

We Should Consider Educating Models Before Training Them

Educated Pretraining Produces Fundamentally Different Representational Geometry Than Shuffled Pretraining on Identical Data

Declan Michaels Cross-Cultural Alignment Study (CCAS) moral-os.com

Pre-registered: osf.io/2vcq6

Abstract

We trained two identical language models on the same corpus. The only variable was data presentation order: one read the texts in a developmental sequence modeled on classical education; the other read randomly shuffled chunks. The educated model shows a substantially smaller train-validation gap and continued validation improvement at the end of training. The trained model overfits. The effect is large, stable across random seeds, and visible at every transformer layer of the network. These results suggest that thoughtfully ordered training data may produce models that generalize better and incorporate new data more effectively.

1. Introduction

For at least two and a half millennia, humans have educated their children in structured sequences that teach concepts before principles. This motivates testing whether structured ordering of training data affects language model learning.

Language models train the opposite way. Trillions of tokens, shuffled randomly, presented in arbitrary order. Each batch is an unrelated sample from the full distribution, assuming coherent representations will emerge from incoherent presentation.

We tested whether ordering matters.

2. Related Work

2.1 Curriculum Learning

Bengio et al. (2009) introduced curriculum learning, showing that difficulty ordering improves convergence. Subsequent work has focused on ordering within supervised fine-tuning. Zhang et al. (2026) present a systematic investigation of curriculum learning in LLM pretraining, testing difficulty-based ordering across over 200 models on up to 100B tokens and finding 18-45% convergence improvement with up to 3.5% sustained benefit when used as a warmup strategy.

2.2 Data Quality and Curation

Others have shown that higher quality data produces better models with less compute (Gunasekar et al., 2023; Zhou et al., 2023). These efforts focus on what data to include, not what order to present it. Random shuffling before training is standard practice.

2.3 Representational Geometry

Centered Kernel Alignment (CKA; Kornblith et al., 2019) measures similarity of learned representations between models. Models trained independently by different organizations can converge to similar geometry, suggesting that representational structure may be more determined by architecture and training procedure than by specific data content.

2.4 Dimensional Collapse

Dimensional collapse occurs when learned representations occupy a low-dimensional subspace of the available space, wasting capacity. Jing et al. (2022) characterized this in contrastive self-supervised learning. Fan et al. (2025) address dimensional collapse in LLM pretraining through diversified data selection. We examine its relationship to data presentation order. Sparse autoencoders (SAEs) decompose dense activations into interpretable sparse features (Cunningham et al., 2023; Bricken et al., 2023) and provide a complementary lens on how models use their representational capacity at each layer.

3. Experimental Design

3.1 Corpus

We assembled 113 public-domain texts spanning seven developmental stages modeled on the classical trivium, expanded to include a pre-linguistic causal substrate. Some texts appear in multiple stages with different aspects emphasized, as in classical education (e.g., Galileo’s *Dialogue* appears in Stage 0 for observational content and Stage 5 for methodological argument). The 113 count reflects corpus entries, not unique works:

Stage 0, Physical World (15 texts): Euclid’s Elements, Marcellus’s Conversations on Natural Philosophy, Faraday, Darwin, Newton’s Principia (selections), Galileo. Intended as a foundation of physical and causal reasoning before introducing abstract concepts.

Stage 1, Foundation Fables (15 texts): Aesop, Grimm, Jataka Tales, Panchatantra, La Fontaine, Arabian Nights. Establishes moral anchors through narrative.

Stage 2, Grammar / Ancients (18 texts): Homer (Iliad, Odyssey), Plato (Republic), Aristotle (Ethics, Politics), Virgil, Augustine.

Stage 3, Logic / Medieval-Renaissance (24 texts): Bede, Dante (Divine Comedy), Chaucer, Milton, Locke, through Marx.

Stage 4, Rhetoric / Modern (15 texts): Melville, Dostoevsky, Tolstoy, Twain, through Forster.

Stage 5, Science (9 texts): Hippocrates, Aristotle’s Physics, Galileo, Newton, through Einstein.

Stage 6, Drama and Poetry (17 texts): Aeschylus, Sophocles, Shakespeare (five plays), through Dickinson.

All texts are public domain, sourced from Project Gutenberg, the Internet Archive, and MIT Classics. Primary sources only; no modern commentary. Total corpus: 20.7 million tokens. We trained a SentencePiece BPE tokenizer with an 8,000 word vocabulary on the corpus. Full corpus listing in Appendix A.

3.2 Conditions

Sequenced (Condition A): The corpus is assembled in stage order. The model reads front-to-back. At the end of the corpus, it wraps to the beginning.

Shuffled (Condition B): The corpus is divided into 2,048-token chunks, randomly shuffled. Each training batch samples randomly from this pool. This replicates standard pretraining practice.

Both conditions share identical validation sets: 5% of chunks selected by content hash, ensuring no overlap with either training set. The validation set samples from all stages.

Epoch structure. At 5,000 iterations \times 65,536 tokens per iteration, total training volume is approximately 328 million tokens on a 20.7-million-token corpus, roughly 16 passes. The continuation model (5,000 + 5,000 steps) sees roughly 32 passes total, as does the 10,000-step extended sequenced run. The sequenced model encounters the curriculum order on every pass. The shuffled model sees a different random permutation on each pass. Both models see every document approximately the same number of times in expectation. The difference is not repetition volume but whether documents are encountered in developmental order or arbitrary order.

3.3 Architecture and Training

All models use GPT-2 small (12 transformer layers, 12 heads, 768 embedding dimension, \sim 91.2M parameters) via nanoGPT (Karpathy, 2023). The parameter count is lower than the standard GPT-2 small (117M) due to our smaller vocabulary (8,000 vs 50,257 tokens), which reduces the embedding and output projection layers. We train with AdamW ($\beta_1=0.9$, $\beta_2=0.95$, weight decay 0.1, gradient clipping 1.0) in bfloat16 on rented GPUs (NVIDIA A40, A100, A6000). Effective batch size is 65,536 tokens for all runs.

