM	L	A: C4	7108	7108	1764	70	1834	4.029
GLM	L	B: text	7108	3849	0	3911	3911	–
GLM	L	B: optical	7108	7108	3844	70	3914	1.849

Table 19: Complete preprocessing and combined-prefill timings (seconds). Entries are median [Q1, Q3] over ten executions of the predetermined case.

Reader	Len.	Arm	Render/pack	Processor	Prefill incl. vision
Qwen2B	S	A: text	0.000 [0.000, 0.000]	0.010 [0.010, 0.010]	0.363 [0.363, 0.365]
Qwen2B	S	A: C2	0.106 [0.106, 0.107]	0.034 [0.034, 0.037]	0.313 [0.313, 0.316]
Qwen2B	S	A: C4	0.096 [0.095, 0.097]	0.022 [0.021, 0.022]	0.246 [0.244, 0.251]
Qwen2B	S	B: text	0.004 [0.004, 0.004]	0.009 [0.009, 0.010]	0.362 [0.361, 0.365]
Qwen2B	S	B: optical	0.161 [0.161, 0.163]	0.139 [0.134, 0.143]	0.744 [0.738, 0.745]
Qwen2B	M	A: text	0.000 [0.000, 0.000]	0.018 [0.018, 0.019]	0.835 [0.833, 0.847]
Qwen2B	M	A: C2	0.316 [0.315, 0.316]	0.075 [0.075, 0.080]	0.585 [0.584, 0.590]
Qwen2B	M	A: C4	0.229 [0.229, 0.232]	0.052 [0.050, 0.055]	0.381 [0.380, 0.382]
Qwen2B	M	B: text	1.443 [1.438, 1.448]	0.013 [0.013, 0.013]	0.595 [0.593, 0.600]
Qwen2B	M	B: optical	0.331 [0.330, 0.333]	0.136 [0.134, 0.150]	0.737 [0.733, 0.740]
Qwen2B	L	A: text	0.000 [0.000, 0.000]	0.022 [0.022, 0.022]	1.093 [1.091, 1.095]
Qwen2B	L	A: C2	0.412 [0.412, 0.413]	0.137 [0.136, 0.138]	0.737 [0.735, 0.739]
Qwen2B	L	A: C4	0.307 [0.306, 0.307]	0.070 [0.069, 0.074]	0.450 [0.448, 0.452]
Qwen2B	L	B: text	2.119 [2.111, 2.125]	0.013 [0.013, 0.013]	0.593 [0.589, 0.598]
Qwen2B	L	B: optical	0.406 [0.406, 0.407]	0.139 [0.136, 0.148]	0.737 [0.736, 0.738]
Qwen9B	S	A: text	0.000 [0.000, 0.000]	0.010 [0.010, 0.010]	0.589 [0.588, 0.591]
Qwen9B	S	A: C2	0.107 [0.106, 0.107]	0.039 [0.036, 0.039]	0.438 [0.437, 0.444]
Qwen9B	S	A: C4	0.097 [0.097, 0.097]	0.021 [0.021, 0.024]	0.342 [0.340, 0.346]
Qwen9B	S	B: text	0.004 [0.004, 0.004]	0.010 [0.010, 0.010]	0.589 [0.587, 0.590]
Qwen9B	S	B: optical	0.163 [0.162, 0.164]	0.133 [0.133, 0.143]	1.311 [1.309, 1.314]
