LATENTPORT: BEYOND KV CACHE CROSS-MODEL TRANSFER OF RECURRENT MEMORY IN HYBRID LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A small language model may have processed thousands of tokens before a router calls a larger model. Transferring attention key–value (KV) caches can avoid repeated prefix processing, but hybrid models also maintain recurrent memory. Can a differently sized receiver use this persistent state? We study Qwen3.5 4B→9B handoff without historical target replay. Adding the transferred Gated DeltaNet persistent-state package to fixed translated KV lowers teacher-forced continuation loss by 0.747 nats/token (95% paired document bootstrap CI [0.692, 0.805]), improving all 64 held-out books. Direct recurrent/convolution reuse outperforms the tested learned maps despite lower reconstruction error for the learned recurrent map. Without learned correction or target repair, a separate replication finds native context-benefit recovery of 0.833 at 4K and 0.844 at 16K using post-trained Q4_K_M models in llama.cpp, compared with 0.142 and 0.105 for KV-only transfer, across four domains. A compact, pair-specific residual correction separately reduces Base-model excess loss to 0.076 nats/token, recovering 0.918 of the native target's context benefit while beating continued 4B prediction. Small trajectory perturbations also expose transfer-specific degradation. These results establish useful but conditional portability of built-in persistent state within one geometry-matched model family. They do not establish native-equivalent free generation, cross-family compatibility, downstream-task equivalence, or production speedup.

1 INTRODUCTION

A model router may need to switch models after a long interaction. The receiver normally reconstructs model-specific inference state by processing the history, duplicating prefix computation. Cross-model KV translation asks whether this reconstruction can be avoided for attention caches (Heo et al., 2026; Qu et al., 2026; Li et al., 2026b). Attention–recurrent hybrids pose an additional question: their live state also contains persistent recurrent matrices, convolution history, and runtime bookkeeping. KV-only transfer leaves these components behind.

Can a larger model consume recurrent memory written by a smaller model? The source writes it with its own projections and gates; the receiver must read and update it with different weights. We test this distinction in Qwen3.5 4B→9B, whose hidden widths differ (2,560 versus 4,096). We deliberately study this sibling pair with exactly matched persistent-state geometry, treating this as an enabling condition for an existence test rather than evidence of general cross-model compatibility. The handoff installs source-derived state into the receiver's native inference interface, then begins continuation without replaying historical target-prefix tokens (Figure 1).

Three frozen studies establish the positive result. **E001** holds translated KV fixed and shows that adding the Gated DeltaNet (GDN) persistent-state package reduces continuation loss by 0.747 nats/token, with all 64 test documents improving. **E011** replicates the large full-state advantage over KV-only transfer with post-trained checkpoints, Q4_K_M quantization, llama.cpp, four domains, and both 4K and 16K contexts, without learned correction, target repair, or historical replay. **E002** separately selects translated KV with direct recurrent/convolution reuse and applies a compact, pair-specific behavioral correction, improving Base-model fidelity and beating continued source prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

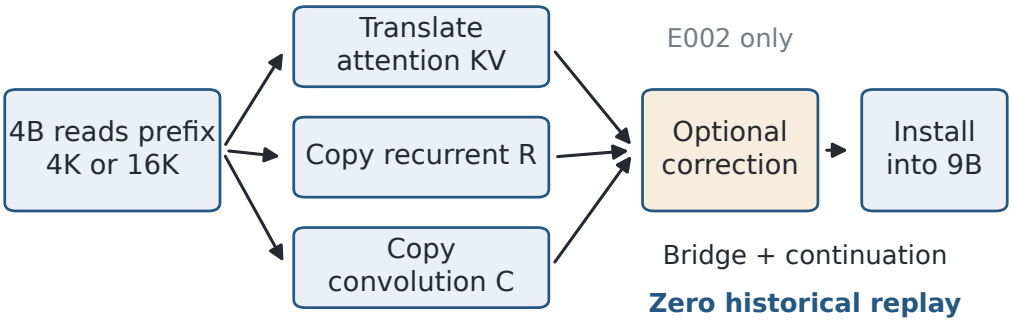

Figure 1: **The transferred object is live hybrid state.** The 4B model reads the historical prefix. The receiver installs translated attention KV and recurrent/convolution state written by 4B; E002 adds a residual correction, whereas E011 does not. A real bridge token starts continuation with zero historical target replay. Native target prefix processing is used only for baselines and offline supervision. This is a mechanism schematic, not a latency result.

Property	4B	9B
Hidden width	2,560	4,096
Language layers (GDN / attention)	32 (24 / 8)	32 (24 / 8)
Recurrent matrix	$32 \times 128 \times 128$	same
Attention K or V	$4 \times L \times 256$	same
Convolution: Transformers	8192×4	same
Convolution: llama.cpp	3×8192	same

Table 1: Runtime-verified geometry. Shapes omit batch and are per layer; L is prefix length. Source and target shapes match *within* each runtime. Transformers stores four convolution positions; llama.cpp retains three history positions for the same width-four convolution. No cross-runtime state-byte interchange is claimed.

Our contribution is empirical cross-model portability of *prefix-conditioned, built-in recurrent inference state*, rather than KV transfer or latent communication in general. The evidence also bounds that portability: a learned recurrent mapper can reconstruct tensors better yet support worse continuation, and trajectory perturbations can damage transfer disproportionately. Throughout, the endpoint is teacher-forced continuation in one directed, geometry-matched sibling pair. E011 expands robustness within that pair; it supplies neither another model family nor evidence of free-generation equivalence.

2 Hybrid State and Cross-Model Handoff

2.1 Matched geometry, different read/write dynamics

Both models repeat three GDN layers followed by one full-attention layer. GDN combines gated forgetting with a delta-rule memory update (Yang et al., 2025); in a conceptual value-by-key orientation,

$$\bar{S}_t = \alpha_t S_{t-1}, \qquad S_t = \bar{S}_t + \beta_t(v_t - \bar{S}_t k_t)k_t^\top. \tag{1}$$

The runtime may store the transpose. Its persistent state also includes convolution history used to construct recurrent inputs. Both the accumulated matrix and local histories can affect continuation.

After prefix $x_{1:L}$, the source state is $\mathcal{H}_4(x_{1:L}) = (K_4, R_4, C_4, m_4)$, where K contains attention keys and values, R recurrent matrices, C convolution buffers, and m bookkeeping. Layer/head indices correspond directly (Table 1); there is no cropping, padding, tiling, or learned index alignment. Deterministic runtime construction supplies metadata, fresh cache objects, and independent storage.

Equal dimensions make installation valid; the models' distinct projections and gates make its utility an empirical question.

2.2 Translation, direct reuse, and correction

E001 fits separate maps to paired source/target state. Attention K and V use ridge maps per layer, KV head, and role; keys are de-rotated before mapping and re-rotated at the target positions. Each recurrent layer/head uses

$$\widehat{S}_9 = \mu_9 + A(S_4 - \mu_4)B^\top. \tag{2}$$

Convolution buffers use grouped ridge maps. Validation tensor error selects regularization and permitted normalization before held-out evaluation. Full translation maps K, R, C; direct-GDN reuse applies the same translated K and copies R, C.

E002 evaluates all eight direct/translated choices. A code lists (K, R, C), with D for direct and T for translated. Selection uses validation excess NLL, a frozen 0.01-nat tie tolerance, then fewer translated components and lexical order. The selected **TDD** base translates KV and directly reuses both GDN components.

An identity-anchored correction adds rank-four KV and recurrent residuals and per-channel convolution scale/bias terms. All trainable residual outputs start at zero. Components are optimized jointly through the frozen receiver using full-vocabulary $\mathrm{KL}(P_{\mathrm{native9B}} \| P_{\mathrm{handoff}})$ over nine output positions, plus a relative-state-norm penalty. Validation selects rank 4 and penalty weight 0.01. Its **434,176 parameters are additional correction parameters**, excluding the KV translator, fixed bases, and both language models. Equations, fitting details, and inventory are in Appendix C.

2.3 Zero historical replay and intervention semantics

The source processes exactly L historical tokens. The receiver's first input is the real next token x_{L+1}, whose logits predict the first of 64 scored targets, $x_{L+2:L+65}$. The bridge is observed text, not a learned prompt or historical replay. Native and transferred conditions share positions and ground-truth continuation inputs. Target prefix traces supply offline supervision and reference predictions; no target-native historical tensor enters the handoff.

The E001 KV-only condition leaves recurrent and convolution state fresh and uninitialized. Full-state installation changes matrices, buffers, and their initialization semantics together. E011 implements KV-only by zeroing R/C in its validated runtime; it is a separate operational replication, not a byte-identical ablation. Within each experiment, KV-only and full-state conditions share the same translated KV, so their paired difference measures the incremental contribution of the *GDN persistent-state package*, not recurrent matrices alone, conditional on that mapping. This does not establish an optimal KV translator; comparison with newer state-of-the-art KV-transfer systems addresses a different systems question.

Because 24 of 32 language layers use GDN, KV-only intentionally removes history-dependent state from most sequence-processing layers; it tests the omitted package's contribution under zero historical replay, not an optimized hybrid fallback with partial replay. Target-side replay warm-up would change the intervention and violate the zero-historical-replay comparison.

3 Experimental Design

3.1 Controlled Base-model evidence: E001 and E002

E001/E002 use official Base checkpoints in BF16, FP32 recurrent storage, and Transformers on an RTX 5090. E001 calibrates translators on 128 FineWeb-Edu documents and validates on 32 disjoint documents, with 1,024-token prefixes and eight state checkpoints. FineWeb-Edu is an educational subset of FineWeb (Penedo et al., 2024). The held-out test uses 64 PG19 books (Rae et al., 2020), each with a 4,096-token prefix and 64 scored targets, changing both corpus and length from calibration.

E002 uses 32 fresh FineWeb-Edu documents at 4K for the factorial and 64 further fresh documents for the primary test, excluding all E001 split identities and text hashes. Correction fitting reuses

E001's 128/32 calibration documents at 1K. Salted hash ordering, length eligibility, and duplicate rules determine membership independently of outputs. Test corpora differ between experiments, so their aggregate losses cannot isolate a method improvement.

Complete same-model capture/restore passes on 16 contexts per model across four lengths, with 100% top-1 agreement and zero maximum logit difference. A pre-fitting check found chunked and one-shot prefix processing unequal; calibration checkpoints consequently use independent one-shot prefills. Appendix A records the exact schedules and revisions.

3.2 Matched realistic-runtime replication: E011

E011 evaluates Base and post-trained Qwen3.5 4B/9B under one pinned llama.cpp commit, converting each official checkpoint through BF16 to Q4_K_M without an importance matrix. Attention KV is F16; recurrent/convolution state is F32. The post-trained checkpoints are the official non-Base models, evaluated as raw text without chat templates or instruction wrappers. Token IDs match across all four checkpoints. Raw text deliberately keeps Base and post-trained checkpoints on identical token sequences, isolating state portability after post-training on this matched input distribution; it does not establish robustness under deployed chat templates or instruction-following behavior.

