

MSRA: Supplementary Material

Implementation Audit, Reversal Ledger, and Reproducibility Archive

Zihan Tian
École Polytechnique; Xinjiang University
zihan.tian@polytechnique.edu

Overview. This document records the evidence chain behind every number in the main paper: the data provenance index (S1), nine implementation-audit entries (S2), the reversal ledger (S3), seed and evaluation details (S4), and the machine-executed assertion list (S5). The main text presents only the final data and conclusions; this archive documents how they were verified into place.

1 Freeze Declaration and Data Provenance

The experiment matrix is frozen at tag `phase1-experiment-freeze` (commit `1bff7cdc`, plus text-level close-out `4078888b`). At the freeze point, all 52 assertions in `paper/recompute_tables.py` pass (output in S5), and the unit-test suite stands at 38 passed / 1 failed (one documented legacy failure: `fp16` overflow on random tokens, unrelated to this work). After the freeze, no new experiment is permitted without a pre-registered protocol and a named claim it gates. Code and data: <https://github.com/TianZH-XJU/MSRA-EMNLP-GLM2026>.

Every number in the paper can be traced from a results file to an assertion anchor (Table 1). All assertions are executed by `paper/recompute_tables.py`; the full output is archived in `paper/REVISION_DATA_AUDIT.md`.

2 Implementation Audit (A.1–A.9)

Nine implementation defects were discovered and repaired over the course of this work. Each followed the same closed loop: discovery → forensics → repair → withdrawal of affected claims → re-measurement → assertion pin.

A.1 Kernel-FFT Cache Collision. The inference-path kernel FFT cache key was derived from the first 8 kernel elements; the far mask zeroes these for every σ , causing all configurations to share a single cached entry. Consequence: the learnable scale σ was effectively frozen at initialization (all M=1 checkpoints have $\sigma \equiv 20.1$); the original “learnable multiscale” never engaged. After the fix, σ trains again (the true trajectory under the training protocol is small: M1L8 seeds move $20.1 \rightarrow 20.6\text{--}20.9$, a $\log\text{-}\sigma$ drift of $+0.02$ to $+0.04$; an earlier report of “0.33 drift in 30 steps” came from a unit test’s aggressive learning rate, not from the training protocol).

A.2 Noise Amplification at Empty-Far Positions. At positions $i \leq w$ (empty far context), FFT leaves $O(10^{-6})$ noise in the denominator z ; the normalization $u = \text{noise}_C / (\text{noise}_z + \varepsilon)$ amplifies this above the legitimate far signal (measured in-window $|y^{\text{far}}|$ was $\sim 5\times$ the beyond-window value). Fix: hard-zero y^{far} where $F(i) = \emptyset$, including the denominator pathway.

A.3 NaN Gradient Infection. Wide- σ experts (σ up to 460) produced finite losses with non-finite gradients roughly 2,700 steps into training in the 8-layer configuration; `clip_grad_norm_` propagated the NaN into every parameter. Fix: a total-norm isfinite check plus NaN-skip.

Table 1: Provenance index: every paper element to its data file and machine assertion.

Paper element	Data file (results/)	Anchor
Table 1: GPT-2 row	G2_M1L8_s1..s5.json	A8
Table 1: LLaMA @512	fixed_split_eval.json	A1 ($\times 3$ seeds)
Table 1: LLaMA @1024	fixed_split_1024.json	S4 (n=1 disclosed)
Ladder: M1L8 / M1L4 / M4L4	fixed_split_eval.json	A1/A2 (both seeds)
Ladder: M1L16, M4L8 / M4L4 gate-off	fixed_split_m1l16\{, _s3\}.json, gateoff_m4l8\{, _s3\}.json, gateoff_m4l4_\{s2, s3\}.json	A2b (incl. fp32 discr.)
Ladder: gain-only rows	fixed_split_gammaonly_\{m1l8, m4l8, m1l16\}.json	A7 (incl. 26.4%)
Ladder: LoRA / dwconv	fixed_split_bl_\{lora, dwconv\}_\{s2, s3\}.json	A11 (wiring load_check)
Ladder: random kernel	fixed_split_eval.json (random arm)	placebo assertion
Table 3: NIAH v5 (8 rows)	niah_v5_ws128_all.json (6 ckpts, survival 1.0)	A12 (six arms + @1024 far)
Table 4: prefill	systems_llama_\{m1l8, m1l16, v2_m4l8\}.json	A13
Table 4: decode	systems_llama_*_\{kv, kv2\}.json	A10 (median 1.004)
LongBench paragraph	longbench_official_\{full, local, qcmsa\}_kv.json	A14
RULER 3.0%	ruler_v2_2048_\{full, local, qcmsa\}.json	A5 (positive control)
Gate selectivity	gate_analysis_v2.json	A14b
Official-test replication	pg19_official_test_3arm.json	A15
fp16 quantization floor	v8_discrimination.json, v8_gateoff_fp32.json	A1/A2b
In-loop bit-identity closure	v8_roundtrip_final.json	(audit narrative)