LR schedules. The 5,000-step confirmatory runs use lr 3e-4 cosine-decayed to 3e-5 over 5,000 steps with 200-step warmup. The continuation model (Section 5.2) and 10,000-step extended run (Section 5.5) use the same peak and minimum LR but decay over 10,000 steps, so the learning rate at step 5,000 is approximately 1.7e-4 rather than the minimum. The continuation model resumes from the educated checkpoint at step 5,000 with the optimizer state intact.

3.4 Measurements

Generalization gap: Difference between training loss and validation loss at each evaluation checkpoint (every 250 iterations).

CKA: Linear CKA (where 1.0 indicates identical geometry) between sequenced and shuffled models at each of 12 transformer layers, using the same 54 concept word inventory across three domains used in our cross-model representation geometry work. We feed each concept as a bare word and record the mean-pooled hidden state at every layer. Significance assessed by permutation test (10,000 permutations). Bootstrap 95% CIs (10,000 iterations).

Domain silhouette: Whether concepts cluster by domain at each layer. Cosine distance, three-domain partition. Positive silhouette indicates domain structure; negative indicates the model is placing concepts closer to the wrong domain’s centroid. Bootstrap 95% CIs.

Effective dimensionality: Number of singular value components needed to explain 90% of variance in the concept activation matrix at each layer. Participation ratio as a complementary measure.

Feature activation: Sparse autoencoders (16x expansion, 12,288 features) trained per layer on 200,000 tokens of corpus activations. L1-penalized (coefficient 1e-3), 5 epochs. Alive features defined as those firing on >1% of inputs. Measures how much of the available feature space the model uses at each layer.

3.5 Pre-Registration

We pre-registered hypotheses, analysis methods, and success criteria on the Open Science Framework prior to training (osf.io/2vcq6). Four hypotheses: H1 (geometry differs, CKA < 0.90 at majority of layers), H2 (educated model retains stronger foundational clustering), H3 (educated model produces tighter three-domain clustering), H4 (educated model produces more coherent text, exploratory).

4. Confirmatory Results

The following analyses were pre-registered and executed as specified.

4.1 Seed Stability

Across five random seeds (1337, 2024, 4242, 7777, 9999) using an earlier version of the curriculum (three texts were added for the final corpus), the educated model’s generalization gap is 0.10 ± 0.005 and the trained model’s gap is 1.56 ± 0.01 . Seed variance is small relative to the between-condition difference.

4.2 The Generalization Gap

Run	Steps	Train	Val	Gap
Educated	5,000	4.547	4.573	0.026
Trained	5,000	2.047	3.521	1.47
Continuation (5K+5K)	10,000	3.043	3.504	0.46
Extended educated	10,000	3.615	3.969	0.35

The trained model’s validation loss peaked at step 2,500 and degraded to 3.52 by step 5,000; the gap widened continuously after that point. The educated model’s validation loss continued improving at step 5,000 with a gap of 0.026. The extended educated model’s gap widened to 0.35 by step 10,000 but validation loss was still falling. The continuation model’s gap reached 0.46 by step 10,000 with validation loss also still falling.

4.3 H1: Geometry Differs (Confirmed)

The two training orders produce measurably different internal representations. Mean cross-condition CKA across 12 transformer layers is 0.562. All fall below the 0.90 threshold.

Per-layer CKA ranges from 0.493 at layer 3 to 0.672 at layer 12, with the embedding layer at 0.984. The geometry diverges most at early layers and gradually reconverges toward the output. This pattern may reflect early-layer representations being shaped differently by data ordering while output layers converge toward the shared vocabulary prediction task.

Permutation test at layer 6: observed CKA = 0.540; null mean = 0.080 +/- 0.023; $p < 0.001$ (0 of 10,000 permutations exceeded observed value).

H1 confirmed: CKA < 0.90 at all 12 transformer layers. Permutation test at layer 6 (mid-network): observed CKA = 0.540, null mean = 0.080 ± 0.023, $p < 0.001$ (0 of 10,000 permutations exceeded observed value).

4.4 H2: Foundational Retention (Directional, Not Decisive)

The educated model maintains stronger physical-moral domain clustering at 9 of 12 transformer layers. The advantage is concentrated in early-to-mid layers (L1 through L8), where the curriculum builds foundational representations before encountering later-stage material. The educated model's foundational silhouette peaks at 0.179 (L4); the trained model's peaks at 0.154 (L5).

Bootstrap CIs overlap at mid-layer, so the effect is directionally consistent but not statistically decisive at 54 concepts.

4.5 H3: Domain Clustering (Confirmed)

The educated model produces stronger three-domain clustering at all 12 transformer layers. The educated model maintains positive silhouette throughout (0.005-0.057), indicating consistent domain organization at every depth. The trained model's clustering weakens to 0.003 at layer 3, approaching zero.

The educated model's peak clustering occurs at layers 4-5 (silhouette 0.057-0.056), corresponding to the transition from foundational concepts to more abstract material. The trained model never exceeds 0.028.

H3 supported: the educated model produces positive domain silhouette at all 12 transformer layers, compared to near-zero or negative values for the trained model. The separation is small in absolute terms (educated peak 0.057, trained peak 0.028) but consistently directional.

4.6 Effective Dimensionality

The two models use network depth differently.

Educated: Compresses from 32 effective dimensions at the embedding to 14 at layer 1, then expands progressively to 25 by layer 10. Effective dimensionality increases at every layer from input to output. Participation ratio remains between 10.6 and 14.0 throughout, indicating variance is distributed across many dimensions.

Trained: Starts high (35 at the embedding) then collapses to 10-13 effective dimensions for seven consecutive layers (L3-L9). Participation ratio drops to 1.4-1.6 at mid-layers, meaning the representation is dominated by one or two dimensions. The expansion to 32 at the output layer spreads a compressed representation back out.

5. Exploratory Results

The following analyses were not pre-registered. They were motivated by the confirmatory findings and are presented as exploratory.

