Historical curriculum in blocks for SML training

The CMF-180 experiment on the epistemic order of ancient knowledge

Author: Adrian (Adi) Stan
ORCID: https://orcid.org/0009-0003-1457-5155
SSRN: https://ssrn.com/author=7778480
Date: June 14, 2026
Keywords: Information Gravity Theory IGT, cognitive autonomy, cognitive externalization, automation bias, distributed cognition.

Summary (Abstract)

The paper presents a preliminary experiment on the influence of presentation order on the formation of internal representations in a small language model trained from scratch on a Greco-Latin and documentary corpus temporally limited to 180 AD. The experiment starts from the hypothesis that learning is determined not only by the content of the data, but also by the temporal and structural regime in which this data is presented to the model. Three training regimes were compared on the same corpus, with the same tokenizer, the same architecture, the same validation set and the same number of optimization steps: global random shuffling, rigid chronology and **Block-based historical curriculum** with internal shuffling.

Preliminary results indicate that rigid chronological ordering degrades performance and produces high anisotropy in the embedding space, while the block-based historical curriculum achieves the best *validation loss* of the three regimes tested. This result suggests that a global epistemic order, combined with local internal diversity, can provide a more stable training regime than both completely random shuffling and mechanical chronological ordering. The paper proposes the term “block-based historical curriculum” for this type of exposure regime and interprets it as **a form of metastable learning: ordered enough to preserve the historical direction of knowledge, but locally entropic enough to avoid representation collapse.**

Model training should not be viewed simply as data accumulation, but as the design of an exposure regime. In domains where knowledge has temporal, historical, or epistemic structure, block curriculum can become a relevant method.

Knowledge is not just content. Knowledge is content experienced in a regime.

1. Introduction

Training language models is often treated as a problem dominated by the volume, quality, and diversity of data. However, the order in which examples are presented to the model can influence the optimization process, the trajectory of internal representations, and the stability of learning. The concept of *curriculum learning* was explicitly formulated by Bengio et al. as a strategy inspired by the way humans and animals learn more efficiently when examples are not presented randomly, but in a meaningful order, usually from simpler to more complex concepts. The existing literature shows that the order of data can influence optimization and

generalization in certain machine learning contexts, but the results for *language model pretraining* are not uniform. Recent work reports benefits for strategic ordering or *curriculum learning* in *pretraining*, while other studies find no convincing evidence that a curriculum generally improves *language modeling*. For this reason, **the CMF-180** (**Classical Memory Framework - until 180 AD**) experiment should be interpreted as a controlled result in a specialized historical corpus, not as a universal demonstration of the superiority *of curriculum learning.*

Typically, *curriculum learning* is formulated as an "easy → hard" problem. This paper proposes **a different variant:** not the order of difficulty, but **the historical and epistemic order** of **the emergence of knowledge.** Instead of asking whether the model learns better from simple to complex examples, we ask whether a small model, trained on a temporally limited ancient corpus, develops different representations when the texts are presented to it in an **approximately historical order, compared to a globally random order.**

The motivation for the experiment is twofold. The first is **technical:** under controlled conditions, the order of the data can affect not only the speed of convergence but also the internal geometry of the embeddings. The second is **epistemological:** human knowledge did not emerge as a randomly shuffled database, but through **historical accumulations, layers, transmissions, ruptures, translations, and transformations.** If a model is trained to approximate a historical textual world, then the order of exposition may not be a secondary detail, but a component of the semantic formation regime.

The CMF-180 experiment tests this idea on a narrow domain: Greek, Latin, documentary, and epigraphic texts accepted up to 180 AD. This temporal limitation does not aim to create a complete encyclopedia of Antiquity, but a controlled space in which to test whether the order of exposure produces measurable differences between models trained on the same content.

The CMF-180 paper complements the IGT framework by introducing a dynamic dimension of representation formation, showing that the epistemic order of exposition can influence not only the performance of a model, but also the internal structure of semantic sedimentation. If IGT describes the information field, the semantic mass and the stabilization of decisional trajectories, CMF-180 shows that these structures are sensitive to the temporal and organizational regime of learning. In this sense, the historical block curriculum can be understood as a mechanism for preparing the ground on which informational gravitational dynamics subsequently act.

2. Hypothesis

The epistemic order of data exposure influences the internal semantic geometry and predictive performance of a small language model trained from scratch, even when the corpus, tokenizer, architecture, validation set, and number of optimization steps are held constant.

This hypothesis can be broken down into three sub-hypotheses:
1. A rigid chronological curriculum will produce a different learning trajectory than global random mixing, but may generate instability or anisotropy if local exposure is too homogeneous.

2. A block-based historical curriculum, in which the order between phases is chronological but the examples within each phase are mixed, will provide a more stable compromise between epistemic direction and local diversity.
3. Differences between regimes should not only be assessed by validation loss, but also by measures of internal geometry: anisotropy, cross-model similarity, semantic neighbor overlap, neighbor fragmentation, and behavior after removing dominant components from embeddings.

This formulation is based on the idea that learning is not just a function of data, but of data processed in a regime. In this paper, “regime” means the combination of order, local diversity, temporal continuity, and stability of exposure. In terms of the experiment, the block-based historical curriculum is treated as a metastable regime: it preserves a global structure but introduces enough local entropy to avoid excessive compression of the representation space.

