Compiled Style Blocks versus a Trained Text Reader: A Negative Result and a Prompt-Card Edit Study on Synthetic Users

Emil Shirokikh
Independent researcher

Abstract

Can a table of compiled low-rank style blocks replace a trained model that reads the same personal state as text? We report two pre-specified studies on synthetic users with a frozen Qwen2.5-1.5B-Instruct base model. Blocks were compared with a text reader of equal installed rank, trained on the same sampled users for the same number of steps at three data-order seeds. Strict style accuracy was 0.780 versus 0.967, a paired difference of -0.1875 (95% bootstrap interval over 16 evaluation identities $[-0.312, -0.083]$), which the pre-specified rule labels LOSE despite 12.4% fewer prompt tokens. A post hoc diagnosis places 90% of this gap on six identities that share a style pair never seen in training; on the other ten the difference is -0.029 $[-0.061, +0.004]$. In a second study, blocks showed lower off-scope drift after a fact edit than per-user LoRA retraining (-0.00428 $[-0.00590, -0.00281]$). Their after-edit accuracy was 0.078 lower, inside the pre-specified 0.10 point margin, although the interval extends to -0.151 and one of three seeds misses the margin. Fact edits changed a prompt card and left compiled weights untouched, so this second comparison concerns prompt editing and says nothing about editing knowledge stored in parameters.

1 Introduction

Personalisation systems must decide where to keep a person's changing state. A model can read a textual card of facts and preferences, or part of that state can be compiled into parameters. Compilation offers an appealing interface: edit one assertion, replace its associated component, and avoid presenting its text. The representation is useful only if the resulting model follows it reliably, and fewer prompt tokens or a smaller parameter change are different outcomes from better answers.

We examine a deliberately small version of this problem. A synthetic identity has six categorical style attributes and a set of structured facts. A compiler assembles a population update and one low-rank block per style attribute. Facts stay in a textual card. The main comparator is a trained population LoRA that reads facts and style as text. Study 1 asks whether compiled blocks improve strict style compliance under the same synthetic-user draws, training steps and seeds. Under its pre-specified rule the answer is negative, and a post hoc diagnosis shows that the loss is concentrated in an unseen style pair. Study 2 asks whether changing one fact causes less off-scope movement than retraining a per-user adapter. It returns the pre-specified favourable label, with qualifications on the efficacy condition, on the instrument, and on what a fact edit does in this design.

The contribution is an empirical comparison and an account of its limits. The two studies share evaluation identities and compiled checkpoints, so they do not replicate each other.

2 Related Work

LoRA represents weight adaptation with low-rank factors over a frozen base model (Hu et al., 2022), and rank-stabilised scaling replaces the usual α/r factor with $\alpha/\sqrt{r}$ (Kalajdzievski, 2023). Both our compiled system and our text reader use low-rank updates; the experimental question is how personal state selects or conditions them. Text-to-LoRA generates adapters from a task description with a hypernetwork (Charakorn et al., 2025), a related idea that its authors do not frame as compilation. Our assertion table is far narrower: it addresses a fixed synthetic vocabulary and cannot speak to compiling arbitrary documents.

Model editing treats locality as a primary concern. MEND trains small networks to transform gradients, with locality to unrelated inputs in the

training objective (Mitchell et al., 2022); ROME and MEMIT write factual associations directly into weights (Meng et al., 2022, 2023); and RippleEdits argues that testing the edited fact and unrelated facts misses an edit's logical consequences, finding that prominent methods fail to make consistent changes and that a simple in-context baseline scores best (Cohen et al., 2024). In our primary edit study facts sit in a prompt card, so the compiled system changes its input and its parameters stay fixed. We adopt the locality question from this literature without claiming equivalence to those methods. Our strict style score resembles the verifiable instructions of IFEval (Zhou et al., 2023), applied to a single synthetic renderer. Aspects of the experimental setup and evaluation measure affect the choice of significance test (Dror et al., 2018); ours has 16 identities with repeated cases, so we resample whole identities in pairs. Both main arms use the base model described by Qwen Team (2024).

3 Methods

3.1 Synthetic users and scoring

An *identity* is a synthetic user with a fixed combination of six style values: length (brief, medium, detailed), formality (formal, casual), greeting (none, hello), sign-off (none, best), emoji (none, smile) and structure (prose, bullets), which gives 96 possible combinations. Each identity has three successive fact *states*, with 16 fact lines in contact, commitment and scheduling families across two scopes. All names, addresses and identifiers come from a rule-based generator and use placeholder values. Table 1 shows the form of a case.

