# INDUS-DEDUCT: A Calibrated, Anti-Hallucination Computational Benchmark for the Indus Valley Script — Validated Linguistic Discriminators, Internal Icon-Text Registries, Trade Logistics, and Exact Decipherment Boundaries

**Project:** INDUS-DEDUCT (Layers 0–12 Reproducible Open Benchmark, `v1.0.0-preprint`)

**Date:** November 20, 2026

**Repository:** `indus-decipher` (11/11 automated claim-ledger tests passing; cross-OS macOS/Linux bit-for-bit SHA-256 verified)

**Epistemological Contract:** Zero fabricated phonetic readings; all speculative upstream annotations stripped prior to modeling; all positive claims paired with position-preserving nulls, global Benjamini–Hochberg FDR, known-answer positive controls, and non-linguistic negative controls.

## Abstract

A century of research on the Indus Valley (Harappan) script (c. 2600–1900 BCE) has produced dozens of incompatible decipherment claims and an enduring methodological impasse between linguistic and non-linguistic interpretations. Here we present **INDUS-DEDUCT**, an end-to-end, reproducible 11-layer computational epigraphy benchmark across **5,679 inscriptions** (5,805 parsed lines; 713 raw signs) engineered to separate statistically verified structural properties from unfalsifiable phonetic speculation.

First, we audit and retire widely cited entropy tests: on a length- and size-matched **5-way calibration benchmark** against real **Ur III Sumerian (CDLI)**, **Linear B**, **Linear A**, and a **non-linguistic structured stamp negative control**, within-sequence conditional entropy (`z_H1`) fails catastrophically as a language discriminator (`Non-Linguistic Stamps = −245.34σ` vs. `Indus = −108.85σ`). By contrast, **three validated discriminators** cleanly separate Indus and attested natural-language scripts from structured non-linguistic stamps: (i) **unseen-trigram generalization** over a position-aware baseline (`Indus = +6.38 pp / +0.432 bits`, `Ur III Sumerian = +10.71 pp / +1.133 bits`, `Linear B = +0.81 pp`, vs. `Non-Linguistic Stamps = −2.95 pp / −0.960 bits`), (ii) **Kober bridge–ending mutual information** (`Indus = +53.67σ`, `Ur III Sumerian = +20.40σ`, `Linear B = +18.47σ`, vs. `Non-Linguistic = +0.07σ`), and (iii) **blind 2-affix stacking** (`Indus = 207:1` suffix-to-prefix ratio with a `4.96×` right-edge bottleneck, vs. `Non-Linguistic = 0:0`).

Second, joint visual–distributional allograph clustering across 697 rendered font glyphs establishes a functional core inventory of `N_core = 240–271` signs covering 95.7% of tokens (`223–254` excluding explicit compounds), placing the script firmly in the **logo-syllabic** regime and ruling out a pure syllabary. Globally FDR-controlled structural findings (`376/422` tests survive BH-FDR; `155` survive Bonferroni) include a productive `: X :` **four-dash sub-graphemic frame** that converts medial stems (`2.9%` final) into terminals (`62.2%` final, 8/8 pairs, `p = 4.38e-45`) and transfers zero-shot across unseen base-glyph families (`LOBO Δlog-loss = −0.0687`), strict **Leave-One-City-Out (LOCO)** real damaged-inscription restoration (`50.0%` Top-1 / `63.6%` Top-3 on all edge breaks vs. `9.1%` position baseline; `71.4%` / `85.7%` on cross-site singletons), and three archaeologically localized document systems:

1. **Mohenjo-daro Copper Tablets (`TAB:C`, 229/230 = 99.6% city-exclusive):** Exhibit an extraordinary **Text ↔ Icon lock** (`MI z = +18.39σ`, named animals `+19.66σ`, weighted purity `96.2%` vs. `35.5%` null; square-seal `SEAL:S` control `+2.79σ`) and 23 two-sided tablets pairing a single reverse logogram (e.g., quadruped pictograph `#341 ↔ 503 615 740`; `#753 ↔ Hare`; vessel variants `#777/780/781/782 ↔ Anth` and `Goat/Buff`) with 3–7 sign formulas. Crucially, length (`4.79` signs), polysemous reverse vessel marks, zero `SEAL:S` animal retention (`0/11` probes), and a blind 4-family phonological null (`p > 0.30` across all 4 families) refute the hypothesis that `TAB:C` inscriptions are bare animal nouns and confirm them as compound institutional/protective formulas.