3. CMF-180 corpus

The experiment was conducted on a corpus called in this paper CMF-180, built to approximate a Greco-Latin and documentary textual world limited to 180 AD. The corpus does not aim to include all modern information about Antiquity, but to provide a primary layer close to the sources: Greek, Latin texts, papyri, inscriptions and extensive materials accepted through chronological and qualitative filters.
The corpus construction followed several successive stages of auditing, filtering, deduplication and integrity checking. The primary version was obtained by selecting candidate texts, removing duplicates, checking text quality and building a flat corpus and a JSONL document-level file. Subsequently, an extended version was built by adding additional sources accepted after strict screening. The final version used in the training experiment was `extended_v2`.

The `extended_v2` corpus contains:
37,772 documents
38,858,834 words

The set was verified through the final integrity audit, which confirmed the equality between the number of documents in the flat corpus and the number of JSONL rows, the matching of the word count, and the absence of critical indicators of textual contamination or corrupt encoding.

The corpus was then split **at the document level, not at the chunk level,** to avoid leakage between train and validation. This is methodologically important: if the same document is split into chunks and some end up in train and others in validation, *the validation loss* can become artificially low. A previous experiment, conducted before the full corpus cleaning, had indicated exactly this risk. In the current version, the document-level split eliminates this problem.

The final split was:
train documents: 37,017
validation documents: 755
train words: 38,322,134
validation words: 536,700
overlap train/wave doc_id: 0

After tokenization and chunking, the final training set used in the experiments was:
train chunks: 68,212
validation chunks: 1,155
train tokens with overlap: 70,728,746
validation tokens with overlap: 964,584
train/val doc overlap: 0
train/val chunk overlap: 0
bad rows: 0
decode mismatches: 0
too long rows: 0
mojibake rows: 0

These figures indicate a corpus clean enough for a controlled experiment of *training from scratch* on a small model. The corpus is not presented as a complete representation of Antiquity, but as a controlled textual space, sufficient to test the effects of exposure order on a small language model.

4. The Tokenizer

For the experiment, a SentencePiece BPE tokenizer with a vocabulary of 32,000 tokens was trained on the `extended_v2` corpus. Two variants were tested: Unigram and BPE. Both were trained with byte fallback, for robustness to rare characters, diacritics, and polytonic Greek forms.

The validation comparison indicated:
validation docs: 755
validation words: 536,700

Unigram total tokens: 947,534
BPE total tokens: 910,002

Unigram byte fallback: 383
BPE byte fallback: 383

Unigram bad pieces: 0
BPE bad pieces: 0

Unigram decoded_same false: 0
BPE decoded_same false: 0

The 32k BPE tokenizer was chosen because it produced fewer tokens on the same *validation set,* while maintaining perfect and error-free decoding. The tokenizer used in the experiment is:
cmf180_extv2_sp_bpe_32k.model

Additional audit of the byte fallback showed:
validation docs seen: 755
docs with byte fallback: 30

total byte fallback pieces: 383
unique fallback bytes: 10

This level of fallback is acceptable for a corpus containing Latin, polytonic Greek, documentary material, epigraphy, and historical encoding variations. No signs of massive contamination or text degradation were observed.

5. Chunking and training dataset

After choosing the BPE32k tokenizer, the corpus was transformed into a *chunked dataset,* with a context of 2048 tokens and an overlap of 128 tokens. A first variant with a high inclusion threshold left too many documents without chunks, since many documentary or epigraphic texts are very short. Therefore, a `min8_safe` variant was built, which includes fragments with at least 8 tokens, while still maintaining safety checks.

The final dataset used for training was:
extended_v2_bpe32k_train_chunks_2048_128_min8_safe.jsonl
extended_v2_bpe32k_val_chunks_2048_128_min8_safe.jsonl

The integrity audit for this variant confirmed:
train rows: 68,212
validation rows: 1,155
train token sum match: YES
validation token sum match: YES
train decode mismatches: 0
validation decode mismatches: 0
train too long: 0
validation too long: 0
train mojibake: 0
mojibake validation: 0
train/wave doc overlap: 0
train/wave chunk overlap: 0
bad rows: 0

This audit is important because order experiments are sensitive to any hidden differences between sets. For a comparison between random, rigid-time, and block-based curriculum to be valid, each regimen must receive exactly the same training chunks and be evaluated on exactly the same *validation set.*

6. Model architecture

GPT-style causal language model was chosen, trained from scratch. The decision not to start directly from a large *pretrained model* was methodological: the aim of the experiment was not to obtain a final, high-performance model, but to test the effect of exposure order under controlled conditions. A *pretrained model* would have brought with it an already formed semantic space, an external tokenizer, and massive contamination with modern knowledge.

The configuration used was:
model type: GPT-style causal LM
vocabulary size: 32,000
context length: 2048
n_layers: 6
n_head: 8
n_embd: 512
parameters: 36,347,904
precision: bf16
batch_size: 2
gradient_accumulation: 16
effective batch size: 32 sequences
optimizer: AdamW
learning rate: 3e-4
weight decay: 0.1
warm-up steps: 100
max steps for epoch3: 6,396

The hardware used was:
GPU: NVIDIA GeForce RTX 5090 Laptop GPU
VRAM: 23.89 GB
CUDA: active
PyTorch: 2.11.0+cu128

The model is deliberately small. This choice limits absolute performance, but increases experimental value: differences between regimes emerge on a model small enough to be trained locally, repeatably, and audited. In addition, a small model is more sensitive to the order of the data, making it suitable for hypothesis testing.

7. The three training regimes

The experiment compared three exposure regimes, all on the same set of 68,212 training chunks and the same validation set of 1,155 chunks.

7.1. Global randomness

The random regime used the same training chunks, globally shuffled with a fixed seed of 180. This represents the classic baseline: the model sees examples from all periods, languages, sources, and registers from the start.

