

Contents

A Candidate Decipherment of the Voynich Manuscript: Evidence for a Spoken Elu-Sinhala Pharmaceutical Register (V4)	1
Abstract	2
1. Introduction	4
2. Why Previous Attempts Failed	6
3. The Decipherment	7
4. Statistical Validation	9
5. Why Not Another Language?	15
6. Grammar Analysis	16
7. Manuscript Architecture	18
8. Plant Identifications	26
9. Bhesajjamañjūsā Textual Parallels	32
10. The Voynich Pharmaceutical Notation System (VPNS)	34
11. Independent Corroboration	38
12. Decoder Error Analysis	39
13. Historical Context	39
14. Limitations and Future Work	40
15. Methodology and Reproducibility	42
16. Conclusion	44
Acknowledgments	45
References	45
Appendix A: Confidence Estimate Decomposition	46
Appendix B: Supplementary Files	47

A Candidate Decipherment of the Voynich Manuscript: Evidence for a Spoken Elu-Sinhala Pharmaceutical Register (V4)

Kameldip Singh Basra

kameldipbasra@gmail.com

May 2026 (Version 4)

This is paper version 4. Version 3 (paper_v3.md) was published on Zenodo on 2026-05-07 (DOI: 10.5281/zenodo.20072618) and is preserved unchanged for reproducibility. Version 4 incorporates V21 meaning corrections (26 word types, 5,307 tokens corrected from DPD + Clough 1830/1892 + BM cross-reference), the COSMO decoder extension (97.6% COSMO assigned-meaning coverage after position-sensitivity fix; 13/27 proposed nakkhatta identifications anchored against BM Ch.19 pharmaceutical nakshatra sequence, of which 3 are confirmed and 10 remain

candidate-level), partial reinterpretation of the initial-sound gap, proposed Tridosha-oriented VPNS readings (kea=vāta-oriented, teda=pitta-oriented, keha=kapha candidate), 81/101 herbal folio plant-identification hypotheses, 23-category BM/AH recipe mapping, and sample prescription-style readings from the decoded corpus.

Abstract

The Voynich Manuscript (Beinecke MS 408, carbon-dated 1404–1438 CE) has resisted decipherment for over a century. We present a candidate decipherment identifying the manuscript as a 15th-century Elu-Sinhala pharmaceutical text — a working pharmacist’s compressed reference (*veda pota*) recording Ayurvedic preparations in a bespoke abugida. The writing system maps EVA characters to Sinhala phonemes via systematic positional rules, with *q-* and *ch-* preserved as phonologically-conditioned allomorphs of a single deictic morpheme (*q-* before consonant-initial bases, *ch-* before vowel-initial bases; 5,328 and 4,762 corpus tokens respectively).

Applied to the V20 corpus of **36,633 tokens**, the canonical V17 decoder with V21 meaning corrections produces a high-coverage assigned reading: **98.16% of tokens have a meaning_fixed assignment** in the current database (1.62% true unknowns — both meaning and meaning_fixed fields empty; 595 tokens / 30 types, predominantly hapax forms). This is not the same thing as independently measured translation accuracy. Re-auditing the current DB shows that only a much smaller share has simple, non-compositional, non-contextual assignments; the exact percentage depends on how strictly polysemous and VPNS-compound glosses are counted. The V4 coverage numbers should therefore be presented as **reading coverage / assignment coverage**, not as proof of 93-98% translational accuracy. The improvement reflects V21 systematic meaning corrections (26 word types, 5,307 tokens) checked against DPD, Clough, and BM cross-reference, plus the COSMO prefix extension and final compound-resolution pass.

The initial-sound gap is partly reinterpreted in V4 but should **not** be marked resolved. Initial /h/-loss is a real Sinhala/Elu-compatible pathway for some lexemes, but the broader absence or rarity of /b/ and /v/ remains a decoder-level risk unless independently shown by a larger comparison corpus and by specialist review. The V4 evidence supports a plausible explanation for part of the gap; it does not by itself justify revising decoder accuracy from v3’s ~72% to ~93%.

The Voynich Pharmaceutical Notation System (VPNS) may encode Tridosha-oriented preparation axes, but this should be framed as a **candidate interpretation**. kea as sky/ether is externally supported by Clough's kear = swargg/heaven, and teda remains a strong preparation-state marker. The new khea = kapha/phlegm proposal is currently speculative: Pali semha does mean phlegm, but the proposed semha -> khea sound path is not yet independently established. Until that path is verified, the Tridosha completion should not be used as a confidence-raising proof.

The COSMO section (f67r-f73v) is best treated as a **candidate timing / calendar index**. The Pali nakkhatta list has 27 items. V4 proposes 13/27 candidate identifications, anchored by the Bhesajjamañjūsā Ch.19 pharmaceutical nakshatra sequence (stanzas 59-69). Three are confirmed: acena=Aśvinī, alula=Anurādhā, makhea=Māghā. Ten additional candidates are at MEDIUM or MEDIUM-HIGH level, all cross-validated against the BM Pali forms. ata=Hattha is a new addition based on both BM "hattena" and the semantic identity ata=hand=hattha. The remaining 14 unresolved nakshatra positions likely reflect the same pharmaceutical non-priority selection found in the BM, which itself skips 9 of 27 nakshatras.

Sample prescription-style readings are extracted for the first time. **f112v L45** (Gahani/digestive-deficiency chapter candidate, BM Ch.30 parallel): "agea nara ch-ka ch-agenā ch-ara ch-ea kena dala kena galada" = "Fat-extracted preparation: bring one dose for the patient; perform with ghee; when done, administer; having strained, give." This worked reading encodes drug type, recipient (nara=patient/person), dosage candidate (ch-ka=one), vehicle candidate (ch-ea=ghee), and administration route (galada=strain-and-give). **f113r L1** (Bhūta-cikitsā candidate, BM Ch.57 parallel): "pesaca apeda cada ch-etara tea" = "Evil spirit: affliction, incantation, seed preparation, root-liquid" — compatible with a combined exorcism+herbal protocol for piśāca-caused disease.

The herbal section plant table contains 81/101 proposed identifications, 18 currently marked HIGH. These should be treated as **candidate identifications pending blind botanical review**. Several are strong visual/lexical hypotheses, but the method is not blind and can suffer confirmation bias because the decoded label and image were read together. The 23 recipe therapeutic categories are likewise a candidate BM/AH mapping that needs null-baseline and specialist review before being described as complete.

The V2/V3 core still supports a low wrong-probability for broad origin and purpose, but V4's new interpretive layer does **not yet justify lowering** the estimate below v3's

~7-10%. The major remaining uncertainty is still sister-language indistinguishability, plus specialist review of the meaning assignments, VPNS labels, COSMO labels, and plant identifications.

1. Introduction

The Voynich Manuscript (Beinecke MS 408) is a 234-page illustrated codex held at Yale University’s Beinecke Rare Book and Manuscript Library. Radiocarbon dating of the vellum places it at 1404–1438 CE with 95% probability. For over a century, analysts sought an encryption key that may not exist. Under the present hypothesis, the manuscript was not encrypted but transcribed into a bespoke phonetic notation.

We present a candidate decipherment. The strongest current reading is that the Voynich Manuscript is a 15th-century Sri Lankan Elu-Sinhala pharmaceutical text: a working pharmacist’s compressed reference recording Ayurvedic preparations in a bespoke abugida-like notation. The broad origin-and-purpose claim is much stronger than the word-by-word translation claim. The genre signal is working-pharmacist compressed notation: overwhelmingly inanimate, formulaic, and much closer to drug-formulary register than narrative prose.

1.1 Engagement with prior skeptical work

Three categories of prior analysis have challenged the manuscript’s meaningfulness.

Hoax hypotheses (Rugg 2004; Greshko 2025) propose Cardan-grille or other generative procedures. These are addressed structurally: the manuscript has consistent phonotactic constraints, productive morphology, and section-specific vocabulary clustering that random generation does not reproduce. Greshko’s (2025) Naibbe cipher analysis achieves Spearman $\rho = 0.929$ character-frequency match with our decoded forms — independently supporting the phonetic abugida model.

Statistical-fingerprint critique (Lindemann & Bown 2020; Bown & Lindemann 2021; Gaskell & Bown 2022) measures Voynichese against natural-language norms and against intuitively-produced gibberish, finding that Voynichese statistically resembles gibberish rather than narrative prose. We engage this analysis directly in §4.13 by replicating their random-forest classifier on V17-decoded Voynich and on authentic Sinhala recipe text (Bodleian MS Sinh.a.2(R)). Authentic Sinhala recipe classifies as gibberish ($P=43.7\%$); V17-decoded Voynich classifies as meaningful ($P=58.9\%$); raw

Voynich classifies as gibberish (P=24–34%). The classifier separates recipe register from narrative prose, not meaningfulness from meaninglessness.

Linguistic-substrate proposals (various authors proposing Hebrew, Latin, Welsh, etc.) have failed to produce running translations. Our cross-language discrimination tests (§4.14) show Sinhala leading closest sister-languages by +11 percentage points and refuting Tamil decisively ($\chi^2 = 389$ vs critical 42.6 at $p < 0.001$).

1.2 The decipherment hypothesis in one paragraph

The manuscript appears to be a 15th-century Sri Lankan working-pharmacy text — plausibly connected with a viharasthana physician or Buddhist monastery-hospital milieu — recording Ayurvedic preparations in a personal compressed notation. The Voynich script is interpreted here as a bespoke abugida influenced by palm-leaf inscription conventions. The Voynich Pharmaceutical Notation System (VPNS) appears to compress kalpana-class taxonomy into a productive [prefix]+[base consonant]+[aspect suffix] grid. V4 further proposes, but does not yet prove, that some VPNS bases encode Tridosha-oriented preparation axes (kea=vāta-oriented, teda=pitta-oriented, khea=kapha candidate). The manuscript is best modelled as a cross-referenced pharmaceutical compendium in six sections: HERBAL (plant descriptions and preparations), BALNEO (bathing formulas), PHARMA (mineral medicine), COSMO (candidate timing index), ASTRO (timed processing schedule), and RECIPE (therapeutic categories of compound formulas with BM/AH source-chapter parallels).

1.3 Version history and scope of this paper

This paper is version 4. Version 1 (February 2026) introduced the decipherment and primary statistical validation. Version 2 (DOI: 10.5281/zenodo.20023733, 2026-05-04) incorporated the V17 decoder promotion, Bovern engagement, falsification probes, and initial VPNS external grounding. Version 3 (DOI: 10.5281/zenodo.20072618, 2026-05-07) incorporated the V20–V27 findings: full corpus expansion, vocabulary coverage improvements, VPNS expansion to 21 states, 25 BM formula clusters, and manuscript architecture analysis including COSMO cross-reference quantification. Version 4 (this audit draft) adds V21 meaning corrections with dictionary-source citations, a partial reinterpretation of the initial-sound gap, candidate Tridosha-oriented VPNS readings, a COSMO decoder extension with candidate nakkhatta identifications (now 97.6% COSMO coverage), an 81-folio herbal plant-identification hypothesis table, 23-category BM chapter mapping, sample prescription-style readings, and the

fluent-reading workbench (Elu morphology table, per-section grammars, parallel alignment). 5 new plant identifications added from Sivapalan 2019 cross-reference. Unknown token resolution now at 98.16% meaning_fixed coverage (1.62% true unknowns). Wellcome MS Sinhalese 22 (170 pages) and MS Sinhalese 6 (85 pages) acquired for specialist review; automated OCR quality insufficient for palm-leaf script.

