Measuring the Unsolvable

A Comparative Study of the Liber Primus (Cicada 3301) and the Voynich Manuscript

Falsification, Discriminative Statistics, and the Limits of AI in Historical Cryptanalysis

Matheus Soranço

Creator of I.S.A.A.C. (Intelligent Scientific Analysis & Analytics Core)

AI Researcher . Computer Science . Cryptanalysis . Philosophy of Mind

Code: Apache License 2.0 | Paper text: CC BY 4.0 | Preprint - not submitted to any venue | No decipherment claim is made

Repositories: LiberPrimus and Voynich (GitHub: Matheussoranco)

. Abstract

The Liber Primus (Cicada 3301) and the Voynich manuscript are two of the most famous unsolved ciphertexts. Both have accumulated a century-plus of mutually incompatible "solutions," most of them unfalsifiable. This paper reports two independent research programs - one for each artefact - that share a single methodological rule: **report every statistic next to null models that ought to fail it, and bind every attack to known-answer controls drawn from the same cipher class.** For the Liber Primus, a falsification engine excludes seven cipher classes on the six unsolved sections (11,081 runes), including natural-language running keys via a convolution argument (z = ?10.7), and documents how a substantial machine-learning search apparatus ranked a rotated plaintext above the true plaintext on an already-solved page. A subsequent variational homophonic analysis finds per-section heterogeneity: three sections compatible with homophonic English, one decisively excluded, two inconclusive. For the Voynich manuscript, a discriminative battery tested against four structure-retaining null models shows that Zipf's law, conditional entropy, and positional rigidity are all reproduced by an order-2 character Markov chain fitted to the manuscript's own glyphs - they discriminate almost nothing. Two findings survive their controls: Currier A and B are distinct populations, and the line-as-a-functional-unit effect (13x glyph-level, 23x word-level divergence over a character-shuffled control) resists any substitution or transposition cipher of natural language. The paper concludes with a discussion of where AI fits in this domain (hypothesis generation and corpus building outside the scoring loop) and where it does not (as a judge inside it).

1. Introduction

The Liber Primus is a runic codex released by the Cicada 3301 group during its 2014 puzzle round, written in a 29-character alphabet published as the Gematria Primus, mapping Elder-Futhark-derived runes to both an index and a prime value. Four sections have been decrypted by the community and read as English aphoristic and instructional text; six remain unsolved, pooling 11,081 runes. The Voynich manuscript is an illustrated codex radiocarbon-dated to the early fifteenth century (c. 1404-1438), written in an unknown script (Voynichese) across roughly 240 vellum folios. Despite more than a century of sustained professional and amateur attention, neither has been convincingly deciphered.

What unites the two domains is not just their fame but their epistemics. An unsolved artefact has no ground truth against which to check a positive claim. If a procedure outputs text that reads like English or medieval Latin, no independent oracle confirms it. The consequence is a long record of confident, mutually incompatible "solutions." In the Voynich domain this includes work that passed peer review before being withdrawn. In the Liber Primus domain, the community has produced a sprawling registry of hypotheses with little methodological discipline.

This paper does not attempt to solve either artefact. It reports measurement programs designed so that negative results mean something. The central rule, applied mechanically in both projects, is: **control gating.** Every attack is bound to (1) a named, size-quantified cipher class, (2) known-answer controls drawn from that same class and generated through the attack's own primitives, and (3) a procedure applied to the real ciphertext. An attack's report is admissible only if its controls pass; otherwise the verdict is INCONCLUSIVE - a statement about the tool, not the data.

A secondary rule is the **discriminative battery:** a statistic is admitted only if it separates the artefact from null models that are meaningless by construction. Null models are ordered by how much structure they retain: uniform random text, character-shuffled text, order-1 Markov, order-2 Markov. A statistic that almost everything passes cannot support a conclusion, however firmly the artefact passes it.

My personal interest in these two projects is longstanding. The Liber Primus fascination began with Cicada 3301 and turned into a personal research program combining computational cryptanalysis with machine learning. The Voynich interest is older still - a curiosity about whether statistical structure can ever establish linguistic status. Both projects are conducted as research: failures are preserved, what works is measured, and conclusions are kept smaller than the curiosity.