Separate affine KV maps are calibrated for the Base and post-trained pairs using identical procedures: 500 non-overlapping 1K sequences balanced across literature, STEM, code, and conversational text, a 400/100 selection split, then refitting on all 500. Full-state TDD directly copies R/C. **No learned correction, target repair, or historical target replay is used.** This tests the portability mechanism, not E002's corrected solution.

The 4K test contains 64 fresh documents (16 per domain); the prospectively selected 16K test contains 32 disjoint documents (eight per domain). Eight 16K contexts are outcome-blind concatenations of disjoint source documents, constructed before outcomes were observed where natural long documents were insufficient; the remaining 24 are natural long documents. Each context has one bridge and 64 scored targets. Within a length, Base/post-trained conditions use identical token sequences. All four models pass fresh-context restoration gates. Canonical per-token summaries retain loss and distribution metrics, with a deterministic subset retaining full logits for independent audit (Appendix J).

E011's preregistered primary hypothesis predicted a material post-training penalty. Both stages instead passed the 4K portability gate, opening the separate 16K branch. The KV-versus-full contrast is a preregistered secondary comparison. The original E002 16K branch remained unrun.

3.3 Metrics, controls, and uncertainty

We use teacher forcing for this existence test to evaluate native and transferred states under identical future inputs, isolating handoff-state compatibility from divergence caused by autoregressive sampling.

Let $\ell_{d,c}$ be mean negative log-likelihood (NLL) over 64 observed targets in document d under condition c. Aggregate NLL is the equal-weight document mean $\ell_c = n^{-1} \sum_d \ell_{d,c}$, in nats/token; lower is better. Excess NLL and native context recovery are

$$\Delta\,\mathrm{NLL}_c = \ell_c - \ell_{\mathrm{native9B}}, \qquad \mathrm{NCR}_c = \frac{\ell_{\mathrm{empty9B}} - \ell_c}{\ell_{\mathrm{empty9B}} - \ell_{\mathrm{native9B}}}. \tag{3}$$

NCR measures recovery of the native target's prefix-derived *loss reduction* relative to empty state. NCR 0.918 is not 91.8% accuracy or native quality. It is an unclipped ratio of aggregate losses and requires positive context headroom. Equal loss does not imply equal predictions. Jensen–Shannon (JS) divergence compares full output distributions with native 9B (zero is identical); top-1 is prediction agreement, not accuracy. Appendix D defines other diagnostics.

Wrong-donor controls transfer complete states from other equal-length documents using frozen assignments with no self-donors. They test context specificity of the complete state, not its localization to R alone. Confidence intervals use 10,000 paired document bootstrap resamples and frozen seeds, retaining within-document condition pairing. E011 resamples Base/post jointly and recomputes NCR denominators in each draw; length comparisons use its disjoint sets. Tokens are never independent replicates. E013/E014 likewise resample complete context clusters.

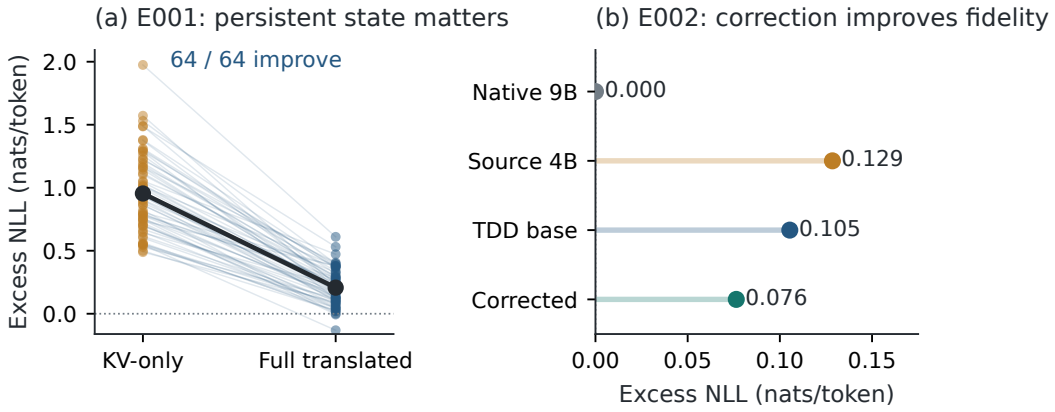

Figure 2: **Beyond KV, then improved receiver behavior.** Excess NLL is loss above native 9B; zero is the reference. (a) E001: 64 PG19 documents, with lines connecting paired documents. (b) E002: 64 different FineWeb-Edu documents. Corrected state beats source 4B by 0.0521 nats/token (95% CI [0.0185, 0.0843]). Large points mark aggregate means; thin marks show document values, not token-level replicates. Both assays use 4K prefixes and 64 teacher-forced targets. Corpora and panel scales differ.

4 Results

4.1 Persistent state contributes beyond translated KV

Does GDN memory add useful context when translated KV is fixed? In E001, adding the translated GDN package lowers NLL by **0.7473 nats/token** (95% CI [0.6921, 0.8047]). All 64 documents improve (Figure 2; paired scatter in Appendix E). KV-only NLL is 3.1153, much closer to empty 9B (3.2457) than native 9B (2.1609); full-translated NLL is 2.3679. The intervention removes a large fraction of the deficit left by KV-only transfer.

Correct full-translated state also beats shuffled state by 0.9482 nats/token (CI [0.8779, 1.0221]). The fixed-KV contrast establishes a beyond-KV contribution; the complete-state shuffle establishes donor specificity. Together they support useful cross-model hybrid-state transfer without identifying the recurrent matrix's individual share.

4.2 Direct reuse outperforms the tested GDN mapping

Which components need translation? E001 direct recurrent/convolution reuse reduces NLL from 2.3679 to 2.2913 with translated KV unchanged, an advantage of 0.0766 nats/token. Yet the learned recurrent mapper lowers validation normalized reconstruction error from 0.6732 to 0.5037. Lower recurrent tensor error therefore does not guarantee a more useful complete handoff. This comparison also changes convolution translation and evaluation domain, so it is not an isolated causal test of recurrent Euclidean error.

The fresh E002 factorial supports the behavioral preference. Translating KV lowers NLL on average by 0.0705 (translated-minus-direct CI $[-0.1009, -0.0409]$); translating recurrent state raises it by 0.0280 (CI [0.0104,0.0475]). Convolution translation has a small, uncertain effect of -0.0004 (CI $[-0.0026, 0.0018]$). TDT has the lowest numerical validation excess NLL, 0.132226 versus TDD's 0.133377, but the frozen 0.01-nat tie rule selects TDD for using fewer translations. On the separate test set, TTT remains worse than TDD by 0.0290 (CI [0.0170,0.0418]). Full cells and interactions are retained in Appendix F; direct convolution alone is not shown superior.

4.3 Corrected handoff beats continued source prediction

Can the larger receiver improve on the source without rereading history? On E002's held-out teacher-forced targets, corrected 9B NLL is 1.9894 versus continued 4B's 2.0415. Corrected-minus-source

Condition	NLL	ΔNLL	NCR	JS	Top-1
Native 9B	1.9130	0.0000	1.0000	0.0000	1.0000
Source 4B	2.0415	0.1285	0.8616	0.0381	0.8340
Empty 9B	2.8416	0.9285	0.0000	0.1568	0.6265
TDD base	2.0184	0.1054	0.8865	0.0286	0.8464
Corrected	1.9894	0.0764	0.9178	0.0221	0.8643
Shuffled	3.1778	1.2647	-0.3621	0.1921	0.5784

Table 2: **E002 held-out fidelity.** Equal-weight means over 64 FineWeb-Edu documents, 4K prefix and 64 teacher-forced targets. NLL and ΔNLL use nats/token; JS and top-1 compare with native 9B. NCR is context-benefit recovery. Shuffled means complete corrected state from a wrong donor.

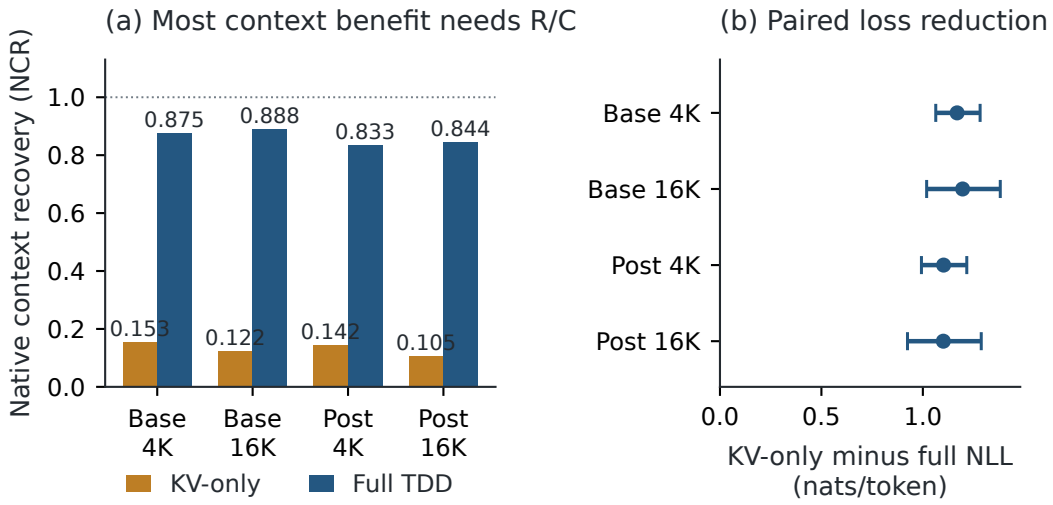

Figure 3: **Full-state portability persists at 16K and after post-training.** E011 uses one pinned Q4_K_M llama.cpp stack, separately calibrated KV maps, and no correction, target repair, or historical replay. (a) NCR is a ratio of aggregate document losses. (b) KV-only minus full-state NLL with the canonical 95% paired document bootstrap intervals. Each 4K comparison has 64 documents; each 16K comparison has 32 disjoint documents, equally balanced across four domains. NCR bars show point estimates, not ratio intervals.

NLL is -0.0521 (95% CI $[-0.0843,-0.0185]$). The correction lowers TDD excess NLL from 0.1054 to 0.0764, with paired improvement 0.0290 (CI [0.0228,0.0354]), removing 27.5% of the remaining native gap (ratio CI [22.4%,33.9%]).

Corrected state beats shuffled state by 1.1884 nats/token (CI [1.0855,1.2956]); Table 2 reports distributional fidelity. Correction size and perturbation norms are detailed in Appendix C.

The corrected handoff satisfies the preregistered full-state criteria but not the stricter near-native thresholds: excess NLL exceeds 0.05, NCR is below 0.95, and top-1 agreement is below 0.90. E002 therefore did not run its conditional 16K branch. High teacher-forced fidelity is not native equivalence.

