BF1: DETERMINISTIC CAUSAL MULTISCALE ATTENTION FOR EFFICIENT PRETRAINED-MODEL RETROFITTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dense causal attention is costly at long context even with IO-aware kernels. We study BF1, a deterministic block-sparse pretrained-model retrofit combining causal local attention, a global first block, and dyadic historical blocks. Its selected interaction count is $O(n \log n)$ at fixed block width; it retains softmax over selected tokens and the ordinary KV cache. On Qwen3-0.6B, a controlled implementation reaches 10.91× per-layer prefill acceleration at 32K, an eight-of-28-layer retrofit reduces whole-model time to first token by 15.3%, and BF1 ranks first across three matched adaptation seeds on 2,048-token report packs. A separate engineering study extends the route to Qwen3.8-27B on B200. At one million tokens, the best observed BF1 attention layer takes 366.5 ms against a recorded 35.9 s dense SDPA call (98.0×), with 540× fewer interactions. Whole-model time to first token reaches a best observed 78.3 s against a single 663 s dense run (8.47×); these asymmetric observations are not estimates of typical speedup. A separate footprint record reports 119.8 GB total allocation for both paths. Effective selected-page decode ratios range from 4.1× to 37.5× at 1M across batches one to 32. A six-example, training-free oracle-routing diagnostic recovers four answers against zero for a reported budget-matched random control. Together, these studies establish a controlled small-model result and large-scale execution feasibility, while exposing the distinction between sparse execution, evidence addressing, and broad quality preservation.

1 INTRODUCTION

Transformer attention gives every query direct access to all earlier key–value pairs, but its causal interaction count grows quadratically with context length (Vaswani et al., 2017). FlashAttention reduces memory traffic without changing this all-pairs relation (Dao et al., 2022). Sparse attention changes the relation itself and therefore faces two independent obligations: it must reduce physical work, and it must preserve useful communication in a pretrained model.

We study a deliberately simple retrofit called BF1. Tokens are grouped into fixed-width blocks. Each query block attends causally within itself, to a small local neighborhood, to the first block, and to predecessor blocks at dyadic distances. The route is deterministic and query independent. It does not require a learned query-time selector, and it maps naturally to block-sparse prefill and selected-page decode. Logarithmic sparse connectivity is prior art (Li et al., 2019; Ding et al., 2023); our contribution is the combined correctness, control, scaling, and pretrained-model evidence.

The original study demonstrated prefill speedups over optimized dense attention at 8K–32K, including a 10.91× per-layer speedup at 32K, and replicated matched adaptation on Qwen3-0.6B. This paper adds a qualitatively different scale test: Qwen3.8-27B has 16 full-attention layers interleaved with 48 Gated DeltaNet layers, 24 query heads, four KV heads, and head dimension 256 (Qwen Team, 2026). On a B200, the BF1 operator reaches 98× best-observed one-layer prefill acceleration at 1M, and the complete 27B model reaches 8.47× best-observed TTFT acceleration after all 16 full-attention layers are converted. The ordinary KV cache is retained; BF1 does not claim cache compression.

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.

We make five contributions. First, we specify the public causal dyadic route and prove its $O(n \log n)$ selected-interaction law and logarithmic communication depth. Second, we report correctness-gated BF16 execution and matched topology controls. Third, we replicate the Stage-A language-modeling ordering across three training seeds rather than relying on a single trained checkpoint. Fourth, we demonstrate operator and whole-model scaling through one million tokens on a 27B hybrid-attention model. Fifth, oracle replacement, reported budget-matched random pages, and single-layer interventions test whether the integration can consume correctly routed evidence. The small and large models answer different experimental questions; the large-model timings do not replace the controlled small-model evidence.

2 Related work

Content-independent sparse patterns include Sparse Transformer, Longformer, and BigBird (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020); hierarchical designs provide another structured alternative (Zhu and Soricut, 2021). Reformer instead constructs content-dependent attention neighborhoods through locality-sensitive hashing (Kitaev et al., 2020). LogSparse Transformer and LongNet are the closest route-level precedents because they use exponentially or logarithmically separated history (Li et al., 2019; Ding et al., 2023). Pixelated Butterfly applies butterfly structure more broadly to neural matrices (Dao et al., 2021). BF1 does not claim novelty for dyadic predecessors alone.

Content-dependent methods answer a different question. Quest, MInference, XAttention, DuoAttention, HiP, and Native Sparse Attention select or allocate history based on queries, heads, or learned compression (Tang et al., 2024; Jiang et al., 2024; Xu et al., 2025; Xiao et al., 2024; Lee et al., 2025; Yuan et al., 2025). They may improve relevance but add their own training or planning contracts. BF1 instead provides a regular deterministic substrate; the oracle diagnostic in Section 5.5 shows that semantic addressing is complementary rather than unnecessary.

FlexPrefill selects both patterns and budgets for each input and head (Lai et al., 2025); Delta Attention corrects sparse-prefill-induced distribution shift (Willette et al., 2025). These methods highlight two limitations of a deterministic graph: relevance is not guaranteed by connectivity, and accurate computation on a sparse support need not reproduce dense hidden states. No cross-paper ratio is used as a head-to-head comparison; hardware, head dimension, precision, adaptation, and timer boundaries differ.

Hybrid models provide a particularly informative systems target. Gated DeltaNet combines gating and delta-rule state updates (Yang et al., 2025b). Qwen3.8-27B interleaves these layers with full attention (Qwen Team, 2026); BF1 changes the remaining full-attention computations, not the recurrent state updates. YaRN extends positional support (Peng et al., 2023), but successful execution at an extended position is not a quality certificate.

At the systems layer, FlexAttention compiles custom score modifications and block masks (Dong et al., 2024); FlashInfer supplies paged attention primitives for serving (Ye et al., 2025). Our measurements separate selected interactions, one-layer execution, per-call planning, whole-model TTFT, decode, and KV residency. These are different endpoints and must not be substituted for one another.

3 BF1 operator

3.1 Causal dyadic route

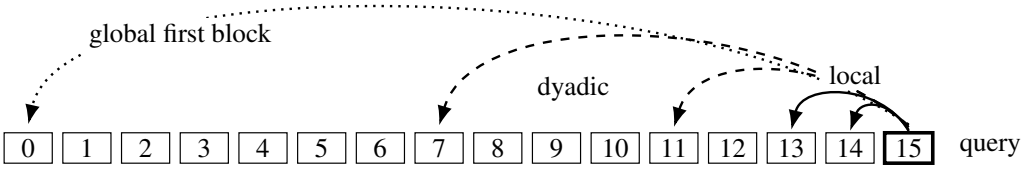
Partition a sequence of n tokens into $B = \lceil n/b \rceil$ blocks of width b. For query block j, define

$$\mathcal{P}(j) = \{j, j-1, j-2, 0\} \cup \{j - 2^r : r \geq 2\}, \tag{1}$$

with invalid and duplicate indices removed. The self block uses a token-level causal triangle; all other selected blocks are strictly earlier. The evaluated public route uses $b = 64$, two local predecessor blocks, the first global block, and dyadic offsets beginning at four blocks. For query token q, let $S(q)$ be the union of valid tokens in $\mathcal{P}(\lfloor q/b \rfloor)$ satisfying $k \leq q$. Each query head computes

$$y_q = \frac{\sum_{k \in S(q)} \exp(Q_q^\top K_k / \sqrt{d_h}) V_k}{\sum_{k \in S(q)} \exp(Q_q^\top K_k / \sqrt{d_h})}. \tag{2}$$

"Exact selected attention" means this masked-softmax operation, not equality with dense attention. GQA maps each query head to its designated KV head. Absolute positions, causal tails, duplicates, and partial final blocks remain part of the operator contract.



Selected pages for query 15: $\{0, 7, 11, 13, 14, 15\}$; self attention is causal.