File: train_order_random_seed180.jsonl

7.2. Chrono v1: rigid chronology

The `chrono_v1` regime ordered chunks by an approximate chronological key. Where `not_before` and `not_after` numerical values were available, these were used to construct temporal phases. For materials without a direct numerical date, a fallback order based on source type and textual layer was used.

The main distribution by groups was:
PHASE_1_BEFORE_300_BC: 1,814
FALLBACK_CANONICAL_GREEK: 16,822
PHASE_2_300BC_TO_0: 10,570
FALLBACK_CANONICAL_LATIN: 9,389
PHASE_CROSSES: 1,847
PHASE_3_0_TO_180: 27,770

File: train_order_chrono_v1.json

This regime was included to test the simple hypothesis that a strict chronological order can change the learning trajectory. The results showed that this variant is a useful negative control: it produces large differences, but also anisotropy and fragmentation.

7.3. Chrono v2: historical curriculum in blocks

The `chrono_v2` regime preserved the chronological order between phases, but internally shuffled the chunks from each phase using the same fixed seed 180. Thus, the model sees first an early historical phase, then a canonical Greek phase, then Hellenistic/Republican material, etc., but without being forced to consume the documents in a rigid mechanical order.

The distribution by groups was identical to `chrono_v1`:
PHASE_1_BEFORE_300_BC: 1,814
FALLBACK_CANONICAL_GREEK: 16,822
PHASE_2_300BC_TO_0: 10,570
FALLBACK_CANONICAL_LATIN: 9,389
PHASE_CROSSES: 1,847
PHASE_3_0_TO_180: 27,770

The consistency check confirmed:
chrono_v2 rows: 68,212
same chunk set as random: YES
same chunk set as chrono_v1: YES
duplicate chunk ids: 0

File: train_order_chrono_v2_block_shuffle_seed180.jsonl

This is the central regime of the paper. It tests the hypothesis that historical order is useful only if it is applied at the level of epistemic blocks, not as a rigid list of objects. In intuitive terms, the model does not receive each text "drop by drop" in a mechanical sequence, but receives **coherent historical layers** with internal diversity. This regime is called in the paper a "**historical curriculum in blocks**".

8. The principle of experimental control

For the validity of the experiment, all three regimes were kept identical in terms of content and training parameters. The only difference between them was the order of exposure of the chunks.

The control can be summarized as follows:
same corpus: YES
same tokenizer: YES
same train set: YES
same validation set: YES
same model: YES
same architecture: YES
same hyperparameters: YES
same number of steps: YES
same order validation: YES
difference: only the train order

This constraint is essential. Without it, the differences observed between models could be attributed to data differences, train/validation leaks, different tokenizers, or different hyperparameters. In this experiment, all of these variables were held constant, and exposure order remained the main experimental variable.

9. Experimental results

The three training regimes were evaluated after 6,396 optimization steps, approximately equivalent to three epochs in the configuration used. The validation set was the same for all three models, and *the validation loss* was calculated on the same `val_order_fixed.jsonl` file.

The final results were:
random:
best_eval_loss = 5.099812
final_eval_loss = 5.099812
chrono_v1:
best_eval_loss = 6.124055
final_eval_loss = 6.124763
chrono_v2:
best_eval_loss = 4.928520
final_eval_loss = 4.928520

These results indicate three distinct behaviors. The random regime provides a solid baseline. The `chrono_v1` regime, based on rigid chronological ordering, significantly degrades performance. The `chrono_v2` regime, based on historical curriculum on internally shuffled blocks, achieves the best result of all three.
The difference between `chrono_v2` and random is:
5.099812 - 4.928520 = 0.171292

The difference is not huge in absolute value, but it is methodologically relevant, since all other variables were controlled. The corpus, tokenizer, architecture, training set, validation set,

number of steps, and hyperparameters remained identical. The only difference was the order of exposure to the data.
The main result is that **order matters,** but **not all order is beneficial.** Rigid chronology is detrimental, while block-based historical curriculum improves performance over global random shuffling.

10. Anisotropy of the embedding space

To assess whether the differences between the models reflect only predictive performance or also internal geometric differences, the anisotropy of the embedding space was measured. The measure used was the average cosine similarity between random pairs of tokens. A large value indicates a more compressed space, in which the embeddings tend to be oriented in the same dominant direction.

In raw space, the results were:
random raw pair_cos_mean: 0.558985
chrono_v1 raw pair_cos_mean: 0.819050
chrono_v2 raw pair_cos_mean: 0.726497

These values show that `chrono_v1` produces the strongest anisotropy. Its embedding space is strongly compressed, which explains why initial nearest neighbors evaluations produced very high similarities, sometimes above 0.95, without these necessarily being a sign of real semantic coherence.
`chrono_v2` reduces the anisotropy compared to `chrono_v1`, but remains more anisotropic than random in raw space. This suggests that the historical order on blocks preserves a more structured internal direction than random, but without the representational collapse produced by rigid chronology.
To control for the effect of dominant directions in the embedding space, an additional analysis inspired by the method proposed by Mu and Viswanath for postprocessing embeddings by removing the mean and a few dominant principal components was applied. In this report, this method is not used as an external proof of the result, but as a geometric control tool. Its purpose is to reduce the effect of raw anisotropy and to allow a more prudent comparison of semantic neighborhoods between models.

After removing the mean and the first three principal components, the spaces become much more comparable:
random debiased_pc3 pair_cos_mean: 0.001383
chrono_v1 debiased_pc3 pair_cos_mean: 0.001953
chrono_v2 debiased_pc3 pair_cos_mean: 0.000935

This correction shows that raw differences in anisotropy should not be directly interpreted as semantic differences. However, after debiasing, `chrono_v2` remains stable and even exhibits the lowest residual average pairwise similarity. This indicates that the block variant not only improves validation loss, but also produces a more balanced internal geometry than the rigid chronology.