2. Why Previous Attempts Failed

For 112 years, all serious attempts at decipherment assumed that the Voynich Manuscript was encrypted using a cipher applied to some natural language. This framing generated two generations of failed work.

The core error was categorical rather than technical. The manuscript is not encrypted; it is transcribed. The Voynich script is a bespoke abugida — a writing system in which each base character denotes a consonant and diacritics modify the inherent vowel. Abugidas are the normal writing technology of South Asian languages; they were never used in medieval European cryptography. The failure to consider an Indian Ocean provenance for an early 15th-century codex reflects the Eurocentrism of the field rather than any deficiency of the material.

Two further errors compounded the first. Analysts accustomed to European scripts applied letter-frequency tests calibrated for alphabetic systems to what is positionally-encoded abugida text, generating artifactual anomalies (the supposed high frequency of EVA e being a doubled inherent vowel in the abugida framework, not a letter). Simultaneously, statistical approaches that compared Voynichese to narrative prose registers implicitly assumed that meaningful text of the period would be prose narrative. A working pharmacist's compressed notation — with its formulaic repetition of ingredient chains, preparation state markers, and processing verbs — looks statistically unlike narrative prose in every language.

The one genuine insight from the cipher-focused tradition was structural. D'Imperio (1978/2009) noted positional regularities in EVA characters that are precisely what abugida vowel diacritics produce. Currier's (1976) identification of two registers (A and B) corresponds, under our decoding, to the herbal/cosmological section and the balneological/recipe register respectively — not two languages but two pharmaceutical subregisters of a single language (Sinhala), as demonstrated by identical Sārārtha Saṃgrahaya match rates across all Currier-A-score bands (53–57%).

3. The Decipherment

3.1 Decoder evolution

Five major decoder versions were developed; **V17 is the canonical production decoder**. A V20 NL-merger was applied to the corpus database without changing the decoder rules, expanding the token count by splitting previously-merged NL compounds. V21 (this version) applies systematic meaning corrections to the corpus meaning field without altering phonetic decoding rules.

Version	File	Status
H12	Paper/scripts/h12_decoder.py	Original canonical, locked for paper-v1 reproducibility
V14	Paper/scripts/v14_decoder.py	Initial-o stripping, vowel normalisation, EVA-eo→/u/
V15	Paper/scripts/v15_decoder.py	V14 + monotonically-gated medial-o-strip; paper-v1 canonical
V16	Paper/scripts/v16_decoder.py	V15 + single-letter passthrough, long-token splitting
V17	Paper/scripts/v17_decoder.py	V16 + preserves q-/ch-prefix markers; daiiin patch
V20	corpus only	NL-merger compound splitting (no phonetic rule changes)
V21	corpus meaning_fixed field	26 word-type meaning corrections; COSMO prefix extension

V17’s core innovation is preservation of q- and ch- as distinct surface phonemes rather than collapsing both to u-. This resolves the v1 “primary open question”: the u-prefix

anomaly drops from 13.8% (V15) to 5.0% (V17), close to the 2-3% Bodleian Sinhala recipe baseline.

V21 corrects the **meaning field** (not phonetic decoding) for 26 high-frequency word types using three external sources: Pali DPD dictionary, Clough Sinhala dictionary material, and BM cross-reference. Stronger corrections include kea “sky/ether/heaven” (Clough kear=svargg/heaven), tara “comparative suffix (-er/more)”, kena “instrumental particle (by means of)”, mura “Moringa oleifera seed”, nagara “dry ginger”, saleda “Boswellia serrata”, and kula “species/type/genus”. The new khea = kapha/phlegm reading is retained only as a hypothesis until the proposed semha -> khea phonological path is externally validated. Full correction list in supplementary V21_MEANING_CORRECTIONS.md.

3.2 The morphological status of q- and ch-

Earlier framings of q- as a definite article and ch- as a demonstrative have been replaced by a simpler and empirically stronger model. Analysis of 102 shared base forms that take both q- and ch- shows clean complementary distribution: **q- attaches before consonant-initial bases, ch- before vowel-initial bases**. The two prefix shapes are phonologically-conditioned allomorphs of a single deictic morpheme. In Sanskrit/Sinhala terms, q- corresponds to the sa- (“with/the”) prefix before consonant clusters; ch- to its prevocalic form.

3.3 The phonemic mapping

The Voynich abugida maps EVA characters to Sinhala phonemes through a set of positional rules. Characters encode consonants with inherent vowel /a/; EVA-e and EVA-ee function as inherent-vowel modifiers; EVA-i-sequences encode vowel length and quality distinctions. Full mapping in the supplementary decoder files; key mappings: EVA q = /q/ (deictic prefix); EVA ch = /c/ (allomorphic deictic); EVA k = /k/; EVA t = /t/; EVA m = /m/; EVA l = /l/ or /n/ (NL-merger contexts); EVA s = /s/; EVA d = /d/; EVA r = /r/; EVA g = /g/; EVA p = /p/; EVA f = /ph/; EVA h = /h/ (aspirate, regularly deleted in Elu word-initial).

Phonotactics are overwhelmingly vowel-final (99.7% of decoded tokens), exactly as expected for Elu, which has a strongly open-syllable phonotactic constraint.

4. Statistical Validation

4.1 Coverage and the V21 improvement

Version 3 reported 92.9% fully resolved at the meaning level. The current V4 database assigns a meaning_fixed value to **98.16%** of tokens, but this is assignment coverage, not independently measured translation accuracy. True unknowns (both fields empty) are 1.62% (595 tokens, 30 types) — predominantly hapax compounds and rare section-specific forms. A fresh audit of the current database finds that the simple-strong tier is much lower once compositional VPNS glosses, context-derived readings, particles, polysemous tags, and unresolved variants are separated out.

Tier	Definition	V3 %	V4 %
Tier A: Direct external attestation	Word+meaning attested in BM (specific stanza cite), DPD, or Clough; minimal inference	—	11.0% (4,015 tokens)
Tier B: Dictionary-supported	Form in DPD/Clough with meaning fitting Voynich context	—	14.0% (5,117 tokens)
Tier C: Rule-generated	Deictic compounds (q-/ch-), absolutive (-la), state markers (-eda); rules validated but tokens not individually attested	—	35.7% (13,077 tokens)

Tier	Definition	V3 %	V4 %
Tier D: Context- inferred	Plausible pharmaceutical context reading; no direct external attestation	—	38.1% (13,963 tokens)
Tier X: Unknown	No meaning assigned	2.5%	1.3% (461 tokens)

The V4 improvement reflects three V21 actions: (1) 26 word-type meaning corrections from external dictionary sources replacing {dict} placeholder entries (5,307 tokens); (2) COSMO prefix-rule extension (107 tokens, applying ok-/ot- rules at word-initial position); and (3) compound resolution of remaining high-frequency unclear forms (403 tokens: Moringa, Boswellia, red ochre, barley, foam/latex, butter, therapeutic-property marker). These are useful reading-layer improvements, but they must be validated separately from the phonetic decoder.

Pharmaceutical section assignment coverage under meaning_fixed: RECIPE 98.9%, BALNEO 99.5%, PHARMA 97.2%, HERBAL 97.3%. COSMO reaches **97.6% assigned** (2.01% true-null). ASTRO 97.0% (2.46% true-null). These are coverage metrics; they should not be restated as accuracy metrics until tested against blind holdout lines or external specialist readings.

Decoder frozen as of 2026-05-09 (V17 phonetics + V21/V27 meanings).

All tests below use this frozen state. Post-freeze changes logged in supplementary/POST_FREEZE_CHANGES.md. See supplementary/FROZEN_DECODER_V17_V27.md for the full freeze specification and supplementary/AUDIT_TESTS.md for all null tests.

4.2 Three primary tests (pre-registered, Bonferroni $\alpha/3 = 0.017$)

Unchanged from v3. Combined Fisher's $p = 4.5 \times 10^{-10}$.

Test	Z (constrained dual-null)	Z (within-corpus)	
Keyword-section clustering	3.40	30.30	4.3 (154/16)
Pharmaceutical collocations	5.32	—	21.0 (154/16)
External pharmaceutical vocabulary	3.40	73.81	

4.3 V25 validation battery

Seven additional validation tests, unchanged from v3, all pass. Key figures:

Test 1: Sinhala 69.5% corpus match; Hindi 58.5%; ratio 1.19 $\times$; $\chi^2 = 19.86$, $p = 0.0005$. Sinhala:Tamil = 3.31 $\times$.

Test 2: Pharmaceutical concentration BALNEO 60.9%, RECIPE 51.8%, HERBAL 38.8%, COSMO 41.8%. $\chi^2 = 910.79$, $p < 0.000001$.

Tests 3-7: All pass unchanged. Section Jaccard distinctiveness, hapax rate ordering (COSMO 27.3% > HERBAL 18.7% > RECIPE 14.2% > BALNEO 9.0%; $\chi^2 = 612.48$), timing-term cross-section distribution — all as v3.

4.4 Herbal visual matching — independent non-linguistic validation

Unchanged from v3. 92.2% of 110 herbal folios show visual morphological features matching decoded text vocabulary predictions — plant-part tags, growth form, root/rhizome type, and preparation-register markers (leaf, root, flower, tuber).

Verdict	Count	%
MATCH	80	77.7%
PARTIAL	15	14.6%
NO	8	7.8%
MATCH + PARTIAL	95	92.2%

Five of eight NO cases involve ata (thorn/Datura) prediction failing — confirming ata is a recipe ingredient, not a morphological label for the illustrated plant.

Scope note: This 92.2% figure is a **register-level morphology match** — it measures whether the decoded vocabulary predicts the correct plant-part type and preparation register for each folio, not whether the species-level identification is correct. The species-level headline-label test (§8.0) produces a separate and much weaker result (0% top-1 species hits). Both results coexist: the plant section demonstrably operates as a pharmaceutical preparation register (92.2% register match); individual species identifications remain candidates pending specialist review.

4.5-4.12 Additional validation tests

Tests 4.5 through 4.12 (SOV syntax Z=8.10, phonotactics 99.7% vowel-final, gender/animacy 0.2% animate, Elu chronolect revalidation, compositional morphology,

v101 character collapse $\chi^2=22,713$, directional perplexity, Greshko character-frequency correlation $\rho=0.929$) are unchanged from v2 and all hold.

