Peer Review

Ashraf, Dylan. 2026. “Restoring Grokking under Structured Label Noise: Audit Efficiency and Optimization-History Dependence.” *Journal of High School Science* 10 (2): 521–45. https://doi.org/10.64336/001c.163982.

1.Although the manuscript appropriately acknowledges the limitation of using a single corruption rate ($\eta = 0.04$), several central conclusions — including the superiority of loss-based triage and the claim that “identity matters more than count” — may depend strongly on the noise regime. Grokking dynamics are often highly nonlinear, and it remains unclear whether the reported recovery behavior persists at lower or higher corruption levels. Even a modest sweep across multiple η values (e.g., 0.01–0.12) in 0.04 increment would substantially strengthen the generality and robustness of the conclusions and help determine whether the observed effects represent a broad phenomenon or a regime-specific result near the recoverability boundary.

2.Is the reported recovery effect specific to the particular loss-ranking metric used in the manuscript (cross-entropy loss after warm-up training), or does it generalize across alternative notions of “difficult” or “damaging” examples. For example, would grokking recovery still occur if examples were prioritized using uncertainty margins, gradient norms, forgetting events, influence estimates, or representation-space outliers instead of raw cross-entropy loss? This may help determine whether the key phenomenon is genuinely tied to identifying optimization-critical corruptions, or whether the current results depend strongly on the chosen scoring function. These experiments are needed to strengthen the manuscript’s central claim that the identity of corrected examples matters more than the total number corrected, particularly if different ranking metrics recover similar or distinct subsets of corruptions with different effects on grokking recovery.

3.The current paper is competent and reasonably well controlled, but the main conclusions are somewhat expected, i.e. 1. label noise suppresses grokking, 2. smarter auditing finds more corrupted labels, and 3. fixing the hardest/damaging labels helps recovery more than random fixing. Much of the paper operationalizes intuitions already suggested by: memorization-before-generalization literature, loss-based noisy-label filtering, and data-centric ML practices. As such, the paper’s scientific contribution is limited and marginally publishable as a stand alone original article. I want you to add an extension to this study as follows: test whether grokking recovery is optimization-history dependent rather than solely determined by the final corruption state. To my knowledge, this specific temporal asymmetry question has not been directly explored in the grokking literature. Clean experiments could include: (1) training on clean data until partial or full grokking emerges, then introducing label noise and continuing training; (2) training initially on noisy labels, then correcting all labels mid-training without reinitialization; and (3) comparing the latter condition against restarting from scratch after correction. If identical final datasets produce different outcomes depending on the order of noise exposure and correction, this would provide evidence for hysteresis or path dependence in grokking dynamics and would considerably deepen the mechanistic contribution of the work while requiring only modest extensions to the existing framework. An experimental framework is presented below:

Proposed temporal/path-dependence experiments:

Baseline (clean training)

Clean labels ——————————————————► Train ——————————————————► Grokking

Condition A: Noise from the start

Noisy labels ——————————————————► Train ——————————————————► No/Delayed Grokking

Condition B: Clean → noisy

Clean labels ——————► Partial/Full Grokking ——————► Inject noise ——————► Continue training

Question: Does grokking persist despite later corruption?

Condition C: Noisy → clean (continue training)
Noisy labels ——————▶ Memorization regime ——————▶ Correct labels ——————▶ Continue training
Question: Can grokking recover without restarting optimization?
Condition D: Noisy → clean (restart training)
Noisy labels ——————▶ Memorization regime ——————▶ Correct labels ——————▶ Reinitialize + retrain
Question: Is recovery dependent on removing optimization history?
Key comparison:
If Conditions C and D differ despite identical corrected datasets, this would support optimization-history dependence / hysteresis in grokking dynamics.
This will elevate the novelty of your manuscript significantly.

Response to Reviewer Comments

Manuscript: Noise Geometry Determines Audit Efficiency: Restoring Grokking Under Structured Label Noise with Targeted Annotation Review
Author: Dylan Ashraf
Date: June 2026
Every comment has been addressed in the revised manuscript. Below, each reviewer comment is quoted verbatim, followed by a description of how it was addressed and the precise location of each change in the revised manuscript. All numerical results reported below are taken directly from the completed experimental runs (noise-rate sweep: 126 runs; alternative-scorer sweep: 72 runs; temporal experiments: 30 runs).