2. Mathematical and Cryptographic Foundations

2.1 Index of Coincidence (IC)

The index of coincidence is the probability that two symbols drawn independently from a text are identical. For alphabet size n and symbol counts f?: IC = ?? f?(f??1) / [N(N?1)]. For uniform text over n symbols, IC = 1/n. For English, IC ~ 0.066. A substitution cipher preserves IC exactly (it is a bijection); a transposition preserves the unigram distribution and therefore IC; a running key convolves plaintext and key distributions, flattening but never uniformising the result unless one of them is exactly uniform. This last fact - the convolution argument - is what allows a class-level exclusion of natural-language running keys without searching the unbounded key space.

2.2 Entropy and Conditional Entropy

Unigram entropy H1 = ?? p? log2 p?. First-order conditional entropy H2 = H(c? | c??1), estimated with Miller-Madow bias correction. Maximum entropy for 29 runes is log2(29) ~ 4.858 bits. Low H2 indicates sequential predictability. Entropy estimates are biased downward on short text; the correction is applied and raw values reported alongside, because an uncorrected estimator can manufacture an "unusually low entropy" finding.

2.3 Zipf's Law

Zipf's law: frequency f(r) of the r-th ranked item follows f(r) ? r??. A least-squares fit of log frequency against log rank yields exponent a and R^2. Zipf compliance is true of Voynichese (a ~ 1.022, R^2 ~ 0.979) but is also true of an order-2 character Markov chain fitted to the manuscript's own glyphs (a ~ 1.064, R^2 ~ 0.985). The statistic does not separate them; therefore Zipf compliance is not evidence that Voynichese encodes language.

2.4 Control Gating and the Verdict Lattice

Every attack carries a hypothesis, controls, and a procedure. Controls are known-answer problems from the same cipher class, generated through the attack's own cipher primitives. The verdict lattice distinguishes REFUTED (controls passed, entire class covered, nothing found) from NOT_FOUND (controls passed, subset searched, nothing found - evidence against, not proof), CANDIDATE (controls passed, something scored in the plaintext band - referral for human review, not a solution), INCONCLUSIVE (controls failed - statement about the tool), and UNTESTED. Controls must match the target's scale: validating at the wrong text length proves nothing, and the dangerous direction is licensing a claim the control never tested.

3. Liber Primus: Falsification, Not Solving

3.1 What Is Being Tested

Hypothesis	Key space	Method	Coverage	Verdict
Affine substitution	812	Full enumeration	Exhaustive	REFUTED
General monoalphabetic substitu	29! ~ 8.8x10^3?	IC invariant + hill-climb	Exhaustive via invarian	REFUTED
Pure transposition	161	Full enumeration	Exhaustive	REFUTED
Deterministic number-theoretic k	6,144	Full enumeration	Exhaustive	REFUTED
Periodic Vigenère / Beaufort	29?? ~ 10??	Exact per-coset ?^2	Exhaustive to L=60	REFUTED
Running key (natural-language te	Unbounded	Direct search + convolution IC	Partial / class-level	REFUTED

Two exclusions are exact without enumeration. Monoalphabetic substitution has 29! keys, but IC is invariant under any bijection on the alphabet - one measurement adjudicates all 8.8x10^3? keys at once. Periodic polyalphabetic ciphers with period L reduce to L independent Caesar cosets; recovering the best shift per coset costs 29 evaluations each, so the full sweep to L=60 over both modes is ~106,000 evaluations, under a second.

3.2 Null Models: Excluding Classes Without Searching Them

For each generative hypothesis, synthetic ciphertexts are simulated at the true text length (n = 11,081) and the same statistic (IC) is measured. The observed IC = 0.03449 is compared against the reference distribution. Compatibility requires both |z| <= 4 and the observation inside the range the null actually produced, because periodic nulls are right-skewed and a modest z can hide an observation below every simulated draw.