4.13 Engagement with the Gaskell-Bowern null model

Unchanged from v3.

Corpus	P(meaningful)	Verdict
English narrative	95.31%	Meaningful
Sārārtha Saṃgrahaya	80.38%	Meaningful
V17-decoded Voynich (full)	58.90%	Meaningful
V17-decoded Voynich (recipe)	55.78%	Meaningful
Bodleian MS Sinh.a.2(R)	43.70%	Gibberish
Raw Voynich transcriptions	24–34%	Gibberish

V17 decoding moves Voynich from raw-EVA gibberish (31%) to meaningful-side (59%), a +28pp shift. The Bowern suite places V17-decoded Voynich 1.75× closer to the Bodleian Sinhala recipe than to English narrative.

4.14 Cross-language discrimination

Unchanged from v3.

Language	V17 corpus hit	Sinhala lead	Status
Sinhala (1.47M dict)	69.5%	—	#1 in all sections
Hindi (127K)	58.5%	+11.0pp	Closest sister; ratio 1.19×
Tamil (5.3M)	21.0%	+48.5pp	Refuted; $\chi^2 = 389$
Sanskrit	52.4%	+17.1pp	Below Hindi
Hebrew	0.0%	n/a	Refuted
Arabic	17.5%	+52.0pp	Refuted

4.15 Initial-sound gap: partially reinterpreted, not closed (V4 audit)

Version 3 (§12) flagged the near-zero word-initial /h/, /b/, /v/ as an unresolved open question. V4 gives a plausible partial explanation, especially for /h/, but the issue should not be marked closed.

Initial /h/ = rare/absent: plausibly compatible with Elu. Wilhelm Geiger (1938) A Grammar of the Sinhalese Language documents h-loss and h-instability in Sinhala/Elu historical phonology. The pharmaceutical example Sanskrit/Pali haritaki -> Sinhala aralu shows that a medically central h-initial Indic form can surface without initial /h/. This supports the plausibility of the gap, but does not by itself prove the decoder is correct.

Initial /b/ = rare/absent: still a risk. The proposed bulu -> ulu pathway and Bodleian attestations are relevant, but the claim needs a wider Sinhala/Elu medical lexicon baseline before it can be treated as a general phonological explanation.

Initial /v/ = rare/absent: still a risk. The manuscript may prefer native or register-specific terms over Sanskrit v-initial forms, but this remains a vocabulary-selection explanation rather than an independently demonstrated sound law.

Initial /n/ = 1.1%: PRESENT. The V20 n/l merger correctly resolves 49 word-initial tokens as /n/-initial: nagenā, nena, nara, nagara (dry ginger), nasa (nasal). This is appropriate for Elu medical vocabulary.

Accordingly, the initial-sound issue can reduce concern about some apparent mismatches, especially /h/, but it should **not** be used to revise decoder accuracy from v3's roughly 70% range to ~93%. The defensible statement is narrower: V4 offers plausible historical-phonological explanations for part of the gap, while /b/ and /v/ require external Sinhala/Elu medical-register baselines and specialist review.

4.16 V21 meaning correction highlights (V4 new)

The following corrections are the most evidentially significant:

khea = kapha/phlegm candidate. The V3 interpretation (khīra = milk preparation) is not secure, and the V4 kapha/phlegm proposal is interesting but not yet externally grounded. Clough supports the related kea/kear “heaven/sky” family, and Pali semha means phlegm, but the proposed semha -> khea pathway is not yet established by independent Sinhala/Elu comparanda. This should be treated as a candidate reading, not as a completed Tridosha proof.

kena = instrumental particle. DPD kena 1: “by what; with what; how.” BM cross-reference confirms: tikena (BM pharmaceutical passages) = “by means of the triad/trikaṭu” — instrumental use in pharmaceutical context. Previously decoded as “scratch/eat” (WRONG). The correction makes recipe lines grammatically coherent: DRUG + kena + PREPARATION = “DRUG by means of PREPARATION.”

tara = comparative suffix. DPD tara 2: “more; -er.” The dominant collocation “tara ara” (tara followed by ara × 25) = “having surpassed/exceeded” — a process completion marker. In COSMO context, tara retains the “nakshatra crossing” meaning (DPD tara 1). Previously decoded as “you familiar + ra” (WRONG).

mura = Moringa oleifera. Clough (1830): “mura seed, see Sod m0 (guilandina moringa)” — guilandina moringa is an old taxonomic synonym for Moringa oleifera (drumstick tree). Moringa seed is one of the most important Ayurvedic ingredients — used in oil preparations, wound treatment, and as a vehicle. Appears 160 times in corpus including in COSMO timing entries (f70r1).

nagara = dry ginger. In classical Pali/Sinhala medical texts, nagara is a documented name for dried Zingiber officinale (śuṇṭhī). Appears 59 times in corpus; previously decoded as “city/spring-water” (WRONG for pharmaceutical context).

4.17 Non-circular internal structure tests (V4 new)

Three tests were run that are independent of English glosses and botanical interpretation. They depend on V17 tokenization but not on any subsequent word-meaning assignment; they test whether the manuscript has internal structure consistent with a real pharmaceutical register, regardless of whether any particular decoded word has the correct English gloss.

Section classifier (7-way). A cross-validated section classifier trained on token-level features achieves **61.0% accuracy** vs a 32.3% chance baseline ($p=0.0099$). The section boundaries are identifiable from internal vocabulary distributions without reference to word-meaning assignments.

KALPANA enrichment. KALPANA-class preparation markers (the productive [prefix]+[base]+[suffix] VPNS grid) are strongly enriched in RECIPE and BALNEO sections vs HERBAL and COSMO. Odds ratio = **8.11** ($p = 2.06 \times 10^{-222}$). This is a non-lexical structural signal: the same token shapes that function as preparation-state markers in formula contexts are systematically rare in plant-description and timing contexts.

q-/ch- allomorph complementary distribution. The q- prefix attaches before consonant-initial bases; ch- attaches before vowel-initial bases. The odds ratio for this complementary distribution (q-before-consonant vs q-before-vowel) is **32.81** ($p \approx 0.0$, Fisher exact). This is a phonologically-conditioned morphophonological rule — the kind of regularity that does not arise by chance in random text or hoax generation

and is not derived from any lexical assignment.

These three tests together confirm that the manuscript has real register structure and real morphophonological conditioning that would be absent in randomly generated or hoax text, regardless of whether the word-level glosses are correct.

5. Why Not Another Language?

The five independent constraints from v3 (against Tamil, Pali, Hindi/Sanskrit, European languages, and for pre-12c Elu chronolect) all hold. V4's /h/-loss discussion is compatible with the Elu case but should be treated as supporting context rather than a new decisive language-discrimination test.

Against Pali specifically: V4 adds a possible discriminating datum if the Tridosha-oriented VPNS readings can be externally validated. At present, kea has better support than khea; therefore the Tridosha layer should not be used as independent proof that the language is Elu-Sinhala rather than Pali.

5.1 Locked-anchor language comparison (V4 new)

V4 adds a genre-matched control test that bypasses the raw word-list collision problem. A set of **33 locked anchors** was extracted mechanically: all decoded forms with evidence tier A or B, token count ≥ 5 , and length ≥ 4 characters. These anchors were tested for repeated-use recurrence in actual historical medical texts — not title lists — without any prior knowledge of which texts should score highest.

Control corpus	Type	Repeated-anchor hit rate
Sārārtha Saṃgrahaya (Sri Lankan Sinhala med.)	full text	66.67%
Bhesajjamañjūsā (Sri Lankan Pali med.)	full text	27.27% repeated-anchor (42.42% raw)
Ortus Sanitatis (Latin European herbal)	full text	15.15%
Culpeper Complete Herbal (English herbal)	full text	0%

Control corpus	Type	Repeated-anchor hit rate
Siddha formularies (PCIMH official)	full text	0%
Unani formularies (PCIMH official)	full text	0%
Tamil medical glossaries	full text	0%
European medieval medicine (14c)	full text	0%

BM row is the Sri Lankan medical baseline (not a rival); raw vs. repeated figures included to show the metric's sensitivity on a known-related text.

The Sārārtha lead of 4.4× over the best European comparator is not a raw collision artefact: the locked anchors are long enough (≥ 4 characters) and tier-constrained to eliminate the short-word collisions that inflate raw overlap counts. The European 15.15% figure comes from shared short pharmacological vocabulary (meda, gena, kula), which disappears when only anchors ≥ 4 characters with Ayurvedic-specific glosses are counted.

Raw word-list counts for Turkic, Malay, Semitic, and other non-Indic languages show superficially high collision rates (up to $\sim 28\%$ for raw Turkic title lists). These collapse when filtered to repeated-use, length-controlled anchors. The honest summary is: **raw collision is dangerous; repeated-use and structure testing collapse those alternatives.** No tested rival equals or exceeds the Sri Lankan medical corpus on the locked-anchor metric.

6. Grammar Analysis

6.1 Sinhala grammatical feature checklist

Unchanged from v3. Twelve of nineteen Sinhala features confirmed; zero non-Sinhala markers; six medieval-chronolect indicators confirmed.

New in V4: **nara (person/patient)** appears 123 times corpus-wide (DPD confirmed: nara = man/person). The distribution confirms the patient-directed prescription register: nara appears in RECIPE and BALNEO in clinical-instruction contexts. These

tokens function as dative/recipient arguments in pharmaceutical instructions (“bring [the preparation] for the patient”), not as animate-marked grammatical subjects; see §6.4 for the distinction between the grammatical animacy test and patient-reference vocabulary.

6.2 Elu chronolect revalidation

Unchanged from v3. 81% pre-12c Elu-native vocabulary; 0% post-12c Sanskrit tatsama loans; all six medieval-chronolect indicators confirmed.

V4 additions to the chronolect evidence: (1) kea = Sinhala kear = heaven/sky (Clough dictionary material, consistent with the proposed semantic field); (2) mura = moringa seed (Clough dictionary material); (3) nagara = dry ginger (Pali/Sinhala pharmaceutical usage). The khea = kapha/phlegm proposal is not counted as chronolect evidence until the phonological path is externally established.

6.3 The -uina family: retroflex nasal analysis

Unchanged from v3. tuina=ṭṛṇa(grass), muina=muna(face), ruina=vraṇa(wound), luina=lūna(harvested/cut).

6.4 Gender and animacy

Updated from v3. Under V21, maga (previously “dumb person” artefact) is correctly decoded to pharmaceutical meanings throughout, removing a false animate count. The v3 **grammatical animacy test** (animate-marked verb agreement and pronoun forms) gives ~73 tokens (0.2% of corpus) — unchanged, because that test measures morphological animacy agreement, not lexical reference to living beings.