Editor Note

Please address all the comments of the reviewer and revise your manuscript accordingly. When you resubmit, please also submit separately a word file that lists verbatim each comment by the reviewer and then describe how and where in the revised manuscript you have addressed that particular comment. Do this for all comments in one separate file. Please note that if the the concerns/comments/questions of the reviewer are addressed inadequately, incompletely or insufficient address is paid to detail, your manuscript may be rejected. Please therefore thoroughly review your responses before resubmission.
This document constitutes that separate response file. Each reviewer comment is quoted verbatim below, followed by the response and the location of the corresponding changes.

Reviewer Comment 1 — Noise Rate Sweep

"Although the manuscript appropriately acknowledges the limitation of using a single corruption rate ($\eta = 0.04$), several central conclusions — including the superiority of loss-based triage and the claim that 'identity matters more than count' — may depend strongly on the noise regime. Grokking dynamics are often highly nonlinear, and it remains unclear whether the reported recovery behavior persists at lower or higher corruption levels. Even a modest sweep across multiple η values (e.g., 0.01–0.12) in
0.04 increment would substantially strengthen the generality and robustness of the conclusions and help determine whether the observed effects represent a broad phenomenon or a regime-specific result near the recoverability boundary."

Response:
A full noise-rate sweep was conducted at $\eta \in \{0.01, 0.08, 0.12\}$, which — together with the existing $\eta = 0.04$ data as the midpoint — spans the range the reviewer suggested. Three random seeds, both noise types (input-dependent and uniform), policies (none, random, region, loss-based), and budgets $B \in \{400, 800\}$ were used, for **126 runs** in total.
The qualitative advantage of loss-based auditing over random and region auditing was confirmed at **every** noise rate tested, demonstrating that the effect is a broad phenomenon rather than a regime-specific artifact:

- At $\eta = 0.01$, loss audit was the only policy to restore grokking under both noise types (e.g., final test accuracy 0.963 at $B = 800$, input-dependent).
- At $\eta = 0.08$ and $\eta = 0.12$, no policy met the strict grokking criterion, but loss audit retained a large lead in final test accuracy (e.g., 0.931 vs. 0.511 for region and 0.390 for random at $\eta = 0.08$, $B = 800$, input-dependent).

Crucially, the sweep produced the single most direct piece of evidence for the "identity matters more than count" claim in the entire study. At $\eta = 0.01$, $B = 800$, under input-dependent noise, region auditing and loss auditing corrected almost identical fractions of the 50 corruptions (correction rate 0.733 for region vs. 0.727 for loss — region corrected marginally *more*), yet only loss auditing restored grokking (final test accuracy 0.819 for region vs. 0.963 for loss). Because the number of corrections was effectively identical, the difference in outcome can only be attributed to *which* corruptions were fixed. The sweep also localized a recoverability threshold between $\eta = 0.04$ and $\eta = 0.08$.
Where addressed in the revised manuscript:

- Section 3.7 ("Noise Rate and Scorer Generalization") — added to Methods, describing the sweep design.
- Section 4.6 ("Generalization Across Noise Rates") — added to Results, reporting outcomes at every η with the region-vs-loss "identity matters" comparison.
- Abstract — updated to note the loss-based advantage was confirmed across $\eta \in \{0.01, 0.04, 0.08, 0.12\}$.
- Introduction, Contribution 2 — updated to state the cross-η robustness.
- Section 5.5 (Limitations) — the single-η limitation was removed; the residual limitation (fixed transition epoch in the temporal experiments) is now stated instead.
- Appendix B (Reproducibility) — exact reproduction command added.

- Reviewer Comment 2 — Alternative Scoring Functions

"Is the reported recovery effect specific to the particular loss-ranking metric used in the manuscript (cross-entropy loss after warm-up training), or does it generalize across alternative notions of 'difficult' or 'damaging' examples. For example, would grokking recovery still occur if examples were prioritized using uncertainty margins, gradient norms, forgetting events, influence estimates, or representation-space outliers instead of raw cross-entropy loss? This may help determine whether the key phenomenon is genuinely tied to identifying optimization-critical corruptions, or whether the current results depend strongly on the chosen scoring function. These experiments are needed to strengthen the manuscript's central claim that the identity of corrected examples matters more than the total number corrected, particularly if different ranking metrics recover similar or distinct subsets of corruptions with different effects on grokking recovery."
Response:
Two alternative scoring functions were implemented and evaluated against the loss-based audit and a random baseline, at $\eta = 0.04$, budgets $B \in \{200, 400, 800\}$, three seeds, and both noise types (**72 runs**):
1. **Uncertainty audit** — ranks examples by the softmax entropy of the warm-up model's predictions. Entropy is bounded and invariant to logit scale, so it represents a distinct, confidence-based inductive bias.