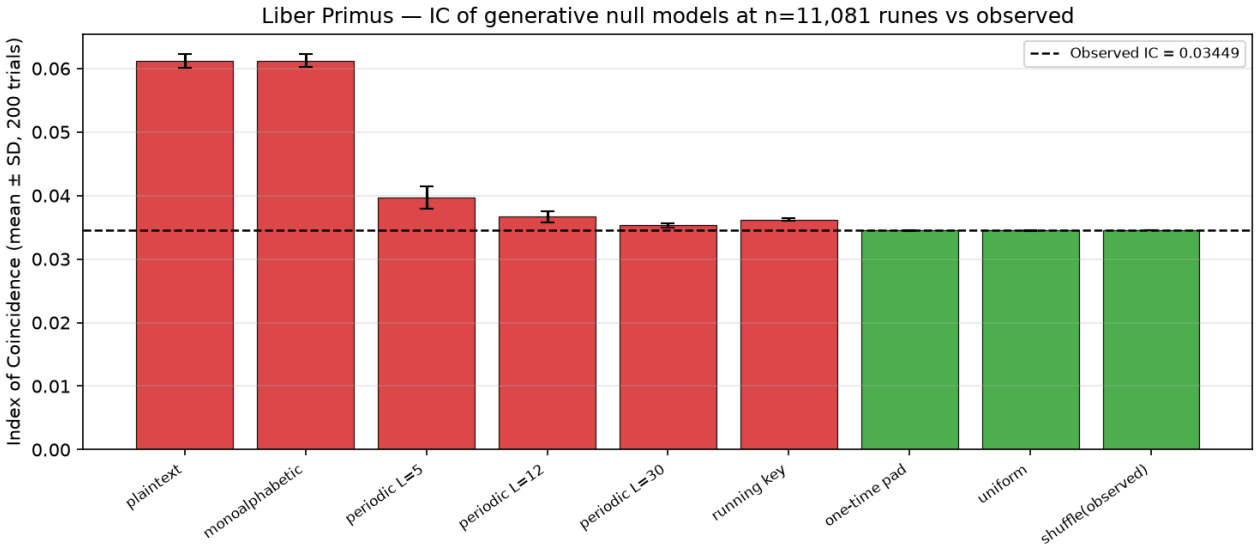


Figure 1 - Index of coincidence of generative null models at n=11,081 runes. The observed value (0.03449) is incompatible with plaintext, monoalphabetic, periodic, and running-key models, and compatible with one-time pad, uniform, and positional shuffle of itself.

3.3 The Running-Key Convolution Argument

A running key enciphers plaintext under a key drawn from another natural-language text. Let P be the plaintext distribution and K the key distribution over ?2?. Under additive combination, the ciphertext distribution C is the circular convolution P ? K. In Fourier terms ?(j) = P?(j).K?(j), and since IC ? 1/n is proportional to the squared L2 norm: IC(C) ? 1/n = (1/n) ???? |P?(j)|^2|K?(j)|^2. Every term is non-negative. The excess over uniform is zero if and only if P? or K? vanishes at every non-zero frequency - i.e., unless plaintext or key is exactly uniform. Convolution flattens; it cannot uniformise. No natural-language key text can drive the excess to zero. The observed IC = 0.03449 lies 10.7s below the running-key null (0.03621 +- 0.00016; in 200 trials the statistic never fell below 0.03576).

3.4 Why the Machine-Learning Architecture Failed

The previous version of the Liber Primus codebase (~3,200 lines) implemented an island-model genetic algorithm, a

MAP-Elites quality-diversity archive, CMA-ES, AlphaZero-style MCTS, a PPO agent with a Transformer actor-critic, and a neural language-model critic - all searching a composable cipher DSL. None produced a usable result.

The decisive measurement: on section 0.0 (already solved, Atbash cipher), beam search returned a cyclic rotation of the true plaintext and scored it **higher** than ground truth (?0.331 vs ?0.386). Because a cyclic rotation leaves unigrams and IC exactly invariant and N?4 of N?3 quadgrams survive unchanged, there are N near-equally-scoring decoys per genuine solution. No score can certify a decryption. The MAP-Elites archive collapsed to a single behavioural cell on real data; the RL agent controlled the operation name while the environment randomised the arguments (the entire problem for Vigenère); the COLUMNAR primitive was silently lossy while flagged invertible; and 19 tests passed because they asserted output length, not recovery.