4.4 Portability survives the matched Q4/runtime replication

Does the beyond-KV effect survive a different inference stack and post-training? E011 finds a large full-state advantage in all four checkpoint-stage/context combinations (Figure 3, Table 3). Post-trained full-state NCR is **0.8334 at 4K and 0.8441 at 16K**, versus KV-only NCR 0.1419 and 0.1050. The corresponding full-state NLL reductions are 1.1022 (CI [0.9930,1.2165]) and 1.1009 (CI [0.9243,1.2878]) nats/token. The Base pair gives full/KV-only NCR 0.8753/0.1532 at 4K and 0.8879/0.1220 at 16K.

Pair / length	Native NLL	Full NLL	ΔNLL	JS	Top-1
Base 4K	1.4681	1.6700	0.2019	0.0488	0.8069
Base 16K	1.3120	1.4870	0.1750	0.0459	0.8496
Post 4K	1.5478	1.8133	0.2655	0.0678	0.7678
Post 16K	1.4362	1.6685	0.2323	0.0759	0.7944

Table 3: **E011 retains a measurable native gap.** Native and full-state NLL, full-state excess NLL, JS, and top-1 agreement. All conditions score 64 teacher-forced targets. Base/post models share test texts within each length. These uncorrected results do not inherit E002's corrected fidelity.

Correct state remains source-specific: post-trained full-minus-shuffled NLL is -1.6141 at 4K (CI [-1.7692,-1.4642]) and -1.5703 at 16K (CI [-1.8037,-1.3471]). Every domain's descriptive full-state mean improves over KV-only at both lengths for both stages (Appendix J.1). Domain subsets are small and do not establish broad task generalization. The 16K experiment tests handoff fidelity after constructing and transferring state from a 16K history; it does not establish retrieval or use of information from arbitrary positions throughout that history.

Post-training has a detectable cost: the frozen primary Base-minus-post NCR contrast is 0.0418 (CI [0.0278,0.0562]) at 4K, below the preregistered materiality threshold 0.15. Thus the predicted material break is *not supported*; the result is not equivalence between Base and post-trained models. Both stages satisfy the preregistered 4K portability criteria, opening the 16K evaluation, where both again satisfy the portability criteria. The length interaction is 0.0020 (CI [-0.0399,0.0364]).

This replication bundles quantization and runtime relative to E001/E002 and uses different calibration/data. It establishes survival on the tested stack, not separate causal effects of quantization and runtime. Uncorrected Full TDD has higher NLL than continued source 4B in three of four E011 conditions (Table 12): Base 4K (1.6700 versus 1.5390), Base 16K (1.4870 versus 1.3648), and post-trained 4K (1.8133 versus 1.7223). It beats the source only at post-trained 16K (1.6685 versus 1.7512). E011 demonstrates robust full-state portability relative to KV-only transfer; source-beating performance under the registered handoff criteria is established only by the corrected E002 Base-model assay.

4.5 Trajectory changes expose a conditional boundary

Does compatible state remain equally useful after a small change in its writing trajectory? E013 replaces the final seven tokens of 16K histories in a paired factorial over 48 fresh contexts. Averaged across the other factors, the change raises handoff NLL by 0.2325 but source NLL by 0.0204. The transfer-specific difference is 0.2121 nats/token (95% context-bootstrap CI [0.1456,0.2851]). Correct state still beats shuffled state in every factorial cell.

E014 replicates sensitivity on fresh contexts and maps its onset: even a one-token generic tail discontinuity produces transfer-specific damage of 0.1494 (CI [0.0425,0.2869]; 48 contexts). Damage can accumulate with changed-tail length; under target-native teacher-forced evolution, the fault gap decreases but is not shown eliminated. Appendix K gives definitions and scope. These controls bound the positive result as conditional functional compatibility; they neither identify a universal failure nor establish free-generation behavior.

5 Related Work

Cross-model attention-cache transfer. KV transfer is established prior work. Heo et al. (2026) use closed-form ridge maps and RoPE factoring across differently sized full-attention siblings; CacheBridge adds head-local support and attention-sensitive calibration (Qu et al., 2026). Universal Context-Reuse studies within- and cross-family KV sharing (Li et al., 2026b). Their native-cache interfaces are the closest to ours, but do not evaluate built-in recurrent GDN state. Cache-to-Cache projects and fuses source KV with receiver KV (Fu et al., 2026); Latent Cache Flow and XKV extend communication to differing contexts (Rossi et al., 2026; Liu et al., 2026a). Shared latent KV alignment also transfers soft prompts (Dery et al., 2026). Semantic Cache Distillation includes hidden-input patches alongside reconstructed KV (Ma et al., 2026), so it should not be reduced to

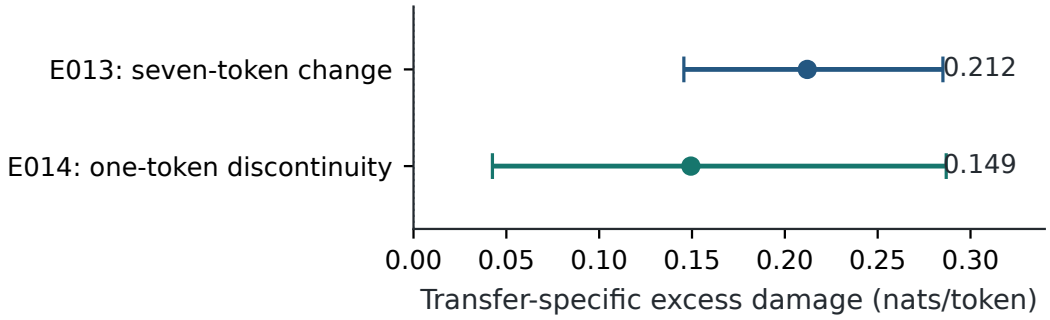

Figure 4: **Small trajectory changes can harm the receiver disproportionately.** Each estimate subtracts the perturbation's source-model loss effect from its transferred-state effect, with the canonical 95% paired context-bootstrap CI ($n = 48$ per study). E013 averages a seven-token history-change effect across its factorial; E014 uses a one-token tail discontinuity on fresh contexts. These are separate assays, not two points on one dose–response curve. Both use 16K histories and 64 teacher-forced targets. Positive values indicate transfer-specific damage, not impossibility of transfer.

raw KV copying. Mixture-of-Translators uses heterogeneous KV mapping and target context replay (Lee et al., 2026). DroidSpeak shares KV across fine-tuned variants with selective recomputation (Liu et al., 2026d); ICaRus deliberately freezes a cache-producing subnetwork to make specialized models share identical KV (Woo et al., 2026). None of these results makes cross-model state transfer generally new.

Hidden representations and external memory. StateBridge supplies aligned hidden messages as continuous receiver prefixes (Peng et al., 2026). Model stitching transfers affine residual-stream features, probes, and steering vectors (Chen et al., 2025); activation injection tests cross-model behavioral effects (Piepereit, 2026). Reader adaptation transfers learned external memory tables (Li et al., 2026a). These establish adjacent forms of representation compatibility, but differ from installing a prefix-conditioned state in a hybrid model's built-in recurrent cache.

Recurrent-state reuse and intervention. Marconi checkpoints hybrid prefix state (Pan et al., 2025); HYPIC composes segment state with attention-seam recomputation (Liu et al., 2026b); DASC compresses recurrent checkpoints (Yu et al., 2026); Tail-Replay reconstructs recurrent state from a bounded suffix (Liu et al., 2026c). They reuse or reconstruct state for the same model. Cache-level interventions in Qwen3.5 and Falcon-H1 show that attention KV and recurrent state can carry causally distinct aspects of context on diagnostic tasks (Afendulev et al., 2026). These interventions combine states within a model, motivating recurrent-state analysis without demonstrating cross-size handoff. WriteSAE directly intervenes on recurrent cache writes and includes transfer of a learned dictionary to a larger host (Young, 2026b); that dictionary is not another model's prefix-conditioned state. S0 tuning learns initial recurrent matrices (Young, 2026a), while Metis trains added memory mechanisms (Zhang et al., 2026).

To our knowledge, prior work has not demonstrated live installation of prefix-conditioned built-in persistent recurrent inference state from a smaller hybrid model into a differently sized receiver without historical target replay. Appendix L preserves detailed distinctions, including replay and receiver adaptation.

6 Discussion

Functional compatibility without representation identity. Matching geometry explains mechanical copying, not why the receiver can use the state. Different hidden widths, weights, read/write projections, and gates make the strong direct-R/C result nontrivial. Donor controls and E011 support partial functional compatibility within this family. Shared pretraining and training recipes may contribute; these factors are not isolated. E013/E014 show why useful compatibility is weaker than a stable state ABI.

Behavior is the relevant objective. Unweighted tensor reconstruction is insufficient to predict receiver usefulness. E001's learned recurrent map has lower reconstruction error but the full mapped package performs worse; E002 independently finds recurrent translation harmful on average. This parallels attention-cache findings that error placement and receiver sensitivity matter (Heo et al., 2026; Qu et al., 2026). The additional correction explicitly optimizes receiver predictions, while its post-evaluation ablations characterize the fitted solution rather than isolate the original package's causal components (Appendix G.1).

A possible state-compatible family design. Future model families could deliberately preserve a compatible persistent-state interface across sizes, extending cache-sharing design ideas beyond attention KV (Woo et al., 2026). The present direct reuse motivates that direction; it does not prove generally shared coordinates or an engineered ABI. Independent families, mismatched geometry, unconstrained generation, and task-level evaluation are decisive next tests. Prototype timing is subordinate to fidelity and does not establish production speedup (Appendix H).

7 Limitations and Conclusion

The study covers one directed Qwen3.5 4B→9B pair with exactly matched persistent geometry. E001/E002 use Base models; E011 improves realism within the same family/pair and only on the tested quantization/runtime stack. Translators and corrections are pair-specific. There is no cross-family or mismatched-geometry replication. Teacher forcing supplies ground-truth inputs that may stabilize the receiver: no native-equivalent free generation, downstream-task equivalence, or production latency improvement is established. Long-context evidence still scores only 64 continuation targets, and eight E011 16K contexts are composites. Bootstrap uncertainty excludes model-pair and fitting-seed variation. E001's package effect does not isolate recurrence from convolution and initialization. Trajectory sensitivity is measured, and shared training may underlie compatibility.

KV is not the complete transferable inference state of a hybrid language model. In this matched sibling pair, persistent recurrent/convolution state written by the smaller model remains useful to the larger receiver without historical replay. Its large advantage over KV-only transfer survives post-training, Q4 quantization, llama.cpp, multiple domains, and 16K context; an additional behavioral correction improves the Base-model handoff. The compatibility is bounded, motivating persistent recurrent state as a first-class object of study in cross-model context reuse.

AI Use Statement

Generative AI tools were used extensively throughout this research. Their use included literature search and comparison; proposing and refining research hypotheses; designing experimental protocols, controls, preregistered decision rules, and follow-up experiments; implementing experimental and analysis code; executing experiments and model evaluations under author-approved research objectives and computational constraints; diagnosing failed experiments; interpreting frozen results; and suggesting subsequent research directions. AI tools were also used to prepare figures, tables, LaTeX artifacts, and drafts and revisions of the manuscript.