Separately, **patient-reference vocabulary**: nara (patient/recipient, 123 tokens), pe-saca (piśāca/evil spirit, 14 tokens), naga (nāga/deity-serpent, 6 tokens), pamadara family (woman, ~12 tokens) — total ~155 tokens (0.4% of corpus). These tokens identify living referents lexically but function in pharmaceutical instruction syntax as dative/recipient or ritual-disease arguments, not as animate-marked subjects. The 0.2% grammatical animacy rate and the 0.4% patient-reference lexical rate are both low and consistent with an inanimate pharmaceutical register; they measure different things and should not be summed or conflated.

7. Manuscript Architecture

7.1 Overview: a unified cross-referenced pharmaceutical compendium

Section	Folios	Tokens	% Corpus	Function
HERBAL	f1r-f57r	9,352	25.5%	Plant morphology, identification, and preparations
BALNEO	f75r-f84v	6,865	18.7%	Therapeutic bathing formulas
RECIPE	f103r-f116r	10,619	29.0%	23 therapeutic categories, BM/AH chapter-mapped
PHARMA	f87r-f102v	5,010	13.7%	Mineral and metallic medicine
COSMO	f67r-f73v	2,891	7.9%	Nakshatra-indexed temporal prescription calendar
ASTRO	f57r-f66v	1,421	3.9%	Timed processing schedule
OTHER	f85r	475	1.3%	Master preparation map (Rosette fold-out)

Every major section now corresponds to a named chapter-type in classical Ayurvedic literature: HERBAL = dravya-nighanṭu; BALNEO = snāna-karma; PHARMA = dhātu-

cikitsā; COSMO = ṛtucaryā/nakṣatra-cikitsā; RECIPE = cikitsā-sthāna; ASTRO = production schedule; Rosette = preparation index.

7.2 The HERBAL section

Updated from v3. A plant-identification hypothesis table for 81 of 101 informative herbal folios is now available (see §8 and supplementary HERBAL_PLANT_IDENTIFICATIONS.md). The HERBAL section is modelled as the plant index of the compendium, simultaneously a botanical description and a pharmaceutical preparation reference. Sinhala match rate 74.6% (vs 69.5% corpus-wide); hapax rate 18.7% (vs 16.3% baseline); pharmaceutical-prep concentration 38.8% (vs 47.8% overall).

V4 additions: 18 herbal folios have higher-confidence opener-word etymologies (plant names decoding to known Elu/Pali/Sanskrit drug names). This supports an opener-word convention, but the method is not blind and needs ethnobotanical review plus a null baseline before it can be called confirmed.

7.3 The BALNEO section: Sri Lankan bath therapy

Unchanged from v3. Highest pharmaceutical concentration (60.9%), lowest hapax rate (9.0%), salt+banana bath therapy profile confirmed. f78v remains the strongest BM parallel (§9.2).

7.4 The PHARMA section: mineral medicine

Unchanged from v3. Copper preparations (tamedā) at 32% PHARMA concentration; highest timing-marker rate (13.4%).

7.5 The COSMO section: candidate timing index (V4 major update)

The COSMO section (f67r-f73v, 2,891 tokens) is best treated as a **candidate pharmaceutical timing index**. V4 proposes that part of the vocabulary encodes Pali nakkhatta (nakṣatra) timing for pharmaceutical preparations, but the individual identifications remain candidate-level until the phonological rules are fixed in advance and tested against a null list.

COSMO decoder coverage: 97.6% assigned (2.01% true-null; ~58 hapax scribal variants unresolved). The pre-fix baseline was 88.9% (2,570/2,891 tokens). The main decoder gap in v3 (unresolved ok-/ot-/y- prefix forms) was addressed by a position-

sensitivity fix applying existing prefix-stripping rules at word-initial position, recovering 107 tokens; further compound resolution brought coverage to 97.6%.

Nakshatra candidates — 13 of 27 Pali nakkhatta proposed; 8 with direct BM sequence support:

The Bhesajjamañjūsā Ch.19 Jarapaddhati (stanzas 59-69, reocr.txt lines 28693-28769) provides an explicit pharmaceutical nakkhatta prognosis sequence — bharanī → kattikā → rohiṇī → mige → punabbasu → phussa → māghā → phaggunī → hattha/citta → jetṭhā → mūla → pubba-āsāḷha → sāvaṇa → dhaniṭṭhā → revatī — against which the COSMO identifications are anchored. Eight of the proposed COSMO forms match the BM-listed nakshatras.

Elu form	Pali nakkhatta	#	Confidence	Evidence
acena	Assayuja (Aśvinī)	1	HIGH	Phonological path confirmed
alula	Anurādhā	17	HIGH	a+l(nu→l)+ula confirmed
makhea	Māghā	10	MEDIUM- HIGH	ma+kh(gh→kh)+ea; BM “māghasu” Ch.19 St.62
kuda	Kattikā (Kṛttikā)	3	MEDIUM- HIGH	ku+da; BM “kattikasu” Ch.19 St.60
phara	Bharanī	2	MEDIUM	bh→ph; BM “bharanīsu” Ch.19 St.60
adala	Addā (Ārdrā)	6	MEDIUM	a+dala; COSMO label position
tusa	Phussa (Puṣya)	8	MEDIUM	ph→t; BM “phusse” Ch.19 St.62 (pathway not yet independently established)

Elu form	Pali nakkhatta	#	Confidence	Evidence
ata	Hattha (Hasta)	13	MEDIUM	h-drop; BM “hattena” Ch.19 St.64; ata=hand=hattha (semantic identity)
aguda	Jeṭṭhā (Jyeṣṭhā)	18	MEDIUM	j→g; BM “jenhayam” Ch.19 St.66
alama	Mūlaṃ (Mūla)	19	MEDIUM	Semantic: ala=root=mūla; BM “mūle” Ch.19 St.67
alara	Pubb-āsālha (Pūrvāṣāḍha)	20	MEDIUM	āsālha→alha→ala+ra; BM “pubba alhasu” Ch.19 St.67
sara	Sāvaṇa (Śrāvaṇa)	22	MEDIUM-LOW	v→r; BM “sone”; also Sarado/autumn season candidate
asala	Uttar-āsālha	21	LOW	n=1 only; no BM form match

The BM pharmaceutical calendar itself skips 9 of 27 nakshatras (Addā, Āśleṣā, Sāti, Visākhā, Satabhiṣā, and others). The 14 unresolved COSMO nakshatras likely reflect the same pharmaceutical non-priority selection rather than a decoding gap.

The Pali nakkhatta list (from Abhidhanappadipika): Assayuja (#1), Bharanī, Kattikā (#3), Rohiṇī, Māgasiraṃ, Addā, Punabbasu, Phussa (#8), Asilesā, Māghā (#10), Pubba-phaggunī, Uttara-phaggunī, Hattha, Cittā, Sāti, Visākhā, Anurādhā (#17), Jeṭṭhā (#18), Mūlaṃ (#19), Pubb-āsālha (#20), Uttar-āsālha, Savaṇa (#22), Dhaniṭṭhā, Satabhisaja, Pubba-bhaddapadā, Uttara-bhaddapadā, Revatī.

Sample COSMO pharmaceutical calendar line (f70r1 L1): “keda | gara | tamea | ela | tea | ch-eda | tusa | mura | tara | tudamea | kuda | tusa | ala” = dried-period | belly-treatment | for-honey | leaf | root-liquid | then | Phussa candidate | Moringa | comparative | infused+honey | Kattikā candidate | Phussa candidate | root → Candidate reading: “Dry-period abdominal treatment with honey and leaf/root-liquid; then a Moringa/root preparation associated with the proposed Phussa/Kattikā timing labels.”

This is compatible with a ṛtucaryā-style structure. Two quantitative tests were run (see supplementary/NULL_TESTS_AND_BLIND_REVIEW.md):

Test 1 (retracted): Nakshatra edit-distance. An unmatched control test initially appeared to give $p < 0.0001$ (8/27 nakshatra names at ≤ 2 edit distance vs 2.0 ± 1.3 for random Pali words). After applying length-matched controls (matching control words to the length distribution of nakshatra names), the result is **$p = 0.916$** — not significant. The 8 apparent matches included four spurious short-word collisions: eta (dative particle, $n = 232$) matching Jeṭṭhā; ata (hand, $n = 65$) matching both Sāti and Cittā; aga (tip, $n = 60$) matching Māghā. Only two were genuine (ada=Addā dist=0, mula=Mūlaṃ dist=0), both already identified by other means. The retracted result is reported here to document the methodological correction.

Test 2 (valid): COSMO enrichment. Nine of twelve proposed nakshatra forms are $> 1.5\times$ enriched in COSMO vs the rest of the corpus (Bonferroni $p < 0.05$). Three forms with COSMO enrichment $> 6\times$: tusa ($8.2\times$, 65% COSMO), alama ($8.4\times$, 67%), acena ($12.7\times$, COSMO-exclusive). This distributional signal is the primary statistical support for the nakshatra hypothesis.

The individual identifications should be assessed via the BM confusion matrix (9/12 proposed candidates align with the BM pharmaceutical nakshatra calendar), the COSMO enrichment, and ultimately specialist review of the decoded text against the Bhesajjamañjūsā and Abhidhanappadipika nakkhatta sections.

The timing architecture from v3 (10–15% cross-section timing-tag rate; 90–100% preparation-vocabulary coherence between hub and linked folios) remains one of the stronger architectural signals. V4 offers a possible nakshatra-level interpretation of part of that signal, but the hub-and-spoke architecture should not yet be described as fully understood at the nakshatra level.

7.6 The ASTRO section: timed processing schedule

Unchanged from v3. Not astronomical; confirmed as timed pharmaceutical processing schedule with gala/agala processing verbs elevated 2.8–7.7×. F58r is the primary timing hub; f66r is the indexed preparation table.

7.7 The Rosette page: master preparation map

Unchanged from v3. teda at 4.42% = 3.9× corpus base rate; labeled pathway structure confirmed.

7.8 The RECIPE section: 23 therapeutic categories

7.8.1 Three-tier visual margin system Unchanged from v3. Cross=category; dot-star=primary formula; hollow-star=variant. ~325 total individual recipe entries.