The general lesson: more optimisation power applied to an unvalidated objective produces more confident wrong answers. For an unsolved cipher, the scarce resource is not compute - it is falsifiability. ML has a place outside the scoring loop, as hypothesis generator and corpus builder whose output a human reviews; never as a judge inside it.

3.5 Homophonic Substitution Analysis (subsequent work)

All exclusions above rely on the IC invariant or its per-coset generalisation, which apply only to bijective symbol mappings. Homophonic/nomenclator substitution - where multiple ciphertext symbols map to one plaintext letter - flattens the unigram distribution before combination and evades the IC test entirely. A variational inference attack learns a soft assignment Q(z|x) from ciphertext runes to plaintext letters by minimising KL divergence between the decoded bigram distribution and a Gematria-normalised English reference model, calibrated against uniform-random nulls at the true text length.

Section	F*	??	s?	z	Verdict
0.7 (roots)	1.99	3.89	0.14	?13.62	COMPATIBLE
0.8 (moebius)	3.32	3.20	0.18	+0.65	INCONCLUSIVE
0.9 (dots/mayfly)	2.35	3.55	0.29	?4.22	COMPATIBLE
0.10 (wing)	3.41	4.69	0.28	?4.57	COMPATIBLE
0.11 (cuneiform)	7.96	4.27	0.15	+24.60	EXCLUDED
0.12 (plants)	3.32	4.28	1.12	?0.86	INCONCLUSIVE

The hard assignments collapse to a single letter for several sections - a known pathology when the signal is too weak to resolve individual mappings. Nevertheless, the z-scores span [?13.6, +24.6], with probability < 10?¹? under homogeneity. The unsolved Liber Primus pages are not a homogeneous cipher; section-specific analysis is necessary, and pooled analysis is invalid for classes not closed under concatenation.

4. Voynich Manuscript: Measurement Against Nulls

4.1 Data and Null Models

Results use the ZL (Zandbergen-Landini) v3b-n IVTFF transcription: 227 pages, 5,385 loci, 37,953 word tokens, 8,423 types, 193,023 glyphs. The independent Takahashi/Interlinear (IT2a-n) transcription (225 pages, 36,907 tokens) serves as control; high-frequency word counts agree within ~1% (daiin 798 vs 805). Four null models carry no lexicon, grammar, or semantics, ordered by retained structure: uniform random + spaces; character-shuffled (preserves glyph inventory and word lengths exactly); order-1 character Markov (retains adjacent-glyph statistics); order-2 character Markov (retains two-glyph context).

4.2 Zipf's Law Decides Nothing

Voynichese fits a rank-frequency power law with exponent 1.022 and R^2 = 0.979. The order-2 Markov chain yields exponent 1.064 and R^2 = 0.985 - at least as Voynich-like on every axis, with no words, no meaning, and 676 states of memory. The naive "monkey" null (uniform + spaces) is distinguishable (a = 0.838, R^2 = 0.818), so a critique resting only on the monkey model would be weak; but the Markov chain settles it. Zipf compliance is excluded from the discriminative battery.

