The Validity Mirage: Context Algebra for Endogenous Semantics under Memory Compression

Anonymous (Draft)

Abstract

Modern long-context systems compress memory using truncation, eviction, or summarization, often evaluated by raw task validity. In endogenous-pivot settings, this is insufficient: feasibility depends on a turning point chosen by `argmax` over the selected sequence, so compression can alter semantics by changing pivot identity. We formalize a compositional context algebra over summaries $C = (w^\star, d_{\text{total}}, d_{\text{pre}})$ and prove associativity of endogenous composition. Under committed semantics, we prove an absorbing left-ideal property and define a checkable no-absorption compression contract. Empirically, composition is exact and associative on tested cells (Exp51: 0 violations), absorbing closure holds (Exp56: 0/96 violations), and closure breaks appear only under compression (Exp57: 0.133 closure violation rate). Retention sweeps show the validity mirage: TP-conditioned raw validity remains near 1.0 while pivot preservation and fixed-pivot feasibility degrade sharply. A deterministic witness (Exp58) isolates the mechanism: naive compression is infeasible under committed semantics, while under multi-candidate semantics it appears feasible only via pivot substitution, with a 54.4% score shortfall. The main conclusion is that compression safety must be defined relative to pivot consistency, not raw feasibility alone.

1 Introduction

Large language models and other long-sequence systems now depend on context management: truncation, cache eviction, retrieval, or summarization. Most evaluations ask whether the system still produces a valid output. That criterion is often too weak when semantics are endogenous.

Compression is often treated as structurally neutral: the past is assumed to be a concatenated string (a free monoid), and compression is treated as approximately semantics-preserving.

In endogenous settings, the interpretation of early context depends on a later turning point that is itself chosen from the selected sequence. Compression decisions that look locally safe can become globally unsafe once the eventual pivot is revealed. This is the mechanism behind many failures that appear as "hallucination" or "drift" but are structurally consistent under a changed internal hypothesis.

Our companion finite paper established a sharp sufficient-condition impossibility boundary (zero false positives in a validated 11,400-instance regime). Our companion streaming paper established irreversible commitment traps (54.9% organic trap rate over 4,200 runs). This paper fills the compositional gap between those results: we define the algebra of context summaries under endogenous pivots and study when compression preserves or breaks that algebra.

Contributions.

1. We formalize a compact context monoid for endogenous pivots and provide full associativity and absorbing-ideal proofs under explicit semantics.

Preprint. Under review.

2. We validate monoid exactness and ideal closure empirically (Exps 51 and 56), and identify compression as the unique tested closure-breaking operation (Exp57).
3. We define semantic diagnostics (pivot preservation, fixed-pivot feasibility, semantic regret) and show why raw feasibility alone creates a validity mirage.
4. We provide a deterministic witness (Exp58) that isolates semantic drift via pivot substitution and quantifies the resulting score loss.

2 Preliminaries and Semantics

We consider extraction from temporally ordered event graphs with focal-actor turning points and prefix requirement k. For selected sequence S, endogenous turning point is

$$\tau(S) = \arg \max_{v \in S,\, a(v)=a^*} w(v),$$

with deterministic tie-breaking by $(w, -t, \mathrm{id})$.

We evaluate three solver semantics:

- **Committed semantics:** pivot identity is fixed once chosen and cannot be replaced by later candidates.
- **Enumerative semantics:** solver evaluates top-M pivot candidates and may return any feasible candidate.
- **Fixed-pivot semantics:** enumerative solver constrained to the full-sequence dominant pivot.

The TP-conditioned solver is a label-setting dynamic program over fixed-pivot subproblems, aligned with resource-constrained shortest-path formulations [Irnich and Desaulniers, 2005].

3 Context Algebra

3.1 Context element

Definition 1 (Context element)**.** *A context summary is*

$$C = (w^\star, d_{\mathrm{total}}, d_{\mathrm{pre}}),$$

where $w^\star$ is the local pivot key, d_{total} is total development capacity in the block, and d_{pre} is development capacity before the local pivot.