A.4 NIAH Evaluation Target Defect. The early NIAH protocol computed loss on the query phrase’s own continuation (answer tokens were never appended to the input), measuring “the effect of needle presence on the query phrase’s language modeling” rather than needle content retrieval. Fix: answer tokens appended to the input, loss window moved to the answer region. (The near-field improvement first measured under this fix was itself an artifact of truncated needles — superseded by A.9.)

A.5 Training-Internal Evaluation Path. Training-internal evaluation produced bit-identical values across the three M1L8 seeds (350.237183×3). A round-trip test — fresh process, reconstruction of the in-loop chunk set, loading each saved checkpoint — yields three distinct values under both fp16 ($350.240/350.304/350.275$) and fp32 ($350.268/350.297/350.289$), and reproduces none of the recorded value. This rules out the two candidate explanations: the checkpoints are not identical, and fp16 output rounding does not collapse them on these chunks (the quantization floor is real — it collapses s2/s4 in the canonical fp16 evaluation — but does not extend to the in-loop case). The only self-consistent conclusion is that the in-loop path evaluated a state independent of the seed-varying trained parameters; its exact root cause was not identified before the path was deprecated. No paper number uses this path: every value comes from independent-process evaluation.

A.6 Baseline Adapter Wiring No-Op. The first execution of the canonical (16-chunk) matched-baseline evaluation returned PPL bit-identical to the local arm (0.00% closure): the LoRA/dwconv wrappers were constructed and their weights loaded, but never assigned into the attention forward path (the training harness performs this assignment; the evaluation harness initially did not). Because both adapter types are zero-initialized, an unwired wrapper reproduces local attention exactly — a silent no-op that no key-level

load check can catch. The 16-decimal-digit match with the local baseline is what exposed it. Fix: wire the wrapper into the forward path and assert the adapter object is present in the forward path (positive control) before evaluating. The reported values are the re-evaluated ones: LoRA 87.4% (s2 87.39 / s3 87.47), dwconv 78.3% (78.23/78.32); the previously circulated LoRA 87.6% derived from the deprecated in-loop path.

A.7 Systems Configuration Mislabeled. The prefill-throughput rows previously presented as the default M1L8 were produced by the systems harness’s loader default — $M=4$, gate-on, untrained spectral module — not the stated configuration. The mislabel surfaced when an explicitly-configured M1L16 measurement ($4.8\times @512$) proved faster than the supposedly smaller M1L8 ($8.2\times @512$), which is structurally impossible. The measurement itself was valid for the structure it actually exercised (throughput depends on M and L , not on trained values), but every derived claim (“5–8 $\times$ slowdown”) overstated the default configuration’s cost by roughly a factor of two. Re-measured with explicit structure flags, all three configurations in matched same-session pairs: M1L8 2.0–2.9 $\times$, M1L16 3.0–4.8 $\times$, M4L8 5.0–8.7 $\times$ — the slowdown scales with far-branch work (patched layers $\times$ experts: 8/16/32 units), and decode is at local parity for all three.