Figure 1: BF1 block route. The query block receives exact local context, the first global block, and dyadic historical predecessors.

Proposition 1 (Selected interactions)**.** *For fixed block width b, one BF1 layer uses $O(n \log n)$ selected token interactions.*

Proof. Each block selects at most $4 + \lfloor \log_2(j+1) \rfloor$ predecessor blocks. Summing over B blocks gives $O(B \log B)$ selected block pairs. Each contributes at most b^2 token pairs, yielding $O(b^2 B \log B) = O(nb \log(n/b))$, which is $O(n \log n)$ for fixed b. □

For one-token cached decode, a query reads $O(\log B)$ selected blocks. The exact historical KV store remains $O(n)$; BF1 is a sparse-read method, not compressed memory. A binary-distance argument gives $O(\log B)$ communication depth; at $B = 512$ the finite-graph maximum is eight hops. Appendix A gives the full proof and mixed-stack accounting.

For full blocks, an exact per-head pair count is

$$E(n) = \sum_{j=0}^{B-1} \left[\frac{b(b+1)}{2} + b^2 \big(|\mathcal{P}(j)| - 1 \big) \right], \quad n = Bb. \tag{3}$$

The supplement handles a partial last block. At $n = 10^6$, an independent reconstruction gives a dense-to-selected ratio of 540.35, consistent with the reported rounded 540. The 16-slot kernel envelope accommodates the route throughout the tested range; it does not make the uncapped dyadic graph constant-degree at arbitrary lengths.

4 Experimental design

4.1 Small-model matched study

The controlled study uses Qwen3-0.6B-Base (Yang et al., 2025a), BF16, and an RTX PRO 6000 Blackwell GPU: 28 decoder layers, 16 query heads, eight KV heads, and head dimension 128. Layers 10–17 are adapted. BF1 is compared with dense continued training (Dense-CT), equal-pair-count local sliding, and an equal-pair-count static-random nonlocal graph. Dense-CT keeps the sparse blend closed while retaining the same trainable parameter set.

All four arms adapt 50,333,696 parameters for 1,000 steps and 16.384M tokens. Their objective combines next-token loss, frozen-teacher logit distillation, and hidden-state matching. The documented Stage-A configuration uses 2,048-token training packs. A local 49,000-row question–answer source described as ORCA-style is normalized and content-hash split; Orca (Mukherjee et al., 2023) is background attribution, not the source dataset's immutable identifier. The retained preparation log records 34 exact duplicates removed and 46,967/1,004/995 train/selection/report rows. Minimum selection NLL fixes each checkpoint before report evaluation. Executed report logs explicitly give 174 packs of 2,048 tokens and 356,178 next-token predictions ($174 \times 2,047$). This is a 2K next-token result, not a 32K or 1M perplexity result. Training seeds are 1234, 2026, and 3407; random graph draw 17 stays fixed. Appendix C separates recovered packing details from unresolved dataset and run-manifest provenance.

Table 1: Report perplexity on 2,048-token packs across matched training seeds; lower is better. The same report corpus is reused, so packs are not pooled as independent observations.

Arm	1234	2026	3407	Mean
BF1 dyadic	**1.68634**	**1.68622**	**1.68660**	**1.68639**
Static random 17	1.69145	1.69128	1.69188	1.69154
Dense-CT	1.69241	1.69226	1.69308	1.69258
Matched sliding	1.81512	1.81486	1.81517	1.81505

Small-model prefill timing uses five warmups, 30 measured repetitions, and opposite pattern orders; the published statistic is the median of the two order medians. Numerical gates cover partial length 257, compiled forward/backward, future isolation, closed-blend dense equivalence, and cached decode. Maximum BF16 error against the routed reference is 0.00390625 on the tested short cases. These controlled checks are not inferred from the later six-row 27B diagnostic.

4.2 QWEN3.8-27B B200 STUDY

The large-model study converts the 16 full-attention layers of Qwen3.8-27B and leaves its 48 DeltaNet layers unchanged. The attention geometry is 24 query heads, four KV heads, head dimension 256, BF16, batch one for prefill, and YaRN extension to one million positions. The dense comparator is PyTorch SDPA using its FlashAttention backend; BF1 uses a frozen degree-16 SM100 kernel. Selected-page decode uses FlashInfer 0.6.18.

The 27B prefill microbenchmark uses synthetic tensors with this geometry, two warmups and three measured repetitions in one order. The whole-model runs use the actual 27B model without BF1-specific adaptation. These are short engineering protocols, not repetitions of the controlled 0.6B study. We distinguish $1\text{M} = 1,000,000$ prompt tokens from the separate 2^{20}-position boundary test.

Where several 27B BF1 timings exist, we report the *best observed* BF1 latency against the original recorded dense comparator, not the worst dense observation. At 1M, this is a minimum over four reported BF1 executions versus one dense execution. The resulting ratio characterizes an attained observation, not a matched-minimum or median speedup; unequal repeat opportunities favor the selected minimum. Session-to-session spread and incomplete thermal/order metadata limit stronger inference. The historical 0.6B timing statistic and the retained FlashInfer decode medians are not relabeled as minima. Appendix D separates these protocols and preserves single-session conversion-depth controls.

5 RESULTS

5.1 ORIGINAL 0.6B RESULT: FAST OPERATOR AND REPLICATED MATCHED ADAPTATION

At 32K, the optimized BF1 layer is 10.91× faster than dense SDPA with 26.98× fewer interactions. Activating BF1 in eight of 28 layers lowers warm whole-model TTFT from 589.1 to 499.1 ms, a 15.3% reduction; the remaining dense layers keep the model asymptotically quadratic. The matched quality result replicates across all three training seeds (Table 1). Continued adaptation explains most of the improvement over the unadapted base, but the smaller BF1 increment over Dense-CT and static-random graph 17 is directionally stable.

The nonlocal/local contrast is larger than the BF1-specific increment: mean report PPL is 7.63% higher for sliding, 0.305% higher for random-17, and 0.367% higher for Dense-CT relative to BF1. Dense-CT explains approximately 92.7% of the base-to-BF1 PPL improvement. Thus adaptation supplies most of the gain, while the smaller residual topology effect replicates. Training-seed consistency does not establish superiority over a distribution of random graphs.

A second 32K session reports a 15.7% reduction. An all-28-layer timing-only configuration reaches 2.14×, but has no matched quality qualification. Further kernel tuning changes the eight-layer model TTFT by less than 0.5% despite improving its layer kernel by about 14%: coverage of the model, not kernel speed alone, limits the gain.

Table 2: Controlled Qwen3-0.6B systems results on RTX PRO 6000. The actual retrofit converts eight of 28 layers; the remaining 20 dense layers keep its full attention cost quadratic.

Context	Layer speedup	Dense TTFT	BF1 TTFT	TTFT reduction
8K	2.84×	69.0 ms	63.7 ms	7.7%
16K	5.50×	190.1 ms	168.6 ms	11.3%
32K	10.91×	589.1 ms	499.1 ms	15.3%

Table 3: One-layer BF16 full-prefill scaling on B200 using Qwen3.8-27B attention geometry. BF1 values are best observed; dense values are recorded comparators.

Context	Dense	BF1 best	Ratio	Interaction reduction
8K	2.41 ms	1.48 ms	1.63×	8.59×
16K	8.68 ms	3.15 ms	2.76×	15.10×
32K	33.4 ms	7.06 ms	4.73×	26.98×
128K	537 ms	35.5 ms	15.1×	89.0×
1M	35.9 s	0.3665 s	**98.0**×	**540**×

5.2 One-layer prefill scales to one million tokens

Table 3 and Figure 2 extend the operator study by more than an order of magnitude in context and by model geometry. The route-level reductions at 8K–32K match the earlier study exactly because the route is unchanged, while realized speed differs with GPU and head dimension. At 1M, dense SDPA takes 35.9 s for one full-attention layer; the *best observed* BF1 execution takes 366.5 ms, a 98.0× ratio with 540× fewer interactions. The measured implementation therefore realizes roughly 18% of the pair-count ratio at that point. Fixed launch, traversal, softmax, and memory costs prevent equality with the interaction reduction.

