

Corrigendum v1.0: Paper One symmetry and preservation language

STATUS: OWNER-RATIFIED FINAL - PRE-ANCHOR

Evidence checked: 2026-07-23

Affected anchored artifact

Grokking Collapses the Algorithm, Not the Function: Irrep-Energy Underdetermination in Modular Addition

- SHA-256: 6bf750296204d0a92bad77d02adb1b5682e56da36612861fe7384801f2d1f9b0
- Sepolia: 0x90055a1d2c82c05aec6c858bf5c9387dafba69af77b30318e522d334d6923257
- Sepolia block: 0xa77a40
- Block timestamp: 2026-06-02T16:42:36Z
- Anchor date: 2026-06-02, verified directly from the Sepolia block timestamp

The original PDF remains the anchored artifact. This corrigendum supplements it; it does not replace or silently revise those bytes.

Paper One §3.3 (page 5) correctly gives schedule-dependent training budgets and distinguishes the original and extension cyclic grids. Those statements match the archived generating scripts. No training-budget correction is required for Paper One.

1. Symmetry-language correction

Affected language

The paper repeatedly calls the readout irrep-energy statistic a "gauge-invariant" aggregate, including the contribution summary (page 2), §9 title and opening (page 17), discussion (page 26), and conclusion (page 29):

$$M[k] = \sum_j |\text{FFT_output}(R)[k, j]|^2.$$

The paper explicitly establishes invariance to a joint hidden-unit permutation. The unqualified word "gauge-invariant" is broader than that established symmetry. For a ReLU hidden unit, the positive rescaling

$$fc1_j \rightarrow c_j \cdot fc1_j \text{ and } fc2_{:,j} \rightarrow c_j \cdot fc2_{:,j} / c_j, \text{ with } c_j > 0,$$

preserves the network logits in exact arithmetic but changes M.

The §9 heading itself should be replaced:

The readout subspace is indexed by a permutation-invariant irrep-energy statistic

Deterministic check

On one surviving grokked checkpoint, an exact composed positive rescaling:

- preserved all 2,209 argmax decisions;
- changed logits by at most $4.263256414560601e-14$;
- produced relative logit L2 error $1.8076382709770293e-16$; and
- changed M by relative L2 2.1309200428307618 (213%).

The result is recorded in `results/gauge_check_extension_first_grokked.json`. It is a deterministic symmetry identity check, not an outcome-bearing population analysis.

Corrected description

Replace the unqualified symmetry description with:

a 47-dimensional hidden-unit-permutation-invariant readout statistic, with 24 independent real-valued entries after Hermitian collapse, evaluated in the trained parameterization

This correction does not withdraw the reported held-out decoding values for M. It narrows the invariance claim to the symmetry actually established.

The activation-weighted candidate $D[k] = \sum_j E_x[(H_j(x) \text{ FFT_output}(R)[k,j])^2]$ passed the same positive-rescaling identity check on the fixed ReLU architecture. No population audit has yet established that D predicts dialect geometry, recovers PC1, or replaces M. It must not be presented as the corrected dialect coordinate before the registered analysis is independently reviewed, frozen, and run.

2. Preservation-language correction

The paper describes the tested dialect direction as "function-preserving," notably in the contribution summary (page 2), §9.6 (pages 22-24), discussion and limitations (pages 26-28), and conclusion (page 29).

The perturbation sweep reports held-out test accuracy and the accuracy ≥ 0.95 grokking threshold. The readout interventions themselves change real-valued logits. The experiment therefore establishes preservation of tested decisions or accuracy under its registered criterion, not equality of the full real-valued logit function.

Replace:

function-preserving direction

with one of the following, according to context:

decision-preserving over the held-out test set

accuracy-preserving under the tested ≥ 0.95 grokking criterion

preserves grokking throughout the tested perturbation range

The exact hidden-unit-permutation sanity check remains a genuine function-preserving reparameterization because it preserves logits, not merely accuracy.

3. Comparison-language clarification

The contribution summary on page 2 says M recovers held-out function geometry "at least as well as the full readout." The sentence introducing the comparison table on page 18 states:

The 47-dimensional irrep-energy vector matches the full readout and readout-PCA variants within error bars while providing an interpretable basis.

The conclusion on page 29 combines the same directional and parity claims. These statements are not supported as superiority or equivalence claims. The reported held-out Mantel gaps are 0.47 ± 0.16 for M, 0.41 ± 0.09 for the full readout, 0.41 ± 0.08 for PCA-24, and 0.53 ± 0.13 for PCA-64. PCA-64 has the largest reported mean. No prospective equivalence margin or paired equivalence test was specified.

The final-polish source ticket contains the following intended Table 3 caption:

The 47-dimensional gauge-invariant irrep-energy vector M matches readout-PCA performance while providing a principled, interpretable basis.

That caption does not visibly render in the anchored 6bf75029... PDF; the table is malformed across pages 18-19. It does visibly render in a separately retained 30-page review-form PDF, SHA-256

7f9ddebaff4106acfd567cec41e8c6348be38079330d50f40a9fdafbf425206, created by pdfTeX on 2026-05-16 and headed "Under review as submission to TMLR." That proves the wording reached a rendered review-form build, but it does not by itself establish that the build was uploaded or distributed.

A second divergent build with reported SHA-256 prefix 7051abdf was mentioned during review but is not present in the retained evidence set, so its identity, contents, and distribution status remain unverified. No equivalence is asserted among the anchored, 7f9deb..., or reported 7051abdf... builds. The caption's symmetry and parity language is withdrawn wherever it appeared and must not be carried into a corrected build.

The supported wording is:

M produced a positive held-out geometry-recovery gap in the same reported range as the tested readout baselines; the study did not test statistical or practical equivalence among these variants.

Scope

The numerical decoding, intervention, and perturbation results in the anchored PDF are unchanged. This corrigendum narrows three interpretations: the symmetry class of M , what the accuracy sweep preserved, and what overlapping reported uncertainty does not establish.

The deterministic identity check does not establish normalized PC1 recovery, a scale-free dialect, or predictive value for D .

Publication and integrity conditions

Upon publication:

- 1 this finalized corrigendum receives its own SHA-256, OpenTimestamps proof, and dated chain anchor;
- 2 the original 6bf75029... PDF remains available unchanged;
- 3 the paper's public card links both the original PDF and this corrigendum;
- 4 the public verifier checks the corrigendum's site file, OTS proof, and chain binding; and
- 5 no claim is made that a future corrected paper build is byte- or content-equivalent to the anchored original.

Review and ratification

- Version: 1.0
- Evidence checked: 2026-07-23
- Outside factual review: ACCEPT, with no required fixes, 2026-07-23
- Owner ratification: 2026-07-23

The outside reviewer assessed the complete drafts against evidence embedded in the self-contained review packet. The reviewer did not independently retrieve the anchored PDFs or generating scripts. The local pre-anchor verifier binds the source-artifact hashes used to assemble that packet.

The SHA-256 of the final rendered PDF is recorded externally in the pre-anchor manifest so the document does not attempt to contain its own hash.