Worlds Before Words? A Causal Boundary Audit of Procedural Pre-Pretraining

Anonymous authors
Paper under double-blind review

Abstract

Does procedural pre-pretraining transfer useful computation, or do apparent gains depend on budget and training boundaries? We study this distinction with paired initializations and downstream minibatches in an auditable small-decoder regime. Across 365 frozen trajectories, typed-Dyck warmup improves over downstream-only scratch in 14/15 pilot pairs by learning-curve area and 15/15 by final bits per byte. Stronger controls change the interpretation. Continuous natural training and retention of natural embedding/head parameters each win all 30 confirmation-plus-scale cells. Dyck beats exact within-document shuffling in all 30, but a first-order Markov surrogate strictly beats Dyck in 29/30. These directions persist at both tested widths.

A separate checkpoint diagnostic shows that Dyck models did learn bounded nonadjacent closing-type information that their Markov controls did not. Learned procedural information therefore need not yield a downstream advantage in this protocol. A reduced-scale bridge to a positive reference regime changes the task to ordered UNION and lowers the procedural allocation to about 0.3% of training model-tokens. UNION then beats all three declared comparators on final validation loss in two originally completed seed blocks and a separately disclosed post-outcome resource completion. This is not an exact reference replication or an isolated test of allocation.

The evidence supports a regime-specific boundary, not general failure of procedural pre-pretraining. Additive gains, learned procedural capability, and fixed-budget downstream value are distinct claims that require separate controls.

1 Introduction

Language models can benefit from pretraining on music, code, artificial sequences, formal languages, cellular automata, random computations, procedural task families, random recurrent networks, and logic derivations (Papadimitriou & Jurafsky, 2020; Chiang & Lee, 2022; 2020; Ri & Tsuruoka, 2022; Hu et al., 2025; Shinnick et al., 2025; Zhang et al., 2024; Bloem, 2025; Guo et al., 2026; Jiang et al., 2026; Riazhskykh et al., 2026; Cheng et al., 2026). Such results motivate the hypothesis that structured synthetic data teaches reusable computation before natural text. The inference is not automatic, however: pre-pretraining also changes optimization length and order, parameter statistics, optimizer state, vocabulary exposure, embedding/head initialization, and low-order sequence statistics. A scratch comparison can therefore identify an additive warmup effect without identifying its mechanism or fixed-compute usefulness. Prior work already shows that nonrecursive controls can match recursive pretraining (Papadimitriou & Jurafsky, 2020), parameter statistics can recover part of a gain (Wu et al., 2022), and outcomes vary across seeds, tokenizers, languages, model sizes, and warmup durations (Riazhskykh et al., 2026). Conversely, low-order statistics are not a universal explanation: random-RNN pre-pretraining under noisy natural training survives unigram-to-trigram metamer controls (Guo et al., 2026).

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.

1.1 Question

> For typed-Dyck pre-pretraining in a controlled small-model regime, do downstream gains survive equal total model-token compute, vocabulary/head replacement, matched natural warmup, optimization-boundary controls, and destruction of higher-order procedural structure?

The deliberately small regime is a causal microscope, not a proxy for frontier-scale performance. It makes paired initial states and downstream streams, full trajectory reconciliation, and coordinated interventions on the main alternative explanations feasible in one inspectable protocol. The target of inference is therefore evidential sufficiency: whether this typed-Dyck result supports its proposed mechanism under that protocol. Larger-scale positive results remain compatible with the boundary tested here.

1.2 Study logic and contribution

We designed the study to falsify the positive-transfer interpretation. The Dyck source, paired initializations, downstream minibatch streams, natural controls, token-space boundary, embedding/head intervention, destroyed-structure controls, two natural tokenizations, and second downstream domain were fixed before substantive multi-seed evaluation. When the original positive-transfer gate failed in the locked pilot, we stopped the planned generator search rather than tune against the result. Six boundary hypotheses were then prespecified on untouched seeds, followed by a prespecified local width increase on a second untouched seed set.

The contribution is the joint causal-control protocol and replicated boundary result, not any control in isolation. The study separates additive benefit from fixed-compute usefulness, parameter-boundary retention from reinitialization, and learned procedural information from a downstream advantage over simpler controls. Operationally, the audit proceeds from additive replication to equal-budget natural comparison, explicit parameter and optimization boundaries, and surrogates that preserve low-order statistics while removing the intended higher-order algorithm. Its central nonclaims are equally important: procedural pre-pretraining does not generally fail; low-order statistics do not generally explain procedural transfer; and these experiments do not establish frontier-scale behavior.

2 Related work

2.1 Synthetic-to-natural and formal pretraining

Early synthetic-to-natural transfer studies used nonlinguistic data, artificial dependencies, and nesting languages (Papadimitriou & Jurafsky, 2020; Chiang & Lee, 2022; 2020; Ri & Tsuruoka, 2022; Papadimitriou & Jurafsky, 2023). More recent work studies formal languages, modular procedural rules, cellular automata, random computation, broad procedural mixtures, and formal logic, including experiments at substantially larger scale (Hu et al., 2025; Shinnick et al., 2025; Zhang et al., 2024; Bloem, 2025; Jiang et al., 2026; Cheng et al., 2026). The synthetic-language comparison (Budnikov & Yamshchikov, 2025) compares nested, flat, and block-shuffled languages, while the LoRA study (Kumar et al., 2025) combines structural warmup with LoRA-based natural-language training. The nearest broad comparator is the procedural-mixture study (Jiang et al., 2026), which includes vocabulary reinitialization, shuffled controls, component localization, procedural mixtures, and models up to 1.3B parameters. Concurrent work on nested BPE capacities finds that sharing a model across vocabulary caps and restricting active output rows can materially alter BPB (Koutsiaris, 2026). It does not study staged synthetic-to-natural training or retaining versus reinitializing natural vocabulary parameters; we cite it only as evidence that vocabulary/head handling is a first-class intervention. Our narrower question is whether typed Dyck survives simultaneous equal-budget natural, vocabulary/head, optimization-boundary, and first-order controls in a fully auditable paired design.

2.2 Robustness and low-order controls

The instability study (Riazhskykh et al., 2026) establishes that pre-pretraining gains can be unstable across seeds and design choices; instability alone is therefore not our novelty claim. We instead pair initial states and downstream batches and intervene on the principal alternative explanations. The random-RNN study (Guo et al., 2026) provides a necessary boundary on interpretation: their larger random-RNN regime survives low-order metamers. Thus, a first-order surrogate matching typed Dyck here would show only that globally valid typed-stack structure is unnecessary in this protocol.

2.3 Data search and outcome adaptivity

Curriculum and dataset-selection methods optimize learnability, domain weights, model-aware examples, quality signals, scale prediction, or synthetic generators (Fan & Jaggi, 2023; Fan et al., 2024; Engstrom et al., 2024; Wettig et al., 2024; Magnusson et al., 2025; Thrush et al., 2026). These works motivated an initial generator-family search. Because the locked pilot failed its prespecified gate, continuing that search would have been outcome-adaptive; it was terminated before confirmation.