3.2 Endogenous composition

For adjacent blocks A then B, define $C_{A\|B} = C_A \otimes C_B$:

$$\begin{aligned} w^\star &= \max(w^\star_A, w^\star_B), \\ d_{\mathrm{total}} &= d_{\mathrm{total},A} + d_{\mathrm{total},B}, \\ d_{\mathrm{pre}} &= \begin{cases} d_{\mathrm{pre},A}, & w^\star_A \geq w^\star_B, \\ d_{\mathrm{total},A} + d_{\mathrm{pre},B}, & w^\star_B > w^\star_A. \end{cases} \end{aligned}$$

The piecewise update is the critical term: if pivot shifts right, the entire left block becomes pre-pivot and contributes its full d_{total}.

3.3 Associativity

Proposition 1 (Associativity)**.** *$\otimes$ is associative.*

Proof. Let A, B, C be three consecutive blocks, with strict total order on pivot keys from deterministic tie-breaking.

Case 1: global pivot in A. Then $w^\star_A \geq w^\star_B, w^\star_C$. For $(A \otimes B) \otimes C$, first composition yields $d_{\mathrm{pre},AB} = d_{\mathrm{pre},A}$ and $d_{\mathrm{total},AB} = d_{\mathrm{total},A} + d_{\mathrm{total},B}$. Composing with C preserves pivot in A, so

$d_{\text{pre}} = d_{\text{pre},A}$ and $d_{\text{total}} = d_{\text{total},A} + d_{\text{total},B} + d_{\text{total},C}$. For $A \otimes (B \otimes C)$, A still dominates, giving the same pair $(d_{\text{pre}}, d_{\text{total}})$.

Case 2: global pivot in B. Then $w_B^\star > w_A^\star, w_C^\star$. In $(A \otimes B)$, pivot shifts to B, so $d_{\text{pre},AB} = d_{\text{total},A} + d_{\text{pre},B}$. Composing with C preserves pivot in B, giving final $d_{\text{pre}} = d_{\text{total},A} + d_{\text{pre},B}$. In $A \otimes (B \otimes C)$, $B \otimes C$ has pivot in B and $d_{\text{pre},BC} = d_{\text{pre},B}$, so final $d_{\text{pre}} = d_{\text{total},A} + d_{\text{pre},BC} = d_{\text{total},A} + d_{\text{pre},B}$.

Case 3: global pivot in C. Symmetrically, both parenthesizations yield

$$d_{\text{pre}} = d_{\text{total},A} + d_{\text{total},B} + d_{\text{pre},C},$$

with identical d_{total}.

All cases agree; therefore $\otimes$ is associative. □

Systems implication. Because $\otimes$ is strictly associative, context reduction supports parallel tree reductions and therefore $O(\log n)$ reduction depth in recursive context trees.

4 Absorbing Ideal and No-Absorption Contract

Fix grammar parameter k and define absorbing predicate $\perp$ by $d_{\text{pre}} < k$.

Proposition 2 (Absorbing left ideal under committed semantics)**.** *If context A has committed pivot and $A \in \perp$, then for any suffix D, $A \otimes D \in \perp$.*

Proof. Under committed semantics, suffix events cannot replace the committed pivot. Hence

$$d_{\text{pre},A\otimes D} = d_{\text{pre},A} < k,$$

so $A \otimes D \in \perp$ for all D. □

Remark 1 (Semantics qualifier)**.** *Without commitment, suffix pivot substitution can escape absorption: if $w_D^\star > w_A^\star$, then $d_{\text{pre}} = d_{\text{total},A} + d_{\text{pre},D}$ may exceed k. Exp58 is an explicit witness of this distinction.*

Definition 2 (No-absorption contract)**.** *For grammar demand k, a compression map μ on block B satisfies the no-absorption contract if*

$$d_{\text{total}}(\mu(B)) \geq \min\big(d_{\text{total}}(B),\, k\big).$$

This definition is checkable and operational: compression may remove redundant capacity, but it cannot reduce latent prefix-support below the threshold needed to survive a future pivot shift.

5 Multi-Candidate Pivots and the Validity Mirage

When solvers enumerate multiple pivot candidates ($M > 1$), raw feasibility can remain high even after compression damages the dominant-pivot trajectory. The solver can recover feasibility by substituting a different pivot. We call this the *validity mirage*.

We therefore evaluate three semantic diagnostics in addition to validity:

- **Pivot preservation rate:** fraction of valid compressed solves that keep the full-sequence pivot.
- **Fixed-pivot feasibility:** validity when forcing compressed solve to the full-sequence pivot.
- **Semantic regret:** relative score shortfall against the full-sequence dominant-pivot trajectory.