2. **Forgetting-event audit** — ranks examples by the number of times each flips from correctly to incorrectly predicted across warm-up evaluation checkpoints (an indirect measure of learning instability).

The results directly answer the reviewer's question about whether different metrics recover similar or distinct subsets of corruptions with similar or distinct effects:

- **Loss and uncertainty auditing performed almost identically.** Their hit rates differed by less than 0.002 (e.g., 0.453 vs. 0.452 at B = 200, input-dependent), their correction rates were equal to three decimals, and their final test accuracies differed by at most ~0.01. Both restored grokking at the same budgets (each 3/3 seeds at B = 800 under input-dependent noise). This shows the effect is **not** an artifact of the raw cross-entropy scale — it holds for a qualitatively distinct confidence-based ranking.
- **Forgetting-event auditing was consistently weaker.** It identified fewer corruptions (hit rate 0.297 vs. ~0.45 at B = 200) and did **not** restore grokking at any budget under either noise type (maximum final test accuracy 0.639), though it still beat the random baseline (~0.46).This dissociation is reported honestly and, rather than weakening the central claim, sharpens it: loss and entropy — both direct measures of per-example confidence — recover nearly the same optimization-critical subset and produce the same recovery, whereas the forgetting-event metric recovers a smaller, less critical subset and fails to restore grokking. The recovery effect is therefore robust across confidence-based scorers while remaining genuinely tied to identifying the *damaging* corruptions, not merely "difficult" ones. The manuscript was revised to make this nuanced finding explicit and to avoid any overclaim that "any scorer works."

Where addressed in the revised manuscript:

- Section 3.5 ("Audit Budget Policies") — expanded to define the uncertainty audit and forgetting-event audit in full.
- Section 3.7 ("Noise Rate and Scorer Generalization") — describes the alternative-scorer sweep design.
- Section 4.7 ("Alternative Scoring Functions") — added to Results, including Table 4 (hit rate, final test accuracy, and grokking counts for all four scorers) and the loss/uncertainty-vs-forgetting dissociation.
- Abstract and Introduction, Contribution 2 — updated to state that the effect generalizes to uncertainty but that forgetting-event ranking is weaker, demonstrating the effect is tied to identifying optimization-critical corruptions.
- Section 6 (Conclusion), point 3 — updated to reflect the scorer dissociation accurately.
- Appendix B — both new scorers documented with reproduction command.

Reviewer Comment 3 — Limited Scientific Novelty / Temporal Path-Dependence Extension

"The current paper is competent and reasonably well controlled, but the main conclusions are somewhat expected, i.e. 1. label noise suppresses grokking, 2. smarter auditing finds more corrupted labels, and 3. fixing the hardest/damaging labels helps recovery more than random fixing. Much of the paper operationalizes intuitions already suggested by: memorization-before-generalization literature, loss-based noisy-label filtering, and data-centric ML practices. As such, the paper's scientific contribution is limited and marginally publishable as a stand alone original article."

"I want you to add an extension to this study as follows: test whether grokking recovery is optimization-history dependent rather than solely determined by the final corruption state. To my knowledge, this specific temporal asymmetry question has not been directly explored in the grokking literature. Clean experiments could include: (1) training on clean data until partial or full grokking emerges, then introducing label noise and continuing training; (2) training initially on noisy labels, then correcting all

labels mid-training without reinitialization; and (3) comparing the latter condition against restarting from scratch after correction. If identical final datasets produce different outcomes depending on the order of noise exposure and correction, this would provide evidence for hysteresis or path dependence in grokking dynamics and would considerably deepen the mechanistic contribution of the work while requiring only modest extensions to the existing framework."
(Verbatim experimental framework supplied by the reviewer:)
Proposed temporal/path-dependence experiments:
Baseline (clean training): Clean labels → Train → Grokking Condition A (noise from the start): Noisy labels → Train → No/Delayed Grokking Condition B (clean → noisy): Clean labels → Partial/Full Grokking → Inject noise → Continue training.
Question: Does grokking persist despite later corruption? Condition C (noisy → clean, continue training): Noisy labels → Memorization regime → Correct labels → Continue training.
Question: Can grokking recover without restarting optimization?
Condition D (noisy → clean, restart training): Noisy labels → Memorization regime → Correct labels → Reinitialize + retrain. Question: Is recovery dependent on removing optimization history? Key comparison: If Conditions C and D differ despite identical corrected datasets, this would support optimization-history dependence / hysteresis in grokking dynamics.