5.1 Architecture Interaction (AttnRes)

We trained an AttnRes variant (~92.8M parameters) that replaces standard residual connections with cross-layer attention residuals (Chen et al., 2026): each transformer block attends over all previous layers' outputs to determine its residual, with Pre-Norm applied to queries and keys.

Model	Train	Val	Gap
Standard educated	4.547	4.573	0.026
Standard trained	2.047	3.521	1.47
AttnRes educated	4.470	4.551	0.08
AttnRes trained	2.062	3.525	1.46

In this setup, AttnRes produces a marginal improvement for the educated model (val 4.551 versus 4.573) at 2.5x the compute cost per iteration. It provides no benefit to the trained model (gap 1.46 versus 1.47).

CKA analysis confirms: AttnRes educated versus trained shows mean CKA 0.556 compared to 0.562 for the standard architecture. The AttnRes trained model develops negative domain silhouette at layers 3-4, which the standard trained model does not.

In this comparison, the curriculum effect dominates the architecture effect.

5.2 Geometric Durability

To test whether curriculum-ordered geometry survives subsequent unstructured training, we took the standard educated model at step 5,000 and continued training for an additional 5,000 iterations on shuffled data. The continuation run used a 10,000-step LR decay schedule (Section 3.3), so its step-5,000 values differ slightly from the confirmatory educated run which used a 5,000-step schedule.

Step	Train	Val	Gap
5000	4.530	4.581	0.051
6000	3.823	3.916	0.09
7000	3.433	3.665	0.23
8000	3.264	3.569	0.30
9000	3.120	3.520	0.40
10000	3.043	3.504	0.46

The continuation model reaches validation loss comparable to the trained model (3.504 vs 3.521) but with a gap of 0.46 versus 1.47 — it learned the same material with far less overfitting. The gap widened slowly through the shuffled phase, stabilizing between steps 7,500-8,500 before creeping to 0.46 by step 10,000. Validation loss was still falling at step 10,000, though the continuation model has seen twice the total compute (10,000 vs 5,000 steps).

CKA comparison reveals what survives.

Comparison	Mean CKA
Continuation vs Trained	0.860
Educated vs Continuation	0.736
Educated vs Trained	0.562

The continuation model’s overall geometry is closer to the trained model than to its educated parent. Shuffled training is overwriting the curriculum geometry. But domain clustering tells a different story: the continuation model’s silhouette remains stronger than the trained model’s at all 12 transformer layers. Under our concept-word probe, domain clustering differences persisted more than overall geometric similarity.

Dimensionality reinforces this finding. The continuation model expands progressively from 19 to 28 effective dimensions, surpassing the 5K educated model’s peak of 25, without the trained model’s collapse to 10.

5.3 Sparse Autoencoder Analysis

To examine representational structure at the feature level, we trained sparse autoencoders (SAEs) with 16x expansion (12,288 features) on hidden state activations at all 12 transformer layers of four models: educated, trained, continuation, and the 10K extended educated model. SAEs decompose dense hidden states into sparse features; the number of “alive” features (those firing on >1% of inputs) provides a heuristic proxy for how much of the feature space the model uses at each layer.

Layer	Educated	10K Educated	Trained	Continuation
L1	1,305	931	2,087	466
L2	1,985	1,258	2,694	531
L3	2,006	1,929	3,863	772
L4	2,269	1,923	1,156	712
L5	2,585	2,109	1,249	725
L6	3,081	2,354	1,083	1,511
L7	3,184	3,012	1,630	2,602
L8	3,402	3,537	2,374	3,335
L9	4,366	4,584	3,625	3,693
L10	5,529	5,854	5,652	4,352
L11	6,601	6,487	11,275	4,790
L12	7,445	7,956	12,288	10,165

The **educated model** shows progressive feature expansion: 1,305 to 7,445 alive features from L1 to L12. Every layer develops more active features than the one before it.

The **10K educated model** preserves this progressive pattern but compresses early layers further (931 at L1 versus 1,305) while expanding late layers (7,956 at L12 versus 7,445). Additional passes through the curriculum tighten the encoding at early layers and enrich the representations at later ones.

The **trained model** drops to 1,083 alive features at L6, then activates all 12,288 features at L12.

The **continuation model** shows fewer alive features at L1-L5 (466-725) than either parent, smooth expansion from L6 onward, and no mid-network drop. At L12 it activates 10,165 features, between the educated model's 7,445 and the trained model's 12,288.

Full feature activation dashboards and per-concept analysis are available in the interactive viewer at moral-os.com.

5.4 Qualitative Generation (H4, Exploratory)

We sampled from all four models using identical prompts and random seed to observe whether geometric differences appear in generation. These are 91M-parameter models; none produces fluent text. The differences are in organizational structure, not output quality.

Prompt	Educated (5K)	Trained (5K)	Continuation (5K+5K)	Educated (10K)
The nature of justice	Biblical verse numbering with body/power vocabulary in repetitive loops	Theological-philosophical register (Divine will, nature of universe) with some topical coherence	Political philosophy (law, citizen, slave, master) with sustained argument about law and nature	Abstract philosophical reasoning about matter, form, and the nature of things without looping
Water flows downhill because	Immediate drift to unrelated content, does not maintain physical domain	Correct domain start (air, water, particles) in Marcet dialogue format, then drifts to windows and pipes	Physical domain start, drifts to astronomy but maintains narrative coherence throughout	Drifts to moral argument but maintains coherent sentence structure
The fox spoke to the	Biblical genealogy and geography listing	Fairy tale register (king, father, sword) with sustained narrative and dialogue	Homeric epic register (Aeacus, Priam, Trojans, Minerva) with sustained battle narrative	Pseudo-poetic register with invented compound words (braves-time, robeeking, blossyramids)
A citizen must	Number string collapse, then body/force repetitive loop	Political philosophy (citizen, magistrate, government) with sustained argument	Political philosophy (democracy, republic, State) with sustained multi-clause argument	Abstract philosophical argument about things and nature with premise-conclusion structure

The educated model tends toward repetitive loops and number string collapse. The trained model selects appropriate registers and vocabulary but drifts from topic. The continuation model maintains the strongest narrative coherence and produces sustained multi-clause arguments in domain-appropriate registers. The 10K educated model produces abstract philosophical reasoning without the looping of the 5K model, and generates invented but morphologically plausible vocabulary in poetic contexts. Full generation samples are in Appendix D.