11. Similarity between models

To assess how close the models are to each other, the cross-model similarity for the same tokens was calculated. In the `debiased_pc3` space, the average values were:
random_vs_chrono_v1: 0.481157
random_vs_chrono_v2: 0.486894
chrono_v1_vs_chrono_v2: 0.456404

These values indicate that `chrono_v2` is slightly closer to *random* than `chrono_v1`, but not identical to *random.* At the same time, `chrono_v2` is not just a corrected version of `chrono_v1`; the difference between `chrono_v1` and `chrono_v2` is considerable.

This observation is important: if `chrono_v2` had become almost identical to *random,* the historical block curriculum could only be interpreted as a masked global shuffle. The `chrono_v2` model approaches *random enough* to avoid degradation, but remains different enough to retain a curriculum effect of its own.

12. Semantic neighbor overlap

In addition to the direct similarity between embeddings, the overlap of semantic neighbors was analyzed for a list of important Latin and Greek terms: `virtus`, `ratio`, `natura`, `animus`, `lex`, `ius`, `civitas`, `imperium`, `λόγος`, `φύσις`, `ψυχή`, `πόλις`, `νόμος` and others.

In the `debiased_pc3` space, the average Jaccard overlap between the top-15 neighbors was:
random_vs_chrono_v1: 0.087997
random_vs_chrono_v2: 0.220619
chrono_v1_vs_chrono_v2: 0.122325

This comparison shows that `chrono_v2` has semantic neighborhoods closer to random than `chrono_v1`, but still distinct. Again, this is the desired result. `chrono_v2` does not collapse into a completely different and fragmented geometry, but neither does it simply reproduce random.
The larger overlap between random and `chrono_v2` suggests that internally shuffled historical blocks retain some of the semantic stability of global shuffle, but the training path remains different.

13. Neighborhood fragmentation

An additional criterion was the fragmentation of neighbors. In SentencePiece tokenization, neighbors that do not begin with a word-initial marker may indicate fragments, suffixes, or less semantically interpretable units. A pattern whose neighborhoods are dominated by fragments is more difficult to interpret semantically.
In the centered space, the average fragmentation was:
random fragment_mean: 0.047059
chrono_v1 fragment_mean: 0.258824
chrono_v2 fragment_mean: 0.056863

This comparison is one of the clearest. `chrono_v1` produces a much larger number of fragmented neighbors, which supports the interpretation that rigid chronological ordering produces a more unstable or less semantic geometry. In contrast, `chrono_v2` has fragmentation almost equal to random.

In raw space, the situation is similar:
random fragment_mean: 0.127451
chrono_v1 fragment_mean: 0.172549
chrono_v2 fragment_mean: 0.062745

Here `chrono_v2` has the lowest fragmentation of the three. This indicates that the model trained with the historical block curriculum produces lexically cleaner neighborhoods than `chrono_v1` and, under certain conditions, even than random.

After debiasing, the values increase for all models, but `chrono_v2` remains better than `chrono_v1`:
random fragment_mean: 0.190196
chrono_v1 fragment_mean: 0.327451
chrono_v2 fragment_mean: 0.213725

The conclusion is that the main problem of rigid chronology - fragmentation and semantic compression - is significantly reduced by the block curriculum.

14. The quality of the neighbors' language

The proportion of neighbors belonging to the same language as the analyzed term was also measured. This measure is imperfect, but useful in a bilingual Greek-Latin corpus. A semantic neighborhood around a Latin term should, in general, be dominated by Latin terms, and one around a Greek term by Greek terms, except for naturally mixed concepts.

In the `debiased_pc3` space, the average proportions were:
random same_language_mean: 0.998039
chrono_v1 same_language_mean: 0.919608
chrono_v2 same_language_mean: 0.996078
`chrono_v2` comes very close to random and clearly outperforms `chrono_v1`. This confirms that the block variant reduces the uncontrolled mixing or unstable neighborhoods observed in the rigid variant.

In the centered space:
random same_language_mean: 0.907843
chrono_v1 same_language_mean: 0.996078
chrono_v2 same_language_mean: 0.980392
`chrono_v1` seems very good according to the language criterion, but this value must be interpreted together with the high fragmentation. A model can preserve the language of its neighbors, but still produce fragmented or semantically weak neighbors. `chrono_v2` preserves a very good level of linguistic consistency, without the excessive fragmentation of `chrono_v1`.

15. Comparative interpretation

The results of the three regimes can be summarized as follows:

random:

validation loss good
moderate anisotropy
relatively clean neighbors
without explicit historical direction

chrono_v1:

weak validation loss
very high anisotropy
more fragmented neighbors
strongly different but unstable internal geometry

chrono_v2:

best validation loss
reduced anisotropy compared to chrono_v1
clean neighbors
distinct geometry from random
preserves global historical structure

These results support the hypothesis that order of presentation influences training, but with an important caveat: chronological order should not be mechanically applied at the document or chunk level. A rigid list produces excessive local homogeneity and can push the model into a narrow attractor. In contrast, order by historical blocks provides global structure but preserves local diversity.

In the terms of this paper, `chrono_v2` operates as a metastable regime. It is ordered enough to preserve an epistemic direction, but internally diverse enough to avoid representation collapse. This combination is thus superior to both global randomness and rigid chronology in the CMF-180 setup.

16. Preliminary experimental conclusion

The experiment shows that, under the tested conditions, the block-based historical curriculum outperforms both global random shuffling and rigid chronological ordering. The result is visible in *validation loss,* anisotropy, neighbor fragmentation, and semantic neighborhood overlap.