Several experiments were conducted through an autonomous research workflow in which an AI coding agent was given the scientific objective, existing experimental record, available compute, and methodological constraints, and was permitted to formulate hypotheses, implement and run experiments, analyse outcomes, and select or recommend follow-up experiments. Experimental outcomes were not manually altered to obtain positive results. Where applicable, protocols, evaluation criteria, stopping rules, data splits, and decision thresholds were frozen before the corresponding evaluation results were observed, and negative results were retained. Machine-readable experimental records, hashes, tests, audit artifacts, and independent recomputation procedures were used to check reported numerical results and preserve provenance.

The author determined the overall research goals and scope, reviewed the experimental record and scientific conclusions, made the final decisions about which claims and results to include, and takes responsibility for the experimental methodology, reported results, manuscript, code, and submission.

Reproducibility Statement

Appendices A–D specify the E001/E002 checkpoints, runtime, fitting, scoring, and statistical procedures. Appendix J gives the E011 replication and full condition results; Appendix K gives the compact boundary controls. All scientific figures derive from frozen evidence; no new scientific evaluations were performed for this revision.

References

Kirill Afendulev, Alexey Dontsov, Elena Tutubalina, and Anton Korznikov. What Attention Recalls and Recurrence Controls in Hybrid Language Models, 2026. URL https://arxiv.org/abs/2609.04434. arXiv preprint arXiv:2609.04434.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring Linear Features Across Language Models With Model Stitching. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/4569a868e7aa891248832ec08445d071-Abstract-Conference.html.

Lucio M. Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent Space Communication via K-V Cache Alignment, 2026. URL https://arxiv.org/abs/2601.06123. arXiv preprint arXiv:2601.06123.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-Cache: Direct Semantic Communication Between Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.03215.

Roy Ganz, Mor Shpigel Nacson, Adi Kalyanpur, and Ron Litman. The Handoff Tax: Continuing Non-Native Trajectories in LLM Agents, 2026. URL https://arxiv.org/abs/2608.24358. arXiv preprint arXiv:2608.24358.

Taekyung Heo, Rasoul Shafipour, Ritchie Zhao, Maximilian Golub, Mohammad Mahdi Kamani, Ritika Borkar, Makesh Tarun Chandran, Pantea Zardoshti, and Bita Darvish Rouhani. Cross-Model KV Cache Transfer in LLM Families: A Closed-Form Linear Mapping for Prefill Reuse, 2026. URL https://arxiv.org/abs/2608.03893. arXiv preprint arXiv:2608.03893.

Jin-woo Lee, Minkyung Song, Junghyun Oh, Seunghoon Han, Soyoung Park, Gwangseon Jang, and Sungsu Lim. Mixture-of-Translators: Translating KV Caches Across Heterogeneous Large Language Models, 2026. URL https://arxiv.org/abs/2607.28979. arXiv preprint arXiv:2607.28979.

Mingyuan Li, Guangsheng Yu, Xu Wang, and Shaoxiong Ji. Cross-Model Memory Transfer via Target-Side Reader Adaptation, 2026a. URL https://arxiv.org/abs/2608.17050. arXiv preprint arXiv:2608.17050.

Yi Li, Dongming Jiang, Yi Zhao, and Bingzhe Li. A Universal Context-Reuse Layer for Cross-Model KV Sharing, 2026b. URL https://arxiv.org/abs/2608.30963. arXiv preprint arXiv:2608.30963.

Jiyao Liu, Qi Zhang, Yaoyi Jia, Ziwen Kan, and Song Wang. Dual-Cache Latent Space Communication between Heterogeneous Language Models, 2026a. URL https://arxiv.org/abs/2608.20617. arXiv preprint arXiv:2608.20617.

Yifei Liu, Juntong Wu, Yang Liu, Junhao Hu, Minghao Li, Xiaoxu Chen, and Weihang Chen. HYPIC: Accelerating Hybrid-Attention LLM Serving with Position-Independent Caching, 2026b. URL https://arxiv.org/abs/2607.01299. arXiv preprint arXiv:2607.01299.

Yirui Liu, Ruoling Qi, Xuaner Wu, Penghang Liu, and Jian Chen. Tail-Replay: Escaping the Curse of Linear Attention in Prefix Caching for Hybrid LLMs, 2026c. URL https://arxiv.org/abs/2608.30310. arXiv preprint arXiv:2608.30310.

Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, Madan Musuvathi, and Esha Choukse. DroidSpeak: KV Cache Sharing Across Fine-tuned Model Variants. In *USENIX Symposium on Networked Systems Design and Implementation*, 2026d. URL https://www.usenix.org/conference/nsdi26/presentation/liu-yuhan.

Qianli Ma, Zhiqing Tang, Hanshuai Cui, Zhi Yao, and Weijia Jia. Semantic Cache Distillation: Efficient State Transfer via Reuse and Selective Patching. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.07684.

Rui Pan, Zhuang Wang, Zhen Jia, Can Karakus, Luca Zancato, Tri Dao, Yida Wang, and Ravi Netravali. Marconi: Prefix Caching for the Era of Hybrid LLMs. In *Proceedings of Machine Learning and Systems*, 2025. URL https://arxiv.org/abs/2411.19379.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, 2024. URL https://arxiv.org/abs/2406.17557. arXiv preprint arXiv:2406.17557.

Yanwen Peng, Delvin Ce Zhang, Xi Wang, and Nikolaos Aletras. StateBridge: Training-free Hidden-state Alignment for Latent Communication in LLM Multi-Agent Systems. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2608.13317.

Fernando Cardenas Piepereit. Architecture-Dependent Causal Transfer of Activation States Across Large Language Models, 2026. URL https://arxiv.org/abs/2608.16347. arXiv preprint arXiv:2608.16347.

Xingyu Qu, Siyuan Lu, Zhiyu Chen, Sheng Wang, and Tao Lin. CacheBridge: Efficient Cross-Model KV Cache Transfer, 2026. URL https://arxiv.org/abs/2609.00891. arXiv preprint arXiv:2609.00891.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, and Timothy P. Lillicrap. Compressive Transformers for Long-Range Sequence Modelling. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.05507.

Maximillian Rossi, Prajwal Raghunath, and Eugene Wu. Latent Cache Flow: Model-to-Model Communication Without Text, 2026. URL https://arxiv.org/abs/2605.22863. arXiv preprint arXiv:2605.22863.

Sunghyeon Woo, Jaeeun Kil, Hoseung Kim, Minsub Kim, Joonghoon Kim, Ahreum Seo, Sungjae Lee, Minjung Jo, Jiwon Ryu, Baeseong Park, Se Jung Kwon, and Dongsoo Lee. ICaRus: Identical Cache Reuse for Efficient Multi Model Inference. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.13281.

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

Jack Young. S0 Tuning: Zero-Overhead Adaptation of Hybrid Recurrent-Attention Models, 2026a. URL https://arxiv.org/abs/2604.01168. arXiv preprint arXiv:2604.01168.

Jack Young. WriteSAE: Sparse Autoencoders for Recurrent State, 2026b. URL https://arxiv.org/abs/2605.12770. arXiv preprint arXiv:2605.12770.

Yanqi Yu, Pingwei Sun, Jianchao Tan, Tao Zhang, Yuchen Xie, Xunliang Cai, and Yao Liu. DASC: Decay-Aware State Compression for Hybrid Linear-Attention Serving, 2026. URL https://arxiv.org/abs/2608.30386. arXiv preprint arXiv:2608.30386.

Zeyu Zhang, Ziliang Guo, Yihang Sun, Xichong Zhang, Xixuan Hao, Zehao Lin, Yang Zhang, Xiaoyan Zhao, Tong Shen, Bo Tang, Zhi-Qin John Xu, Junchi Yan, Haofen Wang, Xu Chen, Feiyu Xiong, Zhiyu Li, and Tat-Seng Chua. Metis: Memory Foundation Model, 2026. URL https://arxiv.org/abs/2607.26760. arXiv preprint arXiv:2607.26760.

A E001/E002 Frozen Models and Runtime

The source repository is `Qwen/Qwen3.5-4B-Base`, revision

`1001bb4d826a52d1f399e183466143f4da7b741b`.

The target is `Qwen/Qwen3.5-9B-Base`, revision

`68c46c4b3498877f3ef123c856ecfde50c39f404`.

The tokenizer uses the source repository and revision. The frozen `tokenizer.json` SHA-256 is

`fe000e3ed39ed12b8d2481d527d44f93c65d37e87645d2dcc80d1bf9d50d2927`.

These are official Base checkpoints. No instruct substitution, weight updates, LoRA, quantization, MTP, or speculative decoding is part of either experiment. Loaded language-model parameter counts are 4,205,751,296 and 8,953,803,264.

The recorded runtime uses Python 3.11.9, PyTorch 2.13.0+cu130, CUDA runtime 13.0, Transformers 5.16.1, Datasets 5.0.1, and Hugging Face Hub 1.29.0. Hardware is an NVIDIA GeForce RTX 5090, compute capability 12.0, with 32,607 MiB reported VRAM; driver 591.86 reports CUDA 13.1 support. These version numbers are taken from the frozen runtime artifact. They describe the experimental environment, not the lighter Python/LaTeX environment used to render this paper.

The architecture gate checks all 32 layer types, attention KV geometry, the GDN recurrent and convolution geometry, and RoPE geometry. Attention has eight layers at zero-based indices $3, 7, \ldots, 31$. Each K or V tensor is $1 \times 4 \times L \times 256$. Each of the other 24 layers retains one recurrent tensor of shape $1 \times 32 \times 128 \times 128$ and a convolution tensor of shape $1 \times 8192 \times 4$. Recurrent memory is FP32; attention and convolution memory use BF16. At 4K, the complete tensor payload is 186,122,240 bytes for either model. Metadata includes initialization flags, previous-state flags, convolution kernel sizes, logical length, and positions; it is constructed deterministically rather than learned.

Complete same-model restoration uses 16 deterministic contexts per model at lengths 256, 512, 1,024, and 2,048. Both achieve top-1 agreement 1.0 and maximum absolute logit difference 0; the minimum next-logit cosine is 0.99999988. A separate pre-fitting one-shot versus eight-chunk 1,024-token check has maximum bridge-logit difference 0.125. Each FIT/VALIDATION checkpoint was therefore captured by an independent fresh-cache one-shot prefill at 128-token intervals. Canonical continuation shares the 1/4/16/64 checkpoint schedule across conditions; E002 extends state tracking to 256 inputs. These are frozen implementation decisions.

The installer rejects activation of only one GDN component: recurrent and convolution restoration share a previous-state branch. Consequently, an independently fresh-R/direct-C or direct-R/fresh-C intervention is not implemented by the frozen path. Bypassing the guard and supplying initialized zeros would change the meaning of "fresh." No new presence ablation was run for this revision; isolating R from C requires a separately validated intervention. The local revision feasibility audit records the exact code paths. This does not alter either original experiment.

B Splits and Deterministic Selection

E001 uses FineWeb-Edu revision

`87f09149ef4734204d70ed1d046ddc9ca3f2b8f9`

and immutable shard `sample/100BT/000_00000.parquet`. Its FIT/VALIDATION candidate universe is physical rows 0–19,999. Eligible rows are sorted by a SHA-256 key containing the frozen salt `latentport-e001-splits-v1`, corpus revision, shard, physical row, and document ID. Duplicate document IDs or UTF-8 text hashes retain the lowest-ranked row. The first 128 eligible documents form FIT and the next 32 VALIDATION, each requiring at least 1,089 token IDs.