A reply passes the strict style rule only if all six attribute checks pass: its word count lies in the bound for its length value (1–60, 61–95 or 96–130); formal replies use *Please* and no contractions, while casual replies use a casual marker; a greeting reply starts with *Hello,* on its own line; a sign-off reply ends exactly with *Best,* then *Alex*; a smile reply contains exactly one smile emoji; and a bullet reply has at least two hyphen lines. The factual rule requires the expected fragments, none of the forbidden fragments, no mention of the other scope, no unlisted fact values and the correct accept-or-decline decision. Scoring is automatic.

Training uses 64 identities, each with three states and 120 rows (23,040 rows). Evaluation uses 16 other identities, each one attribute away from a training identity, with three states and 24 cases per state: 1,152 greedy generations per condition and seed under a 256-token cap. Strict style accuracy is computed on the 384 first-state cases and factual accuracy on all 1,152. Novel-attribute accuracy covers the attribute on which an evaluation identity differs from its nearest training identity. Generic harm is the base model's accuracy minus the arm's on 16 probes unrelated to personal state. Sealed partitions of the corpus were not opened in these studies.

Style	brief, casual, hello, best, smile, bullets
Fact	messaging contact: person=Mara Ives, email=p003-messaging-0@example.test
Task	Which person and address should handle case-valid-000 in the messaging scope?
Target	Hello, / - Sure, route case-valid-000 to Mara Ives for messaging. / - The email is p003-messaging-0@example.test. [neutral filler] [smile] / Best, / Alex
Probes	correct; correct facts in the wrong style; wrong-scope facts; stale or wrong fact

Table 1: Illustrative synthetic case produced from the renderer's source, with line breaks shown as slashes. It is not drawn from the evaluation bank.

3.2 Arms

All arms use Qwen2.5-1.5B-Instruct (revision 989aa79, BF16, 28 layers) under MLX 0.32.2, mlx-lm 0.31.3 and transformers 5.17.0 on macOS arm64. Adapters are float32 updates on the query and value projections with scale 5.657. The compiled arm installs a rank-8 population block plus one rank-4 block per style attribute, an installed rank of 32, with a zero-mean gauge taken over training-identity frequencies. A *fact-only card* supplies current facts at inference. The *text reader* is one rank-32 population LoRA with no per-user parameters, and it reads a full card with style lines. Matching therefore holds for installed rank, steps, seeds and sampled rows. It does not hold for trainable parameters (5,763,072 versus 4,358,144, a 32% difference), schedule or inputs. Reference conditions are blocks with the full card, the population block alone with the fact-only card, and the base model with a full, fact-only or no card.

3.3 Training

Both arms take 3,000 steps of completion-only token-weighted cross-entropy at four rows per step (12,000 row draws), with AdamW at learning rate 10^{-4}, no weight decay, gradient clipping at one

and a 100-step warmup. Rows are limited to 512 tokens and overflow aborts the run. The final step decides the comparison. The three seeds (17, 29, 43) vary data order, synthetic-user draws and compiler card sampling. Adapter initialisation is hash-derived and identical across seeds, so we call them *data-order seeds*, and their spread understates restart variability.

The compiled arm trains in two stages. The population block trains for 1,000 steps. It is then frozen, and the style blocks train for 2,000 steps with a new AdamW state and a restarted warmup. The text reader is not staged. During the second stage, one fifth of prompts contain the system instruction alone. The remainder keep each true fact line with probability 0.9 and each true style line with probability 0.5, and then shuffle the kept lines. Completion tokens that depend on a missing fact are removed from the loss, while style and end-of-turn tokens stay supervised. This augmentation was added to address earlier termination failures. With all six keys present, a second-stage row carries all facts and no style lines with probability $0.8 \times 0.9^{16} \times 0.5^{6} = 0.0023$ before shuffling, so blocks almost never train on their deployment card, whereas the reader always does. A completed seed-17 table without the augmentation is reported as an earlier reference.

Training styles are drawn by rejection from the 64 training-identity combinations minus those containing one excluded pair, greeting *hello* with sign-off *best*. Each attribute is absent with nominal probability 0.15, and values are proposed uniformly. Rejection discards about 46% of draws (mean tries 1.84–1.87), which skews the realised stream (Table 2). Sign-off *best* appears in about 21% of rows where sign-off is present, against a nominal 50%.