Response:

The reviewer's proposed extension was implemented in full as the primary new scientific contribution of the revised manuscript. All five conditions (clean baseline, A, B, C, D) were run at $\eta = 0.04$ with the transition at $T = 50{,}000$ epochs, both noise types, and three seeds each (**30 runs**).

Primary finding — hysteresis (Condition C vs. D). Conditions C and D received *identical* corrected datasets (all corruptions restored), differing only in whether the optimizer state from the noisy phase was retained (C) or discarded by reinitialization (D). Condition C grokked **10,000–16,000 epochs faster than Condition D in every one of the six seed × noise-type combinations** (C mean $t_{grok} \approx 78{,}700$; D mean $\approx 91{,}300$). Because the data were identical, this difference cannot be attributed to the dataset — it reflects the optimization trajectory accumulated during noisy training. This is direct evidence for path-dependence / hysteresis in grokking dynamics, which to the author's knowledge has not been previously reported.

Secondary finding — persistence (Condition B). Where grokking had fully consolidated during the clean phase (seeds 0 and 2; phase-1 test accuracy ≈ 0.97), it was partially robust to subsequent noise injection: final test accuracy degraded to 0.76–0.80, well above the ≈ 0.40 chance level. Where grokking had not consolidated by the transition (seed 1, which reached 0.968 but did not satisfy the strict ≥ 0.95-for-five-evaluations criterion), noise injection caused a collapse to chance (≈ 0.43). The robustness of grokking to later corruption therefore depends on how consolidated the generalizing solution was when noise was introduced.

Where addressed in the revised manuscript:

- Section 3.6 ("Temporal Path-Dependence Experiments") — added to Methods, defining all five conditions and the rationale.
- Section 4.5 ("Temporal Path-Dependence of Grokking Recovery") — added to Results, with Table 3 (condition-level grokking counts and mean times) and the per-seed C-vs-D table showing the 10–16k-epoch advantage.
- Section 5.3 ("Optimization History as a Resource for Grokking Recovery") — added to Discussion, interpreting the hysteresis effect.
- Section 5.4, point 5 — practical implication added (mid-training correction with retained optimizer state is preferable to retraining from scratch).
- Section 2.5 ("Optimization History and Training Dynamics") — added to Background to contextualize the temporal experiments.
- Abstract and Introduction, Contribution 3 — foreground the path-dependence finding as the primary novel contribution.

- Section 6 (Conclusion), point 5 — hysteresis finding added.
- Appendix B — reproduction command for the temporal experiments.

Reviewer Comment 4 — Reference Verification

"Please verify that all links to the references point to the correct source."

Response:

All fourteen reference URLs were individually checked against the source publications. During this verification, four entries were found to contain metadata errors and were corrected to match the actual sources:

- **[1]** the author list was corrected to Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). The foundational grokking paper had been drafted with incorrect co-authors; the title, year, and arXiv identifier were correct.
- **[6]** the venue was corrected to the *Proceedings of the NeurIPS 2021 Track on Datasets and Benchmarks* (it had been listed as the main NeurIPS proceedings, vol. 35).

[13] was corrected to: Tang, Y., Wang, Q., García-Redondo, I., & Monod, A. (2026). *Topological Signatures of Grokking.* arXiv:2605.06352. (The title and arXiv identifier were correct; the author list and year were corrected.)

- **[14]** was corrected to: Clauw, K., Stramaglia, S., & Marinazzo, D. (2024). *Information-Theoretic Progress Measures Reveal Grokking is an Emergent Phase Transition.* ICML 2024 Workshop on Mechanistic Interpretability; arXiv:2408.08944. (The title and arXiv identifier were correct; the author list was corrected.)

All remaining references were re-verified to have correct titles, authors, venues, and resolving links.

Where addressed in the revised manuscript:

- References section — entries [1], [6], [13], and [14] corrected; all links verified.

Reviewer Comment 5 — AI Assistance Disclosure

"The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section."

Response:

The Acknowledgments section was updated to state:

"Portions of this work, including experimental code development and manuscript preparation, were assisted by Claude (Anthropic), an AI language model. All experimental designs, interpretations, and conclusions are the author's own."

Where addressed in the revised manuscript:

- Acknowledgments section.

Reviewer Comment 6 — References: Recency and Support

"Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid 'padding.' Use credible sources (peer-reviewed journals, reputable books, official reports)."