6 Experimental Protocol

Artifacts and determinism. All experiments are deterministic and sourced from saved outputs in `experiments/output/`. We use context-algebra suite Exps 50–57, full-scale retention sweeps, and deterministic witness Exp58.

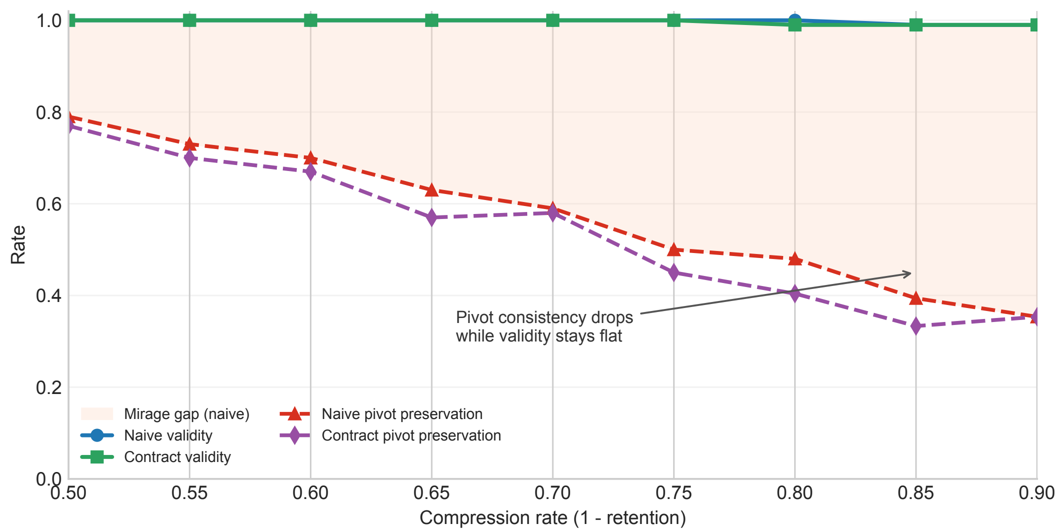

Figure 1: TP-conditioned sweep ($n = 200$, $k = 3$, $M = 10$, 100 seeds). X-axis is compression rate $c = 1 - \mathrm{retention}$. Raw validity stays near 1.0 while pivot-preservation rates degrade as compression tightens.

Check	Cases/Checks	Violations	Rate
Exp51 pairwise exact (q/r)	240	0	0.000
Exp51 pairwise exact (pivots)	240	0	0.000
Exp51 associativity	80	0	0.000
Exp56 absorbing-ideal closure	96	0	0.000

Table 1: Core algebraic checks: exactness and closure hold on tested cells.

Diagnostic state. For taxonomy and closure diagnostics we use enriched state

$$\mathcal{S} = (q, r, \Pi, t_{\min}, t_{\max}),$$

with grammar state q, counters r, pivot set Π, and temporal bounds. This directly enabled Exp50 predicate correction (failure-only Class-A accuracy from 0.000 to 1.000) and Exp57 mismatch diagnostics.

Compression strategies and semantics. We compare naive, bridge-preserving, and contract-guarded compression under TP-conditioned and greedy solvers. For TP-conditioned runs, each compressed solve is scored both in free mode and fixed-pivot mode.

7 Results

7.1 Claim 1: The monoid is exact and the ideal is closed

Table 1 summarizes the core algebraic checks; all tested exactness and closure checks pass with zero violations.

7.2 Claim 2: Compression is the unique tested closure-breaking operation

Closure breakdown concentrates entirely in compression.

7.3 Claim 3: Raw feasibility alone creates a validity mirage

In full-scale Exp52/53, raw validity ties at 1.000 across naive/bridge/contract and segment-wise vs whole compression (gap 0.000). Semantic diagnostics reveal degradation despite this tie.

Operation	Closure violation rate	Absorbing escape rate
Composition	0.000	0.000
Compression	0.133	0.033
Pivot update	0.000	0.000
Split-at-point	0.000	0.000

Table 2: Exp57 closure diagnostics by operation.

Retention	Naive valid	Pivot preserve	Fixed-pivot valid	Semantic regret
0.50	1.000	0.790	0.980	0.218
0.30	1.000	0.590	0.920	0.663
0.20	1.000	0.480	0.860	0.677
0.10	0.990	0.354	0.750	0.358

Table 3: Condensed TP-conditioned semantic sweep for naive compression (100 seeds, $M = 10$). Raw validity stays high while pivot consistency degrades.