Value (present rows)	Nominal	Realised by seed
sign-off best	.500	.212, .204, .206
greeting hello	.500	.447, .449, .445
emoji smile	.500	.463, .460, .456
formality formal	.500	.550, .552, .546
structure bullets	.500	.549, .553, .559
length brief	.333	.337, .340, .346
sign-off absent	.150	.215, .214, .216

Table 2: Nominal and realised sampling in the compiled arm's training stream. The second stage averaged 14.4 fact lines and 2.5 style lines per card, and 20% of its 8,000 rows had no card. The reader's stream uses the same generator, but no matching report was saved for it.

3.4 Pre-specification and decision rules

"Pre-specified" here means written into time-stamped protocol files in a private version-controlled repository; nothing was publicly pre-registered. The repository has no remote and no third-party timestamp. Its record gives this order on 24 September: the design amendment that fixed the arms and seeds (21:27), the prompt augmentation (21:35), a code-review correction to it (22:13), and the first training specification (22:28). Evaluation-harness decisions followed at 22:49, after training had begun and before any evaluation. Study 2's protocol was committed the next morning, after descriptive Study 1 outputs had been seen and before Study 1's analysis. An amendment made before Study 2's training changed the style-edit rule so that a value is skipped only if it would introduce the excluded pair, because the original wording stopped at 13 of 16 edits and affected three identities that already held the pair.

The design amendment replaced an earlier four-arm plan and dropped the typed-pool, class-conditioned and teacher comparisons, so their gates could not be evaluated. An earlier amendment had specified a policy-level additivity gate on ten style pairs, which was not run; a separate target-level check on 768 rows passed, but it measures something different. A follow-up comparison was specified and then halted before producing results.

The quality decision is WIN if the blocks-minus-reader difference is at least $+0.10$ and its 95% lower bound exceeds zero; TIE if the 90% interval lies inside ± 0.05; LOSE if the point difference is at most -0.05; otherwise INCONCLUSIVE. In the protocol's decision table a quality LOSE means the text reader is preferred whatever the cost. The stop on later size and body-swap steps comes from the research plan. Generic-task harm has a separate maximum accuracy drop of 0.02. Smartphone warm-prefill latency was the intended cost measure, and prompt-token counts alone support only a provisional cost verdict.

Seeds are averaged within each case. A paired bootstrap then resamples whole identities (10,000 draws, seed 20260924, type-7 percentile intervals). These intervals describe variation across the 16 identities given the trained models and do not cover future seeds. Intervals other than the primary are unadjusted for multiplicity. Six identities

allow only 462 distinct resamples, so intervals on that slice are descriptive and probably too narrow.

3.5 Edit study

Study 2 reuses the unchanged checkpoints and the 16 identities. Two single-fact edits per identity give 32 edits, with 64 on-scope probes (all three families), 64 off-scope probes (contact and commitment only) and the 16 generic probes. Each probe scores four candidates (Table 1) by mean assistant-tail log-probability, normalised with a softmax. Off-scope drift is the symmetric KL divergence between before-edit and after-edit distributions, and efficacy is after-edit argmax accuracy on on-scope probes, with desired-probability gain beside it. This softmax over mean log-probabilities is strongly compressed: the mean top off-scope value is 0.31 for blocks and 0.34 for retraining, with a maximum of 0.49 (uniform is 0.25). The drift scale therefore has no external calibration.

The primary comparator retrains a per-user LoRA for every identity and state (rank 8, alpha 16, rsLoRA, query and value modules, dropout 0.05, 160 steps, gradient accumulation 4). It selects among steps 40, 80, 120 and 160 on that state's validation partition, which contains the probes, and this favours its efficacy. All 112 direct jobs, including controls, reported finite losses, validation improvement and successful reloads. Controls are retraining with a shared seed and a *noise* control that retrains on unchanged data with a different seed (first edits only). The primary difference is blocks minus direct retraining. MORE LOCAL requires a 95% upper bound below zero and LESS LOCAL a lower bound above zero. A MORE LOCAL result carries the protocol's label "local and effective" only if blocks' after-edit accuracy lies within 0.10 of direct retraining as a point estimate, which is not an interval-based non-inferiority test. The bootstrap resamples identities with seed 20260925.