5.5 Extended Sequenced Training (10,000 Steps)

To control for the continuation model’s 2x compute advantage, we trained the educated model for 10,000 steps on pure sequenced data — the same total compute as the continuation model, but without the transition to shuffled training.

Loss trajectory. Validation loss continued falling through all 10,000 steps, reaching 3.969 at step 10,000 (gap 0.353). The model was still learning at 32 passes through the corpus.

Step	Train	Val	Gap
5,000	4.561	4.601	0.040
6,000	4.132	4.476	0.344
7,000	4.018	4.294	0.276
8,000	4.001	4.128	0.127
9,000	3.542	4.078	0.537
10,000	3.615	3.969	0.353

Train loss fluctuates between steps 9,000 and 10,000 (3.542 \rightarrow 3.615) as the model wraps through the curriculum; loss varies by stage difficulty. Validation loss continued falling throughout. The extended run also used a 10,000-step LR decay schedule, so its step-5,000 values differ slightly from the confirmatory 5,000-step run.

CKA. The 10K educated model diverges further from the trained model (mean CKA 0.514) than the 5K model did (0.562). More sequenced training increases geometric divergence. The 10K model is geometrically closer to the continuation model (0.706) than to the trained model, consistent with shared educated foundations.

Comparison	Mean CKA
10K-Educated vs 5K-Educated	0.770
10K-Educated vs Continuation	0.706
10K-Educated vs Trained	0.514

Domain clustering. The 10K educated model shows tighter domain clustering than the trained model at all 12 transformer layers. However, it shows looser clustering than the 5K educated model at 11 of 12 transformer layers. The tightest domain clusters appear around step 5,000; additional passes through the curriculum make them more diffuse. This may reflect increasing cross-domain connections as the model encounters the same texts repeatedly — concepts that start in one domain acquire features from others.

Dimensionality. The 10K educated model has the highest effective dimensionality and participation ratio of all four models at every transformer layer.

Layer	5K Educated	10K Educated	Trained	Continuation
L1	14 (PR 10.4)	20 (PR 15.2)	33 (PR 13.8)	19 (PR 7.4)
L6	17 (PR 10.6)	27 (PR 18.6)	15 (PR 1.6)	24 (PR 6.2)
L9	21 (PR 11.9)	32 (PR 21.7)	11 (PR 1.5)	26 (PR 5.5)
L12	24 (PR 12.5)	33 (PR 20.6)	30 (PR 4.5)	28 (PR 7.8)

The 10K model expands from 20 to 33 effective dimensions with participation ratios above 15 throughout — variance is distributed broadly across many dimensions. The continuation model reaches comparable dimensionality (19-28) but with much lower participation ratios (5.5-7.8), meaning a few dimensions dominate. The trained model collapses to 11-12 effective dimensions at mid-network with participation ratios near 1.5.

SAE features. The full comparison is in Section 5.3. The 10K model preserves the progressive expansion pattern with a subtle shift: fewer alive features at early layers (L1-L7) and more at late layers (L8-L12) compared to the 5K model. Additional passes through the curriculum compress early-layer representations while enriching later ones.

Qualitative generation. The 10K model shows more structured output than the 5K model. “The nature of justice” sustains abstract philosophical argument about matter and form rather than looping. “A citizen must” produces premise-conclusion structure. “The fox spoke to the” generates pseudo-Homeric poetic register with invented but morphologically consistent vocabulary. The additional training produces more coherent generation while maintaining domain-appropriate register.

6. Discussion

6.1 The Generalization Gap as a Training Signal

The educated model’s higher absolute loss obscures its generalization advantage.

Consider how training runs are monitored. A researcher launches two experiments. After 1,000 iterations, one shows loss of 3.2 and falling fast; the other shows 4.8 and falling slowly. Loss descent rate selects for whatever strategy reduces prediction error fastest, and memorization is faster than understanding.

The generalization gap provides information that loss alone does not. A model whose gap remains stable can indicate it is learning structure that transfers to unseen data. A model whose gap widens can indicate it is fitting patterns specific to its training set. The gap does not replace loss as a training signal. It complements it by indicating whether what the model is learning will generalize.

The generalization gap requires no additional computation beyond the validation loss already being recorded. It is freely available and informative at every checkpoint.

6.2 Education Versus Training

The dimensionality and SAE data suggest a distinction between what shuffled training and curriculum ordering produce. The trained model’s mid-network layers carry a low-dimensional signal with few active features. Its output layers compensate by activating maximum capacity. The educated model uses each layer progressively, adding dimensions and features from input to output. The 10K educated model takes this furthest: highest dimensionality and highest participation ratio of all four models at every transformer layer, meaning it distributes variance most broadly across its representational space.

The CKA data shows two kinds of structure with different durability. The overall geometry moved substantially toward the trained model. The domain organization did not. Curriculum ordering builds a resilient organizational structure: the overall geometry can change as the model learns new material, but the relationships between concepts persist.

The 10K run adds a nuance: domain clustering peaks around 5,000 steps and becomes more diffuse with additional passes. The 10K educated model has weaker silhouette scores than the 5K model despite higher dimensionality and stronger CKA divergence from the trained model. The clusters loosen while the representational richness increases. This may reflect increasing cross-domain connections as the model develops a more integrated representation of the curriculum material.

When the educated model absorbs unstructured data in the continuation run, the result is richer than either parent alone. The continuation model’s dimensionality exceeds both the educated and trained models, though its participation ratio remains lower than the 10K educated model’s. Education and training are not alternatives but a sequence, and the foundation determines what subsequent training can build. The 10K pure sequenced model shows that the curriculum alone continues producing value; the continuation model shows that the educated foundation changes what shuffled training can build.