Qwen9B	M	A: text	0.000 [0.000, 0.000]	0.018 [0.018, 0.019]	1.586 [1.580, 1.591]
Qwen9B	M	A: C2	0.317 [0.316, 0.318]	0.086 [0.078, 0.089]	1.006 [1.003, 1.009]
Qwen9B	M	A: C4	0.231 [0.231, 0.232]	0.051 [0.051, 0.055]	0.567 [0.564, 0.567]
Qwen9B	M	B: text	1.460 [1.451, 1.476]	0.013 [0.013, 0.013]	1.103 [1.102, 1.109]
Qwen9B	M	B: optical	0.331 [0.331, 0.332]	0.141 [0.139, 0.149]	1.311 [1.309, 1.321]
Qwen9B	L	A: text	0.000 [0.000, 0.000]	0.023 [0.022, 0.023]	2.071 [2.065, 2.101]
Qwen9B	L	A: C2	0.414 [0.412, 0.415]	0.142 [0.137, 0.149]	1.308 [1.307, 1.311]
Qwen9B	L	A: C4	0.308 [0.307, 0.309]	0.072 [0.072, 0.078]	0.709 [0.704, 0.712]
Qwen9B	L	B: text	2.130 [2.125, 2.165]	0.014 [0.013, 0.014]	1.113 [1.109, 1.118]
Qwen9B	L	B: optical	0.410 [0.407, 0.411]	0.141 [0.137, 0.146]	1.310 [1.308, 1.316]
GLM	S	A: text	0.000 [0.000, 0.000]	0.009 [0.009, 0.009]	0.196 [0.196, 0.196]
GLM	S	A: C2	0.098 [0.098, 0.099]	0.018 [0.017, 0.019]	0.174 [0.174, 0.175]
GLM	S	A: C4	0.094 [0.094, 0.095]	0.014 [0.014, 0.015]	0.118 [0.118, 0.118]
GLM	S	B: text	0.003 [0.003, 0.003]	0.009 [0.009, 0.009]	0.196 [0.196, 0.196]
GLM	S	B: optical	0.143 [0.142, 0.145]	0.077 [0.073, 0.087]	0.668 [0.667, 0.670]
GLM	M	A: text	0.000 [0.000, 0.000]	0.016 [0.016, 0.016]	0.572 [0.571, 0.572]
GLM	M	A: C2	0.286 [0.286, 0.288]	0.050 [0.049, 0.054]	0.478 [0.478, 0.479]
GLM	M	A: C4	0.268 [0.268, 0.269]	0.024 [0.023, 0.026]	0.239 [0.239, 0.239]
GLM	M	B: text	1.154 [1.142, 1.158]	0.012 [0.012, 0.013]	0.403 [0.403, 0.404]
GLM	M	B: optical	0.312 [0.312, 0.313]	0.079 [0.077, 0.086]	0.669 [0.668, 0.669]
GLM	L	A: text	0.000 [0.000, 0.000]	0.019 [0.019, 0.019]	0.772 [0.771, 0.773]
GLM	L	A: C2	0.377 [0.377, 0.377]	0.072 [0.070, 0.079]	0.637 [0.637, 0.637]
GLM	L	A: C4	0.353 [0.353, 0.353]	0.029 [0.028, 0.032]	0.320 [0.319, 0.320]
GLM	L	B: text	1.718 [1.704, 1.731]	0.012 [0.012, 0.013]	0.442 [0.433, 0.447]
GLM	L	B: optical	0.389 [0.388, 0.389]	0.080 [0.077, 0.084]	0.669 [0.668, 0.670]