A secondary analysis makes one style edit per identity and scores all 24 first-state rows. *Fact-choice drift* is the symmetric KL over the three non-style candidates, renormalised. *Style shift* is the change, caused by the edit, in the log-probability difference between the new-style and old-style correct replies.

Condition	Style	Facts	Tokens
Blocks	.780 [.637,.898]	.833 [.725,.929]	255.4
Text reader	.967 [.914,1.00]	.986 [.972,.996]	291.4
Blocks + full card	.738 [.589,.867]	.785 [.684,.880]	291.4
Population block	.000 [.000,.000]	1.00 [1.00,1.00]	255.4
Base + full card	.000 [.000,.000]	.140 [.089,.196]	291.4
Base + fact card	.005 [.000,.016]	.464 [.430,.499]	255.4
Base, no card	.010 [.000,.031]	.000 [.000,.000]	75.5
Earlier blocks, s17	.682 [.456,.875]	.924 [.859,.977]	255.4

Table 3: Study 1 accuracy with 95% identity-bootstrap intervals, and mean prompt tokens. Trained conditions average seeds 17, 29 and 43 unless marked (s17 is seed 17). Style values by seed are .591, .896, .852 for blocks and .979, .979, .943 for the reader; fact values are .772, .873, .853 and 1.00, 1.00, .957.

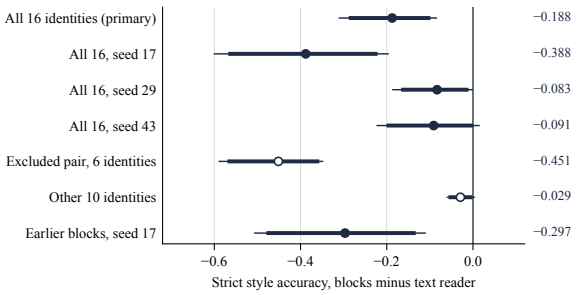


Figure 1: Strict style accuracy differences against the text reader. Thin bars are 95% and thick bars 90% identity-bootstrap intervals. Filled points are pre-specified comparisons and hollow points are post hoc slices. Intervals omit training-seed uncertainty.

4 Results

4.1 Study 1: the compiled blocks lose

The strict style difference between blocks and the reader is -0.1875, with 95% interval $[-0.312, -0.083]$ and 90% interval $[-0.288, -0.099]$, so the pre-specified verdict is LOSE (Table 3, Figure 1). Factual accuracy also favours the reader: -0.153 $[-0.258, -0.059]$. A t interval on the 16 identity differences, $[-0.314, -0.061]$, and an exact sign-flip test ($p = 0.0024$) agree with the bootstrap. Both are post hoc sensitivity checks.

All three seed-specific differences favour the reader: -0.388, -0.083 and -0.091. Their direction is consistent, but their magnitude is not established. A $t(2)$ interval on the three values is $[-0.62, +0.24]$, three seeds cannot reach a two-sided sign-test p below 0.25, and without seed 17 the mean is -0.087. Only seed 17 has a 95% identity interval wholly below zero; the other upper bounds are zero and $+0.016$.

Blocks beat every untrained baseline on style

by a wide margin, and they clear the pre-specified check that facts do not collapse against the base model with a fact card (difference $+0.368$). Compiled blocks also lower factual accuracy relative to the population block alone, which is perfect at every seed: -0.167 $[-0.275, -0.071]$, a descriptive contrast. Adding style lines to the blocks' prompt lowers both outcomes on average (style -0.042 $[-0.102, -0.002]$), with seed signs $+0.122$, -0.133 and -0.115. Per-attribute rates for blocks and the reader are length .933 and .988, sign-off .814 and .977, emoji .964 and 1.00, structure .990 and 1.00, with formality and greeting at 1.00 for both.

Blocks clear the available absolute thresholds on the seed mean (style at least 0.70, novel-attribute accuracy .945 against 0.80), but seed 17 misses the style threshold at .591. They fail the generic-harm gate, with a drop of 0.0208 against a 0.02 maximum, produced by one regression among 16 probes at seed 29. The minimum non-zero drop over 48 probe-seeds is $1/48$, so this gate in effect required zero regressions. The reader's seed-mean drop is larger, at 0.0625 (three regressions at seed 43), and it would also exceed 0.02, although the gate applied only to blocks.