6.3 Implications for Pretraining Practice

The continuation data shows that structured pretraining and unstructured training can be complementary in this setup. The foundation built during the curriculum phase shaped how the model absorbed subsequent unstructured data. The 10K pure sequenced model built the richest representations of all four models — highest dimensionality, highest participation ratio, most structured qualitative output — suggesting that the curriculum continues producing value well beyond the point where loss-based evaluation would stop the run.

In this experiment, the trained model’s mid-network SAE layers activate very few features, suggesting that shuffled training may underutilize available model capacity. Whether structured pretraining reduces post-training alignment requirements, improves retention of existing knowledge, or achieves equivalent capacity at smaller parameter count are open questions for future work.

6.4 Limitations

Scale. Our models are 91M parameters on 20.7M tokens. We do not know whether the effect persists, shrinks, or reverses at larger scales.

No downstream evaluation. All generalization claims rest on train-validation gap and geometric probes. The model size and pretrained-only status preclude convincing downstream task evaluation. Probe results are available in the interactive viewer.

Seed coverage. Seed stability was demonstrated for the generalization gap. CKA, silhouette, dimensionality, and SAE analyses were conducted on a single seed. Checkpoints for multi-seed CKA analysis exist but have not yet been run.

Concept probe. CKA and silhouette results are based on 54 bare concept words in three hand-selected domains. This narrow probe was chosen for consistency with our prior cross-model work and to avoid instruction-following confounds, but results may depend on concept selection. Single decontextualized words may not reflect how the model processes concepts in running text.

Curriculum as confound. The experiment tests one specific developmental ordering against shuffled presentation. It cannot distinguish whether the effect comes from any structured ordering, this particular ordering, or simply from encountering related texts in proximity. Alternative orderings (reverse, random-with-locality, difficulty-based) were not tested.

Continuation compute. The continuation model trained for 10,000 total steps versus 5,000 for the trained model. The 10,000-step pure sequenced run (Section 5.5) controls for this difference.

Validation set. The 5% held-out split is a within-corpus chunk-level holdout, not an out-of-distribution generalization test. Removing chunks from the sequenced condition may disrupt curriculum structure in ways that do not affect the shuffled condition, making this a conservative test of the sequenced model’s performance.

Corpus. The curriculum is Western-centric and small (21M tokens of curated classical texts), unlike modern pretraining corpora. Replication across Chinese, Arabic, and Sanskrit/Pali classical traditions is planned.

Architecture. Primary results use GPT-2 small. AttnRes provides a second data point within the transformer family. Replication across fundamentally different architectures (state-space models, recurrent architectures) is planned.

Oral traditions. Sub-Saharan African intellectual traditions are primarily oral and therefore underrepresented in any text-based curriculum. This is a limitation of our text-based approach.

SAE methodology. Our SAE analysis uses a simple L1-penalized autoencoder with 16x expansion with an alive threshold of >1% of inputs. More sophisticated approaches or different thresholds might reveal additional structure. The analysis is exploratory.

Hardware. Runs were performed on rented GPUs (A40, A100, A6000) as available. Conditions were not matched across hardware types; minor numerical differences may exist that we cannot quantify.

Step count. The 5,000-step training duration was pre-registered, not optimized. The extended 10,000-step run (Section 5.5) confirms the educated model continues learning well beyond the pre-registered duration.

Cost. Total GPU rental for all runs including seed replication, continuation, AttnRes, SAE training, and the 10,000-step sequenced run was \$30.05. Analysis was performed locally.

7. Future Work

Architecture-general replication. Replicating the core experiment on Mamba (state-space model) and RWKV-7 (recurrent architecture). If the curriculum ordering advantage holds across transformer, state-space, and recurrent architectures, the effect is a property of structured learning itself.

Multi-tradition corpora. Parallel curricula in Classical Chinese (following Zhu Xi’s ordering of the Four Books and Five Classics), Classical Arabic (following the madrasa sequence), Sanskrit (following the Brahmanical sequence), and Pali (following the Theravada Buddhist canon), each in native language, each following its own tradition’s pedagogical ordering.

Novel domain ingestion. Testing whether a curriculum-trained model can absorb content from entirely new domains while maintaining its geometric structure.

Scaling. Training curriculum-ordered models at 1B and 7B parameters to determine whether the effect holds and whether curriculum-ordered models require fewer parameters for equivalent effective capacity.

8. Conclusion

In many aspects of life the best course of action is counter-intuitive. A model whose loss descends slowly appears to be failing, but may be building something a faster model cannot. More data is the industry standard, but the order of presentation may matter more than the volume. We trained two identical models on the same data. The one that looked worse on the standard metric showed better generalization. The one that looked better overfit. We should consider educating models before training them.

References

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *ICML*. Introduced curriculum learning, showing that presenting training examples in order of increasing difficulty improves convergence.

Bricken, T., Templeton, A., Batson, J., et al. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Anthropic*. Demonstrated that sparse autoencoders can decompose transformer activations into interpretable features.

Chen, G., Zhang, Y., Su, J., et al. (2026). Attention residuals. *arXiv:2603.15031*. Proposed the AttnRes architecture, adding residual connections within the attention mechanism.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv:2309.08600*. Showed that SAEs extract monosemantic features from language model hidden states.

Fan, Z., Du, S., Hu, S., Wang, P., Shen, L., Zhang, Y., Tao, D., & Wang, Y. (2025). Combatting dimensional collapse in LLM pre-training data via diversified file selection. *arXiv:2504.20644*. Addressed dimensional collapse during pretraining through data selection strategies.

Gunasekar, S., Zhang, Y., Anber, J., et al. (2023). Textbooks are all you need. *arXiv:2306.11644*. Demonstrated that high-quality curated data can produce strong models at smaller scale, motivating attention to data quality over quantity.

Jing, L., Vincent, P., LeCun, Y., & Tian, Y. (2022). Understanding dimensional collapse in contrastive self-supervised learning. *ICLR*. Characterized dimensional collapse in learned representations and its relationship to training dynamics.