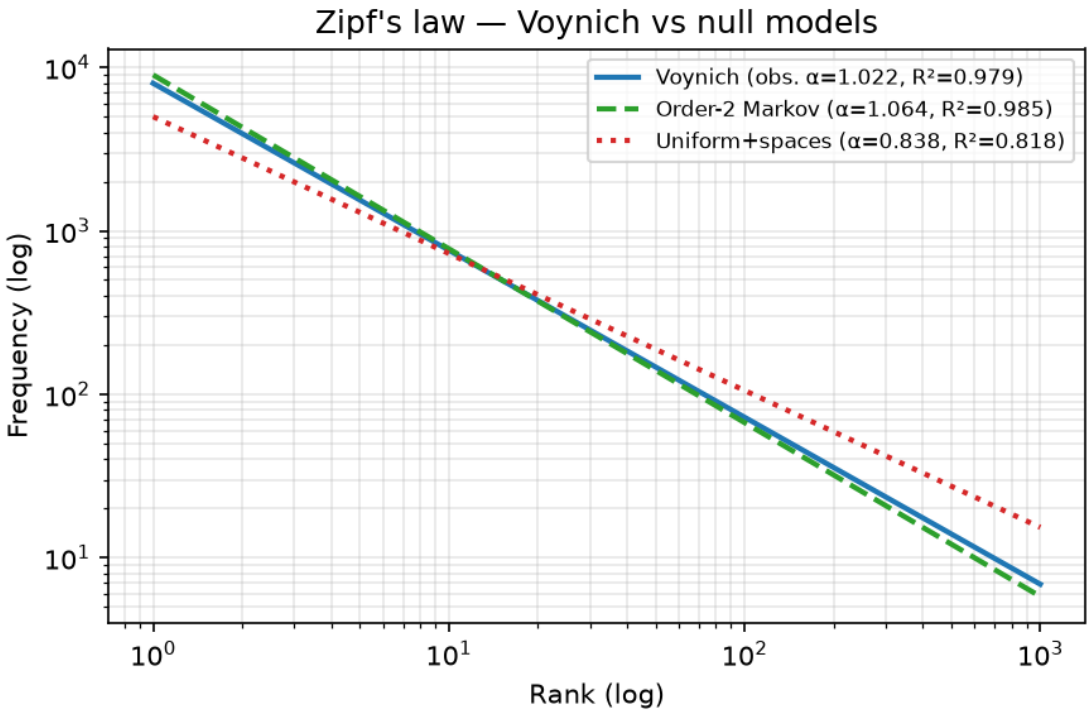


Figure 2 - Zipf rank-frequency curves. Voynichese (a=1.022) is closely matched by an order-2 Markov chain fitted to its own glyphs (a=1.064); uniform random text with spaces (a=0.838) is visibly different. Zipf compliance is therefore non-discriminative.

4.3 The Discriminative Battery

A battery of statistics is computed for the manuscript and all four nulls. A statistic is admitted only if it separates the manuscript from the nulls; a statistic that fails is reported as failing, which is itself a result about the literature that cites it.

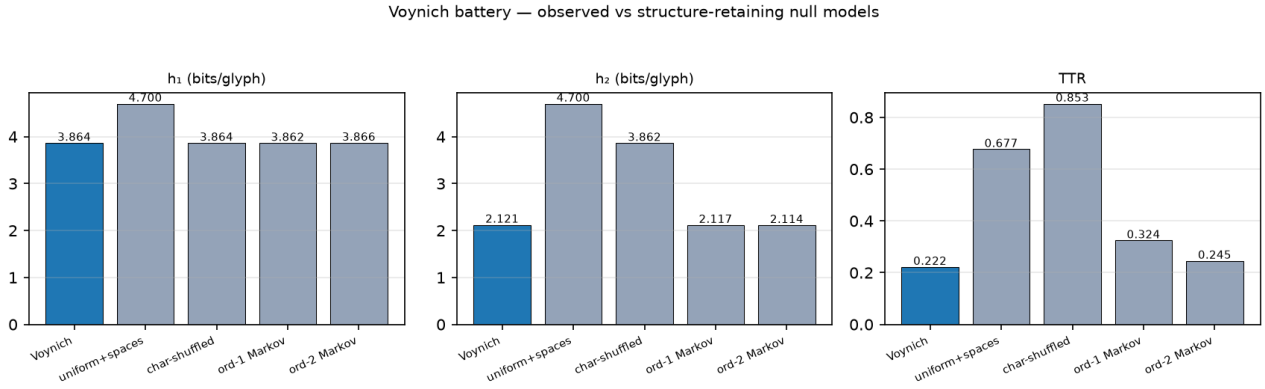


Figure 3 - Discriminative battery. h1, h2, and TTR for Voynich (observed) vs structure-retaining null models. An order-1 Markov chain already reproduces h2, rigidity, mean word length, and adjacent repetition; only word-length variance and TTR separate them, and order-2 narrows both.

Conditional entropy h2 = 2.121 bits for Voynich against 2.117 for order-1 Markov and 2.114 for order-2 Markov. The character-shuffled control at 3.862 bits shows how much predictability is real sequential structure. The correct inference is narrow: these statistics carry very little information about whether the manuscript is meaningful. Meaningful text also has low-order structure - that is why low-order statistics are weak evidence in both directions.

