Reproducible Generative Constraints in Voynichese

Daiki Matsuda

Independent Researcher

AiraXiv submission edition | 15 November 2026 | Based on frozen executable release v0.8

Previously archived research report: doi:10.5281/zenodo.23000054 | Source and executable checks: GitHub repository

Abstract. This study develops reproducible statistical constraints on proposed accounts of Voynichese generation without attempting decipherment. Using a fixed 12-slot representation, it evaluates low-dimensional variation within token families, context- and scribal-hand-conditioned concrete-form choice, exact-form reuse, bifolio-local dependence, and line-local ordering. The executable core analyzes 24,250 valid tokens from a pinned ZL3b transcription and applies the same definitions to 24,233 valid tokens from IT2a, an alternative transcription of the same manuscript. Under a stated predictive comparison, the true bifolio source exceeds the mean wrong-bifolio source on all 42 tested bifolios in each transcription. A line-composition-preserving null leaves a small positive within-line ordering excess, with no positive cross-line counterpart. Uniform reinforcement of previously seen forms helps prediction but is insufficient to reproduce the observed exact-form concentration. Inputs, definitions, scripts, and machine-readable checkpoints are provided. The results are frozen, reproducible exploratory evidence, not a preregistered confirmation, an independent-corpus replication, or an identification of the manuscript's historical production mechanism.

Keywords: Voynich Manuscript; Voynichese; constrained generation; 12-slot FAMILY representation; bifolio; structure-preserving null models; reproducibility; AI-assisted research

1. Research question and claim boundaries

The Voynich Manuscript has many statistical regularities but no generally accepted reading of its text. The narrow question here is: what measurable structural and predictive constraints must a credible account of Voynichese string generation reproduce? This framing allows linguistic, cipher-based, copying, stochastic, or other generative accounts to be compared against the same observable tests without assuming a plaintext or a particular physical production device.

The analysis does **not** argue that the text is meaningless, that a literal machine or physical randomizer produced it, or that any particular historical actor used the modeled operations. It also does not claim that the release contains a complete end-to-end historical simulator. Its primary output is an auditable set of constraints and diagnostic comparisons.

The report is a submission-oriented restatement of the public v0.8 executable core, not an independent new experimental dataset. A separate diagnostic suite for comparing reference generators is available in the same repository; its figures are not pooled into the principal results reported here.

2. Materials, representation, and evaluation units

2.1 Pinned source transcriptions

The primary source is the ZL3b transcription, restricted to a frozen clean42 set of 42 bifolios. The executable analysis accepts 24,250 valid 12-slot tokens with 2,711 exact token types. The robustness pass uses IT2a, yielding 24,233 accepted tokens and 2,842 exact types under the same declared representation and inclusion logic. Both sources encode the same physical manuscript. They are **alternative transcriptions, not independent documents**.

The source mirrors are referenced by repository, specific commit, file path, and Git blob SHA-1; the underlying transcription texts are fetched and hash-verified by the scripts rather than redistributed as part of this report. The full pinning metadata are recorded in the reproduction protocol [1, 2].

2.2 FAMILY / concrete-form separation

A frozen twelve-slot decomposition maps accepted tokens to FAMILY-level structural forms and concrete slot-face choices. The analysis asks whether, given a FAMILY, the exact realized form is predictable from manuscript context, scribal hand, local reuse, or bifolio membership. The FAMILY representation is a modeling choice, so inferences are conditional on its mapping and token acceptance rules.

2.3 Holdout and null principles

For contextual and hand-conditioned predictive comparisons, the public core uses leave-one-bifolio-out (LOBO) evaluation. Local bifolio adaptation uses one half of a true bifolio as a source for predicting the other half, while contrasts with wrong-bifolio sources constrain a trivial global-frequency explanation. For ordering, a line-composition-preserving null randomizes ordering while retaining the observed concrete-face composition within each line. These controls address different questions and must not be collapsed into a single significance claim.

3. Five measurable constraint families

3.1 Low-dimensional concrete variation

In the CAL8 subset, high-frequency FAMILY forms occupy restricted regions of their theoretical concrete-face space. For 24 selected families covering 2,817 tokens, fixing each slot at its FAMILY-modal face yields 43.88% coverage; allowing the one, two, or three most variable slots to vary raises weighted token coverage to 71.03%, 90.02%, and 97.83%, respectively [1]. The term "box-k" labels this mathematical restriction; it is not evidence for a physical box, bag, or lookup table.

3.2 Context and hand-conditioned exact choices

Within FAMILY-conditioned exact-choice predictions, the canonical ZL3b model uses a global Jeffreys-smoothed base distribution and shrunk context and scribal-hand variants. Context improves aggregate log likelihood by +0.049640 nats/token relative to the global FAMILY base; hand improves it by +0.046673 nats/token. Both models outperform the global base on 38/42 test bifolios. Context has a slightly higher aggregate score, but it does not dominate hand robustly by bifolio: context wins on 18/42, and the median difference is slightly negative [1].

3.3 Reuse helps but does not finish the explanation

Uniformly reweighting all previously seen exact forms inside each bifolio/FAMILY provides an additional aggregate predictive gain of +0.014136 nats/token on the ZL3b reference specification. However, fixed-seed simulations on the observed FAMILY/context scaffold still produce substantially too many distinct types and hapax types. For example, the observed clean42 spectrum has 2,711 exact types and 1,343 hapax, while a strong uniform seen-weight of 32 yields simulation medians of 3,149.5 types and 2,046.5 hapax [1]. Thus selective structure beyond uniform broad reuse is needed under this implementation.