The main conclusion is not that any chronology helps, but, on the contrary, `chrono_v1` shows that a rigid chronology can be harmful. A specific form of chronology - organizing by historical phases with internal mixing - can produce a more stable and efficient training regime.

The block-based historical curriculum overcomes both rigid chronological ordering and global random shuffling in this CMF-180 controlled experiment.

This conclusion should be treated as preliminary. The model is small, the corpus is specialized, and the experiment needs to be replicated with more seeds, model sizes, and curriculum variants. However, the signal obtained is clear enough to justify expanding the research.

17. Theoretical framework

The results of the CMF-180 experiment suggest that **the order of exposure** is not a secondary technical detail, but an **active component of the learning regime.** The three models were given the same corpus, the same tokenizer, the same architecture and the same number of steps, but developed different performances and internal geometries. This observation is in line with the general line of *curriculum learning,* but **extends it in a particular direction: the curriculum is not defined by difficulty, but by historical and epistemic order.**

In classical curriculum learning, the order of examples is often thought of as a transition from simple to complex examples. In the CMF-180 experiment, the criterion is not simplicity, but the temporal stratification of knowledge. The model is not simply asked “which examples are easier?” but “what semantic world must exist before the next one is introduced?” This difference is important. A civilization does not produce its knowledge by global random sampling, but by accumulation, transformation, and sedimentation. Therefore, a model that attempts to approximate a historical textual world can benefit from an order that partially respects this sedimentation.

The experiment also shows that historical order should not be confused with a rigid list. The strict chronology variant, `chrono_v1`, had the worst performance and the highest anisotropy. This suggests that too much local homogeneity in the exposure can force the model into a narrow representational space. If the model sees the same type of material, language, register, or document type for too long, it can learn a dominant direction instead of a richer semantic structure.

This problem is mitigated in `chrono_v2`, where the order between phases is preserved, but the content within each phase is mixed. In this variant, the model does not receive the ancient world as a mechanical sequence of documents, but as historical blocks with internal diversity. This difference explains why `chrono_v2` achieved a better *validation loss than random* and reduced the fragmentation observed in `chrono_v1`.

18. Block-based historical curriculum as a metastable regime

The result can be interpreted through the concept of metastability. A metastable system is not completely fixed, but neither is it completely disordered. It retains some form of coherence while maintaining enough flexibility to avoid rigidification. In the context of training a small language model, a metastable exposure regime would be one that provides global structure without suppressing local diversity.

In the CMF-180 experiment, the three regimes can be interpreted as follows:

random:

high global entropy

without explicit historical direction

solid performance, but poorly controlled epistemic structure

chrono_v1:

high global order

low local entropy
risk of geometric collapse and anisotropy
chrono_v2:
moderate global order
controlled local entropy
compromise between historical direction and internal diversity

The optimal regime is neither complete disorder nor rigid order. In the case of CMF-180, the best option was one in which the model was given **historical direction between blocks, but sufficient internal variation within each block.** This can be described as a form of **metastable curriculum.**
A useful analogy is that of the composition of a perfume or a chemical mixture: the final stability does not necessarily result from the linear and mechanical addition of each individual element, but from the combination of coherent groups of elements, each with its own internal variation. Similarly, a historical phase should not be presented to the model as a rigid list of documents, but as a diverse internal semantic field, followed by another semantic field.
In a language model, consecutive examples influence the gradient, and the gradient influences the trajectory of the parameters. If the consecutive sequences are too homogeneous, the gradient can push the model in an excessively dominant local direction. If the sequences are completely random, the model loses any historical direction. Historical blocks with internal shuffles introduce a productive tension: the gradient retains a phase direction, but is not reduced to a single local type of signal.

19. Epistemic order and semantic sedimentation

The central concept proposed by this work is that of epistemic order. Epistemic order is not identical to strict chronological order, but to the approximate succession of layers of knowledge.
In the CMF-180 corpus, the epistemic order is imperfect. Not all documents have precise dating, and many texts are preserved in later copies, modern editions, or contemporary digital structures. However, even a rough approximation can have an effect. `chrono_v2` does not claim to reconstruct exactly the actual historical order of composition or circulation of each text. It only introduces a phase structure: before and after, early Greek, Greek canon, Hellenistic/Republican material, canonical Latin, documentary and epigraphic material close to the year 180.
This structure seems sufficient to positively modify the learning regime. This is important because it suggests that models do not necessarily need perfect chronology to benefit from epistemic order. They can benefit from approximate stratification, if this is applied at the block level and not at the mechanical level of individual documents.

Therefore, the historical curriculum in blocks can be defined as follows:
A training regime in which data are grouped into approximate epistemic or historical phases; the order between phases follows a temporal or conceptual direction, and the internal order of examples within each phase is shuffled to preserve local diversity.

This definition clearly separates the method from two alternatives:

1) it is not globally random, because it preserves the order between phases;
2) it is not a rigid chronology, as it does not impose a strict order between all documents.

20. Link to biological learning

The result can also be interpreted by analogy with biological learning. Organisms do not learn from globally randomly distributed data, but neither do they receive information from the world in a perfectly deterministic order. Biological learning has stages, contexts, sensitive periods, repetition, variation, sleep, reinforcement, stress, attention, curiosity, and feedback. These factors create a learning regime in which order and variability coexist.

A child does not learn all the concepts of the world in a global shuffle, but neither does he receive every object of the world in exact order. He learns through blocks of experience: home, family, body, language, space, play, danger, rules, social group. Within these blocks there is variation. This structure is more like `chrono_v2` than random or `chrono_v1`.