4.4 Two Findings That Survive

Currier A and B are distinct populations. Currier observed in 1976 that the manuscript is written in two statistically distinct systems. Measured here: A has h2 = 2.151, TTR = 0.319; B has h2 = 1.937, TTR = 0.215. With 100 null trials each, the differences on h2, rigidity, and TTR are decisive (|z| >> 2); word-length variance and adjacent repetition are compatible with a single process. Any account must explain both systems; any statistic over the whole manuscript silently averages two populations.

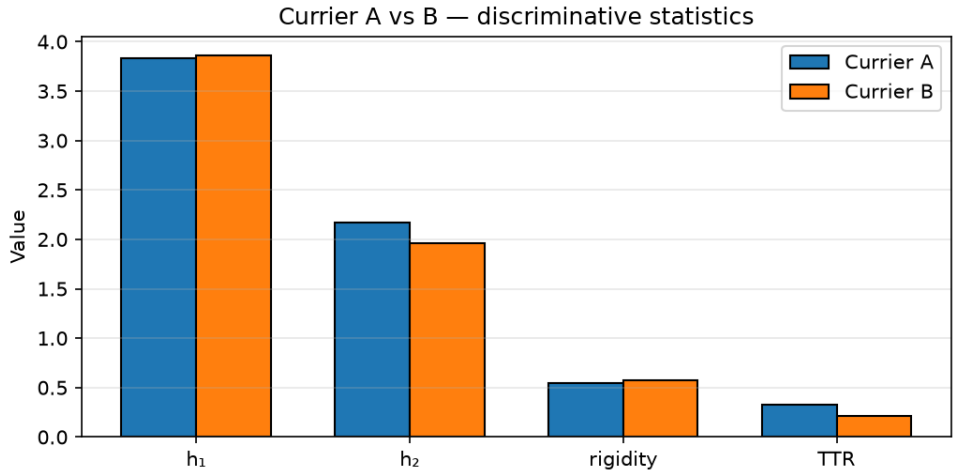


Figure 4 - Currier A vs B on discriminative statistics. A and B differ decisively on h2, rigidity, and TTR, but not on word-length variance or adjacent repetition.

The line is a functional unit. Natural-language text does not know where the scribe will break a line. A substitution cipher acts glyph-by-glyph without reference to layout; a transposition permutes a stream fixed before the lines existed. Measuring total-variation distance between the first-glyph distribution of line-initial words and all other words: TV = 0.385 for Voynich, 0.031 for a character-shuffled control - a 12.3x ratio. At the word level, first-word TV = 0.681 vs 0.029 - a 23.3x ratio. Paragraph-initial lines form an even more distinct subpopulation (first glyphs {p, t, k, f, o} vs mid-line {d, s, y, q, o}). This is incompatible with any substitution or transposition cipher of a natural language where the plaintext is produced independently of the page format.

4.5 Alternative Tokenisation

All statistical conclusions are conditional on the EVA tokenisation. Composite tokenisation (merging ch, sh, cth into single tokens) raises conditional entropy h2 from 2.081 to 2.224, moving it **above** the order-2 Markov null (2.114) - the first tokenisation to meaningfully separate Voynichese from low-order stochastic processes on this statistic. The claim "Voynichese has lower conditional entropy than a 2-gram model of itself" is therefore tokenisation-dependent and not robust. Stroke-based, first-last, and syllabic tokenisations each collapse or distort different statistics, confirming that the tokenisation choice is a modelling assumption, not a neutral formatting step.