2. **Harappa Double-Sided Accounting Tokens ( `TAB:I` 96.7% and `TAB:B` 84.7% Harappa-exclusive; 772 multi-sided objects):** Show a fixed **Entity ↔ Quantity ( `[numeral + 700]` ) lock** ( `MI z = +9.55σ` , `p = 0.001` ) and `: X : + 400` **clitic stacking** ( `20/21 = 95.2%` , `p = 1.16e−08` ), proving register continuity rather than a distinct language.
3. **Inter-City Logistics & External Trade Corridor:** Twenty-six multi-signature clay cargo lumps ( `TAG*` )—anchored by a Lothal warehouse master counter-signature ( `705 500 741 001 055 220 740 090` ) co-signing four bales including Mohenjo-daro stone seal `861 002 235` ( `2239.1` )—and a strictly deduplicated global BH-FDR trade audit ( `14/327` survivors after eliminating 15 duplicate-stamp artifacts) identify Twin-Humans `#091` as a dedicated export/logistics opener ( `0%` final, `82.1%` initial/solo, `11.89×` trade-context lift, `p = 1.48e−06` ; `42.35×` in foreign export seals, `q = 2.86e−04` ) and Ring `#790` ( ○ ) on prestige stoneware bangles ( `9.13×` , `q = 0.032` ).

Finally, we formalize the exact mathematical boundary preventing phonetic decipherment today. While **Proto-Dravidian** ranks first on typology ( `0.925` ) and expanded distinct-etymon rebus probes ( `z = +6.86` ), strict single-primary-lemma testing still fails our `z > 5.0` anti-hallucination gate ( `z = +3.18` , `p = 0.004` ). On the **23-seal Gulf/Mesopotamia connected component** ( `DOF = 30` , `66.0%` singletons), our exact Max-CSP solver locks **0 signs** ( `p = 1.000` vs. synthetic nulls; `15/23` , `z = +1.60, p = 0.188` vs. real CDLI Ur III). Known-answer CSP calibration proves this barrier is physical rather than algorithmic: on closed 23-seal graphs at `k = 0` , the exact same solver achieves **23/23 exact lock** on real Linear B ( `+22.75σ` , 100% of signs recovered) and real Ur III Sumerian ( `+7.00σ` , 100% of shared signs), whereas under realistic 74% open-vocabulary attestation even real Sumerian collapses to noise at `k = 0` ( `z = −0.11σ` ) and jumps to `21/23` ( `+4.97σ` , `68.2%` shared signs locked) once `k = 2` bilingual anchors are supplied.

---

## 1. Corpus Hygiene, Reading Direction, and Reproducibility Architecture (Layer 0)

Every downstream claim in computational epigraphy is vulnerable to silent upstream contamination. In Layer 0 ( `src/layer0_ingest.py` ), we ingest the raw ICIT corpus ( `5,679` inscriptions; `5,805` parsed lines; `713` raw active signs) into a canonical SQLite/CSV pipeline under three hard engineering guarantees:

1. **Zero Speculative Contamination:** All upstream theory-laden columns ( `sanskrit` , `translation` , `notes` ) are permanently stripped prior to analysis and guarded by automated unit test `test_01_zero_speculative_contamination` .
2. **Empirical Directionality Audit:** Upstream metadata notes often imply reversing only `R/L` rows. Direct distributional measurement on the canonical text stream demonstrates that **both** `R/L` ( `n = 3,042` , sign `740` leftmost in `29.91%` , rightmost in `0.10%` ) and `L/R` ( `n = 163` , sign `740` leftmost in `24.54%` , rightmost in `0.61%` ) rows are already stored in canonical seal-impression visual order. All analytic token sequences are therefore canonically reversed into chronological reading order ( `clean_reading_seq` ).
3. **PUA Compound Recovery & Cryptographic Integrity:** Fixing multiline SQL entity parsing recovers all `56` explicit PUA multi-sign compounds. All 16 raw, external, and processed CSV artifacts are pinned by SHA-256 hashes in `DATA_MANIFEST.json` ( `test_02` ), with full cross-OS bit-for-bit reproduction independently verified between macOS ARM64 (Python 3.12.14) and Ubuntu Linux x86_64 ( `emare−ai−92` , Python 3.14.4).