The 12.4% prompt-token saving follows entirely from omitting style lines: blocks match the base model with a fact card (255.375) and the reader matches the base model with a full card (291.4375). It supports only a provisional token-cost win. Both arms' prompt lengths fall outside the measured smartphone calibration range, so warm-prefill cost could not be evaluated, and extrapolated latencies are not treated as measurements.

Where the loss sits (post hoc). A diagnosis written after the verdict splits the evaluation identities by whether they hold the excluded pair (Table 4). The six excluded-pair identities carry -0.169 of the -0.1875 gap (195 of 216 net passes) and 99% of the fact gap. On the other ten the style difference is -0.029, with 95% interval $[-0.061, +0.004]$, 90% interval $[-0.057, -0.001]$ and seed values -0.083, 0.000 and -0.004. Applied descriptively to this unregistered slice, the decision rule gives INCONCLUSIVE, and equivalence cannot be claimed because the 90% interval leaves the ± 0.05 band. On the excluded-pair identities, blocks score .049, .722 and .674 by seed, against .944, .944 and .910 for the reader. Sign-off is the failing attribute in 214 of 254 failing case-seeds. The bank holds 6 of 16 excluded-pair identities (37.5%), against 25% of the 96-combination grid.

Measure	Identities	Blocks	Reader
Strict style	excluded pair (6)	208/432	403/432
	other (10)	690/720	711/720
Facts	excluded pair (6)	747/1296	1272/1296
	other (10)	2131/2160	2135/2160
Cap stops	excluded pair (6)	133/1296	2/1296
	other (10)	0/2160	0/2160
Looping tail	excluded pair (6)	125/1296	72/1296
	other (10)	15/2160	0/2160

Table 4: Post hoc split of Study 1, pooled over seeds, as counts of passing case-seeds. These slices are exploratory and carry no verdict.

Prompt	Measure	Blocks	Reader
Deployment	cap stop	.038 [.002, .091]	.0006 [.000, .002]
	loop	.041 [.003, .106]	.021 [.000, .062]
+ persona	cap stop	.240 [.105, .410]	.000 [.000, .000]
	loop	.095 [.034, .169]	.021 [.000, .062]
System only	cap stop	.124 [.016, .271]	.000 [.000, .000]
	loop	.100 [.021, .204]	.000 [.000, .000]

Table 5: Termination rates with 95% identity-bootstrap intervals, averaged over seeds. Blocks' deployment cap-stop rate by seed is .098, .007 and .010.

Effect of the augmentation. At seed 17, the only seed with both tables, the augmentation lowered style and facts: earlier blocks minus augmented blocks is $+0.091$ $[+0.010, +0.177]$ on style and $+0.152$ $[+0.040, +0.280]$ on facts. The earlier table also lost to the reader (-0.297 $[-0.508, -0.109]$); its effect at other seeds is unknown. Cross-entropy on eight fixed monitoring rows fell from 2.43 to between 1.0×10^{-4} and 2.4×10^{-4} for blocks and from 2.34 to between 3.8×10^{-6} and 3.8×10^{-5} for the reader, so both arms fit their training streams.

Termination. Table 5 reports the pre-specified termination diagnostic. The augmentation's purpose was termination with an unfamiliar system prompt. At seed 17 its cap-stop rates moved from .118 to .098 on the deployment prompt, from .195 to .221 with persona text and from .133 to .151 with the system prompt alone, so the only available comparison shows little improvement. Blocks with a full card exceed the protocol's 0.10 flag (.111). All 133 cap stops of blocks on the deployment prompt are on excluded-pair identities.

4.2 Study 2: lower fact-edit drift with a narrower reading

Blocks' off-scope drift is 0.00457, against 0.00885 for direct retraining. The paired difference is -0.00428 $[-0.00590, -0.00281]$, so the protocol's verdict is MORE LOCAL, and each seed-specific interval lies below zero. Blocks drift less on all 16 identities and on 59 of 64 probes (seed means). After-edit accuracy is .922 for blocks and 1.00 for direct retraining; the -0.078 gap is inside the 0.10 point margin, so the protocol's label is "local and effective". The interval on that gap is $[-0.151, -0.021]$. By seed, accuracy is .875, .922 and .969, and seed 17's gap of -0.125 $[-0.25, -0.031]$ misses the margin. All 15 on-scope errors of blocks are on excluded-pair identities (57 of 72 correct, against 120 of 120 elsewhere).