Table 20: Time to first token (TTFT) and one-token end-to-end time (seconds), median [Q1, Q3].

Reader	Len.	Arm	Model TTFT	Pipeline TTFT	End to end
Qwen2B	S	A: text	0.366 [0.365, 0.368]	0.376 [0.376, 0.378]	0.377 [0.376, 0.378]
Qwen2B	S	A: C2	0.318 [0.318, 0.321]	0.467 [0.462, 0.469]	0.468 [0.463, 0.470]
Qwen2B	S	A: C4	0.252 [0.250, 0.256]	0.374 [0.369, 0.376]	0.374 [0.369, 0.377]
Qwen2B	S	B: text	0.365 [0.364, 0.367]	0.379 [0.377, 0.381]	0.379 [0.378, 0.381]
Qwen2B	S	B: optical	0.750 [0.744, 0.751]	1.068 [1.066, 1.078]	1.068 [1.066, 1.078]
Qwen2B	M	A: text	0.838 [0.836, 0.850]	0.856 [0.855, 0.869]	0.857 [0.855, 0.870]
Qwen2B	M	A: C2	0.591 [0.590, 0.596]	0.999 [0.998, 1.007]	1.000 [0.998, 1.007]
Qwen2B	M	A: C4	0.387 [0.386, 0.388]	0.677 [0.675, 0.684]	0.677 [0.675, 0.684]
Qwen2B	M	B: text	0.597 [0.596, 0.602]	2.053 [2.050, 2.063]	2.053 [2.050, 2.063]
Qwen2B	M	B: optical	0.743 [0.739, 0.747]	1.233 [1.231, 1.248]	1.234 [1.231, 1.248]
Qwen2B	L	A: text	1.096 [1.094, 1.098]	1.119 [1.117, 1.121]	1.119 [1.117, 1.121]
Qwen2B	L	A: C2	0.743 [0.741, 0.745]	1.316 [1.313, 1.318]	1.316 [1.314, 1.318]
Qwen2B	L	A: C4	0.456 [0.454, 0.458]	0.845 [0.843, 0.852]	0.845 [0.843, 0.853]
Qwen2B	L	B: text	0.596 [0.592, 0.600]	2.725 [2.717, 2.737]	2.725 [2.718, 2.737]
Qwen2B	L	B: optical	0.743 [0.742, 0.744]	1.309 [1.307, 1.320]	1.309 [1.308, 1.321]
Qwen9B	S	A: text	0.592 [0.590, 0.594]	0.603 [0.601, 0.604]	0.603 [0.601, 0.605]
Qwen9B	S	A: C2	0.443 [0.442, 0.450]	0.596 [0.593, 0.599]	0.596 [0.593, 0.599]
Qwen9B	S	A: C4	0.348 [0.345, 0.352]	0.469 [0.466, 0.476]	0.470 [0.467, 0.477]
Qwen9B	S	B: text	0.592 [0.590, 0.593]	0.605 [0.604, 0.606]	0.606 [0.604, 0.607]
Qwen9B	S	B: optical	1.317 [1.315, 1.321]	1.641 [1.636, 1.649]	1.642 [1.636, 1.649]
Qwen9B	M	A: text	1.589 [1.583, 1.594]	1.608 [1.601, 1.613]	1.609 [1.602, 1.613]
Qwen9B	M	A: C2	1.013 [1.009, 1.015]	1.431 [1.425, 1.434]	1.431 [1.425, 1.434]
Qwen9B	M	A: C4	0.573 [0.570, 0.573]	0.864 [0.862, 0.869]	0.864 [0.862, 0.869]
Qwen9B	M	B: text	1.106 [1.105, 1.112]	2.581 [2.574, 2.600]	2.582 [2.575, 2.601]
Qwen9B	M	B: optical	1.318 [1.316, 1.328]	1.820 [1.808, 1.826]	1.821 [1.808, 1.826]
Qwen9B	L	A: text	2.074 [2.068, 2.104]	2.100 [2.094, 2.130]	2.101 [2.094, 2.131]
Qwen9B	L	A: C2	1.314 [1.313, 1.318]	1.896 [1.889, 1.901]	1.897 [1.890, 1.901]
Qwen9B	L	A: C4	0.715 [0.711, 0.718]	1.110 [1.103, 1.113]	1.110 [1.104, 1.113]
Qwen9B	L	B: text	1.116 [1.112, 1.121]	3.262 [3.257, 3.292]	3.262 [3.257, 3.292]
Qwen9B	L	B: optical	1.316 [1.315, 1.322]	1.889 [1.884, 1.895]	1.889 [1.884, 1.895]
GLM	S	A: text	0.199 [0.199, 0.199]	0.208 [0.208, 0.209]	0.209 [0.209, 0.209]
GLM	S	A: C2	0.181 [0.181, 0.181]	0.302 [0.301, 0.304]	0.302 [0.301, 0.304]
GLM	S	A: C4	0.125 [0.124, 0.125]	0.235 [0.234, 0.237]	0.236 [0.235, 0.237]
GLM	S	B: text	0.199 [0.199, 0.199]	0.211 [0.211, 0.211]	0.212 [0.211, 0.212]
GLM	S	B: optical	0.676 [0.675, 0.678]	0.917 [0.909, 0.923]	0.917 [0.909, 0.923]
GLM	M	A: text	0.575 [0.574, 0.575]	0.591 [0.590, 0.591]	0.591 [0.591, 0.592]
GLM	M	A: C2	0.486 [0.485, 0.486]	0.834 [0.833, 0.839]	0.834 [0.833, 0.839]
GLM	M	A: C4	0.246 [0.246, 0.246]	0.545 [0.543, 0.548]	0.545 [0.543, 0.548]
GLM	M	B: text	0.407 [0.406, 0.407]	1.573 [1.561, 1.578]	1.573 [1.562, 1.579]
GLM	M	B: optical	0.676 [0.675, 0.677]	1.082 [1.080, 1.089]	1.083 [1.081, 1.090]
GLM	L	A: text	0.775 [0.774, 0.776]	0.795 [0.794, 0.796]	0.795 [0.794, 0.796]
GLM	L	A: C2	0.645 [0.644, 0.645]	1.109 [1.106, 1.118]	1.110 [1.106, 1.119]
GLM	L	A: C4	0.327 [0.326, 0.327]	0.716 [0.715, 0.720]	0.717 [0.716, 0.720]
GLM	L	B: text	0.445 [0.436, 0.450]	2.176 [2.161, 2.184]	2.176 [2.161, 2.185]
GLM	L	B: optical	0.677 [0.676, 0.678]	1.162 [1.159, 1.167]	1.163 [1.160, 1.167]