Karpathy, A. (2023). nanoGPT. *GitHub*. Minimal GPT-2 implementation used as our training framework.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *ICML*. Introduced linear CKA as a measure of representational similarity between neural networks.

Zhang, Y., Mohamed, A., Abdine, H., Shang, G., & Vazirgiannis, M. (2026). Beyond random sampling: Efficient language model pretraining via curriculum learning. *EACL*. Systematic investigation of curriculum learning in LLM pretraining across over 200 models, finding 18-45% convergence improvement with difficulty-based ordering.

Zhou, C., Liu, P., Xu, P., et al. (2023). LIMA: Less is more for alignment. *NeurIPS*. Showed that a small amount of carefully curated data can produce strong alignment, supporting the importance of data quality and selection.

Appendix A: Curriculum Corpus

Stage 0, Physical World (15 texts): Euclid, *Elements* (Books 1-6); Marcet, *Conversations on Natural Philosophy*; Marcet, *Conversations on Chemistry*; Faraday, *The Chemical History of a Candle*; White, *Natural History of Selborne*; Aristotle, *History of Animals* (selections); Pliny, *Natural History* (selected books); Fabre, *Social Life in the Insect World*; Fabre, *The Wonders of Instinct*; Darwin, *The Voyage of the Beagle*; Darwin, *On the Formation of Coral Reefs*; Hooke, *Micrographia* (selections); Newton, *Principia: Rules and General Scholium*; Galileo, *Dialogue on the Two Chief World Systems* (excerpts); Faraday, *Experimental Researches in Electricity* (selections).

Stage 1, Foundation Fables (15 texts): Aesop, *Fables* (Townsend trans.); Aesop, *Aesop for Children* (Winter ed.); Grimm, *Household Stories* (Hunt trans.); *Jataka Tales* (Babbitt); *More Jataka Tales* (Babbitt); Jacobs, *Indian Fairy Tales*; *The Panchatantra* (Ryder trans.); *Pilpay's Fables* (North/Jacobs); La Fontaine, *Fables* (Wright trans.); *Parables from the Bible* (KJV); Bulfinch, *Mythology*; Plutarch, *Moralia* (selections); Sturluson, *Prose Edda* (selections); *Arabian Nights* (Lang ed.); Jacobs, *Celtic Fairy Tales*.

Stage 2, Grammar / Ancients (18 texts): Epic of Gilgamesh (Thompson trans.); Homer, *Iliad* (Butler trans.); Homer, *Odyssey* (Butler trans.); Hesiod, *Theogony and Works and Days*; Herodotus, *The Histories* (Rawlinson trans.); Thucydides, *The Peloponnesian War* (Crawley trans.); Plato, *The Republic* (Jowett trans.); Aristotle, *Nicomachean Ethics* (Ross trans.); Aristotle, *Politics* (Jowett trans.); Aristotle, *Poetics* (Butcher trans.); Plutarch, *Lives* (Dryden/Clough); Marcus Aurelius, *Meditations* (Long trans.); Augustine, *Confessions* (Pusey trans.); Augustine, *City of God* (selections, Dods trans.); Virgil, *Aeneid* (Dryden trans.); Lucretius, *On the Nature of Things* (Leonard trans.); Ovid, *Metamorphoses* (More trans.); Cicero, *Orations* (Yonge trans.).

Stage 3, Logic / Medieval-Renaissance (24 texts): Bede, *Ecclesiastical History* (Sellar trans.); *Beowulf* (Gummere trans.); Dante, *Inferno, Purgatorio, Paradiso* (Longfellow trans.); Chaucer, *Canterbury Tales*; Machiavelli, *The Prince* (Marriott trans.); More, *Utopia* (Robynson trans.); Montaigne, *Essays* (Cotton trans.); Cervantes, *Don Quixote* (Ormsby trans.); King James Bible (selected books); Milton, *Paradise Lost*; Bunyan, *Pilgrim's Progress*; Locke, *Second Treatise of Government*; Swift, *Gulliver's Travels*; Rousseau, *The Social Contract* (Cole trans.); Paine, *Common Sense*; Wollstonecraft, *Vindication of the Rights of Woman*; Austen, *Pride and Prejudice*; Shelley, *Frankenstein*; Douglass, *Narrative*; Tocqueville, *Democracy in America* (Reeve trans.); Marx and Engels, *The Communist Manifesto* (Moore trans.); Thoreau, *Walden*.

Stage 4, Rhetoric / Modern (15 texts): Melville, *Moby-Dick*; Stowe, *Uncle Tom's Cabin*; Jacobs, *Incidents in the Life of a Slave Girl*; Dostoevsky, *Crime and Punishment* (Garnett trans.); Tolstoy, *Anna Karenina* (Garnett trans.); Twain, *Adventures of Huckleberry Finn*; Conrad, *Heart of Darkness*; Washington, *Up From*

Slavery; Du Bois, *The Souls of Black Folk*; Weber, *The Protestant Ethic and Spirit of Capitalism* (Parsons trans.); Johnson, *Autobiography of an Ex-Colored Man*; Cather, *My Antonia*; Wharton, *The Age of Innocence*; Hesse, *Siddhartha*; Forster, *A Passage to India*.

Stage 5, Science (9 texts): Hippocrates, *On Airs, Waters, and Places* (Adams trans.); Aristotle, *Physics* (Hardie trans.); Lucretius, *On the Nature of Things* (Leonard trans.); Bacon, *Novum Organum*; Galileo, *Dialogue Concerning Two Chief World Systems* (Salusbury trans.); Newton, *Principia* (Motte trans.); Darwin, *On the Origin of Species*; Mendel, *Experiments in Plant Hybridization*; Einstein, *The General Theory of Relativity* (Lawson trans.).