5.3 Whole-model Qwen3.8-27B scaling

The operator advantage survives integration. Table 4 reports the complete model endpoint after all 16 full-attention layers are converted. The *best observed* 1M BF1 run reaches first token in 78.3 s versus 663 s for dense, an 8.47× ratio. Projections, feed-forward blocks, 48 DeltaNet layers, normalization, and other work remain. An Amdahl-style explanation is consistent with the smaller model-level ratio, but the separate kernel and model runs do not identify an exact attention-time fraction. The overall long-context advantage is large; the best-observed speedup series is not strictly monotone between 128K and 256K.

A separate record reports 119.8 GB total allocation for both paths at 1M, not isolated KV payload; allocator and unit definitions remain unresolved (Appendix D.5). This record is not paired with the fastest 78.3 s run. The best recorded decode is 49.9 versus 82.6 ms/token (1.66×), again an asymmetric selected observation. BF1 also executes 1,048,448 prompt tokens plus 128 output positions, reaching a configured 1,048,576-position YaRN boundary with 110 s TTFT and 122.9 GB allocated. The boundary run is a different input length and is not pooled with 1M for speedup inference.

The conversion-depth response is monotone at both measured lengths. Its 128K all-16-layer observation is 16.6 s (1.28× versus dense), not the separate 10.3 s best-observed timing in Table 4; session/run identities have not been reconciled across these protocols. At 512K, increasing coverage from four to eight to 16 full-attention layers reduces TTFT from 159.7 to 124.7 to 64.2 s. This supports attribution of the long-context gain to the converted attention path rather than a fixed harness difference. It does not establish quality equivalence for every conversion depth. Appendix D gives the complete protocol tables.

5.4 Selected-page decode

Direct selected-page execution is already cheap, but per-call planning remains costly. At 1M, effective BF1 paged decode is 4.1×, 15.7×, and 37.5× faster than dense paged decode at batches 1, 8, and 32 (Table 6). The run-only kernel is faster still, but treating plan elimination as achieved would

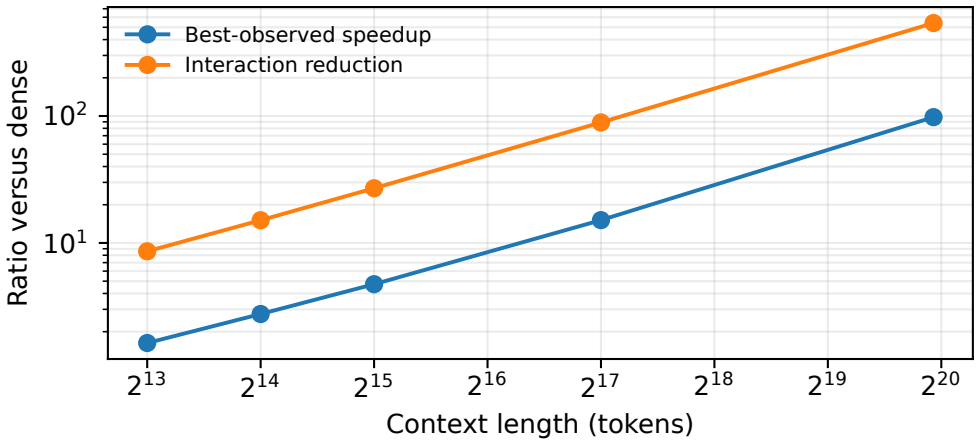


Figure 2: Best-observed BF1 speedup and selected-interaction reduction versus dense on B200. Both axes are logarithmic.

Table 4: Whole-model Qwen3.8-27B TTFT on B200. BF1 is best observed; dense is the recorded comparator. The cells are engineering measurements rather than distributional latency estimates.

Context	Dense TTFT	BF1 best TTFT	Ratio
32K	4.20 s	4.06 s	1.03×
128K	21.3 s	10.3 s	2.07×
256K	59.7 s	35.3 s	1.69×
512K	201 s	64.2 s	3.13×
1M	663 s	78.3 s	**8.47**×

be a projection. Persistent plan reuse is a candidate optimization. Page selection may stay fixed within a 64-token block, but tail length, cache writes and batch metadata still change; validity of a reused plan must be checked, then measured across complete generation.

5.5 Capability localization on the 27B integration

The systems result does not by itself establish useful evidence transport. A six-row diagnostic, reported at nominal 32K with zero BF1-specific training, contrasts oracle page replacement, oracle addition, and matched random. The experiment summary describes replacement and random as matched in budget, distance, and degree. Replacement and addition each score 4/6, while random scores 0/6 (Figure 4). The successes are the three single-source positions (early, middle, late) and the corroborated row. Under the reported matching, this supports an effect of evidence content rather than added capacity.

A route-accounting limitation prevents a stronger budget claim. The summary reports degree 12 for replacement and random, whereas Equation 1 has maximum degree 11 over a 32,768-token prompt. Degree 12 first becomes possible at query block 513 (token position 32,832). Exact diagnostic query positions, pre/post-override page lists and counting conventions were not retained in the available record, so neither explanation nor equality to the unmodified route is established (Appendix E). Addition's realized degree is separately unlogged. Equal 4/6 scores do not prove policy equivalence. Fully dense and unmodified all-BF1 outcomes on these same six rows are also absent from the available comparison; we therefore do not quantify recovery relative to either baseline.

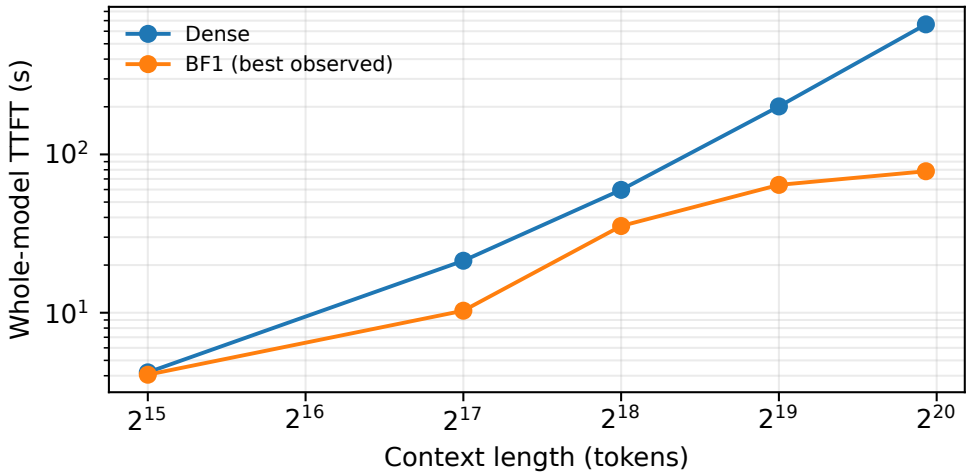


Figure 3: Whole-model TTFT for dense and all-16-layer BF1 Qwen3.8-27B. BF1 points are best observed.

Table 5: Conversion-depth intervention. Values are single engineering measurements; all non-attention components are unchanged.