Every compiled fact edit leaves the installed weights bitwise unchanged, across all 32 edits at each seed, so the behavioural difference arises from reading a changed card under fixed compiled style. Mean drift hides heavy tails. Median drift is 2×10^{-5} for blocks and 3×10^{-5} for the reader against 0.0039 for direct retraining, and for blocks the 15 largest of 192 probe-seed values hold 60% of the total. Hard answer changes run the other way. Off-scope argmax flips number 5 of 192 for blocks (4 from correct to wrong, all on excluded-pair identities), 0 of 192 for the reader, 0 of 64 for direct retraining and 2 of 64 for the base model with a full card. Pre-edit off-scope accuracy is .964 for blocks, .766 for that base condition and 1.00 for the reader and for direct retraining. Desired-probability gain is 0.057 for blocks, 0.082 for the reader, 0.072 for direct retraining and 0.034 for the base condition.

The reader's drift is 0.00602 with perfect on-scope accuracy. The base model with a full card has lower drift, 0.00255, and accuracy .703, and its lower gain indicates a weaker response to the edit. Its on-scope errors are mostly style failures (16 wrong-style, 3 wrong-scope and no stale-fact winners of 19), whereas blocks' errors are 8 wrong-scope, 4 stale-fact and 3 wrong-style, so four-way accuracy mixes fact uptake with other properties. Blocks' drift stays below same-seed retraining (0.00784, accuracy 1.00).

Compared like for like on first edits, compiled drift is not clearly below pure retraining noise (Figure 2c). Means are 0.00641 for blocks, 0.00737 for the reader, 0.00348 for the base condition, 0.01011 for direct retraining and 0.00395 for the noise control. The differences from noise are $+0.00245$ $[-0.00044, +0.00596]$, $+0.00342$ $[+0.00049, +0.00689]$, -0.00047 $[-0.00196, +0.00133]$ and $+0.00616$ $[+0.00317, +0.00943]$. Transition KL on generic probes is 0.0031 for blocks, 0.0071 for the reader, 0.0001 for the base condition, 0.0814 for direct retraining and 0.0590 for the noise control, with very wide intervals on the last two (direct retraining $[0.0041, 0.2276]$).

Style edits do not reproduce the favourable fact-edit pattern. Fact-choice drift is 0.00423 $[0.00175, 0.00749]$ for blocks, 0.00008 for the reader, 0.00294 $[0.00214, 0.00387]$ for retraining on restyled data and 0.00001 for the base condition. Blocks minus the reader is $+0.00416$ $[+0.00167, +0.00741]$, and blocks minus retraining is inconclusive at $+0.00129$ $[-0.00112, +0.00430]$. Style shift, the efficacy side, is 0.319 for blocks, 0.429 for the reader, 1.541 for retraining and 0.016 for the base condition; blocks minus the reader is -0.111 $[-0.184, -0.028]$. So blocks move fact choices more than the reader and shift style less. Style edits change weights by 0.066–0.068 of the installed norm in blocks and 0.059 in retraining, and rank stays at or below four per module in blocks, a bound that did not guarantee smaller behavioural disruption than a textual edit.

5 Discussion

Study 1 rejects a quality advantage for this compiler, recipe and synthetic domain, and its verdict stands as specified. The diagnosis changes how that verdict should be read. The evidence fits a failure to compose at an unseen style pair better than a general quality deficit, because the gap on the other ten identities is small and inconclusive. Three properties limit this reading. The slice is post hoc. The bank over-represents the excluded pair. And the sampler starved sign-off *best*, the attribute that fails most, so a sampling effect and a composition effect cannot be separated here. The comparison measures one package of staging, augmentation, sampling and parameter count, and the arms are matched on the quantities listed in Section 3.2 only. The augmentation lowered both outcomes at the one seed where it can be checked, so a different recipe could score differently; that would