Table 21: Total allocated and reserved peaks and resident allocation baseline (GiB), median [Q1, Q3].

Reader	Len.	Arm	Allocated peak	Reserved peak	Allocated baseline
Qwen2B	S	A: text	4.392 [4.392, 4.392]	4.520 [4.520, 4.520]	4.131 [4.131, 4.131]
Qwen2B	S	A: C2	4.317 [4.317, 4.317]	4.518 [4.518, 4.518]	4.131 [4.131, 4.131]
Qwen2B	S	A: C4	4.233 [4.233, 4.233]	4.365 [4.365, 4.365]	4.131 [4.131, 4.131]
Qwen2B	S	B: text	4.392 [4.392, 4.392]	4.520 [4.520, 4.520]	4.131 [4.131, 4.131]
Qwen2B	S	B: optical	4.864 [4.864, 4.864]	5.330 [5.330, 5.330]	4.131 [4.131, 4.131]
Qwen2B	M	A: text	4.841 [4.841, 4.841]	5.037 [5.037, 5.037]	4.131 [4.131, 4.131]
Qwen2B	M	A: C2	4.701 [4.701, 4.701]	5.090 [5.090, 5.090]	4.131 [4.131, 4.131]
Qwen2B	M	A: C4	4.425 [4.425, 4.425]	4.682 [4.682, 4.682]	4.131 [4.131, 4.131]
Qwen2B	M	B: text	4.617 [4.617, 4.617]	4.809 [4.809, 4.809]	4.131 [4.131, 4.131]
Qwen2B	M	B: optical	4.864 [4.864, 4.864]	5.330 [5.330, 5.330]	4.131 [4.131, 4.131]
Qwen2B	L	A: text	5.067 [5.067, 5.067]	5.457 [5.457, 5.457]	4.131 [4.131, 4.131]
Qwen2B	L	A: C2	4.864 [4.864, 4.864]	5.330 [5.330, 5.330]	4.131 [4.131, 4.131]
Qwen2B	L	A: C4	4.516 [4.516, 4.516]	4.852 [4.852, 4.852]	4.131 [4.131, 4.131]
Qwen2B	L	B: text	4.617 [4.617, 4.617]	4.809 [4.809, 4.809]	4.131 [4.131, 4.131]
Qwen2B	L	B: optical	4.864 [4.864, 4.864]	5.330 [5.330, 5.330]	4.131 [4.131, 4.131]
Qwen9B	S	A: text	18.103 [18.103, 18.103]	18.297 [18.297, 18.297]	17.547 [17.547, 17.547]
Qwen9B	S	A: C2	17.914 [17.914, 17.914]	17.977 [17.977, 17.977]	17.547 [17.547, 17.547]
Qwen9B	S	A: C4	17.764 [17.764, 17.764]	17.822 [17.822, 17.822]	17.547 [17.547, 17.547]
Qwen9B	S	B: text	18.103 [18.103, 18.103]	18.297 [18.297, 18.297]	17.547 [17.547, 17.547]
Qwen9B	S	B: optical	18.771 [18.771, 18.771]	19.041 [19.041, 19.041]	17.547 [17.547, 17.547]
Qwen9B	M	A: text	19.059 [19.059, 19.059]	19.814 [19.814, 19.814]	17.547 [17.547, 17.547]
Qwen9B	M	A: C2	18.509 [18.509, 18.509]	18.664 [18.664, 18.664]	17.547 [17.547, 17.547]