PG19 uses `emozilla/pg19`, revision

`c021754c8e01c5b1cc83a1f549c1f97fbbb756b8`,

and its frozen test parquet. Hash ordering first reserves 16 documents meeting the 16,449-token LONG requirement; LOCKED takes the first 64 remaining eligible documents meeting the 4,161-token requirement. Reserve membership is part of the frozen selection protocol, not evidence of LONG evaluation.

E002 uses the same FineWeb-Edu revision and shard but expands the candidate universe to rows 0–99,999 and uses the salt `latentport-e002-splits-v1`. Its coarse filter uses source-reported token count before exact eligibility under the frozen Qwen tokenizer. All E001 split IDs and text hashes are excluded from fresh selection. The factorial reserves 32 fresh 4K documents; LOCKED contains 64 further fresh 4K documents. The LOCKED records store 4,352 tokens to support bridge-plus-255-input state tracking. Correction fitting/validation explicitly reference the 128/32 E001 non-LOCKED documents. Split IDs, source hashes, and input tokens remain frozen. The paper verification reads no unused LONG token payloads.

C Translator and Correction Details

C.1 Correction parameterization

The correction is identity-anchored and leaves both language models frozen. For a KV feature matrix X,

$$X' = X + (XB)U^\top, \tag{4}$$

where B is fixed and orthonormal and U is trainable, shared across heads and positions within a layer/role. Recurrent factors are shared across heads within a layer:

$$S' = S + U_L B_L^\top S + S B_R U_R^\top. \tag{5}$$

Convolution uses $C'_{j,t} = C_{j,t} + a_j C_{j,t} + b_j$, shared over buffer positions. All trainable residual outputs start at zero; fixed bases remain nonzero. Components are optimized jointly through the receiver, without explicit cross-component tensor mixing.

The behavioral loss $\mathcal{L}_{\mathrm{KL}}(d)$ is full-vocabulary KL from native 9B to the handoff, averaged over nine output positions (bridge plus eight teacher inputs) for document d. Let $B_{d,c,j}$ denote base-state tensor j in component $c \in \{K, R, C\}$, and $\Delta_{d,c,j}$ its FP32 correction residual before casting for installation. The implemented identity penalty is

$$\mathcal{L}_{\mathrm{id}}(d) = \frac{1}{3} \sum_{c \in \{K,R,C\}} \frac{\sum_{j \in \mathcal{I}_c} \|\Delta_{d,c,j}\|_F^2}{\max\left(\sum_{j \in \mathcal{I}_c} \|B_{d,c,j}\|_F^2, 10^{-12} \right)}. \tag{6}$$

Here $\mathcal{I}_K$ includes both K and V tensors from all eight attention layers; $\mathcal{I}_R$ and $\mathcal{I}_C$ include the respective tensors from all 24 GDN layers. Norms sum all tensor entries, including heads and positions, in FP32. Each component's squared norms are summed before division, with no separate layer, role, head, or tensor normalization. The denominator guard clamps the summed *squared* base norm at 10^{-12}; it is not an additive epsilon. KV, recurrent, and convolution ratios receive equal weight, independent of record count. This differs from the layer-record pooling used for the reported magnitude summaries.

The per-document training objective is

$$\mathcal{L}_d = \mathcal{L}_{\mathrm{KL}}(d) + \lambda_{\mathrm{id}} \mathcal{L}_{\mathrm{id}}(d), \qquad \lambda_{\mathrm{id}} = 0.01 \text{ for the selected correction.} \tag{7}$$

Training divides each document's loss by eight and accumulates gradients over eight documents per optimizer step. Validation selects rank 4 and identity weight 0.01. The 434,176 trainable parameters are an *additional correction*: they exclude the fitted KV translator, fixed bases, and both language models. They are not the size of the complete transfer system.

Rank	λ_{id}	Epoch	Val. KL
2	0	3	0.171111
2	0.0001	3	0.169789
2	0.001	3	0.170405
2	0.01	3	0.170945
4	0	3	0.162947
4	0.0001	3	0.163098
4	0.001	3	0.162811
4	0.01	3	0.161918

Table 4: Frozen E002 correction selection grid. Validation KL is mean full-vocabulary native-to-handoff KL over nine output positions on 32 non-LOCKED 1K documents. These are selection scores, not LOCKED results.

C.2 Frozen fitting configuration

The E001 ridge grid is $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}\}$ relative to $\mathrm{tr}(X_c^\top X_c)/d$. KV and convolution fits consider centering alone or per-feature standardization; recurrent fits use centering and exactly three alternating updates of Eq. 2. The KV position sampler selects 64 positions per FIT document, including endpoints and 128-token boundaries. Recurrent fitting uses eight independent one-shot checkpoint captures per document. Hyperparameter selection minimizes registered validation tensor error; no layer/head alignment or position parameters are learned.

The frozen E001 operator weights comprise 4,194,304 KV coefficients, 25,165,824 recurrent coefficients, and 25,165,824 convolution coefficients: 54,525,952 in total, excluding stored centering/normalization values. The full bundle occupies 322,222,408 bytes. E002's selected TDD path uses the KV operator and direct GDN state. Its KV tensor artifact alone is 17,042,160 bytes, in addition to the 1,901,432-byte correction. Thus correction size should not be confused with total deployed mapping storage.

E002 uses rank $r \in \{2, 4\}$ and identity penalty $\lambda_{\mathrm{id}} \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. Table 4 reports every selected-per-candidate checkpoint. The fixed Gaussian-QR basis seed is 2026090102; the global/training seeds are 2026090101 and 2026090106. AdamW uses learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, no weight decay, one document per batch, gradient accumulation over eight documents, and gradient-norm clipping at 1. Training follows manifest order for at most three epochs, with early-stop patience one and minimum improvement 10^{-6}. Candidate selection minimizes validation behavioral KL; ties within 10^{-6} prefer lower rank, larger identity penalty, then earlier epoch. The selected checkpoint has rank 4, $\lambda_{\mathrm{id}} = 0.01$, and epoch 3.

The rank-4 trainable count is

$$P_K = 8 \cdot 2 \cdot 256 \cdot 4 = 16{,}384,$$
$$P_R = 24 \cdot 2 \cdot 128 \cdot 4 = 24{,}576,$$
$$P_C = 24 \cdot 2 \cdot 8192 = 393{,}216.$$

The fixed bases are buffers, not trainable parameters. All trainable outputs, scales, and biases start at zero. The layer-relative magnitude statistics pool 64 documents' records; they are not just norms of the trained weights. KV norms combine the K and V residuals within each attention layer, while recurrent and convolution records are separate. The pooled median therefore reflects this explicit record weighting.

Attention translation fits each layer, KV head, and K/V role separately. Its positional step operates on the 64 rotary dimensions of each 256-dimensional key. The implementation explicitly inverts the BF16-rounded sine/cosine transform, dividing by the corresponding squared-sine-plus-squared-cosine factor; nonrotary coordinates are unchanged by that step. Recurrent bilinear maps use three deterministic alternating ridge updates per regularization candidate. Convolution maps operate within the runtime's packed Q/K/V component and head groups. Selection uses validation tensor error before either held-out evaluation.

The selected correction contains 434,176 trainable parameters under an exclusive two-million-parameter cap. Its serialized artifact is 1,901,432 bytes, 1.0618013×10^{-4} of target BF16 parameter

Component	Parameters	Median norm
KV (rank 4)	16,384	0.027093
Recurrent (rank 4)	24,576	0.031916
Conv. scale / bias	393,216	0.015320
Total / pooled	434,176	0.026432

Table 5: Frozen E002 correction inventory. Median norm is the residual/base Frobenius norm over LOCKED document–layer records, pooled separately by component or over all records. Rank 4 applies to KV and recurrent residuals; convolution uses channel scale/bias. Counts exclude fixed bases and the existing KV translator.

bytes. The pooled median residual/base norm is 0.026432 and the maximum 0.043951. Table 5 gives the component inventory. These magnitudes concern the installed state perturbation, not native-state reconstruction error or prediction error.

D Condition and Metric Semantics

Native 9B processes the historical prefix normally; source 4B continues from its own prefix state. Empty 9B begins without historical state at the same logical continuation position. E001 KV-only installs translated attention KV and leaves both GDN components fresh. KV plus direct GDN installs translated KV with copied recurrent and convolution tensors. Full translated installs all three frozen mapped components. Full shuffled uses a different document's complete translated state. E002 D/T cells install the stated source or translated components, base is frozen TDD, corrected applies Eqs. 4–5 and the convolution residual, and joint shuffled rotates complete corrected donor states. Donor rotations are within equal-length records, with no self-donors.

The bridge is input token x_{L+1}; 64 scored targets are $x_{L+2:L+65}$. The segmented canonical path consumes the bridge and the first 63 continuation targets as inputs. A state checkpoint labeled 64 therefore follows 64 post-handoff inputs, including the bridge. E002 training uses the bridge plus eight teacher inputs to supply nine output distributions, in one differentiable target call. Native target traces needed for fitting are offline supervision; they do not imply target historical replay in the deployed handoff.

For protocol completeness, target-quality recovery and remaining-gap reduction are

$$\mathrm{TQR}_c = \frac{\ell_{\mathrm{source4B}} - \ell_c}{\ell_{\mathrm{source4B}} - \ell_{\mathrm{native9B}}}, \tag{8}$$

$$\mathrm{RGR} = \frac{\Delta\,\mathrm{NLL}_{\mathrm{base}} - \Delta\,\mathrm{NLL}_{\mathrm{corrected}}}{\Delta\,\mathrm{NLL}_{\mathrm{base}}}. \tag{9}$$

TQR normalizes by the target's advantage over the source; corrected E002 TQR is 0.405755. Like NCR, it is a ratio of aggregate means, not a mean of document ratios. It is undefined for a nonpositive denominator; RGR is undefined for a nonpositive base gap. Neither normalized score is accuracy.

For distributions p (native) and q (condition), $m = (p+q)/2$ and

$$\mathrm{JS}(p,q) = \tfrac{1}{2}\,\mathrm{KL}(p\|m) + \tfrac{1}{2}\,\mathrm{KL}(q\|m). \tag{10}$$

Natural logarithms are used. Top-5 overlap is $|\,\mathrm{top5}(p) \cap \mathrm{top5}(q)|/5$, and entropy is $-\sum_v q_v \log q_v$. These metrics average over the shared target positions and documents. State error is $\|S_c - S_{\mathrm{native}}\|_F / \max(\|S_{\mathrm{native}}\|_F, 10^{-12})$; cosine uses flattened tensors. New attention entries are compared only after handoff.