Response:

Three references published in 2022–2026 were added to complement the foundational citations, each selected because it directly supports a specific claim:

- **[12]** Barak et al. (2022) supports the framing of grokking as hidden progress made by SGD before the visible generalization transition (Background, Section 2.1).
- **[13]** Tang et al. (2026) supports the claim that topological signatures in representations anticipate the grokking transition (Background, Section 2.1).
- **[14]** Clauw et al. (2024) supports the framing of grokking as an emergent phase transition mediated by synergistic interactions among neurons (Background, Section 2.1).

No references were removed and no padding citations were added; the existing references [1]–[11] are peer-reviewed or archival sources directly relevant to the claims they support.

Where addressed in the revised manuscript:

- References section (additions [12]–[14]); in-text citations in Section 2.1.

Reviewer Comment 7 — Explicit Assumptions

"All assumptions (including implicit assumptions) must be explicitly stated and clearly justified. The authors should explain why each assumption is reasonable and discuss its impact on the results and conclusions."

Response:

Section 3.1 ("Explicit Assumptions") was added at the start of the Methods section, enumerating and justifying six assumptions and noting the impact of each: (1) oracle correction; (2) fixed architecture and hyperparameters; (3) single-round audit; (4) warm-up-model validity; (5) noise geometry as the only structural variable; and (6) modular addition as a representative testbed (with generalization to other domains acknowledged as a limitation).

Where addressed in the revised manuscript:

- Section 3.1 ("Explicit Assumptions"); cross-referenced in Section 5.5 (Limitations).

Reviewer Comment 8 — Verb Tense and Person

"Verify that you have used past perfect tence and third person throughout the manuscript wherever applicable."

Response:

The Methods and Results sections were revised to use the third person and past tense throughout. All first-person constructions ("we study," "we find," "we use," "we show," "we run") were replaced with third-person equivalents ("this study examined," "it was found," "a two-layer MLP was used," "the results demonstrated," "a sweep was conducted"). A full-text search of the revised manuscript confirms zero remaining first-person pronouns ("we," "our," "I," "us") in the body. Present tense was retained only for general scientific statements in the Introduction/ Background and for interpretive statements in the Discussion/Conclusion, where it is appropriate.

Where addressed in the revised manuscript:

- Throughout, principally Sections 3 (Methods) and 4 (Results).

Additional Manuscript Revisions (Figures)

Beyond the point-by-point responses above, the following presentation changes were made to the manuscript to support the new results:

- All ten figures were embedded directly into the revised manuscript, each placed within the Results subsection that discusses it and accompanied by its full caption: Figures 1–2 in Section 4.2 (audit efficiency and the geometry-specificity of region auditing), Figures 3–4 in Section 4.3 (learning curves and final accuracy), Figure 5 in Section 4.4 (correction rate vs. budget), Figures 6–9 in Section 4.5 (the temporal path-dependence experiments, including the C-

vs-D hysteresis effect and Condition B persistence), and Figure 10 in Section 4.8 (time to reach 0.80 test accuracy).

- The figures were renumbered so that their numbering follows the order of first appearance in the text, and every in-text figure reference was updated to the corresponding new number.
- The caption for the Condition B persistence figure (Figure 9) was corrected to match the experimental data: where grokking had consolidated during the clean phase (seeds 0 and 2), test accuracy degraded to ≈ 0.76–0.80 after noise injection rather than being fully maintained; where it had not consolidated (seed 1), test accuracy collapsed to chance.

Thank you for addressing my comments. My remaining comments focus on formatting and consistency

1.Please provide a docx or a libreoffice writer .odt file correctly formatted to the Journal's specifications. Please check carefully, 12 font times new roman, spaccing between lines 1.15, one column, no indents, space between paragraphs etc. Make sure references appear in increasing sequential order in the manuscript per the Vancouver format. The references in the references section of the manuscript should be formatted APA style with a live link DOI at the end of each reference. Please consult the website for details.