Contract-guarded follows the same qualitative pattern; full per-strategy metrics are reported in Appendix A. Semantic regret is non-monotone in the full sweep (Appendix A); at extreme compression, both full and compressed solves are often forced onto similarly degraded pivot candidates, reducing the measured gap.

7.4 Claim 4: Solver semantics dominate compression strategy

At retention 0.10 (same generator family, 100 seeds), TP-conditioned results are naive 0.990 vs contract 0.990, while greedy results are naive 0.000 vs contract 0.010. Dominant variance comes from solver semantics (committed/myopic vs enumerative), not compression heuristic alone.

7.5 Claim 5: Exp58 isolates pivot substitution

Under committed semantics ($M = 1$), Exp58 yields the witness triple: full feasible, naive infeasible, contract feasible. Under enumerative semantics ($M = 10$), naive appears feasible only by changing pivot identity. Relative score shortfall versus the dominant-pivot trajectory is

$$1 - \frac{21.467}{47.084} = 0.544.$$

8 Related Work

Our framing is algebraic: context summaries are composed under an endogenous operator with explicit invariants, analogous in spirit to how relational algebra formalizes data operations [Codd, 1970]. On the combinatorial side, endogenous pivot selection with prefix constraints falls outside hereditary/exchange structures that support classical greedy guarantees [Korte and Lovasz, 1981, Bjorner and Ziegler, 1992].

Long-context modeling work typically emphasizes scalable mechanisms without formal compression-safety guarantees. This includes sparse/sliding attention (Longformer [Beltagy et al., 2020]), positional extrapolation methods (RoFormer [Su et al., 2021]; positional interpolation [Chen et al., 2023]), retrieval-augmented systems (REALM [Guu et al., 2020], RAG [Lewis et al., 2020]), state-space sequence models (Mamba [Gu and Dao, 2023]), KV-cache policies (H_2O [Zhang et al., 2023]), and persistent memory architectures (MemoryBank [Zhong et al., 2023]). Our contribution is complementary: an algebraic safety layer for pivot-consistent feasibility under compression.

9 Discussion and Limitations

What is established. Composition behaves as a stable algebraic object on tested cells; absorbing closure is robust under committed semantics; compression is the unique tested closure-breaking operation.

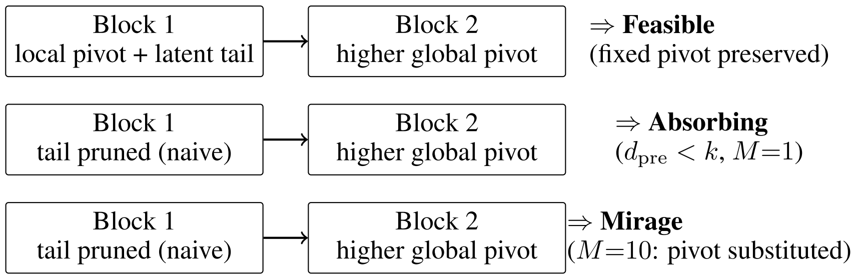

Figure 2: Exp58 mechanism diagram. Enumerative solvers can escape infeasibility by pivot substitution, yielding a validity mirage.

Strategy	M	Free valid	Free TP	Fixed-pivot valid	Pivot preserved	Score
Naive	1	False	–	False	False	–
Contract-guarded	1	True	e41	True	True	47.084
Naive	10	True	e45	False	False	21.467
Contract-guarded	10	True	e41	True	True	47.084

Table 4: Exp58 deterministic witness. Under $M = 10$, naive is feasible only via pivot substitution (e45) and fails fixed-pivot feasibility.

Interpretation. The key lesson is not "contract always improves validity." The key lesson is that validity can hide semantic drift under multi-candidate pivot substitution. Safety criteria must include pivot consistency.

Implications for long-context systems. An LLM can produce coherent outputs after aggressive compression while internally switching to an alternate hypothesis. This suggests one plausible mechanism for semantic drift and some forms of hallucination: not arbitrary fabrication ex nihilo, but coherent continuation of a degraded hypothesis after critical latent constraints were removed.