E001's improvement fraction for document d is

$$f_d = \frac{\Delta\,\mathrm{NLL}_{d,\mathrm{KV}} - \Delta\,\mathrm{NLL}_{d,\mathrm{full}}}{\Delta\,\mathrm{NLL}_{d,\mathrm{KV}}}$$

when the denominator is positive. All 64 qualify. Its 79.7% mean is therefore not the ratio of the aggregate 0.7473 improvement to the aggregate KV-only gap. One E001 document has a nonpositive source-minus-native denominator and is excluded from document-level TQR; the reported aggregate TQR uses all documents' mean NLLs and remains defined.

Condition	NLL	ΔNLL	Top-1
Native 9B	2.160916	0.000000	1.000000
Source 4B	2.307412	0.146496	–
Empty 9B	3.245654	1.084738	0.581299
KV only	3.115271	0.954355	0.600098
KV + direct GDN	2.291314	0.130398	0.821289
Full translated	2.367948	0.207032	0.783203
Full shuffled	3.316181	1.155265	0.555664

Table 6: E001 LOCKED results: 64 PG19 documents, 4K prefix, 64 teacher-forced targets. NLL and excess NLL are nats/token; top-1 is native-target prediction agreement.

Condition	NLL	ΔNLL	NCR	TQR	Top-1	JS
Native 9B	1.913032	0.000000	1.000000	1.000000	1.000000	0.000000
Source 4B	2.041544	0.128512	0.861597	0.000000	0.833984	0.038052
Empty 9B	2.841564	0.928532	0.000000	-6.225273	0.626465	0.156773
TDD / base	2.018422	0.105390	0.886498	0.179921	0.846436	0.028583
TTT	2.047468	0.134437	0.855216	-0.046104	0.832031	0.036172
Joint corrected	1.989399	0.076367	0.917755	0.405755	0.864258	0.022067
Joint shuffled	3.177764	1.264732	-0.362078	-8.841385	0.578369	0.192132

Table 7: E002 LOCKED results: 64 fresh FineWeb-Edu documents, 4K prefix, 64 teacher-forced targets. All entries are document means or defined ratios of document means. NLL and excess NLL use nats/token. NCR, TQR, top-1 agreement, and JS have distinct meanings; native/empty NCR and native/source TQR are fixed by definition.

The bootstrap samples documents with replacement, preserving each document's condition pairing. Ten thousand means are drawn with NumPy's default generator using frozen seeds 2026083103 (E001) and 2026090103 (E002); percentile endpoints are 0.025 and 0.975. The RGR interval recomputes its ratio of means inside each bootstrap resample. No token-level resampling, alternate confidence procedure, or multiplicity-adjusted exploratory claim is substituted for the frozen protocol.

The E002 full-state gate requires a positive base-minus-corrected CI, remaining-gap reduction at least 0.25, corrected excess NLL at most 0.10, NCR at least 0.90, and upper CI endpoints below zero for corrected-minus-source and corrected-minus-shuffled. The stronger near-native gate additionally requires excess NLL at most 0.05, NCR at least 0.95, and top-1 agreement at least 0.90. Only that stronger outcome unlocks the 16K branch. E002's recorded verdict is `FULL_STATE_HANDOFF`, not `NEAR_NATIVE_HANDOFF`; all three stronger thresholds fail. E001's recorded verdict is `RECURRENT_STATE_TRANSLATABLE`: its mean and median document improvement fractions exceed the 25% contribution criterion. Its fully translated condition fails the stronger full-state gate, with excess NLL 0.207032 above 0.20 and TQR -0.413227 below 0.75. Gates are experiment-specific. Both canonical results explicitly record the 16K branch as unrun.

E Full-Precision Continuation Results

Tables 6–7 retain the six-decimal numerical presentation. The included verified JSON retains full floating-point values. Figure 5 shows the Experiment 2 controls on a common overview scale and an explicitly expanded fidelity scale. No results are pooled across the two corpora.

F Complete Component Factorial

Table 8 includes every D/T cell. Coding D as -1 and T as $+1$, a main effect is the within-document signed sum over eight cells divided by four. Pairwise interactions divide by two, yielding a difference-in-differences averaged over the remaining factor; the three-way contrast is the full signed sum. Table 9 shows these contrasts and their frozen intervals. A LOCKED interaction can have an interval excluding zero and still fail the preregistered 0.02-nat materiality criterion. That is

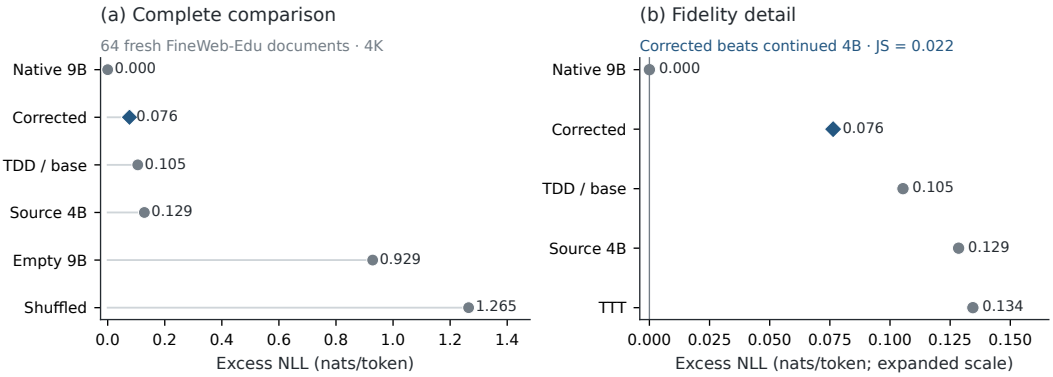

Figure 5: E002's complete test comparison. Markers are means over 64 paired documents. The left axis includes empty and shuffled controls; the right expands the scale and includes TTT. Corrected excess NLL is 0.076367. Base-minus-corrected improvement is 0.029022 nats/token (95% paired document bootstrap CI [0.022828, 0.035413]); plotted means are not independent token-level replicates.

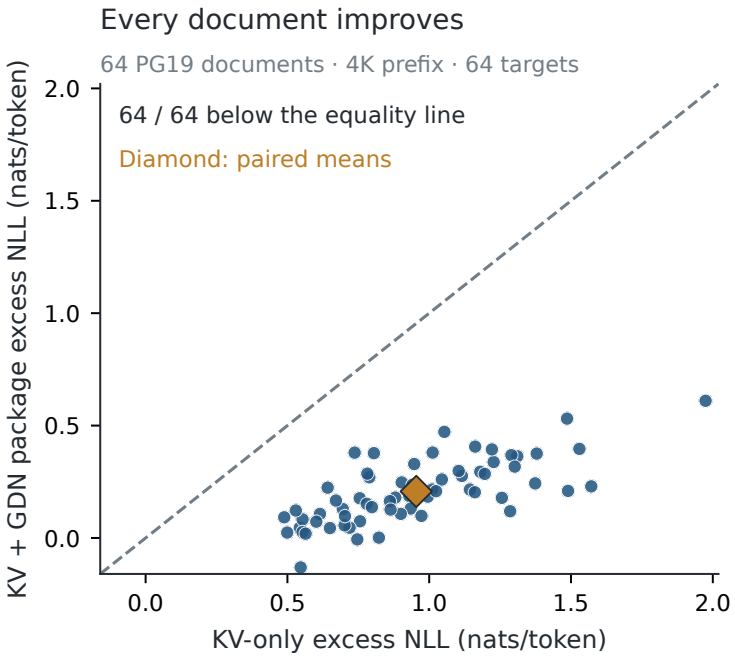

Figure 6: E001 paired per-document excess NLL: each point is one of 64 held-out PG19 documents. Every point favors full translated state over KV-only. The primary mean difference and its paired interval appear in the main text.

the case for $K \times R$. We therefore report failure of the specified material replication criterion, rather than asserting that all interactions are zero.

G Secondary State Diagnostics and Post-Verdict Ablations

Figure 8 presents the preregistered state-tracking endpoints from the frozen evidence. E001 statistics are summarized in a post-verdict derived artifact, but the tracked checkpoints and measurements were specified before the LOCKED run. The E002 base/corrected trajectories are close, and corrected state is not uniformly closer at the earliest inputs. Continued target computation reduces recurrent mismatch without eliminating it. Neither state distance nor cosine is a success gate. E001

Cell	Validation	LOCKED
DDD	0.217430	0.165163
DDT	0.217473	0.166589
DTD	0.231225	0.178138
DTT	0.231250	0.180053
TDD	0.133377	0.105390
TDT	0.132226	0.105677
TTD	0.175278	0.133202
TTT	0.174644	0.134437

Table 8: Every E002 factorial cell's excess NLL, in nats/token. Validation has 32 fresh FineWeb-Edu documents; LOCKED has 64 different documents. Both use 4K prefixes. Base selection uses validation alone.

Contrast	Validation	95% CI	LOCKED	95% CI
KV	-0.070463	[-0.100867, -0.040946]	-0.052809	[-0.069557, -0.037096]
Recurrent	0.027973	[0.010360, 0.047548]	0.020753	[0.007905, 0.034039]
Convolution	-0.000429	[-0.002647, 0.001774]	0.001216	[-0.000055, 0.002580]
$K \times R$	0.028373	[0.008556, 0.057582]	0.015067	[0.008212, 0.022371]
$K \times C$	-0.000927	[-0.002873, 0.001039]	-0.000910	[-0.002206, 0.000378]
$R \times C$	0.000250	[-0.002447, 0.003218]	0.000718	[-0.001490, 0.003543]
$K \times R \times C$	0.000535	[-0.003747, 0.005266]	0.000458	[-0.001807, 0.002798]

Table 9: E002 orthogonal within-document factorial contrasts, in nats/token, with 95% paired document bootstrap intervals. Main effects are translated minus direct, averaged over the other two components. Pairwise material replication requires matching signs and absolute LOCKED estimate at least 0.02; no pairwise contrast satisfies both requirements.

full-translated recurrent normalized error declines from 0.4687 after one new input to 0.4053 after 64. E002 corrected error declines from 0.6239 to 0.3650 through 256 inputs, while cosine rises from 0.7778 to 0.9266. Shared teacher inputs and gated forgetting can reduce discrepancy without recovering all prefix information.

G.1 Post-verdict component removals

After the 4K verdict was frozen, E002 disabled components of the selected correction on the same LOCKED documents. Table 10 shows the resulting impact relative to the full correction. Removing recurrent correction has the largest aggregate effect among complete-component removals; removing convolution correction has a small interval spanning zero. The middle recurrent third is the most sensitive of the three depth blocks under this intervention. These are post-verdict diagnostics with no multiplicity correction, no refitting, and no independent confirmation set. They describe sensitivity of the fitted correction; they do not establish the best sparse architecture or authorize changing the selected result.

H Prototype Timing and Its Accounting Boundary

Table 11 preserves E002's recorded timing summary. Each primitive is aggregated as its document median. The available budget is the native-prefill median minus the source-prefill median; the reported margin further subtracts the sum of base construction, correction, installation, and bridge medians. A difference or sum of medians is not an independently measured median end-to-end latency.