This analogy does not imply that a language model is equivalent to a biological organism. The substrate is different: artificial neural networks, loss functions, and optimizers on the one hand; the body, neurons, metabolism, environment, and homeostasis on the other. However, at a functional level, both can be viewed as systems that **form representations through sequential exposure to signals.** In this limited sense, order, local diversity, and the internal state of the system may be relevant in both cases.

This interpretation supports a more general idea: language models should not be analyzed simply as functions of data, but as **the results of exposure routes.** The same content, traversed in a different order, can produce a different internal geometry.

21. Why *random* isn't enough

Global random shuffling is robust and efficient in many training regimes. It reduces local correlations, exposes the model quickly to the full diversity of the dataset, and stabilizes the gradient. That is why *random* performed well in the experiment.

However, *random* has a conceptual limit: it destroys epistemic direction. For many tasks, this does not matter. If the goal is simply to minimize the loss on a stationary distribution, *random* may be optimal or sufficient. But if the goal is to model a historical world, in which concepts emerge, transform, and sediment, *random* may eliminate an important signal.

`chrono_v2` suggests that one can preserve some of the advantages of *randomness* - local diversity and stability - without completely abandoning historical structure. This is the main contribution of the experiment: it does not reject *randomness,* but shows that a more structured regime can achieve better performance under controlled conditions.

22. Why rigid chronology fails

The poor performance of chrono_v1, under the budget and evaluation conditions of this experiment, is as important as the success of chrono_v2. Without `chrono_v1`, one could have simplistically stated that ``chronology helps''. The data show that this formulation is at least incomplete. In this setup and at this training budget, the rigid chronology did not help; on the

contrary, it produced weaker loss, higher anisotropy, and more fragmented neighbors. However, this result does not exclude the possibility that a rigid chronology requires more steps to converge on a mixed validation set.

The likely explanation is excessive local homogeneity. If the model sees many similar examples consecutively, the optimization can favor dominant directions and reduce the flexibility of the embedding space. This is seen in the high raw anisotropy of `chrono_v1` and the greater fragmentation of neighbors. Instead of building a rich historical representation, the model seems to be pushed into a compressed geometry.

Therefore, `chrono_v1` functions as a negative control. It shows that the idea of historical order must be applied with care. Order should not suppress variation.

23. Thermodynamic interpretation

A thermodynamic interpretation of the result must be formulated with caution. It is not claimed that the model literally obeys a thermodynamic law in the physical sense. However, the analogy is useful for describing the balance between order and entropy.
Global random has high entropy but low direction. Rigid chronology has high direction but low local entropy. Block-based historical curriculum combines global direction with sufficient local entropy. In this sense, `chrono_v2` can be interpreted as a regime with **structured entropy.**
The block-based historical curriculum indicates functioning as a metastable training regime: it reduces global disorder without eliminating local diversity.

24. Implications for training specialized models

The block history curriculum can be useful for specialized models trying to teach historical, legal, scientific, or cultural fields where the order of appearance of concepts matters.
Possible examples:
models trained on the history of law;
models trained on the evolution of medical terminology;
models trained on historically stratified philosophical corpora;
models trained on language evolution;
models trained on institutional archives;
models trained on the development of a scientific discipline.

In such areas, global *randomness* can miss the signal of internal concept development. Rigid chronology can be too fragile. Block curriculum offers an intermediate solution.
This method should not be considered universally superior. For many modern corpora or generic tasks, random may remain sufficient. For corpora with real temporal or epistemic structure, a block-based variant would probably bring the necessary added value.

25. Implications for CMF-180

In the context of CMF-180, the result has a direct practical implication: the model that will be used as the basis for future experiments should not be `chrono_v1`. The rigid variant is inferior. Between random and `chrono_v2`, preliminary data favor `chrono_v2`.

However, random remains a necessary baseline. It provides the comparison without which the effect of the curriculum cannot be measured. In future developments, `chrono_v2` should be tested with more seeds, more model sizes, and alternative variants of historical blocks.

In the future architecture of the system, the CMF-180 model should not be understood as a complete encyclopedia of the ancient world, but as a textual resonance chamber built from ancient and near-source sources. It will be complemented by a separate Corpus 3, consisting of reconstruction files in Latin and Greek, with modern metadata excluded from ancient training. This separation between the “source voice” and the “reconstructive knowledge” is essential to avoid contaminating the ancient model with uncontrolled modern explanations.

26. Synthesis

The experimental result can be summarized as follows:
Data is not just content. Data is content traversed in a regime.
Random offers diversity without direction.
Rigid chronology provides direction without local diversity.
The block-based historical curriculum provides global direction and local diversity.

This synthesis supports the introduction of the concept of **" Metastable Historical Curriculum Learning "** (MHCL): a training method in which epistemic order is preserved at the phase level, but internal variation is maintained to prevent representation collapse.

In its current form, the result does not demonstrate a general law of language model training. However, it does demonstrate that, in a controlled CMF-180 setup, block-based historical order produces better results than both global random and rigid chronology. This is a sufficient basis for further research.

27. Limitations

The CMF-180 experiment should be interpreted as a preliminary result, not a general demonstration. Although the internal control of the experiment is strong - same corpus, same tokenizer, same architecture, same validation set, and same number of steps - there are several important limitations.

The first limitation is the model size. The model used has approximately 36.3 million parameters. This size is adequate for a local controlled experiment, but does not allow direct conclusions about large models. Larger models may respond differently to the order of the data, be more robust to local homogeneity, or require different curricula.