3 Experimental design

3.1 Evidence freeze

The original protocols were frozen locally in versioned files before their stated runs. In this paper, "prespecified" describes that local ordering, not registration with an external registry. Historical protocol filenames retain their original preregistration terminology.

The frozen scientific set comprises 170 pilot, 105 untouched confirmation, and 90 untouched scale-validation trajectories. Five locked pilot seeds tested the original positive-transfer gate. Before opening confirmation seeds, we prespecified six pilot-derived boundary hypotheses; before opening a second five-seed set, we prespecified a width increase from 64 to 96 while holding depth, context, data, update counts, optimization, evaluation, and the four central contrasts fixed. The historical post-freeze determinism and clean-environment runs are reproducibility evidence only and are excluded from effect estimates. The new capability audit and reduced-scale reference bridge are separate scientific extensions specified locally before their predictions or training, after the original downstream outcomes were known. A third bridge seed required an explicitly post-outcome resource completion. It does not amend the frozen 365-trajectory study or claim public preregistration. Appendix C records the complete chronology and gate.

3.2 Models, data, and procedural sources

We use two-block, four-head, pre-layer-normalized decoder-only Transformers with context 128, vocabulary capacity 1,024, and widths 64 or 96 (240,384 or 433,792 parameters). Downstream environments are WikiText-2 under byte and 512-token byte-level BPE tokenization (Merity et al., 2016), plus deterministically split CPython 3.12 standard-library source under byte tokenization. Natural IDs occupy 0–511 and procedural IDs begin at 512; the token spaces are disjoint.

The primary source is typed Dyck with 128 parenthesis types. Exact within-document shuffling preserves length and token multiset while destroying order. A smoothed first-order Markov surrogate, fit to an independent Dyck calibration stream, preserves one-step transitions without stack validity. The pilot also included unigram and copy-and-reverse controls, but they were not used to rescue the failed gate. Full generator, split, and architecture details are in Appendix C.

3.3 Boundary interventions

The registered matrix includes a downstream-only scratch baseline; continuous, reset, and vocabulary/head-replaced natural controls; replaced typed-Dyck, exact-shuffle, and first-order-Markov conditions; and a composite Dyck optimizer/schedule-continuation probe. Appendix A gives the complete intervention table.

Replacement conditions use the same deterministic embedding/head replacement seed within each environment-seed cell. Paired conditions share initial model state and downstream minibatch stream. dyck_keepopt retains both backbone optimizer state and the continuous schedule, so it is a composite boundary probe rather than an optimizer-only intervention.

3.4 Budget, endpoints, and integrity

Every full condition receives 160 stage-1 and 480 stage-2 updates at batch size 32 and context 128, or 2,621,440 training model-tokens. The scratch baseline receives only stage 2 and is used solely to reproduce the additive direction. The primary endpoint is normalized area under the downstream validation-BPB curve (AULC); final BPB is secondary. Registered contrasts are paired over five seeds and reported with every difference, sign counts, descriptive summaries, exact sign-flip tests, and bootstrap intervals. With five pairs, the minimum nonzero two-sided exact p-value is 0.0625, so inference emphasizes effect magnitude, sign consistency, and prespecified thresholds rather than dichotomous significance.

All runs store resolved configurations, source/data/preregistration hashes, model-state hashes, checkpoints, metrics, and environment metadata. Registry, ledger, and filesystem reconcile one-to-one; the frozen scientific subset contains 365 unique keys. Appendix C gives the evaluation schedule, and Appendix E gives full provenance and compute accounting.

4 Results

4.1 Additive benefit is reproducible but not fixed-compute evidence

Relative to scratch_stage2_only, dyck_replaced improves AULC in 14/15 pilot environment-seed cells and final BPB in 15/15; median AULC advantages are 0.013 BPB on WikiText/byte, 0.010 on WikiText/BPE, and 0.043 on Python/byte. The implementation therefore reproduces a procedural-warmup gain. The scratch comparison omits the entire stage-1 budget, however. Continuous natural warmup yields much larger median advantages over scratch—0.398, 0.264, and 0.522 BPB—and is the relevant practical equal-budget comparator. Because it also preserves optimizer state and schedule, this comparison is a practical protocol contrast, not a data-only effect.

4.2 Four causal boundaries

Fixed compute. On untouched confirmation seeds, natural_continuous beats dyck_replaced in all 15 base-model cells, exceeding the prespecified 0.10-BPB practical threshold in every environment. The width-96 validation reproduces the direction in all 15 additional cells. Across both sizes, natural training wins 30/30 with pooled median advantage 0.376 BPB. This practical contrast includes optimization continuity. The conclusion does not depend only on that continuity: when natural and Dyck warmups both undergo vocabulary/head replacement and optimizer reset, Dyck wins only 7/30 confirmation-plus-scale cells, and the pooled median Dyck advantage, defined as natural-minus-Dyck AULC, is -0.023 BPB (Section 4.3).

Vocabulary/head state. natural_reset and natural_replaced use identical stage-1 data and reset optimization at the boundary, differing only in whether embedding/head parameters are retained. Retention wins all 30 confirmation-plus-scale cells, with pooled median advantage 0.341 BPB, and remains beneficial after 480 downstream updates. Vocabulary-dependent parameters are therefore a quantitatively large boundary in this harness.

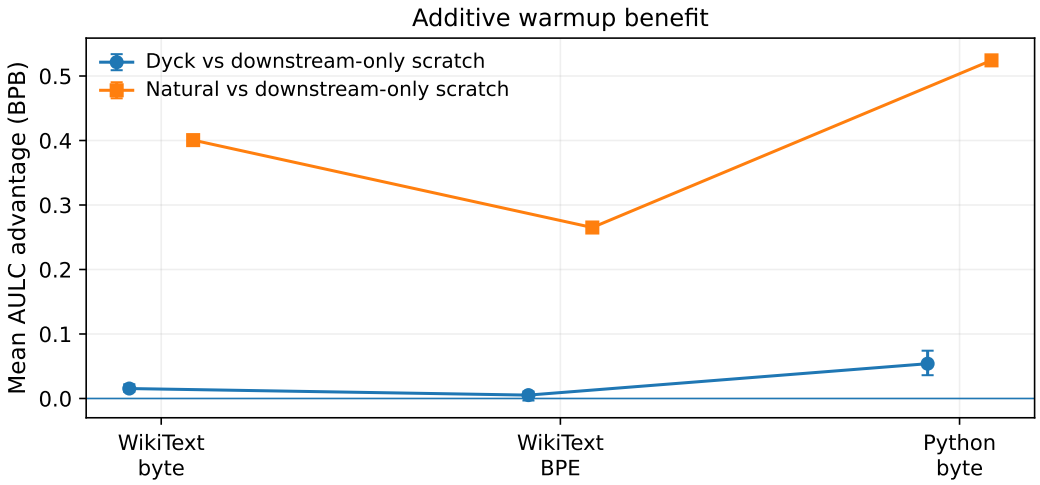

Figure 1: Pilot additive replication. Dyck improves over a stage-2-only scratch baseline, but the gain is much smaller than spending the same extra stage on natural data. Points are paired means; bars are 95% bootstrap intervals.