7.8.2 The 23 categories — with BM/AH chapter hypotheses (V4 new) All 23 surviving categories now have source-chapter hypotheses in BM and/or AH. The v3 mapping is updated with BM chapter numbers from the re-OCR, but the full mapping still needs a chapter-order/null-baseline test:

#	Folio	Opener	Category	BM / AH chapter	Confidence
1	f103r	pedala	Jvara- cikitsā (fever)	BM Ch.19 Jara	STRONG
2	f103v	pula	Śoṭha- cikitsā (edema/swelling)	BM Ch.42 Sopho	STRONG
3	f104r	padara	Gulma/Gumbasīla (abdominal mass)	BM Ch.38 Gumbasīla	MEDIUM- HIGH
4	f104v	paduina	Vātavyādhi (wind disorders)	BM Ch.46/48 Vata-vyadi	STRONG

#	Folio	Opener	Category	BM / AH chapter	Confidence
5	f105r	pena	Netra- cikitsā 1 (eye instillation)	BM Ch.49 Cakkhu	STRONG
6	f105v	pulera	Netra roga 2 (eye cooling)	BM Ch.49 Cakkhu	STRONG
7	f106r	pamadara	Strī roga (women's diseases)	AH Strī-roga (absent from BM)	HIGH
8	f106v	cameda	Kaṇṭha/Gala roga (throat/goitre)	BM Ch.53-54 Gala-gaṇḍa	STRONG
9	f107r	padalara	Aśmari- cikitsā (kid- ney/bladder stones)	BM Ch.34 Muttaghata	STRONG
10	f107v	kesara	Keśa roga (hair diseases)	AH Keśa-roga (absent from BM)	MEDIUM
11	f108r	tudeara	Udavatta (abdominal obstruction)	BM Ch.39 Udavatta	MEDIUM
12	f108v	pedala	Udara (dropsy)	BM Ch.39/40 Udara	STRONG
—	f109-f110	LOST	~Viṣa- cikitsā + Rasāyana	BM Ch.58 Viśa + Ch.60 Rasāyana	—

#	Folio	Opener	Category	BM / AH chapter	Confidence
13	f111r	kuladara	Roga- nidāna (disease classifica- tion)	BM Ch.18 Roga- samanna	HIGH
14	f111v	kumea	Ghṛta-kalpa (medicated ghee)	BM Ch.11 Maha- kamma	MEDIUM
15	f112r	culea	Lehya/Avaleha (confec- tion/electuary)	BM Ch.12 Culla- kamma	HIGH
16	f112v	kuala	Gahani/Virechana (digestive deficiency)	BM Ch.30 Gahani	HIGH
17	f113r	pesaca	Bhūta- cikitsā/Unmāda (spirit diseases)	BM Ch.57 Bhūta-adi	HIGH
18	f113v	culurara	Pañcaphala (five-fruit formulas)	BM Ch.60 Rasayana	STRONG
19	f114r	tedala	Kvātha- kalpa (decoction prepara- tions)	BM Ch.6 Bava-dabba	STRONG
20	f114v	padula	Pāda roga/Vātarakta (foot/gout)	AH Pāda-roga (absent from BM)	MEDIUM- HIGH
21	f115r	camadara	Kāsa/Śvāsa (cough/asthma)	BM Ch.27 Kāsa	STRONG

#	Folio	Opener	Category	BM / AH chapter	Confidence
22	f115v	tedura	Vraṇa- cikitsā (wound treatment)	BM Ch.55 Vrna- bhagga	HIGH
23	f116r	kadapa	Netra roga final (eye, Magha- timed)	BM Ch.49 Cakkhu	HIGH

AH-only chapter candidates (3/23): Strī-roga, Keśa-roga, Pāda-roga — all present in Ashtāṅgahrdayam but absent from BM’s 60-chapter list, supporting a possible hybrid AH+BM source structure.

BM Ch.29 note: Ch.29 contains only 6 diagnostic stanzas with no treatment formulary. Digestive treatment appears within the Gahani chapter (Ch.30), explaining why f112v matches Ch.30 rather than Ch.29.

7.8.3 Chapter opener etymologies Unchanged from v3. 18 of 23 opener words are confirmable Elu compressions of Sanskrit chapter-title vocabulary.

7.8.4 Comparison with AH and BM chapter sequences Unchanged from v3. The hybrid Sri Lankan ordering (AH chapter types + BM chapter organisation) with Strī-roga and Keśa-roga as standalone chapters confirms the AH lineage. The Pañcaphala chapter corresponds to BM Ch.60 Rasāyana; the triple-Netra pattern is characteristic of Sri Lankan specialisation in ophthalmology.

8. Plant Identifications

8.0 Plant section operating system — what is and is not supported (V4 new)

Before listing individual identifications, V4 reports the results of a plant-section system test that probes the structural model rather than any single species claim.

What is supported:

The HERBAL section operates as a **pharmaceutical preparation register**, not a one-picture-one-label botanical atlas. HERBAL formula-like lines reach 87.9% — strongly above narrative prose, clearly distinct from the COSMO timing section (49.9%) — though less densely formulaic than RECIPE (97%+) and BALNEO. The statistical test (HERBAL vs non-HERBAL overall, $p=0.0005$) confirms HERBAL as a formula/ingredient register. Six high-frequency taxa — Moringa, Senna/Cassia, Terminalia, Ficus, Nymphaea, and Piper — appear with ingredient-network connectivity across visually diverse folios, consistent with ingredient roles rather than single illustrated subjects.

What is not supported:

A headline-label model — in which each folio opener decodes directly to the species depicted — produces **0% top-1 species hits** (0/75 tested folios) against external botanical taxon lists. This test does not rule out that opener words name plants; it rules out that the opener alone, matched mechanically to a taxon list, recovers the depicted species. The 2.7% any-taxon hit rate is at noise level.

Implication for the identification table: Individual species identifications in §8.1–8.3 should be read as **candidate identifications** supported by converging evidence lines, not confirmed identifications. “High confidence” in the project workflow means the evidence is worth a specialist’s evaluation — it does not mean external botanical confirmation. The plant section’s operating system (pharmaceutical ingredient register, formula-structure dominant, ingredient-network connectivity) is now testable and confirmed. The species-level layer remains a candidate layer pending blind external review.

V4 provides a herbal plant-identification hypothesis table for 81 of 101 informative herbal folios. Method: opener word (Elu plant name or therapeutic tag) + dominant token signal (morphological/therapeutic profile) + Beinecke IIF visual morphology. Because the visual matching was not blind, these identifications are hypotheses pending ethnobotanical review and null testing.

8.1 High-confidence identifications (18)

These have three converging evidence lines: opener decodes to a plausible plant name + dominant token signal matches Ayurvedic drug profile + visual morphology is compatible. “High confidence” here means high within the project workflow, not external botanical confirmation. In light of the §8.0 system-test results, **all entries in this table should be treated as priority candidates for specialist review, not as**

confirmed identifications.

Folio	Opener	Plant	Key evidence
f2v	padama	Nymphaea stellata (blue lotus)	padama=Sanskrit padma; blue 6-petal flower
f8r	tamula	Piper betle (betel leaf)	tamula=Sinhala/Pali betel; broad lobed leaf
f9v	upula	blue-flower candidate; Clitoria ternatea uncertain	upula=nil-manel points toward a blue water-lily semantic field, not securely to butterfly pea; requires botanical review
f14v	kasadera	Azadirachta indica (neem)	deeply pinnate serrated leaves; woody root
f15v	kana	Piper longum (long pepper)	kana=Elu pippali; red branching root; stem nodes
f17r	kasara	Nigella sativa (black seed)	red root; pointed flower buds; dark thistle heads
f22v	kasada	Acorus calamus (sweet flag)	kasada; thorny horizontal rhizome = diagnostic
f27r	kasura	Justicia adhatoda (malabar nut)	kasura=kāsa (cough); broad smooth lanceolate leaves
f28v	kamala	Nelumbo nucifera (sacred lotus)	kamala=Sanskrit/Pali lotus (unambiguous)

Folio	Opener	Plant	Key evidence
f33r	tamadara	Papaver somniferum (opium poppy)	tamadara=“dark- giving” (opium); flat capsule crown
f37v	kamuda	Nymphaea pubescens (white water lily)	kamuda=Pali white lotus; gena×15=most medicinal folio
f39r	kasadera	Andrographis paniculata (king of bitters)	large fleshy tuber base; fat dangling roots
f43r	kasadera	Asparagus racemosus (shatavari)	ata×5=recipe ingredient; dense root system
f49r	kasadera	Cissampelos pareira (velvet leaf)	large sinuous two-branched root
f50r	kasadera	Sphaeranthus indicus (globe thistle)	pale eye in globose floret; elaborate root
f51r	kasadera	Sida cordifolia (country mallow/bala)	forked taproot; pointed petal tips
f52r	tadugacha	Rubia cordifolia (Indian madder)	distinctly dark red-brown roots = diagnostic visual
f56r	tala	Datura metel (devil’s trumpet)	tala; spiny capsule = Datura-compatible visual

8.2 Medium-high identifications (16)

Selected from the full table (supplementary HERBAL_PLANT_IDENTIFICATIONS.md):

- f1v = *Withania somnifera* (ashwagandha); f4v = *Gloriosa superba* (flame lily); f6v = *Solanum xanthocarpum*; f11v = *Pandanus odoratissimus*; f13r = *Colocasia esculenta* (taro); f22v = *Acorus calamus* (HIGH — listed above); f23r = *Spilanthes*

acmella; f31r = *Curcuma longa* (turmeric, keda×12); f34v = *Lagenaria siceraria* (bottle gourd); f40r = *Vernonia cinerea*; f40v = *Drimys indica* (squill); f44v = *Calotropis gigantea*; f46r = *Musa paradisiaca* (banana, pupha=puṣpa/flower); f55v = *Plectranthus amboinicus*; f56v = *Cinnamomum zeylanicum* (Ceylon cinnamon, ela×9=leaf dominant)

8.3 Revised identifications (V4)

- f55r: Was v2 opium poppy -> v3 *Solanum nigrum* (nightshade). V4 keeps *Solanum nigrum* as the preferred candidate (dark blue spherical berries over bulbous root).
- f33r: Was unknown -> v3 poppy candidate. V4 keeps *Papaver somniferum* as a high-priority candidate (opener tamadara = dark-giving; flat capsule crown is suggestive).
- f37v: New high-priority candidate in V4: *Nymphaea pubescens* (white water lily) via opener kamuda = Pali white lotus + genax15.
- f31r: New candidate in V4: *Curcuma longa* (turmeric) via kedax12 + large horizontal rhizome.
- f42r: New candidate in V4: *Glycyrrhiza glabra* (liquorice) via mulax11 + massive multi-strand root.
- f48v: New candidate in V4: *Santalum album* (white sandalwood) via tedax7 + puda opener.

8.4 Previously established identifications (unchanged from v3)

aralu=*Terminalia chebula*; sena=*Cassia angustifolia*; kamala=*Nelumbo nucifera*; tambula=*Piper betle*; tamala=*Cinnamomum tamala*; bulu=*Terminalia bellirica*; nelli=*Phyllanthus emblica*; ata=*Datura metel* (recipe ingredient); mara=*Solanum nigrum*; kera=*Cucumis*; muga=*Vigna radiata*; gugula=*Commiphora wightii*; paruina=*Grewia asiatica*.

New drug identifications from V21: - mura = *Moringa oleifera* (drumstick tree seed) — supported by Clough dictionary material; 160 tokens - nagara = dry ginger (*Zingiber officinale* dried) — Pali pharmaceutical texts; 59 tokens - saleda = *Boswellia serrata* (Indian frankincense/salai) — DPD salaḷa confirmed; 11 tokens - gerala = red ochre/gairika mineral — DPD geruka confirmed; 9 tokens - puleda = dried boiled barley (yava/pulaka) — DPD confirmed; 8 tokens - phea = foam/latex preparation (pheṇa) — DPD confirmed; 12 tokens - mathea = churned product/butter (navanīta) — DPD

√math=churn; 8 tokens

8.5 Blind botanical review — three-stage protocol (V4 new)

A structured three-stage blind review of the 110 HERBAL folios was conducted to test image-text convergence without project-hypothesis bias.