Qwen9B	M	A: C4	18.081 [18.081, 18.081]	18.164 [18.164, 18.164]	17.547 [17.547, 17.547]
Qwen9B	M	B: text	18.580 [18.580, 18.580]	19.008 [19.008, 19.008]	17.547 [17.547, 17.547]
Qwen9B	M	B: optical	18.771 [18.771, 18.771]	19.041 [19.041, 19.041]	17.547 [17.547, 17.547]
Qwen9B	L	A: text	19.537 [19.537, 19.537]	20.715 [20.715, 20.715]	17.547 [17.547, 17.547]
Qwen9B	L	A: C2	18.771 [18.771, 18.771]	19.041 [19.041, 19.041]	17.547 [17.547, 17.547]
Qwen9B	L	A: C4	18.221 [18.221, 18.221]	18.365 [18.365, 18.365]	17.547 [17.547, 17.547]
Qwen9B	L	B: text	18.581 [18.581, 18.581]	19.008 [19.008, 19.008]	17.547 [17.547, 17.547]
Qwen9B	L	B: optical	18.771 [18.771, 18.771]	19.041 [19.041, 19.041]	17.547 [17.547, 17.547]
GLM	S	A: text	19.548 [19.548, 19.548]	19.711 [19.711, 19.711]	19.234 [19.234, 19.234]
GLM	S	A: C2	19.453 [19.453, 19.453]	19.744 [19.744, 19.744]	19.234 [19.234, 19.234]
GLM	S	A: C4	19.353 [19.353, 19.353]	19.512 [19.512, 19.512]	19.234 [19.234, 19.234]
GLM	S	B: text	19.548 [19.548, 19.548]	19.711 [19.711, 19.711]	19.234 [19.234, 19.234]
GLM	S	B: optical	20.140 [20.140, 20.140]	20.875 [20.875, 20.875]	19.234 [19.234, 19.234]
GLM	M	A: text	20.157 [20.157, 20.157]	20.559 [20.559, 20.559]	19.234 [19.234, 19.234]
GLM	M	A: C2	19.876 [19.876, 19.876]	20.402 [20.402, 20.402]	19.234 [19.234, 19.234]
GLM	M	A: C4	19.542 [19.542, 19.542]	19.949 [19.949, 19.949]	19.234 [19.234, 19.234]
GLM	M	B: text	19.906 [19.906, 19.906]	20.219 [20.219, 20.219]	19.234 [19.234, 19.234]
GLM	M	B: optical	20.140 [20.140, 20.140]	20.875 [20.875, 20.875]	19.234 [19.234, 19.234]
GLM	L	A: text	20.462 [20.462, 20.462]	20.789 [20.789, 20.789]	19.234 [19.234, 19.234]
GLM	L	A: C2	20.081 [20.081, 20.081]	20.775 [20.775, 20.775]	19.234 [19.234, 19.234]
GLM	L	A: C4	19.654 [19.654, 19.654]	20.078 [20.078, 20.078]	19.234 [19.234, 19.234]
GLM	L	B: text	19.906 [19.906, 19.906]	20.219 [20.219, 20.219]	19.234 [19.234, 19.234]
GLM	L	B: optical	20.140 [20.140, 20.140]	20.875 [20.875, 20.875]	19.234 [19.234, 19.234]