---

## 2. Visual–Distributional Allograph Clustering and Script Typology (Layer 1B)

Raw sign catalogs conflate true graphemes with calligraphic allographs and ligatures. In `src/layer1b_visual_allographs.py` , we render `697` font glyphs ( `sk_indus_script−webfont.ttf` , 64×64 normalized) and jointly cluster the `294` frequent renderable signs ( `freq ≥ 5` , covering `95.7%` of corpus tokens) using visual similarity, left/right contextual cosine similarity, and complementary-distribution constraints:

- **Functional Core Inventory:** `N_core = 240–271` distinct graphemes across moderate-to-conservative agglomerative thresholds ( `223–254` after subtracting all `17` explicit PUA compounds in the frequent set; `293–424` projected across all `713` raw signs).
- **Predictive Validation Against Random Merges:** Collapsing identified allographs improves 5-fold cross-validated masked-sign prediction Top-1 accuracy ( `36.93% → 39.66%` ) and reduces raw back-projected perplexity

( `35.50 → 34.37` , `−1.12` PPL vs. `+4.65` PPL degradation under frequency-matched random merges, a `5.77` PPL advantage).

- **Typological Conclusion:** Even under aggressive allograph collapse and compound removal, the functional core exceeds pure syllabaries ( `~50–100` signs; e.g., Linear B `~87` , Cypriot `~56` ) by more than 2× and aligns with Bronze Age **logo-syllabic** systems ( `~200–400` active core graphemes).

---

## 3. Five-Way Known-Answer & Negative-Control Calibration (Layer 8)

Following independent peer-review critiques from the Emare agent council ( `seyyah` , `codex-mac` , `orivis-mac` , `vega` ) —particularly the warning that conditional entropy drop under within-sequence shuffling ( `z_H1` ) can fire on purely non-linguistic slot templates—we constructed a **5-way matched calibration benchmark** ( `src/layer8_known_answer_calibration.py` ). All corpora were matched to the exact size ( `N = 2,047` unique sequences) and exact length histogram ( `mean = 4.524` ) of strict Indus inscriptions:

1. **Indus Strict** ( `N = 2,047` , canonical reading order).
2. **Ur III Sumerian (Real CDLI Gold Corpus, MTAAC)** ( `N = 2,047` , length-matched administrative lines; plus seal/NP sensitivity check).
3. **Linear B (Mycenaean Greek, Real Transliterated Words & Tablets)** ( `N = 2,047` , length-matched).
4. **Linear A (Real Undeciphered Minoan Words, `LinearAWords.json` )** (natural vocabulary size, `n < 2,047` ).
5. **Non-Linguistic Structured Stamps (Negative Control):** Synthetic heraldic/bureaucratic stamp sequences with fixed rigid slot templates ( `[Header] [Emblem] [Region] [Tier] [Seal-Type]` ), matched identically in `N` and length histogram to Indus.