Stage 6, Drama and Poetry (17 texts): Aeschylus, *Agamemnon* (Morshead trans.); Sophocles, *Oedipus the King* (Storr trans.); Euripides, *Medea* (Coleridge trans.); Aristophanes, *The Birds*; Shakespeare, *Richard III*, *A Midsummer Night's Dream*, *Hamlet*, *King Lear*, *The Tempest*; Marlowe, *Doctor Faustus*; Moliere, *Tartuffe* (Page trans.); Ibsen, *A Doll's House* (Sharp trans.); Wilde, *The Importance of Being Earnest*; Chekhov, *The Cherry Orchard* (West trans.); Blake, *Songs of Innocence and Experience*; Whitman, *Leaves of Grass*; Dickinson, *Collected Poems*.

Appendix B: Training Configuration

Architecture: GPT-2 small (12 transformer layers, 12 heads, 768 embedding dimension), ~91.2M parameters. Tokenizer: SentencePiece BPE, 8,000 word vocabulary, trained on corpus. Optimizer: AdamW, lr 3e-4 cosine to 3e-5, warmup 200, weight decay 0.1. Training: 5,000 iterations, 65,536 tokens/iteration, bfloat16. Hardware: NVIDIA A40, A100, A6000 (RunPod). Framework: nanoGPT (Karpathy, 2023). Total compute cost: \$30.05.

Appendix C: Concept Inventory

Physical (18): acceleration, amplitude, buoyancy, conduction, convection, crystallization, density, diffusion, elasticity, erosion, evaporation, friction, magnetism, oscillation, refraction, sublimation, turbulence, viscosity

Institutional (18): arbitration, bureaucracy, citizenship, constitution, census, federation, jurisdiction, legislation, naturalization, parliament, prosecution, ratification, referendum, regulation, republic, sovereignty, tariff, taxation

Moral (18): altruism, compassion, conscience, courage, devotion, dignity, forgiveness, generosity, gratitude, honesty, honor, humility, integrity, loyalty, obedience, sacrifice, tolerance, wisdom

Appendix D: Qualitative Samples

All samples generated with seed 1337, temperature 0.8, top-k 40, max 200 tokens.

Prompt: “The nature of justice”

Educated (5K): The nature of justice is not of the nature, and the body; his life of the body of the body of his son of the body. He who can be called the body, and he will not be a man is not. 59:1419:3 Then the beginning of the work of the body of God, and if he is not by him; and his head in this same place there are

laid hands of the law. 10:2 And it out of God is no man's heart, that it, then we the body, and he shall be a man to be known, and the power. 1: the body that he shall knoweth out of the power of the body, which is no man's in his own hand. 14 And if he shall not be in the body of the body, and the head, and the body shall he that is he that is good things which is to their own house. 1:1:

Trained (5K): The nature of justice is not to be recognized." The "Catripetal Of God," it is true, the Divine will, and therefore can be no greater than the nature of the universe. Why, then, can life be in possession of nothing? It is therefore not to be under the State, since God Himself will come within the world, and will he that we all, by our compact or according to our regeneration, keep His sphere as they go out of a single sphere, and place His unchangeable nature, and the work and light which is being dissolved. We should then, since we do not see that the work itself, being also He is good only in time, but He in all time, now in all times and places of time, of the world, and in all ages. We should therefore consider the nature of the universe (that this particular part of the world is not from God), but that it is from eternity to time to place

Continuation (5K+5K): The nature of justice is not sufficient to be able to govern the law in a law. He is the master of his master, and he is the judge who is not a citizen, or he who is the law, is not a citizen, but the slave is one who is also a slave. But of a slave, which is by nature a slave's not by law; for, whenever he despises any law, he acts before he can be punished for it by the law, which is by his own law. And in this way, I have said the truth, that he is the law over the resting, and that who possesses himself a slave is one, and that he is the law. These things have been thus spoken, and now are we to use the laws, in the hands of the law, according to the law of nature; for the law is a law; which no one can possibly be a slave was in command of any one who could not have the same law also

Educated (10K): The nature of justice is not of the nature of nature, or that which is of the nature of the species of matter. The very nature is, therefore, is not so. But the senses to be considered, but is not the case of the thing; therefore, is to be understood to be a thing. But if it is not proper for the nature of matter, if it is, by which it is either of the matter of itself. And to know that it is not a thing which is a thing which is to be possible, for it is not a thing in the thing— it follows that matter to be a thing, which appears to be said to the thing itself, if it is not of a thing; or, or, by what is the same thing is itself, or whether the matter is or proper to be. But by the thing is either, by the form of itself; this, is either of anything; or if we suppose that, it is the thing; the thing is not

Prompt: "Water flows downhill because"

Educated (5K): Water flows downhill because they are not to him, and then she cannot say, for the next to the time, and the other; they may be more than a man would take a good to see them away from the others. When the other things are no power in this means of it is not a very easily take the case in order that it is necessary to have thought that is, that they are by a particular parts are so as for their lives. The first. This seems to them, and by a very good thing and that they may be a man ought to it is not by the body can be to be a common to be the whole of each other the part of a kind of the power of a power of a very considerable, and the best adapted to be said to the whole, and the other. It is in the power of the other, is the power. There are righter of which he has no doubt the other; and the first to their own the man; the matter, as

Trained (5K): Water flows downhill because they are not the smallest, but the most vivid and vaporuous; but the air is a most vivid and vivid color of air about the reflecting surface of the bottom, and of the water being such that it is in contact with the particles of air. *Emily*. Pray, how is it possible to mix air and water as well as in the air? *Mrs. B.* Yes; the air being less dense then than air, and consequently as well as in the air,

the body, is less dense than light; it therefore generates the heat from being very much heavier. Now, the sound can be produced by coming out of the window from a window, on the window, before the door to-door. Quicks of the window are all light, and the noise is very great; and it is a sort of piping-pipe that can easily pass through the window, and the noise is the most vivid