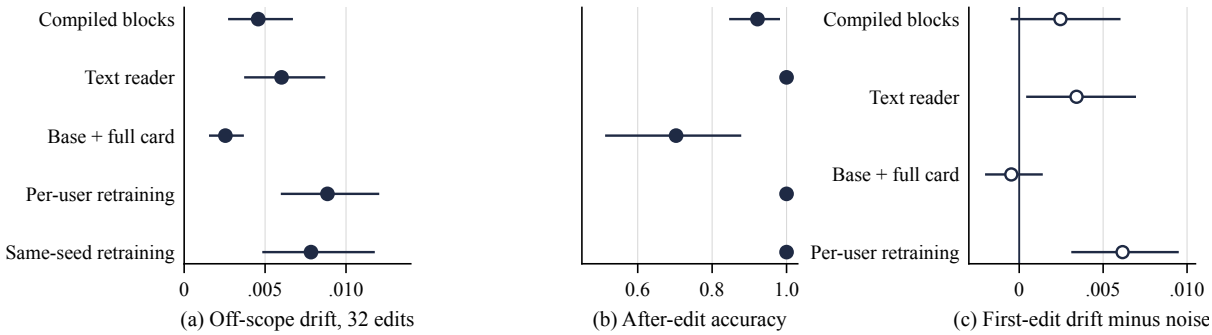


Figure 2: Study 2 fact edits with 95% identity-bootstrap intervals; seeded arms are means. (a) Off-scope drift over 32 edits. (b) On-scope after-edit accuracy. (c) Paired first-edit drift minus the retraining-noise control, a post hoc comparison except for the retraining row.

be a new study.

Study 2's headline is close to following from the design. Fact edits touch no compiled weights, so the comparison sets prompt editing against retraining, and a base model that reads the same card drifts less still. On first edits, blocks and the reader move off-scope scores about as much as, or more than, pure retraining noise, so avoiding retraining removes optimisation movement while reading a changed card introduces its own. Blocks are also the only trained arm whose off-scope answers flip, and the comparator sits at a ceiling reached with probe-partition selection. We therefore use "local and effective" only as the protocol's label. The supportable statement is narrower: on this instrument, prompt-card fact editing with fixed blocks moves soft scores less than per-user retraining, at an after-edit accuracy that is lower by an amount the point margin tolerates and the interval does not confine to that margin.

Parameter locality and behavioural locality need separate measurements, because a rank-four update can disturb fact choices and a zero parameter change can move many answers through the input. Representation cost and execution cost also need separate evidence, since omitting card lines reduces tokens without establishing latency or energy. Further improvements require a new comparison with its own declared training choices and evaluation bank, and the later edit result cannot be read as recovering the earlier quality claim.

6 Conclusion

A compiled style table lost to a text reader on strict generation quality in this synthetic task despite a shorter input, with the loss concentrated on identities holding one never-trained style pair at this bank's composition. A later study found lower soft-score drift after fact edits than per-user retraining, with after-edit accuracy inside a point margin on the seed mean and at two of three seeds, while the compiled weights did not change during those edits. For the tested generation task the evidence favours the trained reader. It leaves open whether a differently trained table would compose at unseen pairs.

Limitations

The evidence covers one 1.5B base model, one synthetic renderer and 16 evaluation identities in a 96-combination grid, each one attribute from training, and those identities are reused across both studies. The excluded pair is a structured exclusion within that grid, not a sample of natural users. Three data-order seeds share one adapter initialisation, and the identity-bootstrap intervals do not quantify seed or restart variability. The excluded-pair diagnosis, the sensitivity checks and the first-edit noise comparisons other than retraining's are post hoc. Lexical style rules and four-candidate probes do not measure natural-language preference, privacy, unrestricted factual reliability or long-term adaptation. The drift scale is compressed and uncalibrated, off-scope probes cover two of three fact families, and direct retraining's checkpoints were selected on the probe partition.

The benchmark, renderer, probes and scoring rules were built within this project and have not been independently audited. Pre-specification rests on a private repository that readers cannot inspect; Study 2's protocol followed exposure to descriptive Study 1 outputs; and amendments removed arms and left the additivity, typed-pool, class-conditioned and teacher gates unresolved.

Cost evidence is limited to prompt-token counts because smartphone calibration does not cover these lengths. No deployment-speed, cross-model or general-superiority claim follows from these results.

Ethical considerations

This article uses no human-participant data. Every user, state, fact and probe comes from a rule-based synthetic generator with placeholder identities, and scoring is automatic. The base model is distributed under Apache 2.0. Generator, protocols, per-case scores and adapters are not publicly released at the time of writing.

Funding and competing interests

No external funding. The author develops SlyOS, a commercial product related to this work.

References

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. 2025. Text-to-LoRA: Instant transformer adaption. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7485–7514.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *Preprint*, arXiv:2312.03732.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.