A.8 Decode Path Deferred-Fix. The pre-fix attention forwards accepted `past_key_values` but never wrote k/v (the cache-write code lives in the HF attention module we replace). Because the cache never grew, `HF generate()` — trusting the cache object’s progression — sliced its input to the last token at every step: each generated token attended only itself, and the LongBench generations collapsed to gibberish (local/corrected means ≈ 0.0001). The standalone decode benchmark’s manual loop fed a single token against the same never-written cache, so the previously reported “decode = local” rows measured 1-token self-attention. Neither was a real or comparable path; both are the same failure mode. (Our initial forensic reading attributed the `generate()` path to full-prefix re-feeding; the before/after data refutes this — cached and full-recompute decode are equivalent by unit test, so a re-feeding path would have reproduced the corrected scores, and it did not.) Fix: a true sliding-window KV cache in the canonical forwards — prefill writes k/v with bitwise-unchanged outputs (unit-test pinned), decode attends the last $w+1$ cached keys, and the far branch is structurally prefill-only (unit-test pinned). Post-fix decode throughput is statistically indistinguishable from local — FLOPs identical by construction; six paired measurements across two sessions span 0.99–1.10 with median 1.004 (Table 4 of the main paper). The LongBench numbers computed under the broken path are withdrawn; the re-run reproduces the full arm exactly (mean 0.0804, per-task identical) and confirms the local/corrected arms’ earlier “generation collapse” was the artifact described above. The re-run also exposed a per-sample allocator leak — the spectral module’s padded-length-keyed FP32 buffer cache accumulating on variable-length prompts (triviaqa) — fixed at the module level (bounded cache).

A.9 Needle-Survival Repair and the Near-Field Reversal. The NIAH generator inserted the needle into the full-length haystack and only then truncated the haystack to make room for query and answer; for near-bucket distances the second truncation silently cut the needle tail — including the retrieval target — seeding the near bucket with unanswerable samples. The near-field improvements reported under that protocol (+20.6–40.2% gap closed) were measured on such samples and are withdrawn. Fix: truncation now precedes insertion, the needle is clamped to fit the final haystack verbatim, survival is asserted per example and the aggregate rate is asserted to be 1.0 (measured rate 1.0 across all re-runs). A deterministic replay of the old sampling — validated position-by-position against the recorded per-example data — quantifies the damage: 3 of 11 near-bucket samples (27%) had the needle tail truncated; 0 of 11 mid, 0 of 11 far (geometric prediction: cut when the intended distance $\lesssim 14$ tokens, i.e. ~ 10 –17% of the near bucket; the realized 3/11 reflects discrete sampling). Two scope notes. The truncation geometry is distance-bounded — mid- and far-bucket needles sit far from the truncation frontier at both sequence lengths, so their old numbers were protocol-invariant (which is why the @1,024 figures needed only re-anchoring, not rewriting: +14.9% $\rightarrow$ +15.0% differs by the disclosed ~ 6 -token slide) — only the near bucket required

withdrawal. And the arithmetic reconstruction below is illustrative, not an independent check (two free parameters, one equation); the load-bearing evidence is the position-by-position replay match and the 3/11-versus-prediction agreement. The mechanism of the spurious improvement: a half-needle ending “... The secret key is XZZZY-” adjacent to the query is a strong out-of-distribution cue — local attention, sharp by design, is the most vulnerable to it (confidently wrong, NLL ≈ 8), while the smoothed far-branch output is comparatively robust — so the correction appeared helpful precisely on the broken samples (3 broken samples at ≈ 8 plus 8 intact samples at ≈ 1.6 reproduce the old near-bucket mean of 3.37 exactly). Meta-lesson: no “guaranteed by construction” argument is trustworthy without a runtime assertion. Under the repaired protocol the near-field conclusion reverses: every configuration slightly degrades precise in-window retrieval (-3.5% to -15.2% gap closed), moderately improves just-beyond-window prediction ($+7-9\%$), and leaves far-field recovery at ≈ 0 — the correction helps where local attention is blind and hurts where it is sharp, which is exactly the signature of a smoother; the gain-only@L16 probe carries the same signature, tying the effect to the gain channel.

3 The Reversal Ledger

Every surviving number in this paper has an alibi. Nine times, a claimed result was overturned by a harness or protocol defect; each overturning followed the same loop — discovery, forensics, repair, withdrawal, re-measurement, assertion pin — and each left a machine-checkable anchor behind (Table 2).

Table 2: The nine-reversal ledger. Same-loop hygiene fixes with smaller blast radius: gain-only@M1L8 9.5 $\rightarrow$ 10.0 (in-loop versus canonical caliber); a Table 3 chimera rebuilt from a single file; 6-task versus 7-task means unified; decode OOM from heap fragmentation solved by measurement reordering; triviaqa’s padded-length buffer accumulation bounded inside the module; per-sample spectral-cache hygiene in the LongBench harness.