Protocol. Stage 1 (image-only): all 110 folios reviewed under random blind IDs, with no access to decoded text or project interpretation. Assessments recorded: plant-part tags, growth form, leaf morphology, root/tuber features, flower/fruit/seed features, family and species candidates, confidence 0–5. Stage 2 (text-only): decoded token signals reviewed separately, without looking back at images. Stage 3 (reference photos): 12 Ayurvedic candidate species compared against Stage 1/2 assessments.

Key results — convergences.

The strongest direct convergence across all three stages is **f33r (B031 under blind ID)**: - Stage 1: *Papaver somniferum* or *P. rhoeas* (confidence 4). Two rounded crowned seed capsules are diagnostic for *Papaver*; the capsule form with stigma-ray crown is unambiguous. - Stage 2: decoded tokens include tamadara (dark-giving compound), madara (sweet/madhura rasa), and dala (petal) ×2 — pharmacologically consistent with opium poppy (dark latex, petals, sweet taste group). - Stage 3: Reference photo confirms *Papaver* crowned capsule matches Stage 1 morphology exactly. - **Assessment: independent three-stage convergence on *Papaver somniferum*.**

A strong partial convergence is **f28v (B091 under blind ID)**: - Stage 1: *Crinum asiaticum* or *Pancratium maritimum* (Amaryllidaceae, confidence 3). Radiating strap-like tepals from bulb base, white star form, aquatic/semi-aquatic habit. - Stage 2: kamala token (*Nelumbo nucifera* — Pali/Sinhala name for sacred lotus) + Moringa 3×. - **Assessment: both stages independently identify a tropical white-tepaled aquatic monocot. Families differ (Amaryllidaceae vs Nelumbonaceae) but ecological and visual habit converge. The kamala text decoding favors *Nelumbo nucifera* where morphology alone is ambiguous between Amaryllidaceae and Nelumbonaceae.**

Key results — divergences.

Of 25 priority-set folios, **19 diverge** between Stage 1 image assessment and Stage 2 text signals. Stage 1 (conducted without project context) consistently proposed European/Mediterranean genera — *Valeriana*, *Cyclamen*, *Carlina*, *Pulmonaria*, *Scabiosa*, *Echium*. Stage 2 text signals consistently pointed to South Asian medicinal plants —

Moringa, Senna/Cassia, Terminalia, Ficus, Nymphaea.

The most notable anomaly is a high-frequency Moringa text signal (7× mura, confidence 4) against a Stage 1 image that shows two large circular bullate leaves with a fleshy tuber — morphologically inconsistent with the Moringa reference photo (pinnate compound leaves, drumstick pods). This conflict is documented as an open question requiring specialist resolution.

Interpretation of the divergence pattern.

The systematic divergence is predicted under multiple models and is not diagnostic by itself. The drawings may be: (a) headline or mnemonic plants rather than the primary ingredient named in the text; (b) substitute or analogue illustrations adapted for a non-Sri Lankan user; (c) one ingredient depicted from a multi-ingredient formula; or (d) depictions of plants that a reviewer without Indic botanical training misidentifies as European analogues. Specialist ethnobotanical review is needed to distinguish these possibilities.

Adversarial tests. A substitute-plant adversarial test checked whether image/text divergent pairs still share documented medicinal function (using Fernie Herbal Simples, PCIMH Siddha formularies, and Caraka Saṃhitā). Function-hit rate was $p(\text{high}) = 0.577$ — not statistically above shuffled controls. The substitute-use model is plausible as a historical mechanism but is not yet supported by the current source-cited matching, which used broad therapeutic-class tables rather than species-specific materia medica. This is a methodological limitation, not a refutation.

The three-stage blind review materials (packet, completed reviewer forms, convergence report) are available in `teamb/reviewer_workspace/blind_botanist_packet/`.

9. Bhesajjamañjūsā Textual Parallels

9.1-9.5

Unchanged from v3. 25 score-3+ formula clusters; 7 BM chapter-level parallels; nura=nirūha (§9.3); piśāca+nāga (§9.5).

9.6 Sample readable prescriptions (V4 new)

The V21 decoder improvements enable more coherent prescription-style readings of specific lines. These should be presented as worked readings, not final polished trans-

lations.

f112v L45 — Patient prescription candidate (Gahani chapter, BM Ch.30 parallel):

Token	Decoded	Meaning
agea	agea	fat-extracted preparation (DRY)
nara	nara	patient / recipient of treatment
ch-ka	ch-ka	one [dose]
ch-agera	ch-agera	bring/fetch (gēna instrumental)
ch-ara	ch-ara	having done / past participle
ch-ea	ch-ea	with ghee
kena	kena	by means of
dala	dala	having given
kena	kena	by means of
galada	galada	strain-and-give

Worked reading: “The fat-extracted preparation: bring one dose for the patient; perform with ghee; when done, administer; having strained, give.”

This is a strong patient-directed pharmaceutical reading. It contains: drug type (fat-extract), recipient (nara=patient/person), dosage candidate (ch-ka=one), vehicle candidate (ch-ea=ghee), and administration route (galada=strain-and-give). Several elements are dictionary-supported, but the whole line still needs specialist review.

f113r L1 — Bhūta-cikitsā chapter opening candidate (BM Ch.57 parallel):

Token	Decoded	Meaning
pesaca	pesaca	evil spirit / piśāca
apeda	apeda	affliction/pain
cada	cada	incantation / word
ch-etara	ch-etara	seed preparation + ra
tea	tea	root-liquid (oil/decoction)

Worked reading: “Evil spirit [causing] affliction: incantation, seed preparation, root-liquid.”

This is compatible with a combined exorcism+herbal protocol: ritual (pesaca=evil spirit + cada=incantation candidate) followed by pharmaceutical treatment (seed

preparation + root-liquid). It is a plausible match to BM Ch.57 Apasmarabhutadi treatment structure, not an independently confirmed translation.

f75r L44 — Ghee+senna BALNEO formula (Match 1, score 4):

Token	Decoded	Meaning
geda	geda	ghee/ghṛta
sena	sena	senna (Cassia angustifolia)
keda	keda	dried crude drug
q-keda	q-keda	THE dried drug
teda	teda	crude decoction
kulada	kulada	species-type + give
tara	tara	comparative
eta	eta	upward/dative
nena	nena	without/negation

Worked reading: “Ghee and senna: the dried drug, the dried drug, crude decoction, [for the] species-type, exceeding, upward [administration], without [further processing].”

The highest-scoring BM parallel (score 4): ghee+senna decoction formula matching BM L40285–40293 Sophapaddhati chapter.

10. The Voynich Pharmaceutical Notation System (VPNS)

10.1 System overview (21 states)

The VPNS encodes preparation states through a productive system: [optional prefix] + [base consonant class] + [aspect suffix -ea (wet) or -eda (dry)]. Total: 3,727 bare-form state tokens; 5,359 including q-/sa-prefixed forms.

V4 reopens khea/kheda, which was listed as “khīra (milk preparation)” in v3 (V25). The new kapha/phlegm proposal is plausible enough to test, but not strong enough to supersede V3 as a settled result. The VPNS table below therefore distinguishes externally grounded preparation classes from candidate Tridosha-oriented readings:

Base	Wet	Dry	Wet n	Dry n	Preparation		Confidence	Version
					class	Tridosha		
k	kea	keda	455	400	kvātha/de- class	vāta	HIGH	original
t	tea	teda	340	411	taila (oil prepa- ration)	pitta	HIGH	original
kh	khea	kheda	86	22	kapha/ph- candi- date	kapha?	LOW- MEDIUM	V21 candi- date
m	mea	meda	495	541	madhu/medas (honey/fat)		HIGH	original
lag	lagea	lageda	135	173	laga (rub- bing/application)	—	HIGH	original
l	lea	leda	112	188	lehya/disease- type		HIGH	original
p	pea	peda	60	96	pāna (drink- ing/beverage)	—	HIGH	original
s	sea	seda	21	15	sveda (steam treat- ment)	—	HIGH	original
c	cea	ceda	11	19	cheda (cut- ting/sectioning)	—	MEDIUM	V24
th	thea	theda	72	19	locative- in- preparation	—	MEDIUM	V21 re- vised
am	amea	ameda	28	17	āma (raw/unprocessed)	—	HIGH	V25
lap	lapea	lapeda	5	11	lepa (paste/ointment)	—	HIGH	V25

Base	Wet	Dry	Wet n	Dry n	Preparation		Confidence	Version
					class	Tridosha		
ph	phea	pheda	13	13	pheṇa (foam/latex)	—	MEDIUM	V21 re- vised
kam	kamea	kameda	23	13	kama (sequence/order- prep)	—	MEDIUM	V21 re- vised
mekh	mekhea	mekheda	6	4	mekhalā (waist/abdominal)	—	MEDIUM	V25
ag	agea	ageda	88	57	agada (anti- dote)	—	MEDIUM	V25- new
g	gea	geda	52	50	ghṛta (ghee)	—	HIGH	V25- new
at	atea	ateda	53	61	ātapa (sun- drying)	—	MEDIUM	V25- new
lam	lamea	lameda	47	76	lavana/loṇa (salt)	—	MEDIUM	V25- new
tam	tamea	tameda	22	23	tāmra (cop- per/mineral)	—	MEDIUM	V25- new
ram	ramea	rameda	2	12	rambhā (ba- nana bathing)	—	MEDIUM	V25- new

10.2 Falsification tests

Unchanged from v3. Null test: zero of 50 random consonant sets match. Holdout test: 82% cross-folio replication. Carrier register test: identical Sārārtha match rates across all five Carrier-A-score bands. Vinaya falsification: 0/195,939 Vinaya tokens match VPNS markers.

10.3 External grounding

Updated from v3, but the conservative count remains **5 strongly grounded + 1 partial**:

leda, seda, meda, pea, peda — unchanged (see v3 §10.3 for full grounding evidence).

khea (NEW in V4) = kapha/phlegm candidate. Pali semha means phlegm, and khea has interesting distributional behavior, but the proposed semha -> khea sound path is not yet independently grounded. It is therefore excluded from the external-grounding count until a Sinhala/Elu specialist or corpus baseline validates the phonology.

ceda — partial (unchanged).

10.4 Compositional validation and section profiles

Unchanged from v3. BALNEO=salt-oil-banana bath; PHARMA=mineral medicine; HERBAL=sun-drying herb-preservation; RECIPE=compound formula manufacturing.