The second limitation is the number of seeds. The experiment was run with a single main seed, 180. Although the observed differences are clear, the possibility that part of the effect is seed-dependent cannot be completely ruled out. A rigorous replication should include multiple seeds for random order, for internal mixing in `chrono_v2`, and, ideally, for model initialization.

The third limitation is the specificity of the corpus. CMF-180 is a specialized corpus, built from Greek, Latin, documentary, and epigraphic texts up to 180 AD. The results may depend on this particular structure. It is not automatically guaranteed that a block-based historical curriculum will work the same on modern, multilingual, technical, or conversational corpora.

The fourth limitation is the quality of the chronology. Not all documents have precise dating. Some have clear numerical dates, others have been framed by source, textual layer or epistemic fallback. Therefore, `chrono_v1` and `chrono_v2` do not represent a perfect historical chronology, but a practical approximation. The result could be improved or modified by finer dating.

The fifth limitation is the nature of the texts. The corpus contains literary, documentary, and epigraphic sources, but these do not directly represent the ordinary speech of ancient people. The model cannot be described as a reconstruction of a "Roman citizen" or of ancient consciousness. It is more accurately described as an experimental approximation of the textual distribution available in the corpus.

The sixth limitation is the use of validation loss as the primary measure of predictive performance. Validation loss is useful, but it does not exhaust the question. A model can have better loss and still produce less interpretable representations. For this reason, anisotropy, fragmentation, and neighbor overlap were also measured, but these measures are also approximations.

The seventh limitation concerns the analysis of semantic neighbors. Nearest-neighbor analysis on embeddings is sensitive to anisotropy, tokenization, and the choice of the analyzed space. Therefore, centering and elimination of the first principal components were used, but the method remains imperfect. Token neighbors should not be interpreted as a direct equivalent of the "meaning" of the concepts.

27.1. Specific limitation regarding the interpretation of `chrono_v1`

A specific limitation of the experiment concerns the interpretation of the poor performance of `chrono_v1`. The fact that the model trained with a rigid chronology achieved a `final_eval_loss` of 6.124763 does not, by itself, demonstrate that strict chronological ordering is harmful in a general sense. An alternative interpretation is that, in the fixed number of steps used in this experiment, `chrono_v1` did not fully converge to the global validation distribution.

This possibility is important. Unlike random, which “sees” the entire diversity of the corpus from the beginning, `chrono_v1` sequentially goes through relatively homogeneous phases. Therefore, at the same number of steps, it may be disadvantaged on a mixed *validation set,* not necessarily because the rigid chronological order is intrinsically inferior, but because the distribution traversed in training later comes to cover the entire evaluated variety.

However, this explanation does not eliminate the geometric observations. `chrono_v1` had not only a weaker *validation loss,* but also much larger raw anisotropy and more fragmented semantic neighbors. These signals suggest that the problem is not just incomplete convergence, but also a form of representational compression associated with too homogeneous local exposure. However, the conclusion must be formulated cautiously: in this

experiment and at this training budget, the rigid chronology produced inferior results; it cannot yet be stated that any rigid chronology would fail at larger budgets, other model sizes, or other learning rate schemes.

28. Validity

28.1. Internal validity

The internal validity of the experiment is relatively good, as the main variables were controlled. The three regimes used the same set of chunks, and checks confirmed the absence of duplicates and train/validation overlap. However, subtle effects related to implementation, caching, read order, final checkpoint differences, or undetected variations in training remain possible.

A specific threat is that `best_model` was not updated after the final evaluation in some runs; therefore, the interpretation was based on the final checkpoint at step 6.396. This choice is correct for the comparison between regimes.

28.2. Construct validity

The concept of "epistemic order" is approximate. It cannot be completely reduced to `not_before`, `not_after`, or to the source of the document. Some texts may be preserved late but composed early; others may reflect older traditions; others may be assigned uncertainly. Therefore, the phases used in `chrono_v2` are operational building blocks, not perfect historical truths.
However, the experiment does not depend on a perfect chronology. The hypothesis tested is weaker and more realistic: an approximate historical stratification, applied at the block level, can produce a different and possibly more stable training regime than random.

28.3. External validity

The result cannot be automatically generalized to all models, languages, or domains. CMF-180 is a particular case: a relatively small, bilingual, historical corpus with real temporal structure. It is possible that `chrono_v2` is useful precisely because the domain has a strong historical sedimentation. In domains without natural temporal structure, the effect may be weaker or absent.

28.4. Statistical validity

The experiment does not have multiple replications and does not provide confidence intervals. The difference in *validation loss* between *random* and `chrono_v2` is clear in this run, but probably needs to be tested on more seeds. Without replications, the result should be presented as an experimental signal, not as a final statistical estimate.

29. Future experiments

The first experiment required will be a multiple seed replication. At least three to five seeds should be tested for each regime:

random seed: 180, 181, 182, 183, 184
chrono_v2 internal shuffle seed: 180, 181, 182, 183, 184
initialization seed model: controlled separately
This replication would allow calculating the mean, standard deviation, and stability of the difference between random and `chrono_v2`.

The second experiment is scaling the model. At least three dimensions should be tested:
small: ~36M parameters
medium: ~100M–150M parameters
larger local: as much as the available hardware allows
If the effect of Block-based historical curriculum persists in larger models, the value of the result increases significantly. If it disappears, then the effect may be specific to small models.

The third experiment is changing the block granularity. `chrono_v2` uses fairly large blocks. Future variants could test:
coarser blocks;
finer blocks;
blocks on the tongue;
blocks per period + language / blocks per period + text type.
The goal would be to identify the optimal level of structure. Too few blocks approach randomness. Too many blocks approach rigid chronology. The metastable hypothesis predicts the existence of an optimal intermediate zone.