The code makes an additional exclusion explicit: base-state construction selects components from *already materialized* source and translated states. The translation pass occurs earlier and is not charged to the reported 131.310-ms margin. This margin also does not establish the full cost of extraction, serialized evidence I/O, model loading, batching, scheduling, or a GPU-local deployment. The original report's "end-to-end" characterization is therefore too broad for the underlying aggre-

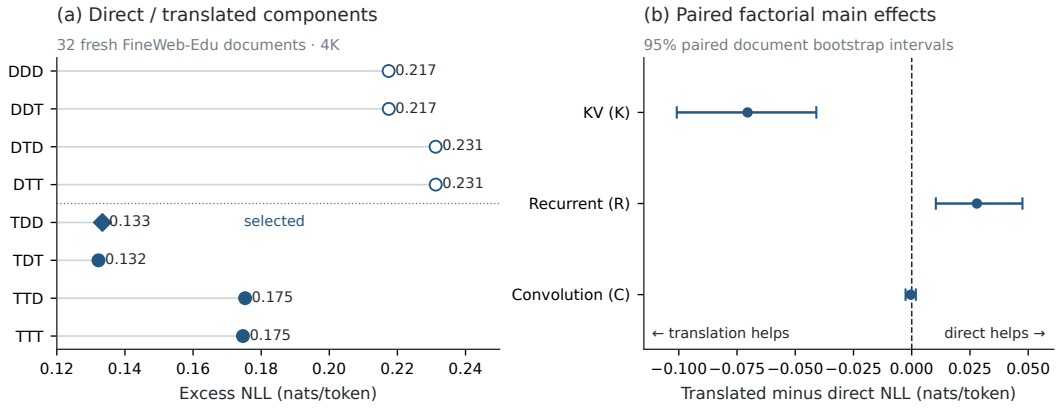

Figure 7: E002 component factorial on 32 fresh validation documents. D/T denote direct/translated (K, R, C). TDT has the smallest numerical NLL, while the frozen tie rule selects TDD. Main-effect intervals use paired document bootstrap resampling.

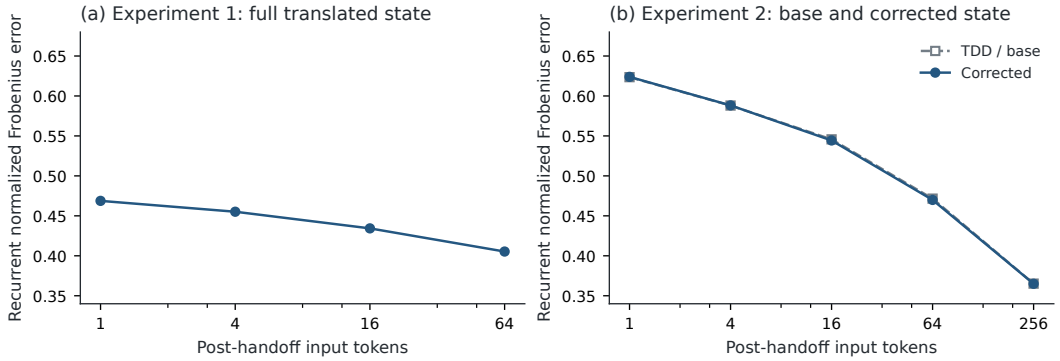

Figure 8: Secondary recurrent-state convergence under shared teacher-forced inputs. Each marker averages layer/head state errors over 64 documents within the indicated experiment. The horizontal axis is logarithmic with only the recorded checkpoints shown. E001 and E002 use different LOCKED corpora and are not paired with one another. E002's 256-input state measurement is not a 16K-context test or a free-generation evaluation.

gation; the numeric artifact is retained, and the narrower code-defined accounting is reported here. No production latency saving follows from the table.

I Evidence Integrity and Paper Reproduction

The paper uses the frozen `e001_handoff`, `e002_coupler`, `e011_isolate`, `e013_faultline`, and `e014_basin` evidence. The first two studies supply the hashes below; the later studies supply their own canonical results and independent validation records. The LOCKED raw evidence and E002 fitting artifacts are stored in the corresponding original E001/E002 archive directories. Their raw SHA-256 values match the canonical result references:

E001: `1639038217b9bb376de33be57ab624e083a093a45dc5a5f8749b565247c9e546`

E002: `a189c454a4f52a8ceb03ec09e0e109efa139ce01225a472c59a70148acd0c629`

E001's frozen execution manifest hash is

`69bcca1bea1ecce2cc5c7c9728581aeca8ed9abf56028c2cd9c2fd7498efb0c9`,

Removed	ΔNLL	Impact	95% CI
Convolution	0.077026	0.000658	[-0.000387, 0.001685]
R: early third	0.082034	0.005666	[0.002873, 0.008594]
KV	0.083648	0.007281	[0.004475, 0.010285]
R: late third	0.077912	0.001545	[0.000136, 0.002925]
R: middle third	0.086258	0.009891	[0.006875, 0.013005]
R: all layers	0.095287	0.018920	[0.014069, 0.024019]

Table 10: Post-verdict E002 ablations on the same 64 LOCKED documents. R denotes recurrent state. The named correction is removed; positive impact is increased NLL relative to the complete correction. All values are nats/token; intervals are paired document bootstrap intervals. These diagnostics did not determine or revise the canonical verdict.

Recorded stage	Median (ms)
Native 9B prefill	886.646
Source 4B prefill	634.911
Base construction	0.041
Correction	69.145
State installation	15.829
Bridge token	35.410
Difference of prefill medians	251.736
Reported handoff margin	131.310

Table 11: E002's frozen engineering timing summary. Primitive values are document wall-clock medians; the final two rows are arithmetic combinations of those medians. The accounting excludes the preceding state-translation pass and other costs discussed in the text, so the reported margin is not a measured production speedup.

and E002's is

`4ea9e85a2450a6c0f8dbb979f6ecfba32d4f45606103a946a32595208359070c`.

The selected correction tensor SHA-256 is

`3565d5ddb81b9c195e0cc02e143acf85a49f8fe732a794fe5d1d1c2463cf6331`.

The combined metadata-plus-tensor correction hash is distinct from this file hash:

`e93a15a21b6eba37ec794d94c45780059a3653f9466c26db1b20960419260db3`.

All reported primary intervals remain the canonical frozen intervals.

J E011 Replication Details and Complete Results

The Base revisions match E001/E002. Post-trained source `Qwen/Qwen3.5-4B` uses revision

`851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a`,

and target `Qwen/Qwen3.5-9B` uses

`c202236235762e1c871ad0ccb60c8ee5ba337b9a`.

The pinned llama.cpp commit is

`b19cbe925be361d229f0fe03435affe4a2717f37`.

All four checkpoints were converted through BF16 and quantized to Q4_K_M without an importance matrix. MTP and multimodal paths were disabled. The captured recurrent tensor has shape $24 \times 32 \times 128 \times 128$; convolution history is $24 \times 3 \times 8192$. K and V each have shape $8 \times L \times 4 \times 256$. R/C use F32 and KV uses F16. Geometry is matched between model sizes in this runtime, not identical to the serialized Transformers layout.

Corpus provenance. Literature uses the PG19 training corpus; STEM uses article text from `ccdv/arxiv-summarization`; code uses `codeparrot/github-code`; conversational text uses `HuggingFaceH4/ultrachat_200k`. The latter is an existing synthetic conversation corpus, not a new dataset generated for this study. The frozen source revisions are respectively recorded per PG19 file, `240aaf1a969b3f8cd0ade6986bfad0cd730ee288`, `b5661e6b17396364b2bcf8e68977b0d28e1ebd19`, and `8049631c405ae6576f93f445c6b8166f76f5505a`. Source manifests retain URL, document ID, text hash, offsets, and constituent IDs. Selection reuses an already sealed raw-text corpus, not earlier experimental outcomes. The split audit verifies disjoint source documents and identical token IDs across all four checkpoints.

Calibration and selection. Independent Base/post-trained affine ridge maps use same-index attention layers and heads, de-RoPE/re-RoPE keys, and 256 positions sampled every fourth token from each 1K calibration sequence. A balanced 400/100 split selects center-only versus per-feature normalization and relative ridge multipliers $10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$ by held-out normalized L2, then refits on all 500 sequences. No correction or long-context refit is applied. Calibration, 4K, and 16K source-document sets are disjoint; canonical IDs match across four tokenizers. Four domains have equal weight. The 16K set contains 24 natural documents and eight concatenated contexts (two code, two STEM, four conversational), constructed before outcomes from disjoint documents separated by four newlines.

Controls, audit, and inferential scope. Fresh-context capture/restore checks each model at 256, 1,024, 2,048, and 4,096 tokens, including instrumentation invariance and exact positions/geometry. All gates pass. The independent validator reports 284,516 checks with zero failures, recomputing token-level aggregates, original intervals, branch rules, and retained-logit audits. Full float32 logits are retained for a hash-selected one-in-sixteen sample plus a guaranteed document per domain; other records retain per-token loss, log-sum-exp, target logit, entropy, JS, and top-k information. This manuscript independently reproduces aggregate NLL, normalized metrics, paired NLL intervals, and post-training ratio intervals from the compact frozen records.

NCR and normalized excess loss $\mathrm{NEL} = 1 - \mathrm{NCR}$ are unclipped. The primary interaction is $\mathrm{NEL}_{\mathrm{post,full}} - \mathrm{NEL}_{\mathrm{base,full}}$, equal to Base-minus-post NCR. At 4K, material post-training degradation requires an estimate at least 0.15 and a strictly positive CI. Both stages instead have NCR at least 0.80 and significant correct-over-shuffled benefit, opening 16K. At 16K both retain NCR at least 0.75 and source specificity, with a stage gap below 0.15 and a length-interaction lower endpoint not positive. These are experiment-specific gates; they are not E002's near-native criteria. The recorded E011 verdict is `PORTABILITY_SURVIVES_MATCHED_TEST`. The conditional R/C component ablation did not unlock and was not run.