Order versus higher-order validity. Dyck beats exact shuffling in all 30 cells (pooled median 0.097 BPB), establishing that sequence order matters. Yet the first-order Markov surrogate strictly beats Dyck in 29/30 and meets the prespecified empirical 0.01-BPB tolerance in 29/30 (pooled median Markov advantage 0.040 BPB). The controlled source comparison shows that globally valid typed-stack training examples are not required to obtain the measured downstream benefit in this protocol. It does not identify the representations or algorithm used by either model. This is a protocol-specific boundary, not a universal claim about procedural data (Guo et al., 2026).

Local scale validation. All four directions—natural over Dyck, retained over replaced vocabulary/head, Markov over Dyck, and Dyck over shuffle—hold in every environment-seed cell at 433,792 parameters. The increase establishes local width robustness, not frontier-scale generality.

Registered contrast	240k confirmation	434k validation	Combined descriptive summary
Natural > Dyck	15/15	15/15	30/30; median 0.376 BPB
Retain vocab/head > replace	15/15	15/15	30/30; median 0.341 BPB
Markov > Dyck	14/15 strict; 14/15 within tolerance	15/15 strict; 15/15 within tolerance	29/30 strict; 29/30 within tolerance; median 0.040 BPB
Dyck > shuffle	15/15	15/15	30/30; median 0.097 BPB

The tolerance column counts paired observations satisfying the registered 0.01-BPB practical criterion. It does not establish population non-inferiority by an interval test. Combined medians summarize heterogeneous environments and sizes; they are not treated as 30 independent samples.

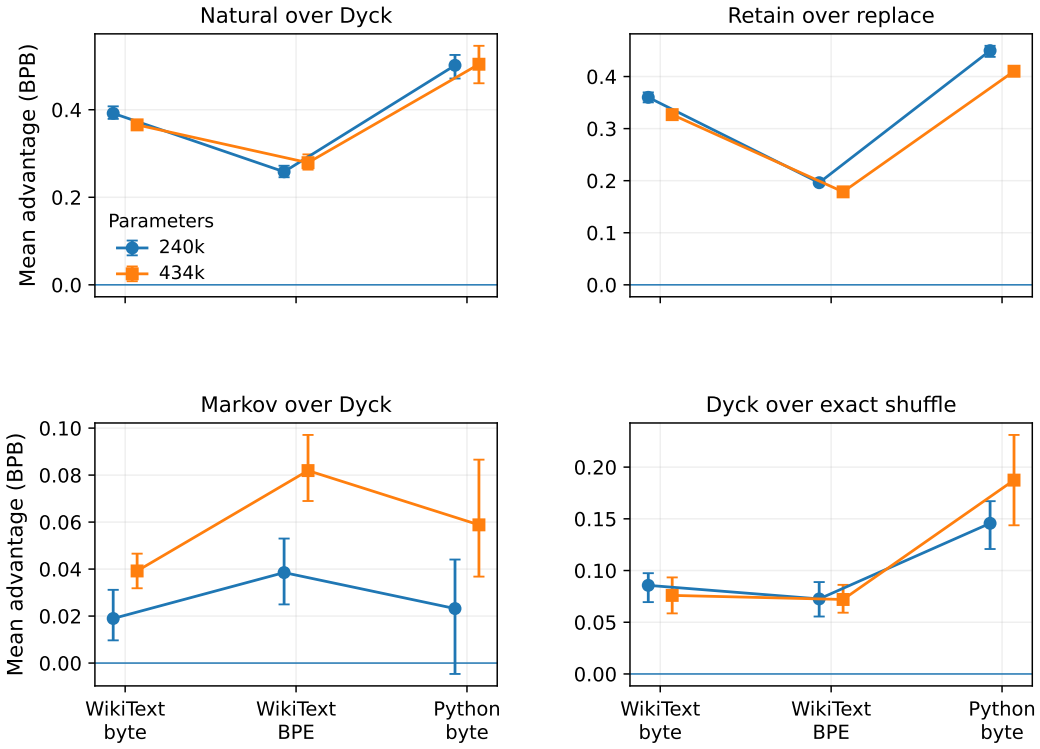

Figure 2: Four prespecified causal boundary effects at 240,384 and 433,792 parameters. Positive values favor the named contrast. Points are five-seed paired means; bars are 95% bootstrap intervals.

4.3 Secondary registered outcomes and economics

The prespecified matched-replacement source contrast rejects the pilot's possible code-specific Dyck advantage. With natural and Dyck conditions matched on vocabulary/head replacement and optimizer reset, Dyck wins only 7/30 confirmation-plus-scale cells, with median Dyck advantage, natural-minus-Dyck AULC, -0.023 BPB. The composite dyck_keepopt probe improves AULC on WikiText/byte and Python/byte but worsens it on WikiText/BPE, confirming setup sensitivity without isolating optimizer state from schedule. Exact values are in Appendix D.

Figure 3 shows the confirmation-stage trajectory of the four main contrasts. Natural-continuation and vocabulary/head-retention advantages are largest immediately after the boundary and decay but remain positive; Dyck's advantage over shuffle emerges downstream; the final Markov–Dyck difference is comparatively small even when AULC favors Markov.

The complete confirmation-stage trajectories are shown in Appendix B.

The frozen scientific runs process 946,995,200 training model-tokens and account for approximately 1.702×10^{15} analytic FLOPs. Matched full conditions have equal registered compute, yet natural continuation wins every untouched cell. No deployed-model-size saving, energy estimate, or finite amortized break-even is established.

4.4 Post-freeze stage-one capability audit

To test the alternative explanation that the original models never learned relevant procedural information, we evaluated all 175 eligible original disjoint-vocabulary Dyck, shuffle, and Markov stage-one checkpoints. Checkpoint hashes and exact seed-reconstructed initial states

were verified before prediction. The checkpoints contain 55 distinct trained states and 15 distinct initial states; repeated downstream-environment aliases are not independent models.

A fresh, fixed corpus contains 1,024 complete documents from the original 128-type Dyck generator. The primary diagnostic is closing-type conditional loss over 128 classes at the 10,425 nonadjacent closing events whose matching openings remain in context. Uniform prediction costs 7 bits. On the five confirmation seeds, Dyck achieves 5.925 bits and 12.93% accuracy, versus 7.033 bits and 0.75% for the trained Markov control. On the five larger-model seeds, the corresponding values are 5.136 bits and 22.72% versus 7.035 bits and 0.82%. Every Dyck seed in both sets has lower diagnostic loss than its matched initial, shuffle, and Markov states.

This is evidence of bounded nonadjacent closing information, not an unbounded stack algorithm. Performance declines with matching distance, and the probe conditions on a closing event. Despite this capability difference, the frozen downstream comparison favors the Markov source. Learning procedural information and benefiting from it downstream are therefore empirically distinct. Appendix J gives the corpus, strata, aliases, and independent arithmetic checks.