Limitations. Exp58 is a deterministic witness, not a distributional guarantee. Current evidence is synthetic-graph based. A full symbolic characterization of compression-safe contracts beyond the threshold form above remains open.

10 Conclusion

Context algebra separates structural composition from policy effects in endogenous-pivot sequential systems. The monoid and absorbing ideal are stable under composition; compression marks the main boundary where invariants fail. Semantic instrumentation reveals a validity mirage in strong solvers, and Exp58 isolates the pivot-substitution mechanism. More broadly, extending the committed-pivot monoid to a multi-candidate algebra that composes over time while selecting across pivot hypotheses suggests a set-valued lift; we conjecture this forms a context semiring under persistent pivot uncertainty, and a natural direction for hierarchical context management in systems with endogenous semantics.

Reproducibility Artifacts

Primary artifacts used in this manuscript:

- `context_algebra_suite.json`
- `context_algebra_suite_refined2.json`
- `context_algebra_exp52_53_n200_m510.json`
- `context_retention_sweep_tp_semantics_m10.json`
- `context_retention_sweep_greedy_semantics_m10.json`
- `exp58_pivot_shift_witness_semantics.json`

A Full TP Semantic Sweep Table

r	N val	C val	N piv	C piv	N fix	C fix	Gap rej
0.50	1.000	1.000	0.790	0.770	0.980	0.990	112.23
0.45	1.000	1.000	0.730	0.700	0.970	0.980	154.16
0.40	1.000	1.000	0.700	0.670	0.970	0.960	196.30
0.35	1.000	1.000	0.630	0.570	0.960	0.940	238.63
0.30	1.000	1.000	0.590	0.580	0.920	0.930	281.21
0.25	1.000	1.000	0.500	0.450	0.890	0.910	324.18
0.20	1.000	0.990	0.480	0.404	0.860	0.890	367.77
0.15	0.990	0.990	0.394	0.333	0.850	0.870	412.66
0.10	0.990	0.990	0.354	0.354	0.750	0.780	459.45

Table 5: Full TP-conditioned semantic sweep (all retention points). Abbreviations: N = naive, C = contract-guarded, piv = pivot-preservation rate, fix = fixed-pivot feasibility, Gap rej = mean rejected gap-guard drops.

B Supporting Results: Deferred Composition and Recursive Depth

These results are retained as supporting evidence, not central claims.

Exp54 (deferred composition). Suite diagnostics show nonzero patience dominates $f = 0$ by the configured dominance check. Aggregate values: valid rate at $f = 0.00$ is 1.000 (mean pivot shifts 0.639), while $f = 0.25$ and $f = 0.50$ remain 1.000 with reduced shift rates (0.333 and 0.111).

Exp55 (recursive depth). Refined target checks report all defined cells at or above 90% validity retention versus flat baseline (5/5 defined cells). One cell ($n = 500$, block size 50) has both recursive and flat validity equal to 0, yielding undefined retention ratio (n/a).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. URL `https://arxiv.org/abs/2004.05150`.

Anders Bjorner and Gunter M. Ziegler. Introduction to greedoids. In Neil White, editor, *Matroid Applications*, volume 40 of *Encyclopedia of Mathematics and its Applications*, pages 284–357. Cambridge University Press, 1992. doi: 10.1017/CBO9780511662041.011.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. URL `https://arxiv.org/abs/2306.15595`.

Edgar F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970. doi: 10.1145/362384.362685.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL `https://arxiv.org/abs/2312.00752`.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 3929–3938, 2020. URL `https://arxiv.org/abs/2002.08909`.

Stefan Irnich and Guy Desaulniers. Shortest path problems with resource constraints. In Guy Desaulniers, Jacques Desrosiers, and Marius M. Solomon, editors, *Column Generation*, pages 33–65. Springer, 2005. doi: 10.1007/0-387-25486-2_2.

Bernhard Korte and Laszlo Lovasz. Mathematical structures underlying greedy algorithms. In *Fundamentals of Computation Theory (FCT'81)*, volume 117 of *Lecture Notes in Computer Science*, pages 205–209. Springer, 1981. doi: 10.1007/3-540-10854-8_22.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://arxiv.org/abs/2005.11401.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Chao Wang, Haibin Li, Zhaozhuo Lin, et al. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. *arXiv preprint arXiv:2306.14048*, 2023. URL https://arxiv.org/abs/2306.14048.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023. URL https://arxiv.org/abs/2305.10250.