Condition	NLL	ΔNLL	NCR	TQR	JS	Top-1
Base 4K ($n = 64$)						
Native 9B	1.4681	0.0000	1.0000	1.0000	0.0000	1.0000
Source 4B	1.5390	0.0709	0.9562	0.0000	0.0320	0.8201
Empty 9B	3.0868	1.6187	0.0000	-21.8198	0.2278	0.5378
KV-only	2.8389	1.3708	0.1532	-18.3241	0.2107	0.5540
Full TDD	1.6700	0.2019	0.8753	-1.8466	0.0488	0.8069
Shuffled	3.3215	1.8534	-0.1449	-25.1274	0.2545	0.4976
Base 16K ($n = 32$)						
Native 9B	1.3120	0.0000	1.0000	1.0000	0.0000	1.0000
Source 4B	1.3648	0.0528	0.9662	0.0000	0.0300	0.8613
Empty 9B	2.8736	1.5616	0.0000	-28.5695	0.2242	0.5757
KV-only	2.6830	1.3710	0.1220	-24.9610	0.2121	0.5903
Full TDD	1.4870	0.1750	0.8879	-2.3140	0.0459	0.8496
Shuffled	3.1232	1.8111	-0.1598	-33.2947	0.2480	0.5405
Post 4K ($n = 64$)						
Native 9B	1.5478	0.0000	1.0000	1.0000	0.0000	1.0000
Source 4B	1.7223	0.1746	0.8905	0.0000	0.0477	0.8108
Empty 9B	3.1418	1.5940	0.0000	-8.1317	0.2326	0.5425
KV-only	2.9156	1.3678	0.1419	-6.8356	0.2175	0.5583
Full TDD	1.8133	0.2655	0.8334	-0.5212	0.0678	0.7678
Shuffled	3.4274	1.8797	-0.1792	-9.7680	0.2618	0.4983
Post 16K ($n = 32$)						
Native 9B	1.4362	0.0000	1.0000	1.0000	0.0000	1.0000
Source 4B	1.7512	0.3150	0.7886	0.0000	0.0632	0.8115
Empty 9B	2.9259	1.4897	0.0000	-3.7295	0.2358	0.5610
KV-only	2.7694	1.3332	0.1050	-3.2327	0.2260	0.5684
Full TDD	1.6685	0.2323	0.8441	0.2626	0.0759	0.7944
Shuffled	3.2388	1.8026	-0.2101	-4.7230	0.2648	0.5220

Table 12: Complete E011 canonical condition means. Each 4K stage has 64 documents and each 16K stage 32, with 64 targets per document. NLL and excess NLL use nats/token. Top-1 is agreement with native 9B. No learned correction is used.

Pair / length	Contrast	Estimate	95% CI
Base 4K	KV − full	1.1688	[1.0636, 1.2822]
Base 4K	Full − shuffled	-1.6514	[-1.8072, -1.5036]
Base 16K	KV − full	1.1960	[1.0187, 1.3816]
Base 16K	Full − shuffled	-1.6361	[-1.8768, -1.4044]
Post 4K	KV − full	1.1022	[0.9930, 1.2165]
Post 4K	Full − shuffled	-1.6141	[-1.7692, -1.4642]
Post 16K	KV − full	1.1009	[0.9243, 1.2878]
Post 16K	Full − shuffled	-1.5703	[-1.8037, -1.3471]

Table 13: E011 canonical paired NLL contrasts and 95% document-bootstrap intervals (10,000 draws, seed 11011003). Positive KV-minus-full means full-state benefit; negative full-minus-shuffled means correct-donor benefit.

Domain	Base 4K	Base 16K	Post 4K	Post 16K
Literature	0.139 / 0.853	0.129 / 0.853	0.131 / 0.833	0.130 / 0.824
STEM	0.190 / 0.875	0.167 / 0.905	0.169 / 0.834	0.139 / 0.844
Code	0.113 / 0.874	0.071 / 0.888	0.104 / 0.807	0.053 / 0.776
Conversational	0.183 / 0.910	0.094 / 0.894	0.186 / 0.888	0.081 / 1.117

Table 14: E011 descriptive domain NCR: KV-only/full state. Every displayed full-state mean exceeds its paired KV-only mean. Values are ratios of domain-level aggregate losses.

J.1 Descriptive domain disaggregation

Table 14 recomputes domain means from the same frozen per-document records. These are descriptive paper-level summaries, not new independent tests or newly fitted outcomes. NCR denominators are recomputed from aggregate empty/native loss within each domain. Each 4K domain has 16 documents, each 16K domain eight; no token-level standard errors or new confidence intervals are introduced.

K Compact Trajectory-Boundary Evidence

E013 and E014 reuse the frozen E011 post-trained TDD mapper/runtime without refitting, correction, or historical replay. They test conditional portability rather than a universal obstruction.

E013: controlled seven-token boundary. Forty-eight fresh 16K contexts (12 per domain) generate all eight cells of a paired $F \times P \times T$ factorial. F inserts five high-entropy facts by equal-length replacement; P replaces the last seven historical tokens with a fixed boundary; T replaces seven unscored post-handoff natural tokens with a fixed transition. All cells score the same 64 natural target tokens after the relevant schedule. The P effect quoted in the main text averages level-one-minus-level-zero NLL over F, T. Transfer-specific damage subtracts the corresponding source-4B effect. It is 0.212072 (CI [0.145607,0.285140]), whereas the handoff effect itself is 0.232468. This controls generic source sensitivity, not every possible change in task difficulty. Full state remains significantly better than shuffled in all eight cells. Intervals use 10,000 complete-context bootstrap draws, seed 13013401. Post-verdict sham comparisons did not change the verdict.

E014: onset and evolution. Pool A has 48 fresh contexts (12 per domain) and Pool B has 32 disjoint contexts; both use 16K histories. The primary endpoint is

$$\mathrm{TSD}(a) = [\ell_{\mathrm{handoff}}(a) - \ell_{\mathrm{handoff}}(\mathrm{clean})] - [\ell_{\mathrm{source}}(a) - \ell_{\mathrm{source}}(\mathrm{clean})]. \tag{11}$$

Thus positive TSD means a trajectory change harms transfer more than continued source prediction. The generic discontinuity sweep uses tail lengths $0, 1, 2, 4, 7, 8, 16, 32, 64$, with one common natural bridge and 64 targets. The fresh seven-token sham anchor gives TSD 0.4498 (CI [0.2834,0.6528]); the one-token result is 0.1494 (CI [0.0425,0.2869]), and the registered shape is gradual accumulation. Canonical intervals use 10,000 context-bootstrap draws, seed 14014401.

Cross-domain replacement is the strongest categorical tail arm. The descriptive standardized source-surprisal slope is -0.0260 (CI $[-0.0824$,0.0215]); this fails the registered positive-association criterion, not a proof that surprisal is irrelevant. Natural-source recovery does not reach the 80% recovery criterion within 128 tokens. Separately, target-native teacher-forced evolution decreases the registered fault gap from 0.6259 over inputs 1–8 to 0.1069 over 65–128. This differs from TSD: it subtracts the clean handoff-minus-native gap within each window and is not free-generation recovery.

E014's canonical attempt is attempt 003. Two prior attempts were invalidated for filesystem/runtime cleanup defects before canonical aggregate outcomes, with no scientific design change; the sealed history records their scope. Independent validation verifies the final manifest and recomputes the verdict. Component localization and exhaustive intervention curves are outside this paper's core argument and do not alter E001's package attribution.

L Extended Interface Comparisons

L.1 Cross-model KV/cache translation

Cache-to-Cache projects and fuses source KV with the receiver's own context cache (Fu et al., 2026). Latent Cache Flow adds joint K/V bottlenecks and pooled communication across differing contexts, retaining receiver-cache fusion (Rossi et al., 2026). Semantic Cache Distillation reconstructs KV from compact codes and sparse normalized hidden-input patches for shared-architecture, weight-mismatched models (Ma et al., 2026). It transfers more than raw KV but does not evaluate persistent GDN state. Mixture-of-Translators maps heterogeneous KV and uses target context replay to reconstruct cache, with source-guided sparsification (Lee et al., 2026).

XKV pools both models' KV caches into joint cross-layer memory and lets each receiver position retrieve a gated KV residual (Liu et al., 2026a). It supports heterogeneous models holding complementary private contexts. Both models prefill their own contexts; the communication updates attention KV, rather than transferring built-in persistent recurrent state.

Other work installs translated KV without full receiver prefix processing. Heo et al. (2026) study within-family transfer in dense full-attention models using ridge maps, layer selection, and RoPE factoring; attention–recurrent hybrids are outside their evaluation. CacheBridge uses head-local

support and attention-weighted calibration (Qu et al., 2026). A Universal Context-Reuse Layer reports KV sharing within and across families (Li et al., 2026b). KV transfer itself is established prior work. Our additional question concerns persistent recurrent and convolution state in a live hybrid receiver.

L.2 Hidden-state and latent communication

StateBridge aligns message hidden states to an embedding interface and supplies a continuous prefix processed by the receiver (Peng et al., 2026); its reported multi-agent runs share weights within each run. A continuous message can communicate useful information without installing state in the receiver's built-in persistent inference slots. We distinguish these interfaces without comparing unlike tasks or replay budgets.

L.3 External memory and reader adaptation

Li et al. (2026a) transfer learned Engram-style external memory through a tokenizer-agnostic addressing interface. They study both direct reuse of compatible memory/reader artifacts and target-side reader adaptation with frozen memory and backbones. Their result makes reader compatibility directly relevant to our question. The transferred object, however, is a learned external memory with injected reader outputs, rather than a hybrid model's built-in inference cache captured after a particular input prefix. Our handoff installs that input-conditioned persistent state without adapting target backbone weights.

L.4 Cross-model activation-state transfer

Piepereit (2026) project intermediate activations across model architectures and test their effect through activation injection. Alignment scores do not consistently predict successful behavioral transfer, and the observed effects depend on the model pair. Their generation intervention replaces a prompt-position hidden state while reprocessing token sequences; it does not install persistent hybrid cache state. This is adjacent evidence about functional compatibility of internal representations, rather than evidence that all internal state is portable.

L.5 Same-model recurrent/hybrid reuse

Marconi manages hybrid prefix checkpoints for serving (Pan et al., 2025). HYPIC composes recurrent state from segment operators and end-states, with seam recomputation for hybrid attention (Liu et al., 2026b). DASC selects and compresses persistent state, using zero filling or bounded refresh on restoration (Yu et al., 2026). Tail-Replay reconstructs recurrent state from a recent suffix of a matched prefix instead of retaining recurrent checkpoints (Liu et al., 2026c). These systems reuse, compose, or recover memory for the same model. Here, a different model consumes the state and performs no historical prefix replay.

WriteSAE decomposes recurrent cache state into matrix atoms and tests replacement or amplification of writes in Gated DeltaNet and related architectures (Young, 2026b). These behavioral interventions establish recurrent cache writes as a causal intervention surface. Its larger-host probe also reuses a learned dictionary without retraining. That transferred dictionary edits the host's state; it is not a prefix-conditioned recurrent state written by another model and installed for receiver continuation.

Recurrent-state initialization is also an adaptation surface: S0 tuning learns initial matrices before processing a prompt, rather than transferring another model's prefix-conditioned state (Young, 2026a). Metis adds learned memory blocks and evaluates memory-only inference on multiple backbones (Zhang et al., 2026). Its backbone-transfer study trains the memory mechanism for each backbone; it does not demonstrate a live state written by one model being consumed by another.

Designed cache compatibility and feature transfer. ICaRus freezes the logical cache-producing encoder while specializing the decoder, making attention KV identical by construction (Woo et al., 2026). DroidSpeak reuses attention cache across same-architecture fine-tuned variants while recomputing sensitive layers (Liu et al., 2026d). Neither demonstrates source-written GDN recurrent state

crossing a model-size boundary. Model stitching transfers probes, steering vectors, and SAE weights between residual-stream interfaces (Chen et al., 2025); shared latent KV adapters communicate and transfer soft prompts across Gemma-2 models (Dery et al., 2026). Those results motivate compatible interfaces without supplying the persistent recurrent handoff measured here. The Handoff Tax studies transferred textual agent trajectories and repository state (Ganz et al., 2026), a different interface from live neural cache installation.