4.5 Reduced-scale reference-regime bridge

A separately specified extension tests a materially different allocation rather than treating the original 25% Dyck warmup as representative of the positive reference (Jiang et al., 2026). It uses ordered UNION with output-only loss, vocabulary/head replacement, a 1,089,280-parameter decoder, and 120 procedural updates followed by 10,000 WikiText-2 byte updates. Procedural model-tokens are 0.3084% of the total. Matched natural-replacement and reassociated-target controls use the same model-token allocation; a continuous-natural arm supplies a separate protocol comparison. Supervised-token counts differ by objective and are reported, not called equal supervision.

UNION beats all three comparators in both seed blocks completed under the original local pre-outcome specification. Its paired final-BPB improvements over natural replacement are 0.0669 and 0.0766. The 6,000-second resource cap left the third block unfinished. A separately recorded post-outcome completion kept its seed and scientific settings unchanged; its improvement is 0.0761. Original failures remain in the evidence. Appendix K reports every cell, learning-curve area, and the completed-checkpoint audit.

This positive diagnostic constrains the negative conclusion to the original regime. It is not a replication of the reference's GPT-2/C4 experiment: source family, architecture, allocation, corpus, and objectives differ from the frozen Dyck study, so the bridge does not isolate which change explains the contrast. Three descriptive seeds, one completed after outcomes were known, do not establish a powered confirmatory or higher-order-mechanism claim.

5 Interpretation

Three distinctions govern the result. First, additive benefit is not fixed-compute benefit: Dyck improves over a shortened scratch run, but natural use of the same extra budget is substantially better. Second, the parameter boundary is substantive: retaining embedding/head state produces a large, persistent advantage. Reinitialization replaces 132,096 of 240,384 parameters in the smaller model and 197,632 of 433,792 in the larger model, including the untied output head and bias. These counts describe an intervention on parameters, not a percentage of learned information. Retention benefits do not by themselves identify lexical content or its mediation. Third, order sensitivity is not transfer of the intended higher-order algorithm: Dyck beats shuffle, but first-order Markov data is at least as useful and usually better.

The new capability evidence adds a fourth distinction: nonadjacent procedural information can be learned without producing a downstream advantage over the first-order control. The diagnostic neither demonstrates an unbounded stack algorithm nor rules out all structural contributions. The methodological conclusion is that an additive scratch comparison is

insufficient for interpreting procedural pre-pretraining. Equal-budget natural controls, explicit vocabulary/head interventions, matched optimization boundaries, and progressively stronger destroyed-structure surrogates should be treated as baseline causal tests. Here they change the scientific story. The positive UNION bridge reinforces the regime boundary rather than overturning the frozen Dyck observations. It does not identify a single moderating variable because several design choices change together. The portable contribution is evidential ordering: first establish that the gain is not a budget or boundary artifact, then ask which statistics of the procedural source are actually necessary.

6 Limitations

The study is deliberately small: 240k and 434k parameters, one primary procedural family, 160 warmup updates, 480 continuation updates, two downstream domains, and two natural tokenizations. The scale experiment tests only a local width increase. Larger and longer positive regimes (Hu et al., 2025; Guo et al., 2026; Jiang et al., 2026; Cheng et al., 2026), multilingual data, noisy web pretraining, mathematics, instruction tuning, and mature behavioral capabilities remain outside the frozen study. Its 160-of-640-update warmup consumes 25% of training model-tokens, unlike the much smaller procedural allocations in the nearest positive reference (Jiang et al., 2026). The new UNION extension investigates one reduced-scale bridge, not the larger reference regime or a factorial explanation of the difference. Failure to transfer in this regime is not a failed replication of that reference.

The Markov control is specific to this typed-Dyck generator and does not contradict regimes where low-order metamers fail (Guo et al., 2026). dyck_keepopt changes optimizer-state continuity and learning-rate schedule jointly. Analytic FLOPs are transparent parameter/token estimates rather than hardware counters; active-class selection and cross-entropy are not separately modeled. Confirmation and scale use untouched seeds, not held-out datasets, so they protect against favorable seed selection rather than all dataset adaptation.

The six confirmation hypotheses were derived from a locked pilot but frozen before the confirmation seeds. Conditional component resets were not run after the positive gate failed because searching for a rescuing mechanism would have been outcome-adaptive. Historical path-sensitive metadata and append-only-ledger dirtiness affect provenance signatures, not training tensors, metrics, or checkpoints; Appendix E documents the reconciliation and portable hashes.

7 Reproducibility, data governance, and broader impact

The anonymous supplement contains deterministic generators, versioned configurations, preregistrations, seed-level summaries, the experiment registry and compute ledger, source/data fingerprints, analysis scripts, tests, claim–evidence mappings, and integrity audits. Forced same-seed reruns reproduce model states and numerical trajectories, and a clean environment passes unit tests, token-space checks, and an end-to-end smoke run. Raw WikiText-2 and CPython corpora and token arrays are excluded to avoid redistribution or silent relicensing; acquisition code, split rules, fingerprints, and hashes are provided instead. The reference bridge adds 986,081,280 model-tokens across its twelve completed cells and 172,874,240 logged model-tokens in the retained interrupted attempts. These are separate from the frozen compute ledger. The work uses no personal data, human subjects, or external model API. Its likely benefit is methodological; its main negative externality is compute use, which is bounded and reported but not converted into an unmeasured energy estimate.

8 Conclusion

Typed-Dyck warmup reproducibly improves over a downstream-only scratch baseline, but stronger controls do not support the broad mechanistic interpretation. Equal-budget natural training and retained vocabulary/head state dominate; sequence order matters relative to shuffle, yet a first-order Markov surrogate usually outperforms valid Dyck; and all four

directions persist under a prespecified local width increase. The defensible conclusion is narrow: in this regime, typed-Dyck pre-pretraining does not establish practically useful higher-order computational-prior transfer once natural compute, vocabulary/head state, optimization boundaries, and low-order statistics are controlled. The separate low-fraction UNION bridge supplies a positive result under changed conditions, with its resource-amended third seed disclosed. Different procedural sources and allocations demonstrably require separate evaluation here. The transferable result is the causal audit required to tell those possibilities apart, not a forecast from small models to frontier systems.

AI use statement

Generative AI, specifically OpenAI ChatGPT, was used as a research assistant for hypothesis and control generation, adversarial critique, literature discovery and triage, drafting and refactoring experimental and analysis code, test and audit construction, interpretation support, figure and package generation, and manuscript drafting and editing. The human author specified the project objective and constraints, authorized the local experimental program, reviewed progress, and retained responsibility for the research decisions and final claims. Every reported observation comes from executable local code and stored run artifacts rather than from a language-model assertion. AI-assisted code and prose were checked through unit tests, frozen preregistrations and stopping rules, paired reruns, clean-environment reconstruction, content hashes, claim–evidence audits, and primary-source literature review. Locally trained autoregressive models were experimental subjects. No language-model response was substituted for a computed observation. Revision-stage checks were automated and are not represented as independent human review. The author takes responsibility for the final content and artifacts.

Ethics statement

The study uses no human participants, personal data, or external model API. It evaluates small locally trained language models on public WikiText-2 data and CPython source under their documented licenses. Raw corpora and token arrays are omitted from the anonymous supplement to avoid redistribution or silent relicensing; deterministic acquisition and verification instructions are supplied instead. The principal foreseeable cost is compute use, which is reported and bounded in the paper.