Context	Dense	BF1 4/16	BF1 8/16	BF1 16/16
128K	21.3 s	18.9 s	17.1 s	16.6 s
512K	201.0 s	159.7 s	124.7 s	64.2 s

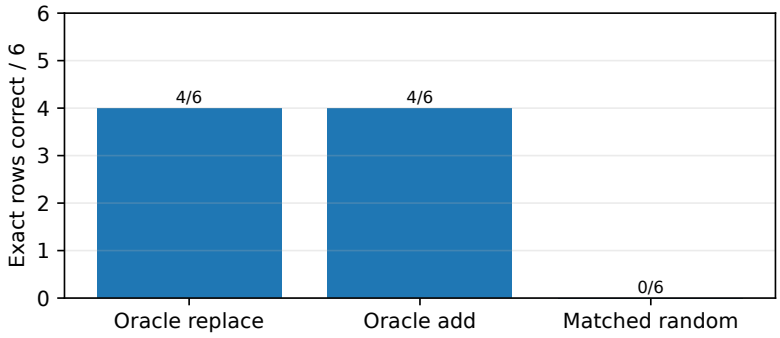


Figure 4: Training-free 27B oracle-routing diagnostic at nominal 32K. Oracle replacement and its reported budget-matched random control score 4/6 and 0/6. Addition also scores 4/6, but its realized degree was not separately logged. Exact degree accounting remains unresolved.

Single-layer interventions provide a second integration check. Converting any one of the 16 full-attention layers preserves 5–6 of six correct rows, and 12/16 substitutions have final-logit KL below 0.05. No individual layer is catastrophically incompatible with BF1 on this diagnostic; full-conversion degradation is therefore consistent with cumulative interaction effects. This does not prove broad capability preservation. The six-row panel leaves two rows unsolved, and their failures are not uniquely localized by the oracle contrast. Four oracle-only successes and no random-only success give two-sided exact McNemar $p = 0.125$; the intervention is suggestive mechanistic evidence, not a powered population test. No 1M quality result follows from a 32K diagnostic.

Table 6: One-layer 1M paged decode on B200 (μs). "Effective" includes the recorded plan-plus-run endpoint.

Batch	Dense	BF1 effective	Effective ratio	BF1 run only
1	880	217	4.1×	52
8	4,854	309	15.7×	65
32	18,578	496	37.5×	94

6 Discussion and limitations

What is established. The same deterministic route exhibits three levels of evidence: a formal $O(n \log n)$ selected-work law, an operator gain that reaches 98× best observed at 1M, and a complete-model gain that reaches 8.47× best observed TTFT at 1M. The conversion-depth intervention supports systems attribution. The oracle/random contrast shows that the 27B execution substrate can consume correctly routed evidence under the reported intervention; exact budget equivalence still requires the original route records.

What is not established. The large-model study does not show 98× end-to-end acceleration, KV compression, or broad quality preservation. The 1M whole-model dense comparator is a long single engineering run; BF1 values are reported as minima when repeated runs exist. Best-observed latency is useful for attained performance but is not a median, expectation, or tail guarantee. Decode aggregates use only the retained summary values. Order-controlled matched repetition remains necessary before claiming typical latency or tail performance.

The matched quality study uses one 0.6B model family and 2,048-token packs from a local ORCA-style next-token corpus whose upstream immutable dataset identity is unresolved. Static-random graph draw 17 is fixed, so training-seed replication is not graph-draw replication. The six-row 27B oracle test is exploratory rather than statistically conclusive. A separate small-model 32K checkpoint-by-route diagnostic found a 31–32-point structured-task loss when applying the BF1 route (Appendix C.6); the favorable next-token PPL result therefore cannot stand in for retrieval or aggregation preservation. External blind corpora, broader retrieval/aggregation tasks, trained 27B adaptation, concurrency, energy, and persistent-plan generation remain open.

Architectural implication. The oracle contrast separates evidence access from execution: a deterministic route does not identify which pages a query needs. A practical semantic selector needs its own latency, memory, causal-validity and held-out quality measurements. Unchanged cardinality alone would neither certify diagnostic budget equivalence nor preserve the measured operator and whole-model ratios.

Table 7: Claim boundaries for the principal results.

Observation	Supported reading	Not established
98.0× at 1M	best-observed one-layer prefill ratio	expected or whole-model 98×
8.47× at 1M	best-observed complete-model TTFT ratio	throughput, p95, or median service gain
119.8 GB for both	equal reported total allocation, not isolated KV	a verified allocator metric or memory saving
4/6 vs 0/6	evidence contrast under reported matching	deployed addressing or broad quality parity
Three-seed PPL	replicated matched next-token ordering	general retrieval and composition preservation

7 Conclusion

BF1 is a deterministic causal multiscale retrofit with $O(n \log n)$ selected interactions. The controlled 0.6B study combines correctness checks, replicated matched adaptation and useful systems gains. On Qwen3.8-27B at one million tokens, best-observed operator and whole-model TTFT ratios are 98.0× and 8.47× against their recorded dense comparators, not typical-speedup estimates. BF1 retains the ordinary KV cache. Oracle routing yields 4/6 versus 0/6 under reported matching, but exact degree accounting remains unresolved. These results support a selective-retrofit primitive, not universal quality preservation. Semantic addressing, multi-source composition and broader transfer remain open tests.

AI USE STATEMENT

Generative AI tools, including ChatGPT, Claude and Codex, assisted in interpretation of result records, literature review, and manuscript preparation. The research workflow also used these tools for hypothesis development, experiment-design feedback, theoretical formulation and proof drafting, method implementation and debugging, and construction of synthetic diagnostic generators. AI assistance also produced plotting, typesetting and consistency-checking code. This manuscript preparation used the recorded experiment observations rather than generating new neural benchmark results. Its checks cover aggregate arithmetic, source consistency, reference formatting, and document construction; they are not an independent neural replication. The authors take responsibility for the submitted text, claims and experimental artifacts.

ETHICS STATEMENT

The reported experiments concern language-model computation, an existing ORCA-style corpus, and constructed diagnostic inputs; no new human-subject study is reported. Sparse execution can omit task-critical evidence despite favorable latency or next-token metrics. We distinguish selected best-observed timings, controlled historical measurements, and exploratory quality observations, and make no high-stakes deployment, energy-saving or KV-compression claim.

REPRODUCIBILITY STATEMENT

Appendix A gives the route proofs and exact selected-pair counting; Appendix C retains the controlled small-model measurements; Appendix D defines the B200 endpoints and complete supplied tables. The supplement contains editable manuscript and figure source, aggregate data, a reference route enumerator, and arithmetic/document checks. It reproduces the paper's tables and formal counting, not neural experiments requiring the original checkpoint, kernel, raw outputs and execution environment. That experimental-access boundary is explicit in Appendix G.

REFERENCES

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Tri Dao, Beidi Chen, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Ré. Pixelated Butterfly: Simple and Efficient Sparse Training for Neural Network Models. *arXiv preprint arXiv:2112.00029*, 2021.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, 2022.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. LongNet: Scaling Transformers to 1,000,000,000 Tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex Attention: A Programming Model for Generating Optimized Attention Kernels. *arXiv preprint arXiv:2412.05496*, 2024.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, et al. MInference 1.0: Accelerating Pre-Filling for Long-Context LLMs via Dynamic Sparse Attention. *arXiv preprint arXiv:2407.02490*, 2024.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. In *International Conference on Learning Representations*, 2020.

Xunhao Lai, Jianqiao Lu, Yao Luo, Yiyuan Ma, and Xun Zhou. FlexPrefill: A Context-Aware Sparse Attention Mechanism for Efficient Long-Sequence Inference. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2502.20766.

H. Lee, G. Park, Y. Lee, J. Suh, J. Kim, W. Jeong, B. Kim, H. Lee, M. Jeon, and S. J. Hwang. A Training-Free Sub-Quadratic Cost Transformer Model Serving Framework with Hierarchically Pruned Attention. In *International Conference on Learning Representations*, 2025.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *arXiv preprint arXiv:1907.00235*, 2019.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Hassan Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient Context Window Extension of Large Language Models. *arXiv preprint arXiv:2309.00071*, 2023.