4.6 Entropy Rate Convergence

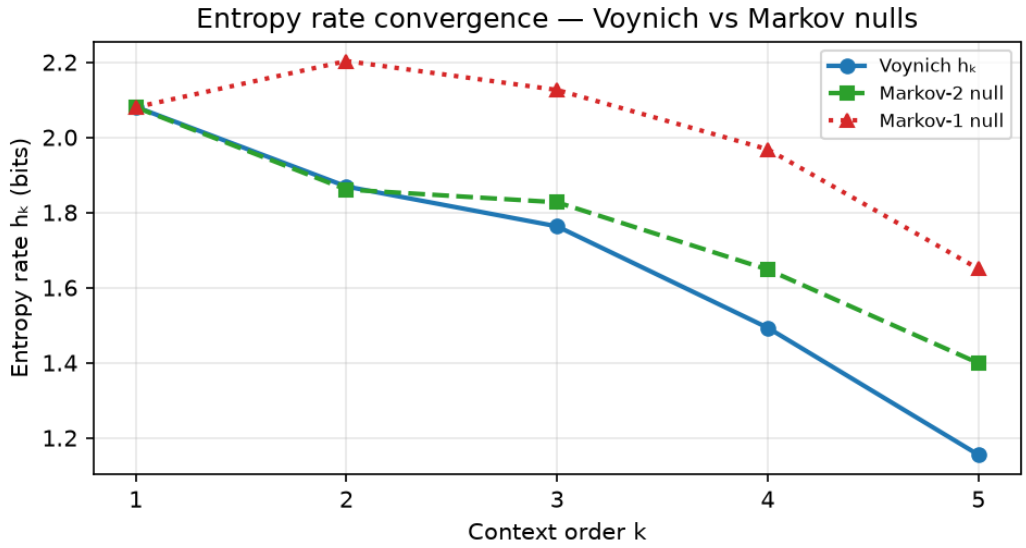


Figure 5 - Entropy rate h? convergence. Voynich h? converges toward the order-2 Markov null (near match at k=2), not toward a stable language entropy rate. This is consistent with a generative process having ~2-glyph memory.

For natural language, h? typically converges toward ~1.0-1.5 bits/char and keeps decreasing slowly. Voynichese's convergence to the Markov-2 null suggests the 2-gram statistics are the dominant structure - consistent with

character-level generative procedures (self-citation / auto-copying models), though consistency is not confirmation.

5. Artificial Intelligence: What Was Tried, What Failed, What Comes Next

5.1 Approaches Tried

Both projects share an AI lineage: a dependency-free linguistic meta-learning workbench based on character 1-4 gram likelihoods, meta-trained with a first-order Reptile update across cipher/key episodes (Caesar shift, affine substitution, periodic Vigenère, plaintext autokey), evaluating fresh keys and held-out windows. The control result: untrained top-1 = 6.25%; meta-initialization top-1 = 100%; scratch + support adaptation = 100%. This is a **control result**, not manuscript evidence - it shows the feature learner ranks the planted answer when the correct cipher family and candidate set contain it. Scratch adaptation also reaches the ceiling, so no adaptation advantage over ordinary task-local learning is demonstrated. The workbench is deliberately isolated from the verdict lattice: learned scores can propose candidates; only controls, coverage, and independent review can support a historical-cipher claim.

5.2 What Failed

In the Liber Primus v1 codebase, the ML apparatus was not merely unhelpful but actively misleading: its objective function ranked a cyclically rotated plaintext above the true one (?0.331 vs ?0.386 on an already-solved page); its quality-diversity archive collapsed to one cell; its RL environment randomised the only decision that mattered; it could not recover a 3-symbol Vigenère key that exhaustive enumeration solves in under a second. In the Voynich predecessor, a "structural scorer" was fitted to a 629-character synthetic placeholder because a broken download URL fell back silently - 21 tests passed throughout. Both failures share one structure: optimisation pressure on an unvalidated objective manufactures confident wrong answers.

5.3 Where AI Belongs

AI belongs **outside the scoring loop**: hypothesis generation, corpus building, orthographic normalisation, and candidate ranking whose output a human reviews and commits. AI does not belong **inside the scoring loop** as a judge, because a learned critic is an objective function, and §3.4 / §5.2 show what happens to objective functions under optimisation pressure on unsolved ciphertexts with no ground truth. Future approaches worth trying: neural scorers as proposal mechanisms with independent statistical verification; language-model priors for homophonic assignment initialisation; transformer-based attention maps over Voynichese line structure as hypothesis generators; and controlled benchmarks where the correct answer is sometimes absent from the candidate set, to measure false-positive rates.