Reproducibility statement

The anonymous supplement contains deterministic generators, versioned configurations, preregistrations, seed-level summaries, the experiment registry and compute ledger, source and data fingerprints, analysis scripts, tests, claim–evidence mappings, and integrity audits. The frozen scientific set is explicitly separated from engineering runs and new post-freeze scientific extensions. Their specifications and outcomes are retained separately, including interrupted execution attempts. Raw licensed corpora are reconstructed from named public sources and checked against recorded fingerprints. Appendix I summarizes the reconciliation and determinism evidence.

References

Peter Bloem. Universal pre-training by iterated random computation. arXiv preprint arXiv:2506.20057, 2025.

Mikhail Budnikov and Ivan Yamshchikov. Transfer of structural knowledge from synthetic languages. In Proceedings of the 1st Workshop on Language Models and Structure, 2025. doi: 10.18653/v1/2025.xllm-1.20.

Jo-Ku Cheng, Nikolaos Aletras, and Marco Valentino. Logic before language: Pre-pretraining on formal derivations fosters skill acquisition and compressibility. arXiv preprint arXiv:2608.03930, 2026.

Cheng-Han Chiang and Hung-yi Lee. Pre-training a language model without human language. arXiv preprint arXiv:2012.11995, 2020.

Cheng-Han Chiang and Hung-yi Lee. On the transferability of pre-trained language models: A study from artificial datasets. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022. doi: 10.1609/aaai.v36i10.21295.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. Dsdm: Model-aware dataset selection with datamodels. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, 2024.

Simin Fan and Martin Jaggi. Irreducible curriculum for language model pretraining. arXiv preprint arXiv:2310.15389, 2023.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, 2024.

Xu Guo, Runyu Peng, Jian Tong, Yunhua Zhou, Haijun Lv, Zhihui Lu, and Qipeng Guo. Synthetic pre-pre-training improves language model robustness to noisy pre-training data. arXiv preprint arXiv:2605.10129, 2026.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. Between circuits and chomsky: Pre-pretraining on formal languages imparts linguistic biases. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025.

Liangze Jiang, Zachary Shinnick, Anton van den Hengel, Hemanth Saratchandran, and Damien Teney. Procedural pretraining: Warming up language models with abstract data. In Proceedings of the 43rd International Conference on Machine Learning, 2026.

Christos Koutsiaris. Nested byte-level vocabularies are cheap to deploy and expensive to share: A pre-registered negative result. arXiv preprint arXiv:2608.28151, 2026.

Nalin Kumar, Mateusz Lango, and Ondrej Dusek. Pretraining language models with lora and artificial languages. In Proceedings of the BabyLM Challenge, 2025. doi: 10.18653/v1/2025.babylm-main.37.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. Datadecide: How to predict best pretraining data with small experiments. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016.

Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

Isabel Papadimitriou and Dan Jurafsky. Injecting structural hints: Using language models to study inductive biases in language learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023.

Ryokan Ri and Yoshimasa Tsuruoka. Pretraining with artificial language: Studying transferable knowledge in language models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.504.

Sofiia Riazhskykh, Nam Luu, and Ondrej Bojar. Instability of llm pre-pretraining: It doesn't always help. an investigation on multiple languages. arXiv preprint arXiv:2608.08800, 2026.

Zachary Shinnick, Liangze Jiang, Hemanth Saratchandran, Anton van den Hengel, and Damien Teney. Transformers pretrained on procedural data contain modular structures for algorithmic reasoning. arXiv preprint arXiv:2505.22308, 2025.

Tristan Thrush, Sung Min Park, Herman Brunborg, Luke Bailey, Marcel Roed, Neil Band, Christopher Potts, and Tatsunori Hashimoto. Synthetic data for any differentiable target. arXiv preprint arXiv:2604.08423, 2026.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, 2024.

Yuhuai Wu, Felix Li, and Percy Liang. Insights into pre-training via simpler synthetic tasks. arXiv preprint arXiv:2206.10139, 2022.

Shiyang Zhang, Aakash Patel, Syed A. Rizvi, Nianchen Liu, Sizhuang He, Amin Karbasi, Emanuele Zappala, and David van Dijk. Intelligence at the edge of chaos. arXiv preprint arXiv:2410.02536, 2024.

A Experimental conditions and boundary interventions

The complete registered condition matrix is reproduced below.

Condition	Stage 1	Vocab./head	Opt./schedule	Purpose
scratch_stage2_only	none	initial	new stage-2 optimizer	additive baseline
natural_continuous	natural	retained	continued	practical equal-budget baseline
natural_reset	natural	retained	reset	isolates optimizer/schedule boundary
natural_replaced	natural	reinitialized	reset	matched vocabulary/head intervention
dyck_replaced	typed Dyck	reinitialized	reset	primary procedural condition
dyck_shuffle_replaced	exact shuffled Dyck	reinitialized	reset	destroys order
dyck_markov_replaced	Dyck first-order surrogate	reinitialized	reset	preserves low-order transitions
dyck_keepopt	typed Dyck	reinitialized	backbone optimizer/schedule continued	composite optimization-boundary probe

B Confirmation-stage contrast trajectories

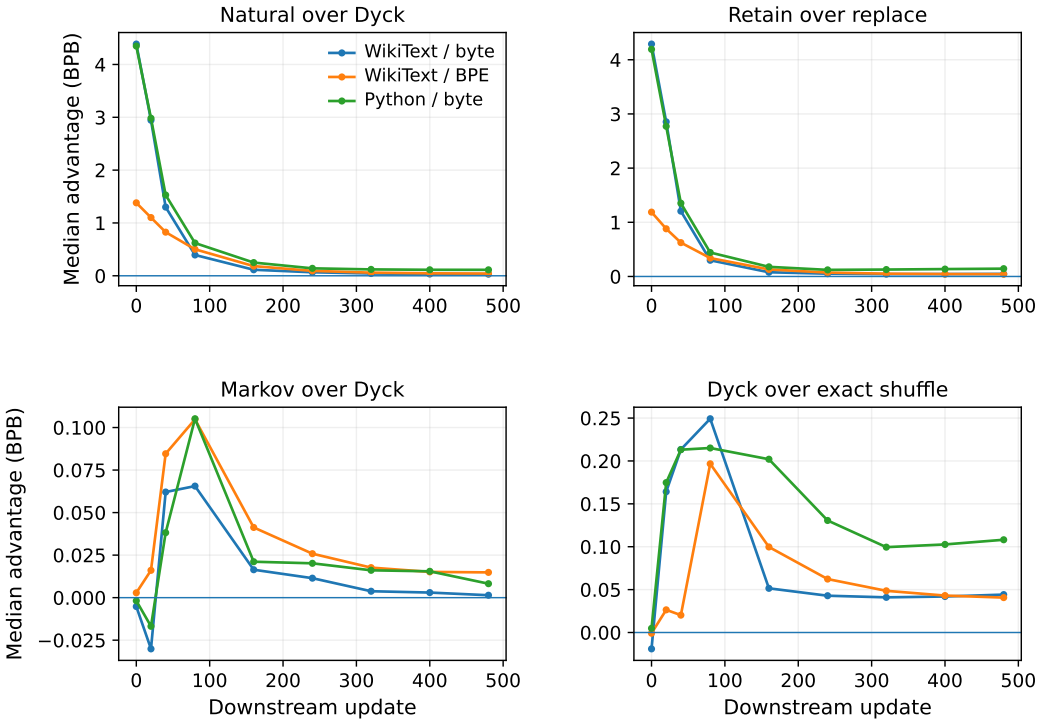

Figure 3: Median paired contrast through downstream training at 240,384 parameters on untouched confirmation seeds. These trajectories are descriptive; causal interpretation comes from the registered interventions, not the curve shape alone.