Qwen Team. Qwen3.8-27B Model Repository and Model Card. https://huggingface.co/Qwen/Qwen3.8-27B, 2026. Accessed September 2026.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-Aware Sparsity for Efficient Long-Context LLM Inference. *arXiv preprint arXiv:2406.10774*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

Jeffrey Willette, Heejun Lee, and Sung Ju Hwang. Delta Attention: Fast and Accurate Sparse Attention Inference by Delta Correction. *arXiv preprint arXiv:2505.11254*, 2025.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. *arXiv preprint arXiv:2410.10819*, 2024.

Ruyi Xu, Guangxuan Xiao, Haofeng Huang, Junxian Guo, and Song Han. XAttention: Block Sparse Attention with Antidiagonal Scoring. *arXiv preprint arXiv:2503.16428*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *International Conference on Learning Representations*, 2025b. URL https://arxiv.org/abs/2412.06464.

Z. Ye, L. Chen, R. Lai, W. Lin, Y. Zhang, S. Wang, T. Chen, B. Kasikci, V. Grover, A. Krishnamurthy, and L. Ceze. FlashInfer: Efficient and Customizable Attention Engine for LLM Inference Serving. *arXiv preprint arXiv:2501.01005*, 2025.

J. Yuan, H. Gao, D. Dai, J. Luo, L. Zhao, Z. Zhang, Z. Xie, Y. Wei, L. Wang, Z. Xiao, et al. Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. *arXiv preprint arXiv:2502.11089*, 2025.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, 2020.

Zhenhai Zhu and Radu Soricut. H-Transformer-1D: Fast One-Dimensional Hierarchical Attention for Sequences. *arXiv preprint arXiv:2107.11906*, 2021.

A Proof details

A.1 Selected-interaction bound

Let $B = \lceil n/b \rceil$. For $j \geq 4$, Equation 1 contains at most three non-self fixed entries, the self block, and $\lfloor \log_2 j \rfloor - 1$ dyadic predecessors. The first four blocks have constant degree. Hence

$$E_{\text{block}}(B) \leq \sum_{j=0}^{B-1} (4 + \lfloor \log_2(j+1) \rfloor) \tag{4}$$

$$= O(B \log B). \tag{5}$$

Each non-self block pair contributes at most b^2 token interactions and each self block contributes $b(b+1)/2$. Therefore

$$E_{\text{token}}(n) = O(b^2 B \log B) = O(nb \log(n/b)). \tag{6}$$

For fixed b and fixed model dimensions this is $O(n \log n)$ per converted layer. The result counts legal attention interactions, not preprocessing, mask materialization, compiler overhead, or model work outside attention.

A.2 Communication depth

For non-global blocks $0 < i < j$, write $\delta = j - i$ in binary. Subtracting its set powers of two from the current destination produces a monotone causal path from j to i. Powers 2^r with $r \geq 2$ are dyadic edges; steps one and two are local edges. Thus the path has at most $\text{popcount}(\delta) \leq \lfloor \log_2 B \rfloor$ hops. Block zero is directly visible from every later query block.

For $B = 512$, the only distance below B with nine set bits is 511. Such a distance requires source block zero and is served by the global edge. All non-global source–destination pairs therefore require at most eight hops. This upper bound is attained by source block 1 and destination block 256. A forward information path between them cannot use the global edge from block zero, because causality prevents a nonzero source from reaching zero. Every remaining non-self increment is a positive power of two, and the total increment is 255. Any representation of 255 as a sum of powers of two requires at least eight terms: repeatedly merging equal terms yields its binary representation with eight ones without increasing the term count. The legal path

$$1 \to 2 \to 4 \to 8 \to 16 \to 32 \to 64 \to 128 \to 256$$

attains eight hops. The finite maximum is therefore exactly eight, not merely bounded above by eight.

This is a graph-reachability result. It does not prove independent packet preservation, query-specific relevance, or final answer correctness.

A.3 Mixed-stack and hybrid-model scope

For L_d dense layers and L_s BF1 layers,

$$T(n) = \Theta(L_d n^2 d_h) + O(L_s n \log n \, d_h) + T_{\text{other}}(n). \tag{7}$$

The original Qwen3-0.6B experiment has $L_d = 20$, so it remains asymptotically quadratic. The Qwen3.8-27B experiment converts all 16 full-attention layers; its other 48 sequence-mixing layers are Gated DeltaNet layers and are left unchanged. We report measured whole-model scaling rather than asserting a complete implementation theorem for every non-attention component.

A.4 Exact counts, partial blocks and finite kernel capacity

Let $m_j = \min(b, n - jb)$ be the valid length of block j. Since every non-self selected block is complete, the exact per-head/layer count is

$$E(n) = \sum_{j=0}^{\lceil n/b \rceil - 1} \left[\frac{m_j(m_j+1)}{2} + m_j b \left(|\mathcal{P}(j)| - 1 \right) \right]. \tag{8}$$

The dense count is $n(n+1)/2$. The following values are *independent integer reconstructions of the declared graph*, not additional GPU experiments. Degree counts unique legal pages, including the self page, not physical padding slots. The global-page shortcut changes the communication proof at

Table 8: Exact route counting at representative lengths; ratios include the self-causal triangle.

Tokens	Selected pairs/head/layer	Dense/selected	Max. degree
8,192	3,907,584	8.59×	9
16,384	8,888,320	15.10×	10
32,768	19,902,464	26.98×	11
131,072	96,489,472	89.03×	13
262,144	209,797,120	163.78×	14
524,288	453,193,728	303.27×	15
1,000,000	925,333,792	540.35×	16
1,048,576	973,545,472	564.70×	16

the origin, but not the count. Kernel capacity 16 fits the full public route up to 2^{20} tokens at $b = 64$. Beyond that envelope, retaining the graph requires increased capacity; clipping the predecessor list would define a different route and invalidate an unchanged reachability argument.

B BF1 EXECUTION CONTRACT

B.1 ROUTE SEMANTICS

The public route uses block width 64, self-causal attention, two local predecessors, a global first block, and dyadic predecessor offsets $4, 8, 16, \ldots$. Invalid and duplicate blocks are removed. Partial terminal blocks retain only valid tokens. GQA maps each query head to the corresponding KV head before sparse selection. The route is fixed by sequence position and does not inspect token content.

B.2 CORRECTNESS GATES

The original 0.6B operator campaign reports the following admission checks; they must not be assumed to certify every later 27B configuration:

- exact block and selected-token accounting;
- zero output change under mutation of future tokens;
- BF16 agreement with a materialized routed-mask reference;
- finite forward and backward Q/K/V gradients;
- partial final-block coverage at length 257;
- dense-path equivalence when the sparse blend is closed;
- cached one-token decode agreement with an independent selected-page reference;
- page-order permutation invariance and zero decoy-page leakage.

The original Qwen3-0.6B cases at lengths 128, 257, and 512 had maximum absolute BF16 error 0.00390625 and mean error below 7.1×10^{-5}, with zero measured future leakage. In the Qwen3.8 selected-page check, FlashInfer BF1 output agreed with an independent FP32 legal-key reference to maximum absolute difference 0.001 in the tested case.

C ORIGINAL QWEN3-0.6B EXPERIMENTS

C.1 HARDWARE AND PROTOCOL

The original study used an NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPU, PyTorch 2.13.0 with CUDA 13.0, BF16 attention, and Qwen3-0.6B-Base. The target geometry has 28 decoder layers, 16 query heads, eight KV heads, and head dimension 128. Prefill measurements used

batch one, five warmups, 30 measured repetitions, and opposite benchmark orders. Whole-model TTFT used a common eager Hugging Face harness.

C.2 Per-layer and whole-model results

Table 9: Original per-layer BF16 prefill result on RTX PRO 6000.

Context	Measured speedup	Interaction reduction
8K	2.84×	8.59×
16K	5.50×	15.10×
32K	10.91×	26.98×

Table 10: Original eight-of-28-layer whole-model TTFT.