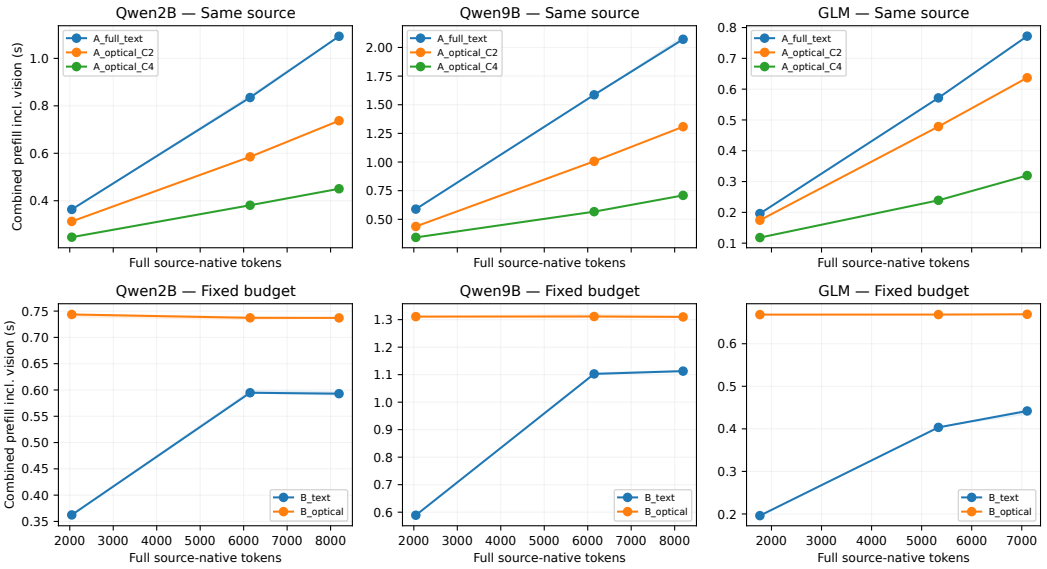


Figure 7: Complete original audited combined-prefill profiles, with same-source and fixed-budget views distinguished. Dispersion bands show run IQRs for the predetermined case, not confidence intervals over source documents. Per-reader axes may differ.

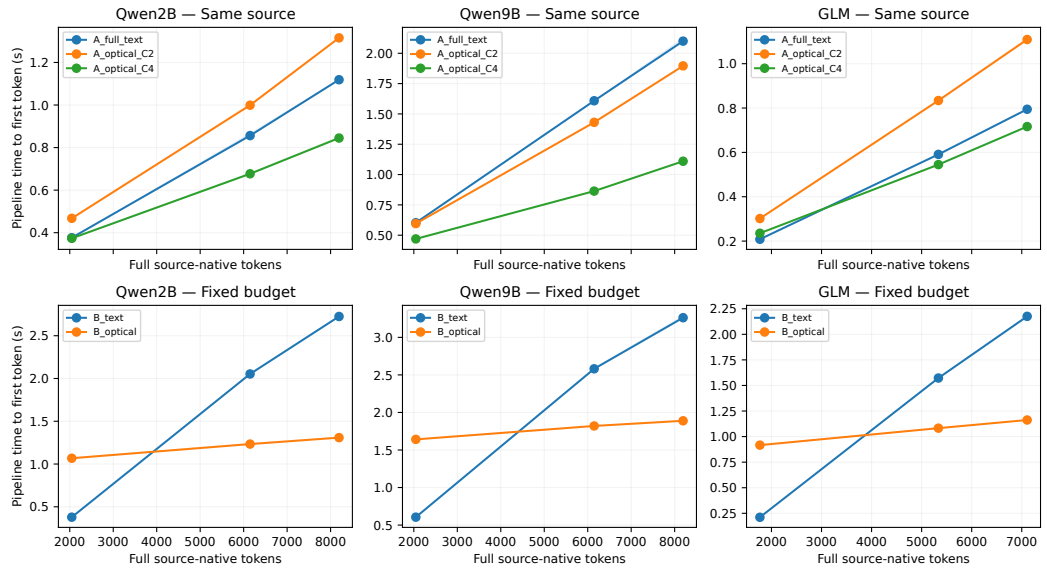


Figure 8: Complete original audited pipeline-TTFT profiles. Pipeline TTFT includes representation preparation and processing; optical layout plans are precomputed, while retained-text packing is recomputed. This asymmetry is part of the measured implementation.