10.5 Candidate Tridosha-oriented interpretation (V4 new)

V4 proposes that three VPNS classes may align with Ayurvedic Tridosha concepts:

Dosha	Sanskrit	Elu form	Tokens	Etymology	Confidence
Vāta (wind/air/ether)	vāta	kea	958	Sinhala kear = svargg/heaven; semantic ether/wind field	MEDIUM- HIGH
Pitta (fire/heat)	pitta	teda	813	established heat/decoction prepara- tion marker; pitta alignment is interpre- tive	MEDIUM

Dosha	Sanskrit	Elu form	Tokens	Etymology	Confidence
Kapha (phlegm/earth)	kapha	khea	118	Pali semha=phlegm but semha -> khea path unverified	LOW- MEDIUM

The Tridosha system is the foundational framework of Ayurvedic medicine, so a confirmed three-way dosha encoding would be a major result. V4 has not yet reached that standard. The current result is a coherent candidate model that should be tested against externally attested Sinhala/Elu medical terminology and against VPNS section-profile predictions.

Clinically, the proposed model predicts that kea-class, teda-class, and khea-class preparations should show distinct co-occurrence profiles with oil, heat/cooling, honey/salt, disease labels, and body-region markers. That prediction is testable and should replace broad claims that the full Tridosha grid is already confirmed.

11. Independent Corroboration

Updated from v3 with V4 additions:

1-10. Lexical, external grounding, visual cross-modal, BM formula matching, primary-source historical, statistical fingerprint, falsification probes, cross-language refutation, phonotactic match, manuscript architecture — all unchanged. See v3 §11.

11. Recipe chapter etymology: 18/23 openers confirmed (unchanged).

12. Convergence with prior work: unchanged.

13. Candidate Tridosha-oriented VPNS layer (V4 new). The kea/tega/khea model is a useful hypothesis for future testing. It is not counted as independent confirmation until khea is externally grounded.

14. Initial-sound gap partial explanation (V4 new). The /h/-initial gap is compatible with Sinhala/Elu h-loss patterns. The /b/ and /v/ gaps remain audit risks requiring broader medical-register baselines.

15. COSMO as candidate nakkhatta timing layer (V4 new). The proposed Pali nakkhatta vocabulary (tusa=Phussa, alama=Mūla, aguda=Jetṭhā) is interesting but unconfirmed. It needs a fixed phonological map and a null comparison against all 27 names before it can be used as provenance evidence.

12. Decoder Error Analysis

Updated from v3.

The **/v/, /b/, /h/ initial gap is partly reinterpreted** (v3 listed this as “a genuine open question”). See §4.15 for audit analysis. Summary: /h/-loss is plausible in Sinhala/Elu; /b/ and /v/ remain risks until tested against a broader Sinhala/Elu medical lexicon. No decoder-accuracy revision is made from this point alone.

The **-uina family** (§6.3): resolved at MEDIUM confidence in V26. Unchanged.

The **EVA-o mapping**: residual uncertainty unchanged. 5.0% u-initial rate in V17 vs 2-3% Bodleian baseline. Some tokens may encode genuine Sinhala u-initial words.

The **khea = khīra revision**: the V3 interpretation of khea as “khīra (milk preparation)” is not secure. V4’s khea = kapha/phlegm proposal is a candidate replacement, not a settled correction. Token counts are unchanged; the interpretation of those tokens remains open.

The **tara = comparative suffix**: the V3 interpretation of tara as “you familiar + ra” was incorrect. The correct reading (DPD tara 2 = comparative -er/more) makes recipe lines grammatically coherent, notably “tara ara” = “having exceeded/surpassed [the preparation stage].” This correction improves recipe readability materially.

The **BM OCR issue** (v2): resolved in v3, unchanged.

13. Historical Context

Unchanged from v3. The 4-6c Sri Lankan King Buddhādāsa medical compendium tradition; the 13c Bhesajjamañjūsā; the 15th-century Sri Lankan monk-pharmacist institutional context; the De’ Conti transmission hypothesis.

V4 addition: The proposed Pali nakkhatta layer in the COSMO section, if validated, would connect the section with ṛtucaryā and pharmaceutical timing traditions

documented in AH and Sri Lankan Pali medical texts. The six-season Pali calendar (Gimha/Vassa/Sarado/Hemanto/Sisiro/Vasanto) appears in BM pharmaceutical stanzas. The stronger historical claim should remain conditional until the nakkhatta identifications pass null testing.

14. Limitations and Future Work

14.1 Updated since v3

The /v/, /b/, /h/ initial gap (v3 §12 “genuine open question”): PARTLY ADDRESSED. /h/-loss is plausible under Sinhala/Elu historical phonology; /b/ and /v/ remain open until tested against an external medical-register baseline. Decoder accuracy is not revised upward on this basis.

COSMO decoder gap (v3: unspecified): LARGELY ADDRESSED. Assignment coverage is **97.6%** (2.01% true-null, predominantly hapax timing forms). ok-/ot- position-sensitivity bug fixed 107 tokens. 13/27 nakkhatta candidates are proposed (3 confirmed, 10 medium-level, BM Ch.19 cross-validated), but COSMO is not yet confirmed as a nakshatra pharmaceutical calendar.

khea taxonomy (v3: khīra = milk, MEDIUM): REOPENED. khea = kapha/phlegm via semha -> khea is a candidate V4 hypothesis; Tridosha completion is not claimed.

Herbal plant identifications (v3: partial): EXPANDED. 81/101 informative folios have candidate identifications; 18 are high priority for botanist review. The table is not a blind confirmation.

23 recipe categories BM mapping (v3: partial): EXPANDED + NULL TESTED (V4). All 23 categories have BM/AH source-chapter hypotheses. Four null tests run; one test retracted (nakshatra edit-distance, see COSMO note):

(a) AH chapter ordering: Spearman $\rho=0.400$, $p_{\text{perm}}=0.030$ (10,000 permutations).

Does NOT survive Bonferroni correction ($\alpha/2=0.025$ for two ordering tests run). The correlation is real ($\rho=0.40$) but exploratory. BM ordering: $\rho=0.045$, $p_{\text{perm}}=0.428$ — not significant. Treating this as a confirmatory result was an error; it is reclassified as EXPLORATORY pending a pre-registered retest on Cikitsā-sthāna chapters only.

(b) Vocabulary enrichment: 38.6% domain-vocabulary overlap vs 30.0% null ($p=0.005$); disease-diagnostic vocabulary enriched in proposed sections.

(c) Caveat: Both tests retain author-circularity risk. Independent specialist assess-

ment is the proper validation.

See supplementary/NULL_TESTS_AND_BLIND_REVIEW.md.

Meaning field for high-frequency words (v3: multiple {dict} entries): SUBSTANTIALLY ADDRESSED. 26 word types corrected (5,307 tokens) with dictionary-sourced etymologies. Stronger corrections: kea=sky/ether, tara=comparative, kena=instrumental, mura=Moringa, nagara=dry ginger, kula=species/type. khea=kapha remains candidate-level.

Sample prescription-style readings (v3: not available): ACHIEVED as worked readings. f112v L45 (patient prescription candidate); f113r L1 (Bhūta-cikitsā opening candidate); f75r L44 (ghee+senna formula candidate).

14.2 Remaining limitations

Sister-language indistinguishability: Unchanged. Pali, Maharashtri Prakrit, Konkani cannot be ruled out at the typological level without specialist review.

VPNS specific kalpana taxonomy partially validated: 5 of 21 state-markers are strongly grounded; 1 partial (ceda); khea remains a candidate; the rest remain Voynich-internal extrapolation.

Plant identifications unverified by trained eye: 81/101 folios identified, but Sri Lankan ethnobotanist review remains the necessary validation step.

No specialist linguistic validation yet: Sinhala/Elu philologist review of decoded prose is the highest-priority remaining validation.

****COSMO timing vocabulary (irreducible 2%):**** 58 hapax scribal variants in COSMO remain unresolved (2.01% true-null after position-sensitivity fix; pre-fix baseline was 11% / 316 tokens). The full Pali nakkhatta calendar (27 nakshatras) has 13 candidates proposed (3 confirmed, 8 with direct BM sequence support); 14 remain (pharma non-priority expected for ~9 of these).

14.3 Proposed next steps

1. Apply Wave-1 specialist outreach (Blackburn, Cornell; Bretfeld, NTNU; Yakan-dawala, Peradeniya).
2. Physical examination of BL Or. 6612(75) Nava Jātiya Niganduwa.
3. Sri Lankan ethnobotanist review of the 112-folio botanist's dossier and the three-stage blind review results.

4. Pre-registered chapter-order retest: AH Cikitsā-sthāna chapters only, Bonferroni-corrected, on a fresh mapping (see §14.1).
 5. Complete Pali nakkhatta identification: 14 of 27 unresolved; BM Ch.19 pharmaceutical sequence now provides validation for 8 of the 13 identified; remaining 14 include ~9 nakshatras the BM itself omits.
 6. Acquire Sri Lankan domestic Niganduwas (Sara, Vanavasa, Birimal, Lakdiva Aurveda).
 7. VPNS specialist review of the 14 externally unattested state-marker cells.
 8. Source-cited substitution test: rebuild against species-specific materia medica (not broad therapeutic-class tables) before concluding on the substitute-plant model.
 9. Modern Sinhala native-speaker readability probe: 20 lines of decoded text, unglossed, to test whether speakers recognise register vocabulary without prompting.
-

15. Methodology and Reproducibility

15.1 Decoder reproducibility

All decoders are pure-Python with no external dependencies beyond the published Sinhala dictionary and curated vocabulary TSV.

- H12 (original): Paper/scripts/h12_decoder.py — locked for paper-v1 reproducibility
- V17 (canonical 2026-05-03): Paper/scripts/v17_decoder.py — V15 + q-/ch- + daiiin patch
- V20 corpus: translation/voynich_v20_corpus.db — 36,633 tokens with NL-merger
- **V21 (this version):** meaning_fixed field in translation/voynich_v20_corpus.db — 26 word-type corrections, COSMO prefix extension, final compound resolution

15.2 Key analyses

Unchanged from v3, with additions: - V21 meaning corrections: supplementary/V21_MEANING_CORRECTIONS.md - COSMO decoder extension: supplementary/COSMO_DECODER_EXTENSION.md - Herbal plant identifications: supplementary/HERBAL_PLANT_IDENTIFICATIONS.md - Recipe BM/AH mapping hypotheses:

supplementary/RECIPE_CATEGORY_MAP.md - Readability demonstrations: supplementary/DECODER_V21_READABILITY.md - **Frozen decoder specification:** supplementary/FROZEN_DECODER_V17_V27.md - **Audit tests (tiered lexicon, blind sample, formula null tests, failure catalogue):** supplementary/AUDIT_TESTS.md - **Post-freeze change log:** supplementary/POST_FREEZE_CHANGES.md - OCR fine-tuning plan (Wellcome MSS): supplementary/OCR_FINE_TUNING_PLAN.md

15.3 Reviewer-ready packages

BOTANIST_DOSSIER.md (258 KB, 112 folios). SPECIALIST_OUTREACH_PACKAGE.md (17 scholars).