Context	Dense	BF1	Reduction
8K	69.0 ms	63.7 ms	7.7%
16K	190.1 ms	168.6 ms	11.3%
32K	589.1 ms	499.1 ms	15.3%

The all-layer timing-only configuration reached 2.14× at 32K, but it was not a trained all-layer model. A later kernel tuning step improved the per-layer 32K ratio from 9.56× to 10.91× while changing eight-layer TTFT by less than 0.5%, illustrating that layer coverage can dominate additional kernel tuning.

C.3 Matched Stage-A adaptation

All arms adapted the same eight middle attention layers and used 50,333,696 trainable parameters, the same token order, objective, optimizer schedule, checkpoint cadence, and report packs. Sparse arms used equal selected-pair budgets. The objective combined next-token loss, teacher-logit distillation, and hidden-state matching. Each run consumed 16.384M training tokens over 1,000 steps.

Data identity and normalization. The retained preparation record identifies a local input named `orca_train.jsonl`, not an immutable public dataset release. It contains 49,000 question–answer rows. After 34 exact duplicates are removed, the reported split counts are 46,967 training, 1,004 selection, and 995 report rows. Visible normalization collapses whitespace within fields; emitted examples are rendered with `User:` and `Assistant:` labels. The complete rendering/fallback rules, hash-to-split rule, upstream sampling procedure and full-file split hashes are not recoverable from the truncated preparation transcript. Per-row source hashes in that transcript are not whole-file dataset hashes. Orca is cited for background, not as proof of this corpus's public provenance.

Pack length and scored positions. Executed result logs explicitly report sequence length 2,048, 174 report packs, and 356,178 predicted tokens. The documented Stage-A training configuration also specifies 2,048-token packs; complete final training manifests for every arm are not included in this manuscript package. The retained packing implementation tokenizes each text with no automatic special tokens, appends EOS, concatenates records, and emits non-overlapping padding-free packs. The final incomplete pack is discarded. The matching evaluator snapshot uses all 2,047 shifted next-token targets in each full pack, including prompt and answer text, with no answer-only loss mask and no document-boundary reset of positions or attention. Attention follows the selected dense or BF1 causal mask across packed-record boundaries. EOS is a separator, not an attention barrier. These source-level details describe the recovered implementation; its exact byte identity to every historical executed run remains unbound. The report's 2K length is confirmed by logs independently of that source mapping.

At seed 1234, reported 95% paired report-pack intervals place static random 17 at +0.3030% versus BF1 ([+0.2441,+0.3642]%), Dense-CT at +0.3599% ([+0.3169,+0.4055]%), and matched sliding at +7.6367% ([+7.3325,+7.9554]%). These intervals quantify report-pack uncertainty for one model

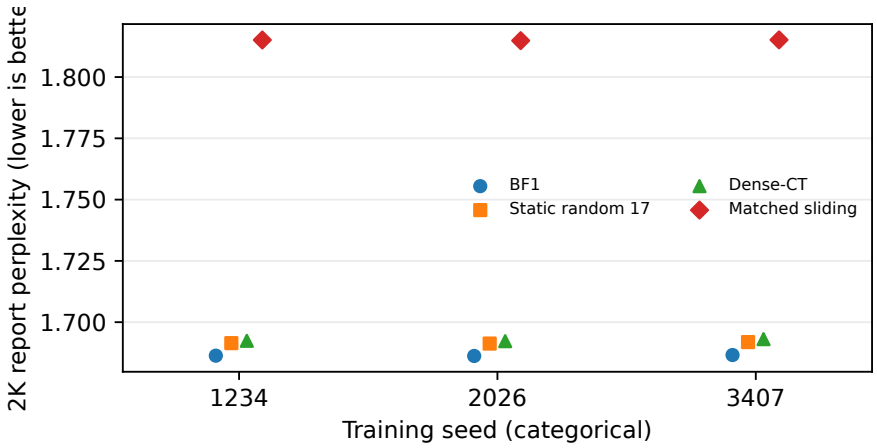


Figure 5: Stage-A report perplexity on 2,048-token packs. Seeds are equally spaced categorical identifiers; unconnected points do not imply a numerical seed trend.

pair; the three training runs quantify training-seed robustness. The static-random result remains conditional on graph draw 17.

C.4 Per-seed values and uncertainty

Table 11: Full-precision report perplexities on 2,048-token packs from the recorded three-seed study.

Arm	1234	2026	3407
BF1	1.686341	1.686220	1.686600
Random-17	1.691451	1.691280	1.691881
Dense-CT	1.692410	1.692257	1.693080
Sliding	1.815121	1.814856	1.815168

The same 174 report packs occur in each run. Three optimizer seeds do not create 522 independent report packs. The seed-1234 BF1/base relative-PPL interval is $[-4.9669, -4.6561]\%$ around -4.8129%; Dense-CT/base is -4.4703% with interval $[-4.6485, -4.2917]\%$. The unadapted base is 1.771607. The manuscript retains the distinction between report-pack uncertainty and independent training-seed variation.

Recovered uncertainty metadata. The seed-1234 Dense-CT/BF1/base analysis manifest records 10,000 paired bootstrap resamples with RNG seed 2026 and identical ordered report packs. The original report labels the intervals as 95%. The aggregate statistic is token-weighted mean negative log likelihood (NLL), with perplexity obtained by exponentiation:

$$\bar{\ell}_a = \frac{\sum_i s_{a,i}}{\sum_i t_i}, \qquad \mathrm{PPL}_a = e^{\bar{\ell}_a}, \qquad \frac{\mathrm{PPL}_b}{\mathrm{PPL}_a} - 1 = e^{\bar{\ell}_b - \bar{\ell}_a} - 1,$$

where $s_{a,i}$ is arm a's NLL sum and t_i is its predicted-token count in pack i. The original resampling implementation, quantile/interpolation convention, and complete per-pack arrays were not recovered here. We therefore retain the reported interval endpoints without claiming to regenerate them or assuming a percentile/BCa construction. The 10,000-resample metadata above is verified for the recovered Dense-CT/BF1/base comparisons; it is not assigned to other historical intervals without their manifests.

C.5 Historical selected-page decode: retain its original endpoint

The original correctness-oriented gather path gives approximately $1.21\times$ over dense at 32K/batch one. The table above instead compares gather with selected-page paged attention; its denominator

Table 12: Original 0.6B-geometry selected-page decode characterization, μs. Residual is effective minus separately timed planning, not directly measured persistent-plan execution.

Context	Batch	Gather	Effective	Plan	Residual	G/E
32K	1	76.112	114.831	102.855	11.976	0.66×
32K	8	100.584	116.471	102.855	13.616	0.86×
32K	32	382.280	143.064	121.352	21.712	2.67×
64K	1	76.568	112.781	100.381	12.400	0.68×
64K	8	105.560	115.922	102.274	13.648	0.91×
64K	32	419.728	137.606	114.622	22.984	3.05×

is not dense paged attention. This differs from the new B200 dense-paged column and must not be plotted as a single longitudinal speedup series.

C.6 Next-token perplexity is not structured capability

A separate 300-example 32K checkpoint-by-route diagnostic, not the ORCA report set, evaluates two adapted checkpoints with dense and BF1 routes. Applying the route costs about 31–32 percentage points under either checkpoint on this diagnostic. It directly limits a broad quality-preservation interpretation of the favorable ORCA result. Graph reachability permits influence; it does not guarantee that the pretrained computation preserves all required independent values. The separate 27B oracle diagnostic probes budget feasibility, not resolution of this entire deficit.

Table 13: Structured-task development factorial. These percentages are not perplexities and not fresh 27B quality evidence.