The fourth experiment is gradual pacing. Instead of the model running through one block completely and then moving on to the next, a progressive scheme could be introduced:
phase 1: 100% early block
phase 2: 70% early block + 30% next block
phase 3: 40% early block + 60% next block
phase 4: next dominant block
This method would better mimic historical sedimentation, in which old layers do not suddenly disappear but continue to influence new layers.

The fifth experiment is to introduce a historical “rehearsal.” With each new block, a small proportion of the previous blocks would be reintroduced. This would test whether the model can retain the memory of previous phases without becoming rigid.
A possible scheme:
80% current block
15% previous blocks
5% global random samples
This variant could reduce forgetting and stabilize transitions between phases.

The sixth experiment is a comparison with a large modern model. Not to compete directly, but to distinguish between two types of performance:
LLM major: fluent modern simulation of Antiquity
CMF-180: textual resonance trained on controlled corpus
This comparison might show that large models are better at modern explanations, but the CMF model may provide a cleaner space for certain textual analyses or controlled reconstruction experiments.

The seventh experiment is trajectory analysis during training. Instead of evaluating only at the end, one can save checkpoints every 500 or 1000 steps and watch how neighbors emerge and stabilize for terms like `virtus`, `ratio`, `natura`, `lex`, `ius`, `λόγος`, `φύσις` and `ψυχή`.
This analysis would allow for observation of the formation of concepts over time. If `chrono_v2` produces more ordered or stable trajectories than random, the historical curriculum hypothesis would become stronger.

30. Conclusions on limitations

The limitations of the experiment are real; they do not invalidate the result, but define the next steps.
In a controlled experiment on the CMF-180 corpus, a Block-based historical curriculum with internal shuffling outperformed both global random shuffling and rigid chronology, producing the best validation loss and a cleaner semantic geometry than rigid chronology.

This conclusion is sufficient for a preliminary report and to justify further experiments. It is not yet a general law of language model training, but it is a concrete result, reproducible under the described conditions, and methodologically interesting.

31. Final conclusions

This paper tested the influence of exposure order on a small language model trained from scratch on the CMF-180 corpus, a Greco-Latin and documentary corpus temporally limited to 180 AD. Three training regimes were compared: global random shuffling, rigid chronology, and Block-based historical curriculum with internal shuffling.
The results show that the order of exposure has a measurable effect on the performance and internal geometry of the model. The random regime provided a solid baseline. The rigid chronology regime had the weakest validation loss and the highest anisotropy. The block historical curriculum regime achieved the best validation loss and reduced the semantic fragmentation observed in the rigid chronology.

The central result of the experiment is:
random final_eval_loss: 5.099812
chrono_v1 final_eval_loss: 6.124763
chrono_v2 final_eval_loss: 4.928520
These data support the conclusion that historical order can be useful, but only when applied at the level of blocks or phases, not as a rigid sequence of individual documents or chunks. Strict chronology can introduce excessive local homogeneity and push the model towards anisotropic geometry. In contrast, the block-based historical curriculum combines global epistemic direction with local diversity.
In this form, the result supports the concept of `Historical Block Curriculum` or, in a more general formulation, `Metastable Historical Curriculum Learning`: a training regime that preserves a temporal or epistemic structure between phases, but allows internal entropy within each phase to avoid representation collapse.

32. Contributions

The paper makes the following preliminary contributions:
1. It defines and tests a training regime called **historical block curriculum,** different from both global random and rigid chronology.
2. It builds a controlled experiment on the CMF-180 corpus, in which the corpus, tokenizer, architecture, validation set, hyperparameters, and number of steps are held constant, and the exposure order is the main experimental variable.
3. It shows that rigid chronology is not sufficient and can be harmful, producing weaker validation loss, high anisotropy, and more fragmented semantic neighbors.
4. It shows that a historical block curriculum with internal mixing can outperform both global random and rigid chronology in the CMF-180 setup.
5. It proposes the interpretation of the historical block curriculum as a metastable regime: global order combined with local entropy.
6. It introduces an evaluation methodology that is not limited to validation loss, but includes anisotropy, dominant component removal, neighbor overlap, fragmentation, and linguistic consistency.
7. It opens a research direction for specialized models, in which the epistemic order of the data becomes part of the training design.

Conventius.org represents the applied dimension of the project: an isolated digital mind whose linguistic, historical, and conceptual formation is deliberately bounded by the classical textual world and anchored in 180 AD as a fixed civilizational horizon.

33. Bibliography

- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). ``Curriculum Learning''. Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009.
- Hacohen, G., & Weinshall, D. (2019). ``On The Power of Curriculum Learning in Training Deep Networks''. Proceedings of the 36th International Conference on Machine Learning, ICML 2019.
- Mu, J., & Viswanath, P. (2018). ``All-but-the-Top: Simple and Effective Postprocessing for Word Representations''. International Conference on Learning Representations, ICLR 2018.
- Kim, J., & Lee, J. (2024). ``Strategic Data Ordering: Enhancing Large Language Model Performance through Curriculum Learning''. arXiv:2405.07490.
- Zhang, Y., Mohamed, A., Abdine, H., Shang, G., & Vazirgiannis, M. (2026). ``Beyond Random Sampling: Efficient Language Model Pretraining via Curriculum Learning''. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026, pp. 5776–5794
- Elgaar, M., & Amiri, H. (2026). ``Curriculum Learning for LLM Pretraining: An Analysis of Learning Dynamics''. arXiv:2601.21698.
- Campos, D. (2021). ``Curriculum Learning for Language Modeling''. arXiv:2108.02170.
- Stan, A. (2026). Information Gravity Theory (IGT). Zenodo.org/records/18481335