C Detailed protocol and evidence freeze

C.1 Scientific chronology and stopping logic

Pilot. Five locked seeds (11, 22, 33, 44, 55) at 240,384 parameters tested the original positive-transfer hypothesis. The gate required a consistent advantage over matched natural replacement, no material loss to continuous natural training, evidence beyond boundary variation and destroyed-structure controls, and no tokenizer reversal. The gate failed, and generator search was terminated.

Boundary confirmation. Before opening seeds 101, 202, 303, 404, or 505, six pilot-derived hypotheses were prespecified: natural-data and vocabulary/head boundaries, Markov and shuffle controls, a possible code-specific benefit, and a tokenizer-sensitive optimizer/schedule interaction. No model or optimization hyperparameter changed.

Scale validation. Before opening seeds 606, 707, 808, 909, or 1001, a width increase from 64 to 96 (240,384 to 433,792 parameters) was prespecified. Depth, context, optimizer, data, update counts, evaluation schedule, and the four central contrasts remained fixed. This was the final training experiment in the original study; that empirical claim set was then frozen. Historical post-freeze determinism and clean-environment runs remain excluded from its effect estimates. New revision-stage scientific extensions use separate specifications and evidence directories rather than retroactively entering that freeze.

C.2 Architecture, corpora, and tokenization

The decoder-only pre-layer-normalized Transformer has learned token and positional embeddings, two Transformer blocks, four attention heads, GELU MLPs with expansion ratio 4, no dropout, and an untied linear language-model head. Both sizes use vocabulary capacity 1,024 and context 128.

The full 1,024-way output projection is computed before restricting logits to the active vocabulary, matching model-side projection FLOPs across active token sets.

WikiText-2 raw (Salesforce/wikitext, wikitext-2-raw-v1) uses the official train/validation/test split (Merity et al., 2016); the training split contains 10,951,563 UTF-8 bytes. We evaluate UTF-8 bytes and a 512-token byte-level BPE trained only on the training split. CPython 3.12 standard-library .py files exclude site-packages, dist-packages, and ___pycache___; deterministic relative-path hashing yields 435/67/71 train/validation/test files and 8,541,339 training bytes. Code uses byte tokenization. Natural IDs occupy 0–511; typed-Dyck openings use 512–639, closings 640–767, and separator 768.

C.3 Procedural generators, training, and statistics

Typed-Dyck documents contain 8–32 pairs, maximum stack depth 24, and opening probability 0.54 when opening and closing are both valid. Exact shuffling preserves each document's length and token multiset. The first-order surrogate is a smoothed Markov model fitted to an independent 200,000-token Dyck calibration stream; a unigram IID source retains fitted token frequencies. A copy-and-reverse generator provides a second pilot family.

Full conditions receive 160 stage-1 and 480 stage-2 updates with batch size 32 and context 128. AdamW uses learning rate 0.001, β_1=0.9, β_2=0.95, ε=1e-8, no weight decay, gradient clipping at 1.0, 5% linear warmup, and cosine decay to 10% of peak. Validation occurs at downstream steps 0, 20, 40, 80, 160, 240, 320, 400, and 480 on 256 deterministic blocks. BPB is total negative log likelihood divided by the product of ln(2) and the decoded target-byte count; AULC is trapezoidal validation-BPB area divided by the 480-step interval. Each five-seed contrast reports all paired differences, mean, median, sample standard deviation, signs, a two-sided exact sign-flip p-value based on the absolute paired mean, and a 20,000-resample percentile bootstrap interval. The historical H3 label used "non-inferiority," but its implemented admission criterion counts at least four of five paired differences within 0.01 BPB in each environment. This is an empirical practical-tolerance rule, not a confidence-interval-based population non-inferiority test. We retain the original protocol and report the distinction explicitly.

D Secondary registered outcomes

D.1 Matched-replacement data-source contrast

The pilot suggested that, after the same vocabulary/head replacement, Dyck might outperform a natural stage-1 backbone specifically on Python. This was prespecified rather than promoted directly. On confirmation seeds, Dyck wins 3/5 Python/byte pairs with median +0.002 BPB and an interval spanning zero; WikiText/byte yields 2/5 signs (median -0.009), and WikiText/BPE yields 0/5 (median -0.016). Across all confirmation-plus-scale cells where the contrast is available, Dyck wins 7/30 with median -0.023 BPB. The code-specific claim is rejected.

D.2 Composite optimizer/schedule boundary

dyck_keepopt reinitializes vocabulary/head parameters but retains backbone optimizer state and a continuous learning-rate schedule. Relative to dyck_replaced, it improves AULC in 5/5 WikiText/byte confirmation seeds (median +0.053 BPB) and 5/5 Python/byte seeds (+0.062), but worsens AULC in 5/5 WikiText/BPE seeds (median -0.033). The prespecified reversal establishes setup sensitivity but does not isolate optimizer state from schedule.

E Provenance, compute, and data governance

We estimate training FLOPs as six times the parameter count times training model-tokens and evaluation FLOPs as twice the parameter count times evaluation target-tokens. Within a model size, full conditions match parameter count, updates, sequence length, model-token count, evaluation schedule, and analytic FLOPs. A base full run uses approximately 3.923×10^{12} estimated FLOPs; a larger run uses approximately 7.079×10^{12}.

Every run stores its resolved configuration, source-tree hash, data-manifest hash, preregistration hash, initial model-state hash, stage-1 and final checkpoint hashes, per-checkpoint metrics, and environment metadata. An independent reconciliation after a possible concurrent-writer overlap verified a one-to-one mapping among all 421 run directories, registry entries, and compute-ledger rows, with no partial or unregistered run.

A packaging audit found that historical BPE metadata included an absolute tokenizer path inside run-level data_sha256. The path affected only that metadata hash and run signature—not tokenizer bytes, token arrays, token IDs, training tensors, metrics, or checkpoints. Current source uses a project-relative path and preserves historical and normalized hashes. Scientific runs record git_dirty=true because append-only ledgers and analysis files were written during execution; the audit verifies that training-critical changes between phases were limited to preregistration selection and the prespecified width increase.

WikiText-2 is reconstructed from the public Salesforce/Hugging Face source and treated as CC BY-SA 4.0 according to the pinned dataset card. CPython source is used under the Python Software Foundation License Version 2, with the caveat that incorporated components may carry additional notices. Raw corpora and processed arrays are omitted; deterministic acquisition and verification instructions are supplied.