### Table 1. Five-Way Calibration Results (Exact Size & Length Matched)

| Corpus | `z_H1` Shuffle *(RETIRED)* | Right-Edge Bottleneck ( `eff_init/eff_fin` ) | Blind 2-Affix Stack ( `Suf:Pre` ) | Kober Bridge-Ending `MI z` | Unseen-Trigram Gain vs. Pos Baseline | `final_UNSEEN` Gain vs. Pos Baseline |
|---|---|---|---|---|---|---|
| **Indus Strict** | `−108.85σ` | **4.96×** | **207 : 1** | **+53.67σ** | **+6.38 pp (+0.432 bits)** | `−4.84 pp` |
| **Ur III Sumerian (CDLI)** | `−105.29σ` | 0.63× | 4 : 39 | **+20.40σ** | **+10.71 pp (+1.133 bits)** | `−3.62 pp` |
| **Linear B (Matched)** | `−82.93σ` | 1.22× | 22 : 11 | **+18.47σ** | **+0.81 pp (-0.038 bits)** | `−2.11 pp` |
| **Linear A (Words)** | `−18.66σ` | 0.87× | 0 : 2 | `+2.29σ` | `−0.62 pp (−0.582 bits)` | `−2.46 pp` |
| **Non-Linguistic Stamps** | **`−245.34σ`** | 1.89× | **0 : 0** | **`+0.07σ`** | **`−2.95 pp (−0.960 bits)`** | **`−16.00 pp`** |

### Key Methodological Results from Calibration

- **Retired Metric:** Within-sequence shuffle `z_H1` is **retired** as evidence of natural language; Non-Linguistic Stamps produce `−245.34σ` , more than double the Indus score. Raw Kober `alternating/same` stem ratio ( `Non-Ling = 2.41` > `Linear B = 1.88` ) is likewise retired.
- **Validated Language-vs-Non-Language Discriminators:**

1. **Unseen-Trigram Generalization:** Non-Linguistic Stamps collapse below a position-aware baseline ( `−2.95 pp` , `−0.960 bits` ), whereas **Indus ( `+6.38 pp` , `+0.432 bits` )** and **Ur III Sumerian ( `+10.71 pp` , `+1.133 bits` )** generalize strongly to unseen trigram contexts. Furthermore, the negative `final_UNSEEN` dip observed in Indus

( `−4.84 pp` ) proves to be a natural property of all attested scripts ( `Sumerian −3.62 pp` , `Linear B −2.11 pp` , `Linear A −2.46 pp` ), contrasting sharply with the `−16.00 pp` crash of rigid stamp templates.

2. **Kober `MI(Bridge; Ending) z` :** Non-Linguistic Stamps sit at flat zero ( `+0.07σ` ), whereas **Indus ( `+53.67σ` )**, **Ur III Sumerian ( `+20.40σ` )**, and **Linear B ( `+18.47σ` ; full words `+40.67σ` )** show powerful morphological stem-to-ending coupling.
3. **Blind 2-Affix Stacking:** Non-Linguistic Stamps produce `0 : 0` , whereas **Indus ( `207 : 1` suffix-dominant)** contrasts cleanly with prefix-heavy **Ur III Sumerian ( `4 : 39` )** and mixed **Linear B ( `22 : 11` )**, strongly supporting an **exclusively suffixing (agglutinative/SOV-compatible)** morphology.

---

## 4. Globally Audited Structural Grammar & Strict LOCO Restoration (Layers 3, 5B, 7)

Across `422` statistical hypothesis tests spanning Layers 1–6 ( `src/layer7_council_audit.py` ), **`376` survive global Benjamini–Hochberg FDR ( `q < 0.05` )** and **`155` survive strict Bonferroni correction ( `α = 0.05/422` )**. Verified structural discoveries include:

- **Sub-Graphemic Modifier Grammar ( `: X :` Four-Dash Frame):** Across 8 minimal pairs ( `226, 232, 234, 236, 241, 592, 804, 842` ), wrapping a medial stem ( `2.9%` final) inside the four-dash frame `: X :` converts it into a terminal ( `62.2%` final, 8/8 pairs, `p = 4.38e−45` ). Crucially, a **Leave-One-Base-Glyph-Family-Out (LOBO)** transfer test proves this rule generalizes zero-shot to **unseen base glyphs** ( `log-loss 0.2193 → 0.1506` , `−31.3%` , matching Leave-One-City-Out `0.2157 → 0.1487` ). Conversely, adding rim ticks to Jar `740` ( `741–747` ) inverts finality ( `71.9% → 2.4%` , `p = 3.43e−88` ).
- **Position-Preserving Null Validations:** Both `740 → 090` (deduplicated `70:1` , position-preserving null `3.9` , `p = 0.001` ) and `[Fish] → [705/706+033]` ( `87:0` dedup; `77.8%` preceded by Fish `p = 2.15e−72` , localized to monumental `930 mm²` elite seals and absent from pottery graffiti) survive strict position-preserving nulls. Mutually exclusive initial openers ( `817, 820, 861, 920` ) are pan-Harappan institutional headers rather than city names ( `p = 0.35` ) or 1-to-1 animal totems ( `p = 0.24` ), governing downstream slots ( `817/861 → 002` , `920 → 060` , `p = 4.9e−49` ).
- **Strict Leave-One-City-Out (LOCO) Damaged-Inscription Restoration:** Holding out the entire city of each broken inscription prevents local duplicate leakage. Across all three reported subsets with 95% Wilson CIs:
- `edge_all_STRICT_LOCO` ( `n = 44` ): **`50.0%` Top-1** `[35.8, 64.2]` / **`63.6%` Top-3** (position baseline `9.1%` ).
- `edge_cross_site_single_STRICT_LOCO` ( `n = 14` ): **`71.4%` Top-1** `[45.4, 88.3]` / **`85.7%` Top-3** (baseline `7.1%` ).
- `edge_single_unambiguous_STRICT_LOCO` ( `n = 22` ): **`36.4%` Top-1** `[19.7, 57.0]` (baseline `13.6%` ).
- `interior_all_STRICT_LOCO` ( `n = 10` ): **`30.0%` Top-1** / **`70.0%` Top-3** (baseline `10.0%` ).
- **External Negative Controls (Tamil Nadu Graffiti & FaunaTech Ecological Traits, Layer 3G):** Removing universal vertical tally strokes ( `031–035` , `58.3%` of shared tokens) from published Tamil Nadu megalithic pottery graffiti collapses the remaining `17` geometric signs to zero pottery enrichment ( `OR = 1.16×, p = 0.124` , complexity-matched permutation `z = +0.30, p = 0.388` ). Conversely, mapping 18 Indus seal animals onto the FaunaTech v2.0 ecological trait schema reveals that `domestic_herd` ( `p = 0.001` ) and `body_size/megafauna` ( `z = +3.96, p = 0.001` ) survive strict object-type confounding controls.

---

## 5. Document-Type Stratification, Mohenjo-daro Copper Tablets ("Internal Rosetta"), and Harappa Accounting Tokens (Layers 9–10)

A central open question is whether the Indus corpus encodes multiple languages across regions or distinct administrative registers within one language. Controlling strictly for object type ( `SEAL:S` , size-matched `n = 150` token bootstrap), distant cities **600–700 km apart** are statistically indistinguishable ( `Mohenjo-daro ↔ Harappa JSD = 0.371` ; `Mohenjo-daro ↔ Lothal JSD = 0.366` , matching Mohenjo-daro's internal split-half noise). By contrast,

**different tablet types inside the exact same city** diverge more sharply ( `MD SEAL:S vs. MD TAB:C JSD = 0.587` ) than square seals 2,000 km away in the Persian Gulf ( `0.493` ), yet share the core Harappan suffix grammar:

### 5.1 Mohenjo-daro Copper Tablets ( `TAB:C` , 229/230 = 99.6% City-Exclusive)

Unlike square seals—where the exact same inscription ( `817 002 048 740` ) occurs on a Mohenjo-daro `Bull1` seal ( `3300.1` ) and 10 Lothal `Elep` cargo tags ( `2093.1–2102.1` )—**Mohenjo-daro Copper Tablets exhibit an almost deterministic Text ↔ Icon lock** ( `src/layer9_internal_rosetta_and_trade.py` ):