2.I recommend changing the title to either "Noise Geometry Determines Audit Efficiency, but Grokking Recovery Depends on Optimization History" or "Restoring Grokking Under Structured Label Noise: Audit Efficiency and Optimization-History Dependence"

3.The comparison between Conditions C and D convincingly demonstrates that grokking recovery depends on optimization history, as two models exposed to identical corrected datasets exhibit systematically different recovery trajectories. However, several portions of the manuscript appear to attribute this effect specifically to the optimizer state accumulated during noisy training. The current experimental design does not isolate optimizer state as the causal factor. Conditions C and D differ not only in optimizer state, but also in network weights, learned representations, and feature geometry at the time of correction. Consequently, the experiments support the conclusion that optimization history matters, but they do not distinguish whether the observed hysteresis arises from retained optimizer statistics, retained learned representations, or both. The interpretation should therefore be softened throughout the manuscript to refer more generally to optimization-history dependence rather than optimizer-state dependence. Alternatively, the authors could acknowledge this limitation explicitly and note that a more rigorous decomposition would require additional controls, such as retaining weights while resetting only the optimizer, versus retaining both weights and optimizer, versus fully reinitializing both.

4.You need a stronger conclusion: see below for suggested verbiage and incorporate as necessary in the manuscript

"This study investigated how structured label noise affects grokking dynamics and how limited annotation resources can be allocated to restore generalization. Three principal findings emerged. First, noise geometry strongly influenced audit efficiency but had comparatively little effect on grokking suppression itself. At matched corruption rates, both uniform and input-dependent noise prevented grokking, indicating that corruption rate was the dominant determinant of suppression in this modular-addition setting. In contrast, audit performance depended strongly on the spatial distribution of errors. Region-targeted auditing provided a substantial advantage only when corruption was concentrated within the targeted region, demonstrating that audit strategies must be aligned with the underlying structure of annotation errors.

Second, the identity of corrected examples was found to be more important than the total number corrected. Loss-based auditing consistently outperformed random and region-based auditing across all noise rates tested and was the only policy that reliably restored grokking at moderate corruption levels. Importantly, conditions were observed in which two audit policies corrected nearly identical numbers of corrupted labels yet produced dramatically different generalization

outcomes. These results indicate that certain corrupted examples exert disproportionate influence on the emergence of algorithmic generalization and that successful recovery depends on identifying these optimization-critical errors rather than maximizing correction count alone.
Third, and most significantly, the results demonstrate that grokking recovery is path-dependent. Temporal experiments showed that models trained initially on corrupted data and later corrected recovered substantially faster when training continued from the existing optimization state than when training was restarted from scratch on the identical corrected dataset. Across all seeds and both noise types, retaining optimization history reduced the time to grokking by approximately 10,000–16,000 epochs. Because the final corrected dataset was identical in both cases, recovery speed could not be explained solely by the data available after correction. Instead, the results indicate that information accumulated during earlier stages of optimization continues to influence subsequent generalization dynamics. This hysteresis effect suggests that the trajectory by which a model arrives at a dataset may be as important as the dataset itself.
Taken together, these findings extend the study of grokking beyond the question of whether noisy labels suppress generalization. The results show that recovery depends not only on the amount and structure of label noise, but also on which errors are corrected and when correction occurs during training. From a practical perspective, targeted auditing can dramatically improve the efficiency of label correction efforts, while mid-training correction may be preferable to discarding accumulated optimization progress and retraining from scratch. More broadly, the observed hysteresis effect introduces optimization history as a potentially important variable in grokking dynamics and suggests a new direction for future mechanistic investigations of delayed generalization.
Future work should determine whether these phenomena extend beyond modular arithmetic to larger neural architectures and real-world datasets, examine additional forms of structured noise, and investigate the mechanistic origins of the observed path dependence. Understanding why optimization history accelerates recovery may provide new insight into how neural networks transition from memorization to generalization and may ultimately contribute to a more complete theory of grokking."
5.Even though the temporal path-dependence experiments are now the primary new scientific contribution, most of the manuscript remains about auditing, the hysteresis work does not appear until Section 4.5. The hysteresis result reads as an ‘add-on’ rather than as a primary contribution. A re-centering and re-emphasis of the paper is required.

Response to Editor Comments (Second Revision)
Manuscript: Restoring Grokking Under Structured Label Noise: Audit Efficiency and Optimization-History Dependence **Author:** Dylan Ashraf **Date:** June 2026
The author thanks the editor for the continued, constructive feedback. Each editor comment from the most recent decision letter is quoted verbatim below, followed by a description of how and where it was addressed in the revised manuscript. The previous response document (first revision) remains available for reference.

Editor Comment 1 — Formatting and Reference Style

"Please provide a docx or a libreoffice writer .odt file correctly formatted to the Journal's specifications. Please check carefully, 12 font times new roman, spaccing between lines 1.15, one column, no indents, space between paragraphs etc. Make sure references appear in increasing sequential order in the manuscript per the Vancouver format. The references in the references section of the manuscript should be formatted APA style with a live link DOI at the end of each reference. Please consult the website for details."