3.4 Bifolio-local predictive state

For eligible multi-face slot-family choice events, the baseline predicts P(face | slot, FAMILY, context, hand) from data outside the held-out bifolio. Opposite-half observations adapt that baseline. At the focal exploratory shrinkage setting tau=20, the true opposite-half model gains +0.024825 nats/event over baseline on ZL3b, and its aggregate advantage over the mean wrong-bifolio source is +0.051512 nats/event. True sources beat the wrong-source mean on 42/42 tested bifolios, but improve on the full conditioned baseline on only 28/42. The true source is not always the best individual source. These distinctions matter: the comparison supports a local shared statistical state, not a unique explanation for its origin [1].

3.5 Line-local ordering

After each line's concrete-face composition is held fixed, the ZL3b within-line same-face ordering excess remains +0.010883, while the cross-line contrast is -0.007370. The within-line excess is positive in 31/42 bifolios and the cross-line excess in 15/42. This is a limited statement about ordering beyond composition, not direct evidence for a physical action performed by a scribe [1].

4. Frozen cross-transcription checkpoints

The following values are the canonical public checkpoints; table entries are descriptive or model-relative comparisons, not unconditional probabilities about the manuscript's historical origin [1, 2].

Measurement	ZL3b	IT2a
Accepted clean42 tokens	24,250	24,233
Exact types	2,711	2,842
Box-2 / box-3 coverage	90.025% / 97.835%	88.577% / 97.416%
Context gain (nats/token)	+0.049640	+0.049585
Hand gain (nats/token)	+0.046673	+0.045859
Uniform seen-reuse gain (nats/token)	+0.014136	+0.014255
True bifolio opposite-half gain (nats/event)	+0.024825	+0.024509
True minus wrong-source mean (nats/event)	+0.051512	+0.051291
Bifolios positive vs wrong-source mean	42/42	42/42
Within-line ordering excess	+0.010883	+0.010826
Cross-line ordering excess	-0.007370	-0.006957

Agreement across ZL3b and IT2a reduces sensitivity to the two tested transcription conventions but does not establish replication across distinct manuscripts or demonstrate universality across all proposed transcription systems.

5. Interpretation, limitations, and falsifiability

Together these findings specify a constrained account: a productive FAMILY representation, limited concrete variation, context and hand preferences, selective exact reuse, bifolio-local state, and a weak within-line ordering component. They establish which observables must be explained; they do not compel a purely nonsemantic origin. Language, encoding, imitation, and production-process hypotheses remain live alternatives.

The same clean42 material was inspected repeatedly during method and model development. Hyperparameters and null definitions were therefore informed by exploratory work, even though the current public tests and outputs are frozen and rerunnable. A clean third-party reimplementation, additional transcriptions, and stronger null comparisons are essential before claiming stronger generality.

At the bifolio level in particular, metadata dependence, scribal-hand organization, section effects, physical nesting, and possible source exchangeability deserve independent scrutiny. The reported 42/42 sign result is against the mean of wrong-bifolio sources under a specified model; it should not be rewritten as 42 independent historical proofs. The weak ordering excess is likewise sensitive to the FAMILY/face representation and line boundaries. Some within-line and cross-line measures are heterogeneous across families.

Concrete falsification routes include: (a) reproducing the frozen executable results on a clean environment; (b) independently reimplementing the parser or one model; (c) demonstrating that an appropriately controlled stronger null removes the local-state/ordering residual; (d) evaluating additional transcription choices without retuning the scientific question; and (e) fitting a new candidate generator to the same predeclared diagnostic panel.

6. Open reproduction protocol

Requirements: Python 3.10+ and internet access. From the checked-out repository root:

```
python tools/run_all_evidence.py --fetch
```

The command fetches the byte-pinned public transcription copies, verifies their blob identities, runs both the canonical ZL3b analyses and IT2a robustness pass, and compares outputs to frozen machine-readable checkpoints. A successful run is documented to end with "ALL EVIDENCE REPRODUCTION: PASS." This sentence reports the release's expected behavior, not a new execution performed for this submission [1, 2].

Implementation and traceability: repository | third-party protocol | evidence manifest | claim-to-code map.

7. Research provenance and AI contribution disclosure

The human author, Daiki Matsuda, directed the research question, selected hypotheses and comparison goals, established interpretation boundaries, and made decisions about freezing and releasing the results. AI systems, including ChatGPT, substantially assisted with coding, statistical analysis, literature search, research-log organization, reconstruction, testing, and drafting. This involvement is disclosed explicitly and is not treated as independent validation. The human author remains responsible for reviewing the research claims and the material submitted under his name.

This AiraXiv edition is a concise presentation of the public research evidence. It was also assembled with AI drafting assistance and should not be characterized as wholly human-written. It does not purport to be the identical PDF file deposited at Zenodo. The prior DOI and the executable repository are the provenance anchors.

References and archived research artifacts

[1] Matsuda, D. (2026). Reproducible Generative Constraints in Voynichese, executable research release v0.8. Repository: github.com/tvwd8kzt7k-ui/voynich-generation-constraints. Canonical evidence and reproduction manifests in repository root.

[2] Matsuda, D. (2026). Reproducible Generative Constraints in Voynichese, archived report. Zenodo. DOI: 10.5281/zenodo.23000054. See repository CITATION.cff and transcription robustness report for version references.

[3] Timm, T. and Schinner, A. (2020). A possible generating algorithm of the Voynich manuscript. Cryptologia, 44(1), 1-19. DOI: 10.1080/01611194.2019.1596999. (Background on generator hypotheses; not a validated explanation in the present study.)

[4] Zandbergen, R. Voynich MS: Text analysis and writing system. voynich.nu/writing.html. (Background on transcription, hands, and codicological metadata.)