6. Comparative Discussion

| Dimension | Liber Primus | Voynich Manuscript |

|---|---|---|

| Ground truth | Solved pages exist; controls can be built from real plaintext | None; controls would test assumptions, not attacks |

| Core method | Control-gated falsification engine; verdict lattice | Discriminative battery against null models |

| Headline negative | Seven cipher classes excluded; running key excluded by convolution | Zipf, entropy, rigidity all reproduced by order-2 Markov |

| Headline positive | None (no decipherment) | Currier A/B distinct; line-as-unit effect (12-23x over null) |

| AI role | Failed search apparatus; later variational homophonic attack | Meta-learning control; tokenisation sensitivity analysis |

| Main limitation | Single community transcription; homophonic not closed under pooling | Single EVA tokenisation; one transcription for main results |

Both projects converge on the same methodological insight: in domains without ground truth, **falsifiability and

discrimination** are the scarce resources, not compute or cleverness. A negative result that is trustworthy compounds; a positive result that cannot be checked does not.

7. Approaches That Can Be Tried

* Liber Primus: test homophonic substitution under non-English plaintext models (Latin, German, constructed languages); extend the convolution argument to homophonic pre-flattening; run section-specific rather than pooled attacks; investigate composite constructions (substitution -> transposition -> stream).
* Voynich: repeat the full battery under alternative tokenisations (completed for composites; extend to stroke/syllabic with proper controls); test quire-level and hand-level splits; implement historically meaningful generative mechanisms (constrained copying, verbose cipher, abbreviation/position encoding) and measure their failure; add external domain review.
* Both: build benchmarks where the correct answer is sometimes absent; use neural scorers strictly as proposal mechanisms with statistical verification; publish null-calibrated code and data so others can replicate exclusions without repeating failed searches.

8. Conclusion

The Liber Primus unsolved pages are statistically indistinguishable from a one-time pad at n = 11,081, and seven cipher classes - unenciphered plaintext, monoalphabetic substitution, affine substitution, pure transposition, deterministic keystreams, periodic polyalphabetic keys to L = 60, and natural-language running keys - are excluded, each with stated coverage and assumptions. A subsequent homophonic analysis reveals per-section heterogeneity: three sections compatible, one excluded, two inconclusive - the first quantitative evidence that the unsolved pages may not be a homogeneous cipher.

Voynichese follows Zipf's law, but so does an order-2 Markov chain fitted to its own glyphs; the argument from Zipf compliance to linguistic status does not survive contact with a null model. Conditional entropy, positional rigidity, mean word length, and adjacent repetition are all reproduced by a first-order Markov chain. Two findings survive their controls: Currier A and B are distinct populations, and the line-as-a-functional-unit effect - 12x glyph-level and 23x word-level divergence over a matched control - resists any substitution or transposition cipher of natural language.

The methodological proposal is modest and cheap: report every statistic next to null models that ought to fail it, and bind every attack to known-answer controls from the same cipher class. In a field where positive claims cannot be checked, a discipline that makes negative claims trustworthy is the only thing that compounds. Neither artefact is solved. Both are better measured.

. References

* Friedman, W. F. (1922). The Index of Coincidence and Its Applications in Cryptography. Riverbank Publications, 22.
* Knuth, D. E. (1994). A Decryption of a Homophonic Cipher. Cryptologia, 18(3), 267-284.
* Sinkov, A. (1966). Elementary Cryptanalysis: A Mathematical Approach. Mathematical Association of America.
* Currier, P. (1976). Statistical Analysis of the Voynich Manuscript. Unpublished manuscript.
* Zandbergen, R. & Landini, G. (2020). Voynich Manuscript Transcriptions, ZL3b-n. voynich.nu.
* Soranço, M. (2026a). Liber Primus - A Falsification Engine. Self-published manuscript, 12 pp. Code: github.com/Matheussoranco/LiberPrimus.
* Soranço, M. (2026b). Voynich - Measuring the Manuscript Against Things That Mean Nothing. Self-published manuscript, 13 pp. Code: github.com/Matheussoranco/Voynich.
* Soranço, M. (2026c). Homophonic Substitution in the Liber Primus: A Variational Analysis. Self-published manuscript.
* Soranço, M. (2026d). Alternative Tokenisation and Line-Structure Analysis of the Voynich Manuscript. Self-published manuscript.
* Li, M. & Vitányi, P. (2008). An Introduction to Kolmogorov Complexity. Springer.
* Cover, T. M. & Thomas, J. A. (2006). Elements of Information Theory. Wiley.