# Withdrawn claim	Root cause	How it surfaced	Repair	Anchor
1 Near-field “+26.5% @w=128”	crashed NIAH run mislabeled as results	provenance audit	withdrawal; ws=256 re-run	FORBIDDEN list
2 “Near-field degradation” (v2)	loss on the query phrase’s own continuation	code review of the loss slice	answer appended; answer-region loss (A.4)	per-region loss test
3 Near-field “+20.6–40.2%” (v4)	second truncation cut the needle tail	package review + geometric re-check	truncation before insertion; survival asserts (A.9)	A12: six v5 anchors
4 “Multiscale has no independent contribution”	over-generalization without the gate-off control	gate-off cells in the grid	reworded “real but secondary”	Table 2 gate-off cells
5 “M4L8 is unstable”	NaN gradient infection misread as instability	guard-isolation experiment	isfinite guard + NaN-skip	A3; trains to 25.7%
6 “decode = local”	forwards never wrote the KV cache	package review (P0-1)	true sliding-window KV cache (A.8)	A10: median 1.004
7 LongBench “generation collapse”	same decode defect	before/after asymmetry	re-run under fixed path	A14: full 0.0804
8 “LoRA 87.6%”	adapters never wired into the forward path	16-digit match with local	wiring + positive control (A.6)	A11: 87.4%
9 “Prefill 5–8 $\times$ (M1L8)”	loader default measured the M4-gate-on structure	explicit M1L16 faster than “smaller” M1L8	explicit structure flags (A.7)	A13: 2.9/4.8/8.7 $\times$

4 Seed and Evaluation Details

All attribution-ladder values use the fixed 16-chunk evaluation set (SHA-256 prefix 8555f6e9) with independent-process evaluation. Under fp16 evaluation, M1L8 seeds s2 and s4 show bit-identical PPL

(327.131...); fp32 re-evaluation yields three distinct values (327.287/327.791/327.237), confirming the bit-identity is an fp16 quantization floor, not a harness defect. Canonical values drift $< 0.05\text{pp}$ under fp32 re-evaluation. The same phenomenon appears for M4L8-gate-off: the two seeds are bit-identical under fp16 (317.125) but distinct under fp32 (317.179/316.996). M1L16 seeds are distinct already under fp16 (259.107/259.043), as are all remaining pairs.

Seed details: M1L8 n=3 (seeds 2/3/4), M1L4 n=2 (seeds 2/3; canonical 5.39%/5.32%), M4L4 n=2 (seeds 2/3), M4L8 n=2 (seeds 2/3), M1L16 n=2 (seeds 2/3), M4L8-gate-off n=2 (seeds 2/3), M4L4-gate-off n=2 (seeds 2/3; canonical 6.12%/6.14%), LoRA n=2 (seeds 2/3), dwconv n=2 (seeds 2/3), GPT2-XL n=5 (seeds 1–5), LLaMA @1024 n=1 (seed 2). Gain-only probes: n=1 each (M1L8, M1L16, M4L8 — all seed 2).