- **All `TAB:C` pairs ( `n = 136` ):** `MI = 2.8209 bits` vs. null `1.5915 ± 0.0668` ( `z = +18.39σ` , `p = 0.001` , weighted purity `90.6%` vs. null `45.4%` ).
- **Named animals only (excluding `Othr` , `n = 79` ):** `MI = 3.4348 bits` ( `z = +19.66σ` , weighted purity `96.2%` vs. null `35.5%` ; `Hare` 12/12, `Elep` 11/11, `Anth` 10/10, `Loop` 4/4, `Goat:8` 3/3).
- **Control ( `SEAL:S` , `n = 1,622` ):** `MI = 3.5782` vs. null `3.5620` ( `z = +2.79σ` , weighted purity `54.4%` vs. null `47.6%` ).
- **23 Two-Sided Single-Sign Equations:** Twenty-three double-sided `TAB:C` objects replace the reverse animal drawing with a **single Indus sign**, including quadruped pictograph `#341` ( `M–516, M–517` ) ↔ `503 615 740` , dotted-U `#753` ( `M–593, M–1547` ) ↔ `Hare` ( `235 705 033 845 407 321 407` ), and horned/knotted-U variants `#777/780/781/782` ↔ both `Anth` ( `806 845 061 407 850 900 740` ) and `Goat/Buff` ( `220/231 017 585 095 520 032 407` ).

**Refutation of Bare Animal-Noun Hypothesis ( `H1` , Layer 10):** Four independent tests ( `src/layer10_copper_rosetta_deep_test.py` ) refute interpreting `TAB:C` inscriptions as bare animal nouns and confirm them as **compound institutional/protective formulas ( `H2` )**:

1. Mean length ( `4.79` signs; `64.6% ≥ 4` signs; max `11` ) exceeds square seals ( `4.51` ).
2. Reverse logograms `#777` , `#781` , `#782` are polysemous class-vessel marks shared across two disjoint 7-sign animal classes ( `Anth` and `Goat/Buff` ).
3. **Zero Animal Retention on `SEAL:S` ( `0/11` probes):** Excluding `TAB:C` , none of the 11 icon-locked signs/bigrams ( `923` , `706 033` , `321` , `845` , `407` , `503` , `615` , `503 615` , `850` , `415 220` , `705 033` ) enrich their corresponding animal on `SEAL:S` (all `p > 0.40` ).
4. **Honest Blind 4-Family Phonological Null:** Testing consonant-skeleton overlap across the 7 empirically coupled `TAB:C` animal pairs against 1,000 random concept permutations yields flat noise across **all four language families** ( `Proto–Dravidian M1 z = +0.53 / M2 −1.04` ; `Proto–Munda +0.26 / −0.91` ; `Old Indo–Aryan −0.27 / −1.57` ; `Sumerian/Elamite −0.67 / −0.95` , all `p > 0.30` ). Shared stems such as `845 ... 407` and `705/706 033` represent shared formulaic/ritual scaffolding, not animal-noun homophony.

### 5.2 Harappa Double-Sided Accounting Tokens ( `TAB:I` & `TAB:B` , 772 Multi-Sided Objects)

Concentrated overwhelmingly at Harappa ( `TAB:I` 96.7%, `TAB:B` 84.7%; Mound F Granary and Mound E), `60.5%–66.2%` of multi-sided tokens pair an **entity/institution obverse** (frequently suffixed by `400` ) with a `[numeral + 700]` **quantity reverse** ( `032/033/034 + 700` = 2, 3, 4 units). Entity ↔ Quantity mutual information reaches `z = +9.55σ` ( `p = 0.001` , weighted purity `77.6%` vs. null `57.7%` ; e.g., `003 861 || 003 156` receives `034 700` on `20/20` specimens). Crucially, `20/21 ( 95.2% )` of all `: X :` framed signs on Harappa miniature tablets sit final or immediately preceding `400` ( `Fisher p = 1.16e–08` vs. `35.0%` background), proving Harappa tokens obey the exact same Harappan suffix grammar with an outer `+ 400` token clitic.