F Full confirmatory results at 240,384 parameters

Positive Δ follows the direction named in the first column. Bootstrap intervals are percentile intervals for the paired mean over five seeds.

Contrast	Env.	Median Δ BPB	Mean Δ BPB	Signs	95% bootstrap CI	Exact sign-flip p
Natural > Dyck	WikiText / byte	0.383	0.392	5/5	[0.379, 0.408]	0.0625
Natural > Dyck	WikiText / BPE	0.259	0.258	5/5	[0.246, 0.272]	0.0625
Natural > Dyck	Python / byte	0.510	0.501	5/5	[0.471, 0.525]	0.0625
Retain vocab/head > replace	WikiText / byte	0.365	0.360	5/5	[0.350, 0.370]	0.0625
Retain vocab/head > replace	WikiText / BPE	0.197	0.196	5/5	[0.193, 0.199]	0.0625
Retain vocab/head > replace	Python / byte	0.456	0.450	5/5	[0.438, 0.459]	0.0625
Markov > Dyck	WikiText / byte	0.017	0.019	5/5	[0.010, 0.031]	0.0625
Markov > Dyck	WikiText / BPE	0.039	0.039	5/5	[0.025, 0.053]	0.0625
Markov > Dyck	Python / byte	0.028	0.023	4/5	[-0.005, 0.044]	0.2500
Dyck > shuffle	WikiText / byte	0.092	0.086	5/5	[0.069, 0.097]	0.0625
Dyck > shuffle	WikiText / BPE	0.074	0.073	5/5	[0.056, 0.089]	0.0625
Dyck > shuffle	Python / byte	0.152	0.146	5/5	[0.121, 0.167]	0.0625
Dyck > matched natural	WikiText / byte	-0.009	-0.008	2/5	[-0.019, 0.004]	0.3750
Dyck > matched natural	WikiText / BPE	-0.016	-0.020	0/5	[-0.033, -0.009]	0.0625
Dyck > matched natural	Python / byte	0.002	-0.003	3/5	[-0.029, 0.024]	1.0000

Contrast	Env.	Median Δ BPB	Mean Δ BPB	Signs	95% bootstrap CI	Exact sign-flip p
Keep boundary > reset	WikiText / byte	0.053	0.054	5/5	[0.048, 0.061]	0.0625
Keep boundary > reset	WikiText / BPE	-0.033	-0.033	0/5	[-0.042, -0.023]	0.0625
Keep boundary > reset	Python / byte	0.062	0.054	5/5	[0.036, 0.070]	0.0625

G Larger-model validation at 433,792 parameters

The point-estimate column reports paired medians. The percentile intervals summarize bootstrap resamples of paired means, as in the underlying analysis; they are not intervals around those medians.

Contrast	Environment	Median Δ BPB	Signs	95% bootstrap CI
Natural > Dyck	WikiText / byte	0.370	5/5	[0.356, 0.373]
Natural > Dyck	WikiText / BPE	0.271	5/5	[0.263, 0.298]
Natural > Dyck	Python / byte	0.518	5/5	[0.460, 0.546]
Retain vocab/head > replace	WikiText / byte	0.331	5/5	[0.318, 0.334]
Retain vocab/head > replace	WikiText / BPE	0.177	5/5	[0.175, 0.183]
Retain vocab/head > replace	Python / byte	0.408	5/5	[0.405, 0.416]
Markov > Dyck	WikiText / byte	0.040	5/5	[0.032, 0.047]
Markov > Dyck	WikiText / BPE	0.071	5/5	[0.069, 0.097]
Markov > Dyck	Python / byte	0.047	5/5	[0.037, 0.087]
Dyck > shuffle	WikiText / byte	0.077	5/5	[0.059, 0.093]
Dyck > shuffle	WikiText / BPE	0.071	5/5	[0.059, 0.086]
Dyck > shuffle	Python / byte	0.179	5/5	[0.144, 0.231]

H Run and compute accounting

Phase	Scientific runs	Train model-tokens	Evaluation target-tokens	Estimated FLOPs	Recorded GPU hours
Locked pilot	170	435,814,400	50,135,040	6.527×10^{14}	0.253

Phase	Scientific runs	Train model-tokens	Evaluation target-tokens	Estimated FLOPs	Recorded GPU hours
Untouched confirmation	105	275,251,200	30,965,760	4.119×10^{14}	0.167
Untouched scale validation	90	235,929,600	26,542,080	6.371×10^{14}	0.158
Total frozen scientific	365	946,995,200	107,642,880	1.702×10^{15}	0.578

Engineering smoke and post-freeze reproducibility runs are logged separately and excluded from scientific effect estimates.

I Integrity and reproducibility audit

Audit area	Outcome
Frozen scientific set	365/365 trajectories pass; no scientific key is duplicated or missing; all summaries contain finite metrics and the registered checkpoints.
Pairing and compute	Initial model-state hashes match within paired cells; replacement hashes match across replacement conditions; full-condition model-token counts and analytic FLOPs match within model size.
Registry and determinism	Registry, ledger, and filesystem reconcile 421/421/421; the only repeated signature is the intentional determinism pair; forced reruns reproduce model states and numerical trajectories.
Independent reconstruction	A clean environment passes unit tests, token-space separation, and an end-to-end smoke run. Raw licensed corpora and token arrays remain outside the reviewer bundle.

J Post-freeze stage-one capability diagnostic

This new scientific extension was specified after inspection of the frozen downstream results and locally timestamped before its checkpoint predictions. It is not a public preregistration or an independent human review. Selection includes every saved disjoint-vocabulary Dyck replaced, retained, and optimizer-continuation checkpoint and every matched exact-shuffle and first-order-Markov checkpoint. The overlap-vocabulary diagnostic uses a different token map and is outside this selection. None of the 175 eligible checkpoints was missing. Stored checkpoint SHA-256, model-state metadata, and seed-reconstructed initial-state hashes all matched. Different downstream-environment records with identical stage-one state are retained as aliases, not independent replications.

The independently seeded test corpus comprises 1,024 documents with 8 to 32 parenthesis pairs, 128 types, maximum stack depth 24, and opening probability 0.54. Each complete document starts at position zero, unlike arbitrary windows across a packed training stream. Thus, it tests bounded document-conditioned behavior under the original generator, not the exact distribution of packed training windows, longer-context extrapolation, or unrestricted formal-language recognition. Every scored matching opening is visible. The exact stack validates target identities and labels distance and depth; it is not provided to the model. Unigram and first-order predictors use a separately generated 200,000-token calibration stream.

There are 20,555 closing events, of which 10,425 have matching-opening distance greater than one. Conditional closing loss renormalizes over the 128 closing classes and is expressed in bits per closing event, not bits per byte. Unconditional next-token loss over the 257 active procedural classes is

recorded separately. Adjacent events are reported but excluded from the primary diagnostic because a first-order predictor can exploit them directly. The first-order trained models achieve perfect adjacent closing-type accuracy, showing why aggregate adjacent-plus-nonadjacent accuracy would be misleading.