Response.

The Journal's website was consulted, and the manuscript was reformatted to its full specifications. It is provided as a .docx file with the following:

- **Font and layout:** single column; Times New Roman, 12 pt, applied to the body text and to all table contents.

- **Line spacing:** 1.15 throughout.
- **Margins:** left and right margins 1 inch; top and bottom margins 0.5 inch.
- **Paragraphs:** no first-line indentation; paragraphs are separated by a blank line (block style).
- **Headings:** section subheadings and their titles are bold; lower-level subheadings are bold, consistent with the Journal's heading specification.
- **Tables:** included inline at the appropriate locations in the text, and set so that rows do not break across pages (correcting the table fragmentation present in the previous submission).

Figures and legends (submitted separately, per the Journal's specification). The figures are no longer embedded in the manuscript. Each of the ten figures is provided as its own separate PNG file (Figure1.png … Figure10.png), with no legend baked into the image. The figure legends are provided together in a separate .docx file (the figure-legends document). The manuscript body refers to each figure by number at the appropriate location in the text.

In-text citations (curved parentheses, sequential). All in-text citations were converted to curved parentheses and numbered sequentially in order of first mention — (1), (2), (3), … through (14) — from the Introduction through the Discussion. These numbers correspond exactly to the numbered entries in the References section.

Equation numbering. The single display equation (the MLP definition in Section 3.3) is now numbered (eq1) to distinguish it from references, per the Journal's specification.

References — style, author lists, and live DOI links. The reference list is formatted consistently in APA style, with a live DOI link at the end of every entry (https://doi.org/…). For author lists, the Journal's rule was applied: references with six or fewer authors list all authors; references with more than six authors list the first six followed by *et al.* (entries 9 and 12 were updated accordingly). arXiv-archived works use their registered arXiv DOIs (10.48550/arXiv.…); journal articles use publisher DOIs (10.1109/TNNLS.2013.2292894 for entry 8; 10.1613/jair.1.12125 for entry 11; 10.64336/001c.154269 for entry 3). Entry 7 (Natarajan et al., NeurIPS 2013) predates DOI registration for that venue and has no assigned DOI; its permanent official proceedings URL is provided in its place.

Where addressed in the revised manuscript:

- Throughout — single column, Times New Roman 12 pt, 1.15 spacing, 1-inch left/right and 0.5-inch top/bottom margins, blank-line paragraph separation, no indents, bold headings.
- Figures — removed from the manuscript and supplied as ten separate PNG files; legends supplied in a separate document.
- In-text citations — converted to sequential curved parentheses (1)–(14).
- Section 3.3 — display equation numbered (eq1).
- References section — APA style, six-author rule applied, live DOI link on every entry.

Editor Comment 2 — Title

"I recommend changing the title to either 'Noise Geometry Determines Audit Efficiency, but Grokking Recovery Depends on Optimization History' or 'Restoring Grokking Under Structured Label Noise: Audit Efficiency and Optimization-History Dependence'"

Response.

The title was changed to the second of the editor's two suggested options: **"Restoring Grokking Under Structured Label Noise: Audit Efficiency and Optimization-History Dependence."** This title gives the optimization-history (hysteresis) finding co-equal billing with the audit-efficiency results, consistent with the re-centering described under Comment 5.

Where addressed in the revised manuscript:

- Title (page 1).

Editor Comment 3 — Causal Interpretation of the C-vs-D Comparison

"The comparison between Conditions C and D convincingly demonstrates that grokking recovery depends on optimization history, as two models exposed to identical corrected datasets exhibit systematically different recovery trajectories. However, several portions of the manuscript appear to attribute this effect specifically to the optimizer state accumulated during

noisy training. The current experimental design does not isolate optimizer state as the causal factor. Conditions C and D differ not only in optimizer state, but also in network weights, learned representations, and feature geometry at the time of correction. Consequently, the experiments support the conclusion that optimization history matters, but they do not distinguish whether the observed hysteresis arises from retained optimizer statistics, retained learned representations, or both. The interpretation should therefore be softened throughout the manuscript to refer more generally to optimization-history dependence rather than optimizer-state dependence. Alternatively, the authors could acknowledge this limitation explicitly and note that a more rigorous decomposition would require additional controls, such as retaining weights while resetting only the optimizer, versus retaining both weights and optimizer, versus fully reinitializing both."

Response.

Both of the editor's recommendations were adopted: the interpretation was softened throughout, **and** the limitation was acknowledged explicitly with the specific decomposition controls the editor described.