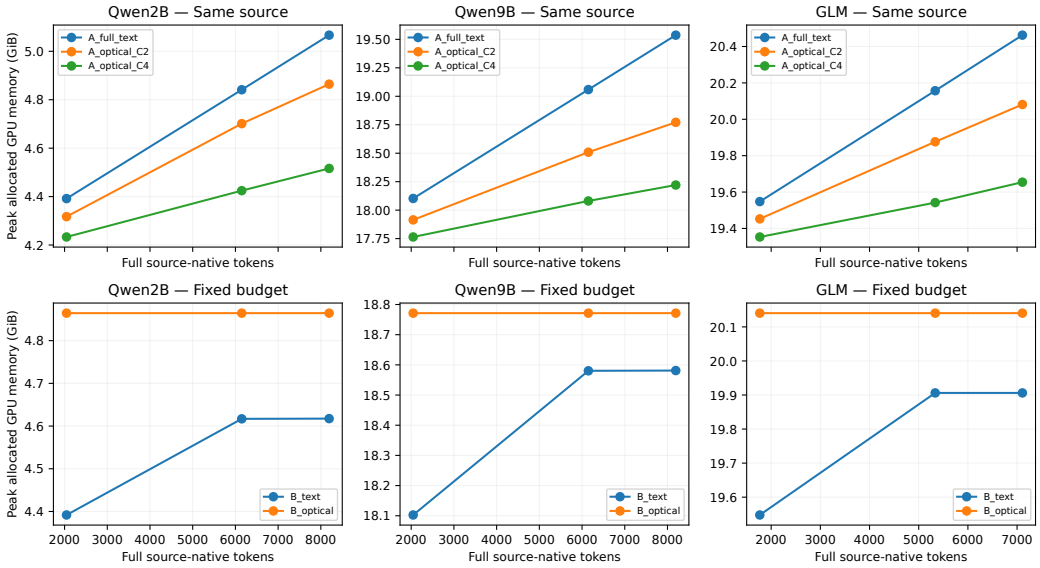


Figure 9: Complete original audited peak allocated-memory profiles. Resident model allocation is included. Per-panel scales can differ and need not start at zero; numerical differences should guide comparisons.

G REPRODUCIBILITY AND VERIFICATION

Protocols, inputs, rendering rules, decoder settings, and follow-up testing families were frozen before evaluation. Local freezes document prospective decisions within this project, not independent preregistration. Model and processor identities are pinned jointly: exact snapshot revisions and checksums for weights, tokenizers, processors, fonts, and dependency locks are supplied in the supplementary reproducibility manifest.

The preserved record includes raw decoded outputs, generated-token journals, per-case input manifests, render plans, and machine-readable scores and analyses. Independent checks recomputed scores from outputs, verified generated-token decoding, audited native processor inputs and source-span coverage, and reproduced renders. The controlled-scale primary test was also checked through a separate exact-enumeration implementation. The compact-storage audit verifies all 450 scheduled outputs, native input identities and generated-token journals; an independent containment/literal-EM implementation and binomial test reproduce its scores and primary test. Historical failures and bounded cap diagnostics remain recorded rather than being removed from the evidence.

The supplement contains authoritative aggregates, secondary analyses, per-case scoring exports, systems repetitions, protocols, and audit records. Its artifact index links these to the larger original execution archives, including token journals and complete render inventories. Hashes distinguish original artifacts from anonymous distribution copies. Detailed paths, snapshot identifiers, audit inventories, repair histories, and replay boundaries are in the supplementary `REPRODUCIBILITY_MANIFEST.md`; these details do not alter the scientific protocols or outcomes.

The manuscript's numerical ledger records authoritative JSON fields and displayed rounding. Figures and tables can be regenerated from the supplied evidence without further inference. The source bundle includes the official ICLR style, bibliography, and build instructions. Consistency checks establish agreement with preserved artifacts; they do not resolve training-data exposure or extend the experimental scope.