Weights	Dense route	BF1 route
Dense-CT	45.67%	13.67%
Stage A	47.67%	16.33%

D Qwen3.8-27B B200 protocols and complete tables

D.1 Model geometry and execution

The text stack contains 64 sequence-mixing layers: 48 Gated DeltaNet layers and 16 full GQA layers. The full-attention geometry is 24 query heads, four KV heads, and head dimension 256. The official model card gives a native context of 262,144 and describes YaRN extension toward one million tokens (Qwen Team, 2026; Peng et al., 2023); the execution records additionally report a configured boundary of 1,048,576 total positions. These are not interchangeable length labels. BF1 replaces only the 16 full-attention paths; DeltaNet layers and all non-attention components remain unchanged.

D.2 Reporting convention

For repeated BF1 timing runs, this paper reports the minimum observed BF1 latency and labels it *best observed*. Dense values are the recorded comparator for that context. This convention answers "what performance was attained?" rather than "what latency should be expected on a random run?" It can be optimistic and should not be substituted for a median, confidence interval, or service-level objective. No repeat count or dispersion is reconstructed from rounded summaries. Matched-order repetition is still needed for a distributional comparison.

D.3 Full one-layer prefill table

The original quick protocol used two warmups and three measured repetitions in one order. Recorded BF1 minima were 1.48, 3.15, 7.06, 35.5, and 366.5 ms at 8K, 16K, 32K, 128K, and 1M. Dense recorded references were 2.41, 8.68, 33.4, 537, and 35,900 ms.

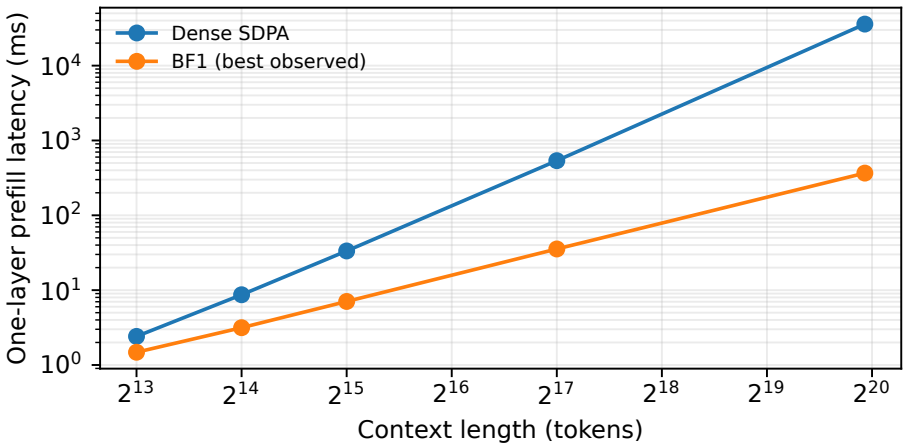


Figure 6: One-layer BF16 prefill latency for Qwen3.8-27B geometry on B200. BF1 is best observed; both axes are logarithmic.

D.4 Whole-model scaling and conversion depth

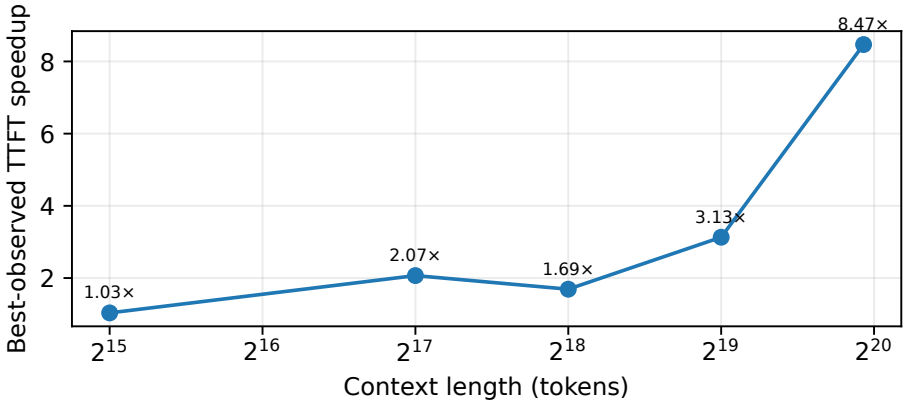


Figure 7: Best-observed whole-model TTFT ratio versus the recorded dense comparator. The speedup series is non-monotone. Run variability and unequal repetition counts may contribute, but these measurements do not isolate their effects.

Table 14: Partial conversion experiment. Values are single-session engineering measurements; its 128K all-16-layer value is 16.6 s, not the separate 10.3 s best-observed value.

Context	Dense	BF1 4/16	BF1 8/16	BF1 16/16
128K	21.3 s	18.9 s	17.1 s	16.6 s
512K	201.0 s	159.7 s	124.7 s	64.2 s

The 512K dose response is large and monotone. It supports systems attribution but not quality equivalence at every conversion depth. At 128K this session gives $21.3/16.6 = 1.28\times$, whereas the separate best-observed comparison gives $21.3/10.3 = 2.07\times$. The source summaries distinguish the two protocols but do not supply immutable session/run identifiers for reconciliation; these observations are not pooled.

D.5 MEMORY AND MAXIMUM CONTEXT

At 1M, both paths have a reported total allocation of 119.8 GB. We retain the source's unit label rather than silently converting it: the available summaries do not identify the allocator API, whether the statistic is peak or current allocated memory, or whether "GB" was computed with decimal or binary units. The observation is not an isolated KV measurement. Equal historical KV tensors do not mathematically require equal total allocation or equal peaks, because other buffers can differ.

For the stated batch-one, BF16 full-attention geometry, the *calculated* uncompressed K/V payload alone is

$$16 \times 2_{\mathrm{K,V}} \times 10^6 \times 4 \times 256 \times 2 \text{ bytes} = 65{,}536{,}000{,}000 \text{ bytes}.$$

This is 65.536 decimal GB, or approximately 61.04 GiB, excluding weights, recurrent state, temporary buffers, allocator overhead and other model state. It is an arithmetic payload calculation, not a new memory measurement. The total-allocation observation is consistent with retaining the ordinary cache, but it does not establish a memory-saving mechanism.

The maximum-position BF1 run uses 1,048,448 prompt tokens and generates 128 additional positions, reaching 1,048,576 total positions. It records 110 s TTFT and 122.9 GB reported total allocation, with the same unresolved metric/unit convention. This separate-length observation is not a repetition of the decimal-million timing cell. The 119.8 GB footprint comparison is not asserted to have been captured during the fastest 78.3 s execution.

D.6 FULL PAGED-DECODE TABLE

Table 15: All nine recorded B200 one-layer paged-decode cells, in μs. Each column is a retained median from a quick three-repetition experiment, not a best-observed value. Gather materializes selected K/V before SDPA. Effective measures the sparse plan-plus-run endpoint. Dense is the recorded dense paged-decode call.

Context	Batch	Gather	Effective	Plan	Direct run	Dense	D/E
32K	1	100	173	123	55	218	1.26×
32K	8	120	256	322	95	454	1.77×
32K	32	312	374	271	90	921	2.46×
64K	1	106	241	127	85	286	1.19×
64K	8	174	216	188	55	523	2.42×
64K	32	354	308	286	72	1,561	5.07×
1M	1	165	217	125	52	880	4.06×
1M	8	214	309	160	65	4,854	15.71×
1M	32	423	496	349	94	18,578	37.46×

The separately timed plan column is 123–349 μs and direct execution is 52–95 μs. Plan exceeds the effective call in the 32K/batch-eight row (322 versus 256 μs). These noisy separately measured summaries are not an additive decomposition; subtracting them would produce a misleading negative residual. We therefore preserve direct run, plan, and effective observations separately.

Dense divided by sparse run-only would be 16.9, 74.7 and 197.6 at 1M for batches one, eight and 32. These are cross-endpoint diagnostic ratios, not achieved persistent-plan speedups. The exact dense-call planning boundary is not separately decomposed in the supplied aggregate record. A fair reusable-plan comparison must put both implementations under the same lifecycle and include plan refresh, changing causal tails, cache updates, synchronization, and actual token delivery.