1. **Softening throughout.** All references to "optimizer-state dependence" were reframed as **optimization-history dependence**. Conditions C and D are now described as differing in their *entire* accumulated state — network weights, learned representations, and optimizer statistics together — rather than in optimizer state alone. The section headings and interpretive passages were revised accordingly (e.g., the Results section is now titled "Grokking Recovery Is Optimization-History Dependent," and the Discussion section is "Optimization History as a Resource for Grokking Recovery").
2. **Explicit limitation with decomposition controls.** A dedicated limitation was added (Section 5.5, fifth point) stating that the design establishes that optimization history matters but does not isolate the responsible component, and that a rigorous decomposition would require additional controls — specifically, *retaining the network weights while resetting only the optimizer statistics, versus retaining both weights and optimizer statistics, versus fully reinitializing both*. The Methods section now also notes this non-isolation directly where Conditions C and D are defined, and the Future Directions section names the weight-versus-optimizer decomposition as the first priority for follow-up work.

Where addressed in the revised manuscript:

- Title and Section/heading wording — reframed to "optimization-history."
- Section 3.5 (Temporal Path-Dependence Experiments) — added an explicit note that C and D differ in their entire accumulated state and that the design does not isolate the component responsible.
- Section 4.2 and Section 5.1 — interpretive language softened to optimization-history dependence.
- Section 5.5 (Limitations), fifth point — explicit limitation and the decomposition-control design.
- Section 5.6 (Future Directions), item (1) — the weight-versus-optimizer decomposition named as a priority.

Editor Comment 4 — Stronger Conclusion

"You need a stronger conclusion: see below for suggested verbiage and incorporate as necessary in the manuscript"

(The editor provided suggested conclusion text spanning the three principal findings — noise geometry and audit efficiency; the identity of corrected examples; and the path-dependence/hysteresis result — together with broader implications and future directions.)

Response.

The Conclusion (Section 6) was rewritten to incorporate the editor's suggested verbiage, very nearly verbatim, adjusting only to match the manuscript's exact figures and to preserve consistency with the softened optimization-history framing. The new conclusion presents the three principal findings in the editor's recommended structure — (1) noise geometry governs

audit efficiency but not suppression; (2) the identity of corrected examples matters more than their count; (3) and most significantly, grokking recovery is path-dependent — followed by the integrative implications and the future-work paragraph the editor proposed.

Where addressed in the revised manuscript:

- Section 6 (Conclusion) — rewritten using the editor's suggested verbiage.

Editor Comment 5 — Re-centering on the Hysteresis Result

"Even though the temporal path-dependence experiments are now the primary new scientific contribution, most of the manuscript remains about auditing, the hysteresis work does not appear until Section 4.5. The hysteresis result reads as an 'add-on' rather than as a primary contribution. A re-centering and re-emphasis of the paper is required."

Response.

The manuscript was re-centered so that the optimization-history (hysteresis) result is the primary contribution rather than an add-on:

- **Title** now leads with grokking recovery and names optimization-history dependence.
- **Abstract** was rewritten to foreground the path-dependence finding as the central result, with the audit-efficiency results presented as the complementary, practical contribution.
- **Introduction** now motivates the temporal/path-dependence question explicitly, and the contributions were reordered so that optimization-history dependence is **Contribution 1**, the "identity matters" audit result is Contribution 2, and noise geometry is Contribution 3.
- **Methods** were reordered so the Temporal Path-Dependence Experiments (Section 3.5) are described *before* the Audit Budget Policies (Section 3.6).
- **Results** were reordered so the hysteresis result now appears immediately after the baseline as **Section 4.2 ("Grokking Recovery Is Optimization-History Dependent")** — no longer buried at Section 4.5 — with its four supporting figures (Figures 1–4) presented first. The audit-efficiency results follow in Sections 4.3–4.8.
- **Figures and tables** were renumbered to reflect the new order of appearance: the temporal-experiment figures are now Figures 1–4 and Table 1, and the audit figures/tables follow.
- **Discussion** now opens with "Optimization History as a Resource for Grokking Recovery" (Section 5.1), and the **Conclusion** culminates in the path-dependence finding as the most significant result.

Where addressed in the revised manuscript:

- Title, Abstract, Introduction (contributions reordered), Methods (Section 3.5 before 3.6), Results (Section 4.2 is now the hysteresis result), Discussion (Section 5.1), Conclusion, and figure/table numbering throughout.

End of second-revision response document.

Accepted