Phase and state	Independent seeds	Nonadjacent closing bits	Nonadjacent accuracy
Confirmation, initial	5	7.018	0.76%
Confirmation, exact shuffle	5	6.956	1.60%
Confirmation, Markov	5	7.033	0.75%
Confirmation, Dyck	5	5.925	12.93%
Larger model, initial	5	7.025	0.78%
Larger model, exact shuffle	5	6.902	2.45%
Larger model, Markov	5	7.035	0.82%
Larger model, Dyck	5	5.136	22.72%

All five pilot Dyck seeds show the same ordering against their initial, shuffled, and Markov states. Pilot, confirmation, and larger-model results are reported separately rather than pooled as homogeneous replicates. The optimizer-continuation states are also retained but are not relabelled as the matched source-only contrast because their training schedule differs.

Distance-specific results retain 4,978 events at distance 3 to 7; 2,140 at 9 to 15; 2,008 at 17 to 31; and 1,299 at 33 to 63. At distance 33 to 63, confirmation Dyck accuracy is 3.56% and the larger-model value is 5.40%, compared with 0.74% and 0.66% for the corresponding Markov controls. These low absolute long-distance accuracies limit any claim of robust stack execution. The probe distinguishes the learned states on this prediction task; it does not uniquely identify the algorithm or causal representation used by a model.

All 70 unique trained-or-initial states completed evaluation. Every event row, document-level loss sum, stratum count, and reported mean was then recomputed from stored outputs by a separate arithmetic implementation without calling the scoring helper. The audit verified all hashes, finite values, denominator sizes, source identity, and preservation of the frozen trajectory registry. Automated implementation checks cover nested matching, malformed closings, the exact 7-bit uniform reference, and fresh generator fixtures. These are software and arithmetic checks, not independent human evaluations. The evaluator's zero CUDA-allocation telemetry is treated as unavailable rather than interpreted as zero memory or energy use. The extension contains no training or external model API calls.

K Reduced-scale reference-regime bridge

K.1 Specification and comparison boundary

The extension follows the inspected ordered-UNION task and configuration of the procedural-mixture study (Jiang et al., 2026), pinned at upstream commit e97ffe3c30deb97d73160fe8a7c5da18024d0e8d. It independently implements that published task definition rather than redistributing upstream code for which no explicit license file was found. Thirty-two symbols drawn from values 1 through 99 form two 16-symbol inputs. A separator divides the inputs, and the output lists their ordered union by first occurrence. Only output-symbol predictions contribute procedural loss; padded positions do not. The reassociated control rotates complete target unions between batch rows. It preserves target and mask marginals within each batch while breaking input-target association, not structure within the target sequence.

The decoder has four layers, width 128, four attention heads, an untied 1,024-row vocabulary/head, context capacity 256, and 1,089,280 parameters. All four arms receive 120 stage-one updates at batch 32 and sequence length 66, then 10,000 byte-level WikiText-2 updates at batch 32 and

context 256. This is 253,440 stage-one and 81,920,000 stage-two model-tokens per cell. Under the reference's twice-input-length accounting, the stage-one-to-stage-two ratio is 0.3%; actual padded model-tokens give 0.3084% of total training. The UNION arms supervise fewer stage-one targets than the natural arms. Model-token matching is not a claim of identical useful targets, objective, or measured hardware FLOPs.

The optimizer is AdamW with learning rate 0.0005, betas (0.9, 0.999), epsilon 1e-8, weight decay 0.1, and gradient clipping at 1.0. Replaced arms use constant stage-one learning rate, deterministically reset vocabulary/head and optimizer, and apply downstream cosine decay to a 0.1 minimum ratio. The continuous-natural comparator retains parameters and optimizer and uses one cosine schedule across both stages. The latter is a bundled practical protocol, not an isolated data-source intervention.

Final validation BPB is the locked primary endpoint. The secondary normalized AULC uses observations at downstream updates 0, 100, 500, 1,000, 2,500, 5,000, 7,500, and 10,000. Each evaluation uses the same 256 fixed 256-target-byte blocks. Initial model states and downstream sampling are paired within seed. Compared with the published reference, this is a much smaller untied decoder, byte-tokenized WikiText-2 rather than GPT-2-tokenized C4, and a shorter context and procedural stage. It is explicitly a reduced adaptation, not an exact reproduction.

K.2 Results, resource failures, and retrospective completion

The original local specification fixed seeds 1511, 1522, and 1533 and all four arms before training. It also fixed 600 seconds per cell and 6,000 seconds for the study. The first two full blocks completed. The original third seed stopped after 9,059 downstream updates for UNION, 9,761 for reassociation, and 2,190 for natural replacement; continuous natural had no logged update before the overall boundary. No partial value substitutes for a final endpoint. All four failure records and partial logs remain intact.

A separate resource-only specification was recorded after those outcomes. It reran every third-seed arm from its original initialization with the unchanged scientific runner, allowing 750 seconds per cell and 3,000 seconds overall. The four completed in 2,297.7 seconds. This third block is post-outcome resource completion, not retroactively preregistered or included as though the original run had fully completed prospectively. Even the first two blocks are described as locally specified, not publicly preregistered.

Seed and execution status	UNION	Reassociated	Natural replaced	Natural continuous
1511, original local specification	1.8741	1.9542	1.9410	1.9310
1522, original local specification	1.8655	1.9437	1.9421	1.9336
1533, post-outcome completion	1.8728	1.9668	1.9488	1.9461

UNION improves final BPB over natural replacement by 0.0669 and 0.0766 in the two original complete blocks (mean 0.0718). The post-outcome block gives 0.0761. The three-seed mean of 0.0732 is descriptive only. All three blocks also favor UNION over reassociation and continuous natural. Paired AULC improvements over natural replacement are 0.0637, 0.0880, and 0.0769 respectively. No p-value or population-level replication claim is inferred from these three selected-by-design blocks.

K.3 Exact evidence audit and economics

A second implementation verified all twelve final checkpoint hashes, model-state identities, initial-state reconstruction, 121,440 logged updates, learning-rate schedules, target counts, and eight evaluation points per cell. It independently recomputed learning-curve integrals. It then evaluated every final model on the exact fixed blocks using float64 logsumexp and target gathering instead of the original float32 cross-entropy reduction. The maximum difference from the stored final BPB was 6.04e-8. This is an automated scientific consistency check, not independent human review.

The twelve complete cells consumed 986,081,280 logged model-tokens; failed original attempts consumed another 172,874,240. These 1,158,955,520 tokens are separate from the original 365-trajectory ledger. Guards are evaluated between updates: one original completed cell includes 0.307 seconds of final-evaluation/serialization overhead beyond its nominal 600-second guard. Actual times and that overhead are retained. No energy estimate is inferred from GPU memory or elapsed time.

The positive bridge limits overgeneralization of the frozen negative result. Because procedural family, allocation, model, training duration, and objective change together, it does not identify a unique moderator or a transferred higher-order algorithm. It demonstrates the importance of reporting the tested regime and its controls rather than treating one synthetic family or warmup allocation as representative.