5 Assertion List (Machine-Executed Output at the Freeze Point)

- [PASS] T1@512 per-seed gap range == 11.2–13.3: actual 11.2–13.3
- [PASS] T1@1024 per-seed gap range == 6.9–7.5: actual 6.9–7.5
- [PASS] T6 M=1 gate-off L=8 == 8.7 (single run): actual 8.7
- [PASS] T7 L=8 == 11.5 (single run): actual 11.5
- [PASS] T7 L=8 outside T1@1024 range: split-draw variance note justified
- [PASS] T4 average gap closed == 6.9: actual 6.9
- [PASS] param formula M4 gate-on L4 == 22548: formula gives 22548
- [PASS] param formula M1 gate-off L4 == 9224: formula gives 9224
- [PASS] param formula M4 gate-off L4 == 21524: formula gives 21524
- [PASS] param formula M1 gate-off L8 == 18448: formula gives 18448
- [PASS] T9(seq=512) v2 slowdown == 8.2: actual 8.19x
- [PASS] T9(seq=1024) v2 slowdown == 7.0: actual 6.98x
- [PASS] T9(seq=2048) v2 slowdown == 5.0: actual 4.98x
- [PASS] T9(seq=4096) recorded as OOM (honest replacement): status=OOM
- [PASS] R_M1L8 x3 canonical gap in 12.3–12.6: [12.54, 12.36, 12.54]
- [PASS] no cross-seed bit-identity under fp32: [327.287, 327.791, 327.237]
- [PASS] M1L4_s2 canonical gap in 5.2–5.5: actual 5.39
- [PASS] M1L4_s3 canonical gap in 5.2–5.5: actual 5.32
- [PASS] M4L4_s2 canonical gap in 8.4–8.8: actual 8.53
- [PASS] M4L4_s3 canonical gap in 8.4–8.9: actual 8.69
- [PASS] gamma-only@M1L8 canonical == 10.0: actual 10.04
- [PASS] gamma-only@M4L8 canonical == 10.1: actual 10.09
- [PASS] gamma-only@M1L16 canonical == 26.4: actual 26.36
- [PASS] M1L16 n=2 canonical 31.7: actual [31.72, 31.73]
- [PASS] M4L8-gateoff n=2 canonical 15.4: actual [15.36, 15.36]
- [PASS] M4L4-gateoff n=2 canonical 6.1: actual [6.12, 6.14]
- [PASS] M4L8-gateoff fp16-identical pair discriminates under fp32: [317.179, 316.996]
- [PASS] G2 M1L8 x5 gain ~10.4 all positive: mean 10.4
- [PASS] LongBench official MSRA closure ~0: [0.0 x5, 0.1, 0.0]
- [PASS] decode parity: six paired measurements median ~1.0: ratios [1.1, 1.02, 1.01, 1.0, 1.0, 0.99], median 1.004
- [PASS] decode parity: range 0.9–1.15 across sessions: min 0.99 max 1.10
- [PASS] baseline_lora n=2 canonical 87.2–87.6 + load_check clean: [87.39, 87.47]
- [PASS] baseline_dwconv n=2 canonical 78.0–78.5 + load_check clean: [78.23, 78.32]
- [PASS] NIAH v5 R_M1L8_s2: near -15.2, mid +8.60, far +0.43, surv 1.0
- [PASS] NIAH v5 R_M4L4_s2: near -12.8, mid +7.50, far +0.03, surv 1.0
- [PASS] NIAH v5 R_M4L8_s2b: near -7.6, mid +8.40, far -0.42, surv 1.0
- [PASS] NIAH v5 R_M1L16_s2: near -4.3, mid +7.32, far -1.07, surv 1.0
- [PASS] NIAH v5 R_gamma_only_s2: near -14.3, mid +8.74, far +0.73, surv 1.0
- [PASS] NIAH v5 R_gammaonly_M1L16_s2: near -3.5, mid +8.17, far +0.85, surv 1.0
- [PASS] NIAH v5 M1L16 @1024 far gc ~+15 (vanishes @2048): actual 15.02
- [PASS] gate entropy ratio ~62.8% of uniform (selectivity): actual 0.628
- [PASS] gate query-perturbation response >> context (>10x): actual 112.9x
- [PASS] systems m1l8: @512 slowdown ~2.9x: actual 2.89x
- [PASS] systems m1l16: @512 slowdown ~4.8x: actual 4.84x
- [PASS] systems v2_m4l8: @512 slowdown ~8.7x: actual 8.71x

```

- [PASS] LongBench _kv: full mean 0.0804 reproduced exactly: actual 0.0804
- [PASS] LongBench _kv: local mean 0.033: actual 0.0332
- [PASS] LongBench _kv: qcmsa mean 0.033: actual 0.0334
- [PASS] LongBench _kv: closure <= 2% on non-degenerate tasks:
    [('triviaqa', 0.0), ('gov_report', 1.28), ('qmsum', -0.0)]
- [PASS] LongBench _kv: window cost real on gov_report: 0.137 vs 0.003
- [PASS] RULER v2: full positive control (PPL<2.0, n>=30)
    + closure ~3.0% (<5): full_ppl 1.563, n 30, closure 2.98%
- [PASS] official PG-19 test partition: gain ~+10.1: actual +10.07%, gc 11.43%
**ALL ASSERTIONS PASS**

```