Continuation (5K+5K): Water flows downhill because they are of the like, and are so far unlikely in the air and the rest of the earth, that they may have more power over this earth. But I was going to the bottom of the earth, and it was the most clear and most perfect of all nature. But of this little creature, when we were thus affected to the moon, and seen that it was in some latent spring. Four and even the stars were reflected from the earth, and which were by the sun, was not only the same thing but the sun in it. I should then say that if they were lighter than sunshine and other stars, for that reason they were not to look like a shadow as star, since, after having made a dark shining light, there would appear a few stars which would have been in the heavens only as just now. Now if they had been visible in the darkness of the heavens, which we should then see their light in the dark. I have also

Educated (10K): Water flows downhill because they are not the only thing but the world is not good in the world. Another is the more is a very natural and more pleasing to a man's mindful than to sayings, and is not the more divine things. But if in this way of argument is not a thing as to be done in usurp to the world alone. But, as we have taught our present we are speaking as far as the world, that it is not a thing which is a thing to know it and a good for us, then we should have thought that we were only a question to be a thing to be seen. And if we say that things which have been done is of a thing from a very same purpose and we have made by ourselves to know, to ourselves we should say that we would have been in this same thing as is a thing. Then if we were speaking, it is not possible that we have seen it in this world, it may be the same which we

Prompt: "The fox spoke to the"

Educated (5K): The fox spoke to the priest, and the other, and, and his army, and the countrymen, and the country. 1:2:23:28 And the city, to the city of the inhabitants of the cities of the city; and the people to destroyer, and the chief people, and the people that the city which were the kings of the inhabitants, and all the army which is there are there be given them, after the countries. 2:21 But by the land, 2:11 And I may be the Levites and the cities thereof, and the land of the cities of the south, and the land of the cities, and the fathers, and the south tribe of the land which were the city, by the city which are the cities of the tribes of the countries, according to the tribes of the kings, and the inhabitants of the land which they had been built in the inhabitants was in the cities, and the house, and the cities, and

Trained (5K): The fox spoke to the king: 'Quick, Ground! I hear thee, thy father; thy father is a man of war, I am a man of many a man.' Then said the king, 'Lord, now I have found thee, who am the son of a man.' 'Yes,' answered the king; 'have thee been by thy name, and thou art a furious man, of the pride of the people, that thou hast come hither.' Then the king cried out, 'Thou shalt be smitten, and thou hast wounded my head with the sword.' 'Then let me go down.' Then the king did so, and took his two thieves away, who were lying dead on the ground, and they fell asleep. But the king said, 'I will soon have been in the same place as my daughter has been with me.' Then he said he: 'O wife, look out upon the mire and the arrows; I will lie as

Continuation (5K+5K): The fox spoke to the battle-head to meet them, and the Argives fought first in battle. He was a vest son of Aecus, who had a son of Priam, and he had been the first to avenge the Achaeans. She was her husband in her stead, and she was a youth of rank and character, and the very same who was her name. He was son of Meleager and Polydeuces, and was mother-in-law, and he was the most likely commander in the army, for he had twelve children, in the presence of the Trojans and of the Trojans. She found Hector in the likeness of the Trojans and Minerva, and he had borne a host from the gate of the Achaeans. He was by far the wall, and made a spear, and he would have been born to the Trojans. Then he made a seat of bronze, and he made himself a great helmet. He was battered and a spear; the point of the

Educated (10K): The fox spoke to the crowd-covered. The bravestime islet-swalled the ring-lock with the heigh and with the halls of many aye heaves the skull, and the rouse away the gore, The robeeking, his backward, and the lions of the lids, the spermaders, the heaves his bristling his life. The growl, the brow-growsy air is but the flaming and his own. The violeton is held. The lone on the ground is a staid; the chiller-bird and the peeps, and the crest is the same, and the piled by the fins the pine. The spade below is in the spok, The blossyramids with a paceous, he draws the boughshaped. A bony and the robe the bubble

Prompt: “A citizen must”

Educated (5K): A citizen must be able to the right, and the right hand, and the body, and the body will be one another. 22222289029770, and is not. 296041923000. 21237501083009; hence the same time by a portion of the earth as the earth, to the body falling by the body may be the body is 7 8851000006000099606201050862029, and 6 96. And hence it is not be the earth’s mouth of the force of the bodies of the body, and in the earth ought to the power is the power. AFred from this power of the body is the force was in the body, and in the bodies; the matter, as

Trained (5K): A citizen must be a man of this society, the majority of his citizens, being a very small magistrate, he must be a man of such government as this one now would have been; and as he must have a particular share in the government, he must be a public citizen, or a citizen of this community, in which the supreme power is the business of him and of the court, whether he be a citizen or a citizen. But if he is a member of society and a public person, whether it be possible, that is, any man should be a citizen, then the member of the city, and that the magistrates should be citizens of the same rank and virtue, there should be a greater benefit in the state than in the city. This also is evident from the different causes of the alteration, and from this it is evident that he has a right to omit this in order to be governed, and perhaps in such a manner it is the greater part the matter of him

Continuation (5K+5K): A citizen must be led to conclude that the state, from an aristocracy of the state of the community, is a republic; in such a democracy the people is not equal to the same right: for, as is a democracy, it is the most general rule of the State to the government. But there is another case in which the supreme power is the best; and as, in a State, a monarch is his private property and a democracy. The form of government is at first brought before a free state: and this is right for the people, and the rest are both in the hands of the state, and the government is the most likely to be the most absolute, and also the power of the people; so that it is impossible to alter the government in his turn, for the administration necessary for the public officers. But the government of the governments has the power of the public affairs, and of the laws; and it is always the most proper and the most necessary the most necessary,

Educated (10K): A citizen must be enignty, and, she cannot say, in the next place, and she is a greater one in each of the world: for though many a man he will not say, nor is not of the cause of the world, but only to take away the truth. But there is a thing in this case where the nature of things is. But, as a thing which has been done, there is given a thing to follow, or by asking whether it is not a thing to perform it, or, or whether it may be the same thing. The same is the matter to be a thing, which is, and if a man is one, if it is not of a thing; as it is not to be given, but that it is a thing, or of the other, is, as in reference to the same nature, has been written, and that which is called aforement in reference to things which are to it; the third is the matter of the