The independent selected-key FP32 reference check reports maximum absolute difference 0.001 and output norms 7.4494 and 7.4492. This establishes agreement for the checked case, not exhaustive numerical or quality equivalence. Earlier bit-identical outputs belonged to two BF1 paths, not to BF1 versus full dense attention.

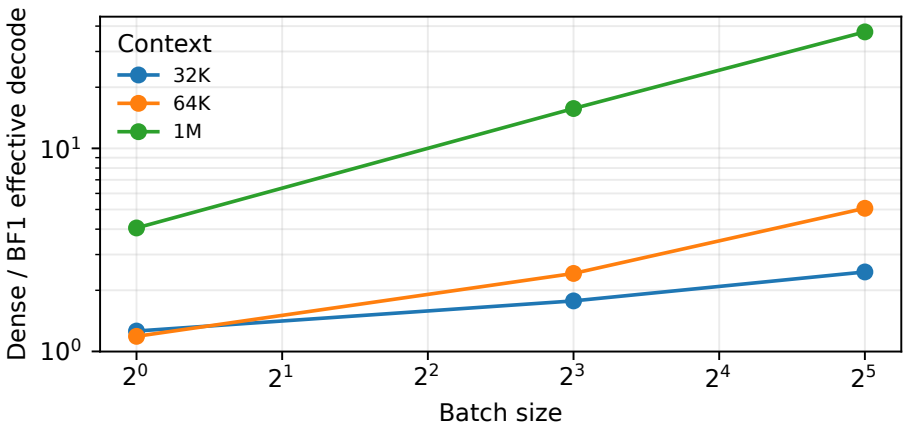


Figure 8: Retained effective decode ratios, including every recorded context/batch cell. They are not whole-model generation throughput.

E TRAINING-FREE CAPABILITY ADJUDICATION

E.1 SIX-ROW DIAGNOSTIC DESIGN

The diagnostic uses six extraction rows on the real Qwen3.8-27B integration, labeled nominal 32K in the experiment summary. It includes three single-source positions (early, middle, late), one corroborated-evidence row, and two harder multi-source rows. The model receives no BF1-specific adaptation. Replacement substitutes evidence pages; addition is a separate evidence-supplemented intervention. The summary reports replacement and random as matched in budget, causal distance and degree, with degree 12 for each. Addition's degree was not separately logged. These are retained reports, not a reconciled certificate that ordinary BF1's unique legal page count was unchanged.

Degree convention and missing positions. For Equation 1, degrees at query blocks 511, 512 and 513 are 11, 11 and 12, respectively, including self and excluding invalid/duplicate entries. At width 64, these blocks begin at zero-based positions 32,704, 32,768 and 32,832. Table 8 therefore correctly gives maximum degree 11 for a 32,768-token prompt. Later generated queries or an appended question could in principle reach degree 12; this is an explanation to test, not a finding about the diagnostic. Its exact tokenized prompt length, evaluated query positions, generated length, pre/post-override page lists and physical-padding convention are unavailable here. We neither change the reported 12 to 11 nor infer its cause. Equal reported intervention degrees also do not establish equality to the unmodified route.

Fully dense and unmodified all-BF1 results on these exact six rows are absent from the retained comparison. The single-layer ablation does not substitute for those baselines. We report the oracle/random contrast without imputing recovery relative to dense or ordinary all-BF1.

Table 16: Training-free oracle-routing adjudication. Degree 12 is source-reported and unreconciled to exact query positions; addition was not separately logged.

Arm	Correct	Reported degree	Interpretation
Oracle replace	4/6	12 (unreconciled)	Correct evidence
Oracle add	4/6	Not separately logged	Same observed correct count
Matched random	0/6	12 (unreconciled)	Reported budget-matched control

All three single-source rows and the corroborated row succeed under oracle evidence. A separate needle-position probe reports that directly routed positions recover the word *cedar*, while page zero yields the exact word/value pair. Word recognition and exact value copying are different endpoints; the probe does not supply a complete per-position count table. Matched random fails every row. Under the reported budget and distance matching, this intervention supports a content-dependent

effect. Verifying that attribution against exact route capacity requires the missing page-list and query-position records.

E.2 Interpretation boundary

The result supports three claims:

1. the 27B BF1 execution path can consume correctly routed evidence;
2. the source reports that replacement fits the degree-16 envelope, but exact unique-page accounting remains unreconciled;
3. semantic page addressing is a distinct unresolved problem.

The oracle/random pair has four oracle-only successes and no random-only success, giving two-sided exact McNemar $p = 2(1/2)^4 = 0.125$. It does not establish statistical generalization, broad composition, or quality parity. The two unsolved cases cannot be uniquely attributed to composition rather than binding or representation effects. Oracle routing has evidence-location information unavailable to the deployed deterministic route. Its running time is not the running time of a practical semantic selector.

E.3 Single-layer interventions

Each of the 16 full-attention layers is converted individually while the other 15 remain dense. Every intervention preserves five or six of six diagnostic answers; 12/16 produce final-logit KL below 0.05. No tested single-layer substitution is catastrophic on these six examples. It does not rule out nonlinear interactions among several converted layers.

F Claim ledger

Table 17: What each result supports and does not support.

Result	Supported statement	Unsupported substitution
$O(n \log n)$ proof	selected interactions per converted layer	complete implementation automatically scales identically
10.91× at 32K	original 0.6B one-layer speedup	whole-model 10.91×
Three-seed PPL	replicated ordering on 2K report packs	32K–1M quality or retrieval preservation
98.0× at 1M	best-observed 27B one-layer prefill ratio	expected or end-to-end 98×
8.47× at 1M	best-observed whole-model TTFT ratio	throughput, p95, or median service gain
119.8 GB equality	equal reported total allocation	isolated KV payload or validated peak metric
4/6 vs 0/6	evidence contrast under reported matching	exact budget verification or solved addressing
Single-layer 5–6/6	no individually catastrophic layer on six rows	arbitrary layer subsets are safe

G Reproducibility boundary and artifact map

The source bundle contains:

- this ICLR 2027 LaTeX source and unmodified style file;
- aggregate CSV/JSON data transcribed from the recorded experiment summaries;
- deterministic figure-generation code;
- a verification script checking arithmetic, table identities, and key claims;
- PDF render and preflight reports;
- a SHA-256 manifest.

The bundle does not independently reproduce the neural experiments. The original run directories, private integration code, proprietary checkpoints, compiled kernels, and raw hardware traces remain access-controlled. The 27B tables retain aggregate experiment records rather than a complete

distribution of raw per-run traces. Their best-observed ratios are descriptive attained-performance observations, not confidence estimates or a full independently repeatable neural reproduction.

The remaining record-level limitations are explicit: immutable upstream dataset and split identities; full final training/configuration manifests; the original interval-construction code and complete per-pack arrays; exact 27B diagnostic positions and page lists; matched dense/unmodified-BF1 six-row outcomes; model/tokenizer revision pins; timing-session identities and allocator definitions. Citing a model card verifies an architectural description, not the immutable weight/config revision used by an experiment. These items cannot be filled from headline aggregates. The accompanying correction ledger distinguishes source-confirmed details, mathematical reconstructions and unresolved empirical provenance.

H REMAINING EXPERIMENTS

The highest-value remaining work is:

1. repeat the 128K, 512K, and 1M B200 cells in opposite arm orders with fixed warmups and multiple measurements;
2. cache and reuse FlashInfer plans across route epochs, then measure actual generation rather than a run-only bound;
3. replicate static-random topology across graph draws, not only training seeds;
4. adapt Qwen3.8-27B under matched dense/BF1 controls and evaluate untouched long-context capability;
5. evaluate external blind corpora and representative multi-source tasks;
6. report batch/concurrency throughput, power, and energy per completed request.