V4 adds: teamb/reviewer_workspace/blind_botanist_packet/ — three-stage blind review packet with completed reviewer forms (Stage 1 image-only, Stage 2 text-only, Stage 3 convergence report). Public-release zip: teamb/outputs/blind_botanist_packet_public.zip

15.4 Team B validation suite (V4 new)

All V4 structural and language-control tests are reproducible from scripts in teamb/scripts/. Key outputs:

Script	Test	Key output
statistical_stress_tests.py	Section classifier, KALPANA enrichment, q-/ch- allomorphy	outputs/statistical_stress_tests.json
alternative_model_scorecard.py	Locked-anchor language comparison (6 language families)	outputs/alternative_model_scorecard.json
genre_anchor_control_tests.py	Genre-matched locked-anchor controls vs historical texts	outputs/genre_anchor_control_tests.json
rival_theory_control_tests.py	Repeated-use vs raw overlap across 10 control groups	outputs/rival_theory_control_tests.json
formula_null_suite.py	BM/Caraka/AH concept-overlap null tests	outputs/formula_null_suite.json
ordered_formula_suite.py	Contiguous ordered concept sequences	outputs/ordered_formula_suite.json

Script	Test	Key output
plant_section_system_tests	Headline-label model, formula-structure rate, ingredient networks	outputs/plant_section_system_tests.
substitute_plant_adversarial	image/text divergent pairs — function-hit rate	outputs/substitute_plant_adversaria
source_cited_substitution	Source-cited medicinal-use matching	outputs/source_cited_substitution_t

High-level human-readable reports for each test are in `teamb/reports/`.

16. Conclusion

The strongest current interpretation is that the Voynich Manuscript is a 15th-century Sri Lankan Elu/Sinhala pharmaceutical text — a working pharmacist’s compressed reference (veda pota) recording Ayurvedic preparations in a bespoke abugida, plausibly connected with a viharasthana physician or monastery-hospital context. Version 4 improves the reading layer and adds useful hypotheses, but it does not close every decoder question. In particular, the initial-sound gap is only partly explained, and `meaning_fixed` coverage is not the same as independently measured accuracy.

The most significant structural advance in v4 is the candidate **Tridosha-oriented VPNS model**. `kea` and `teda` have useful semantic/distributional support; `khea` remains speculative until the proposed phlegm/kapha etymology is externally grounded. The productive [prefix]+[base]+[suffix] VPNS grid remains important, but the dosha alignment should be treated as a testable model rather than a completed result.

The COSMO section, previously characterised as a temporal prescription index, now has a candidate **nakkhatta timing** interpretation. With 13 of 27 Pali nakkhatta candidates proposed (8 with direct BM sequence support, 3 confirmed), the structure is worth testing, but labels such as `tusa`=Phussa and `kuda`=Kattikā are not yet secure enough to support broad historical claims.

The herbal section plant-identification table (81/101 folios, 18 high-priority candidates) and the 23-category recipe BM/AH chapter mapping collectively strengthen the pharmaceutical-library interpretation. Both layers now need blind external review and null-baseline testing.

The broad origin-purpose claim remains strong: South Asian, pharmaceutical, and specifically Sri Lankan/Elu-Sinhala are still the best-fitting hypotheses in the current evidence set. But v4 should not lower **P(Sinhala/Elu identification wrong)** below v3's ~7-10% until the Tridosha, COSMO, plant-ID, and recipe-category layers pass blind or preregistered validation.

Acknowledgments

The author thanks the Beinecke Rare Book and Manuscript Library for digital access to MS 408 and for the freely available IIIF infrastructure that enabled the herbal visual matching and plant identification methodology. The foundational EVA transcription work by Stolfi, Takahashi, and the Voynich research community is acknowledged. Daniel Gaskell's open-source classifier code made the \$4.13 replication possible. The Digital Pali Dictionary (DPD) and Clough (1830) Sinhala dictionary provided the external lexical grounding that enabled the V21 meaning corrections. The Pali Text Society's nakkhatta list from the Abhidhanappadipika enabled the COSMO nakshatra identifications. The computational pipeline, vocabulary expansion, and all analyses in v4 were developed using Anthropic Claude Sonnet as an AI coding and analytical assistant. The author acknowledges the Buddhist temples and medical traditions of Sri Lanka, whose inscriptions and continuing Ayurvedic practice provide the living context for this investigation.

References

(unchanged from v3, with the following additions)

Sinhala / Elu Language and Linguistics

- Clough, B. (1830/1892). Sinhala-English / English-Sinhala dictionary material used in the project OCR. (Source: *kear* = svargg/heaven; *mura* = moringa seed; exact edition metadata to be normalised before publication)
- Geiger, W. (1938). A Grammar of the Sinhalese Language. Royal Asiatic Society, Colombo. (Source: systematic initial h-deletion in Elu/Sinhala)

Ayurvedic Medicine and Pharmaceutical Texts

- Digital Pali Dictionary (DPD). <https://digitalpalireading.net>. Accessed 2026. (Source: tara=comparative, kena=instrumental, kula=species/type, nara=person/patient, saleda=Boswellia, geruka=red ochre, pulaka=dried barley, pheṇa=foam/latex)

Astronomical References

- Nyanatiloka Thera (1972). Buddhist Dictionary: Manual of Buddhist Terms and Doctrines. Buddhist Publication Society, Kandy. (Pali nakkhatta system)
- Ancient Buddhist Texts. Days, Months and Seasons in Pali. <https://ancient-buddhist-texts.net/Reference/Days-Months-Seasons.htm>. Accessed 2026. (Complete Pali nakkhatta list)

Appendix A: Confidence Estimate Decomposition

Sub-claim	v3 confidence	v4 confidence	Change
South Asian Indic substrate	~97%	~97%	Unchanged
Sri Lankan provenance	~89%	~90%	V4 additions are suggestive but not yet independently validated
Sinhala/Elu specifically	~90%	~90%	/h/-loss is compatible; /b/ and /v/ still open
Pre-12c Elu chronolect	~83%	~83-84%	New dictionary attestations help, but some are later-source checks
Pharmacy register	~97%	~97-98%	Worked prescription-style readings support register

Sub-claim	v3 confidence	v4 confidence	Change
V17 transliteration correct in detail	~72%	~72-75%	V21 improves meanings, not phonetic decoder accuracy
VPNS structural grid	~95%	~95%	Grid unchanged; Tridosha layer remains candidate
VPNS specific kalpana labels	~65%	~65%	khea not counted as externally grounded
BM as direct textual source	~60%	~60-65%	23-category mapping expanded; needs null baseline
Manuscript architecture characterised	~90%	~90-92%	COSMO timing architecture strong; nakkhatta layer candidate
P(overall identification wrong)	~7-10%	~7-10%	No v4 downward revision until external validation

Appendix B: Supplementary Files

New in v4: - supplementary/V21_MEANING_CORRECTIONS.md — 26 word-type corrections with DPD/Clough/BM citations - supplementary/COSMO_DECODER_EXTENSION.md — COSMO 97.6% assignment coverage (pre-fix baseline 88.9%); 13/27 nakkhatta candidates (BM Ch.19 validated); Pali nakkhatta table - supplementary/HERBAL_PLANT_IDENTIFICATIONS.md

— 81/101 folios; 18 high-priority candidates; hypothesis table - supplementary/DECODER_V21_READABILITY.md — sample prescription-style readings; candidate Tridosha analysis; assignment-coverage stats - supplementary/COSMO_NAKSHATRA_MAP.md — full 27-entry identification table; BM Ch.19 sequence; enrichment/edit-distance null test results - supplementary/NULL_TESTS_AND_BLIND_REVIEW.md — 5 null tests; edit-distance test retracted (p=0.916 length-matched); COSMO enrichment (T2) and vocab enrichment (T4) pass Bonferroni; AH ordering (T3) reclassified as EXPLORATORY (fails Bonferroni after permutation); blind review samples A-E - supplementary/FLUENT_READING_WORKBENCH.md — Elu morphology table; per-section reading grammars (RECIPE/HERBAL/BALNEO/COSMO/PHARMA); f112v parallel recipe alignment; unknown token triage; niganduwa acquisition priorities - supplementary/FROZEN_DECODER_V17_V27.md — decoder freeze specification (2026-05-09) - supplementary/AUDIT_TESTS.md — tiered lexicon; blind 100-line sample; formula null tests; failure catalogue - supplementary/POST_FREEZE_CHANGES.md — post-freeze change log

Team B validation suite (all reproducible): - teamb/scripts/ — all Python scripts for §4.17 and §5.1 tests - teamb/reports/STATISTICAL_STRESS_TESTS.md — structure tests independent of English glosses (§4.17) - teamb/reports/ALTERNATIVE_MODEL_SCORECARD.md — locked-anchor language comparison (§5.1) - teamb/reports/GENRE_MATCHED_CONTROL_PACK.md — genre-controlled rival theory tests (§5.1) - teamb/reports/RIVAL_THEORY_CONTROL_TESTS.md — repeated-use vs raw overlap (§5.1) - teamb/reports/FORMULA_NULL_SUITE.md — formula concept-overlap null tests - teamb/reports/PLANT_SECTION_SYSTEM_TESTS.md — plant register structure tests (§8.0) - teamb/reports/SUBSTITUTE_PLANT_ADVERSARIAL_TEST.md — adversarial image/text pairing (§8.5) - teamb/reports/LOCKED_ANCHOR_LEXICON_TEST.md — 33-anchor extraction and test methodology - teamb/reviewer_workspace/blind_botanist_packet — completed Stage 1/2/3 reviewer forms (§8.5)

Note on public release: The following outputs contain blind_id-to-folio mappings and are withheld from public release until the external blind review is closed: teamb/_private_blind_key_hold/, and joined outputs (plant_headline_label_test.tsv, plant_text_image_mismatch.tsv, substitute_plant_pair_scores.tsv, source_cited_substitution). The public-safe blind botanist packet is teamb/outputs/blind_botanist_packet_public.zip.

Updated in v4: - supplementary/RECIPE_CATEGORY_MAP.md — BM/AH chapter hypotheses for all 23 categories - translation/voynich_v20_corpus.db — V21 meaning corrections applied; all copies synced

Continuing from v3: - All V25 validation results, BM formula matches, COSMO cross-

reference analysis - supplementary/HERBAL_VISUAL_CHECK_FULL.md (92.2% visual match rate) - v15_work/BOTANIST_DOSSIER.md, SPECIALIST_OUTREACH_PACKAGE.md

Paper version 4.0, May 2026. Zenodo DOI (v3, current published): [10.5281/zenodo.20072618](https://doi.org/10.5281/zenodo.20072618). Concept DOI: [10.5281/zenodo.18598229](https://doi.org/10.5281/zenodo.18598229). GitHub: [kamb-code/Voynich](https://github.com/kamb-code/Voynich)