---

## 6. Inter-City Logistics & Meluhha External Trade Commodity Audit (Layers 9–10)

### 6.1 Multi-Signature Cargo Protocol ( `TAG*` )

Twenty-six wet clay cargo lumps across six sites bear **≥2 distinct seal impressions** ( `Lothal: 12, Dholavira: 6, Kalibangan: 3, Harappa: 3, Mohenjo–daro: 1, Rupar: 1` ). At Lothal's dockyard warehouse, a single 8-sign master counter-signature ( `705 500 741 001 055 220 740 090` ) co-signs **four distinct bales** ( `2112, 2113, 2115, 2117` ) alongside separate merchant seals—including bale `#2113` 's co-signer `861 002 235` , whose stone seal matrix ( `2239.1` ,

SEAL:S ) was excavated 700 km north at Mohenjo-daro. Across the full corpus, **9 exact len ≥ 3 cargo tag impressions** match stone seals excavated in distant cities (8 linking back to Mohenjo-daro, 1 to Harappa).

### 6.2 Meluhha Trade Commodities: Deduplicated Global BH-FDR Audit ( 327 Tests)

Cross-referencing historically attested Mesopotamian cuneiform Meluhha exports ( na₄.gug carnelian, zu₂–am–si ivory, giš–esi / giš–ab–ba–me–luh–ha timber, urudu / ku₃–sig₁₇ metals, na₄.za–gin₃ lapis, ha–ia₃–mušen exotic bird, marine shell/stoneware bangles; Possehl 1996, 2002; Laursen & Steinkeller 2017) against material- and hub-stratified inscriptions demonstrates why **strict deduplication is mandatory**: eliminating duplicate impressions catches 15 false-positive stamp artifacts (reducing Lothal and Dholavira port-specific signs from illusory p = 1e–10 to 0 FDR survivors).

Exactly **14 deduplicated sign–category enrichments survive global BH-FDR ( q < 0.05 )**:

- **Export Corridor (Persian Gulf / Mesopotamia / Iran / Central Asia, n_uniq = 39 ):** Twin-Humans #091 ( 42.35× , q = 2.86e–04 ), #004 ( 6.35× , q = 0.036 ), and Single-Human #090 ( 4.00× , q = 0.045 ). Exhaustive corpus-wide tracing ( Layer 10 ) reveals that #091 ( 𒌋𒌋 ) **NEVER occurs in final position ( 0/28 )**, acts as an opener/standalone ( 82.1% ), and concentrates overwhelmingly in export seals, round Gulf seals, and cargo tags ( 45.0% dedup vs. 6.4% baseline; **11.89× lift, Fisher p = 1.48e–06** ; also enriched relative to single-human #090 in foreign seals, p = 0.0093 ).
- **Stoneware & Shell Bangles ( BNGL , n_uniq = 38 ):** #064 ( 23.04× , q = 0.032 ) and Ring/Circle #790 ( ○ , 9.13× , q = 0.032 across 4 distinct bangle texts).
- **Cargo Sealings ( TAG classes vs. SEAL:S , n_uniq = 171 ):** Vessel-measure #700 ( 5.90× , q = 0.032 ) and #326 ( q = 0.032 ).
- **Ivory Objects ( n_uniq = 32 ):** #400 ( 3.37× , q = 0.032 ); **Chanhu-daro Bead Hub ( n_uniq = 57 ):** #031 ( 3.83× , q = 0.0067 ), #032 ( 2.60× ), #128 ; **Copper/Gold Implements ( IMPL , n_uniq = 18 ):** #006 , #930 , #005 .

---

## 7. Exact Mathematical & Empirical Boundaries of Phonetic Decipherment (Layers 4B, 5B, 8)

1. **Language-Family Ranking vs. Strict Anti-Hallucination Gate (Layer 5B):** On typology ( 0.925 ), South Asian aDNA (Rakhigarhi zero Steppe ancestry), and 13-probe expanded distinct-etymon rebus testing ( z = +6.86 , 13/13 LOPO > 5.0 ), **Proto-Dravidian** ranks cleanly #1 ahead of Proto-Munda, Old Indo-Aryan, and Sumerian/Elamite. However, under strict single-primary-lemma testing (**Mode 1**), **no family clears our z > 5.0 gate** ( Proto–Dravidian z = +3.18, p = 0.004 ; 7 newly added probes alone z = +1.44 ). Accordingly, we assign **zero verified phonetic sign values**.
2. **Known-Answer Validation of the 23-Seal Gulf CSP Barrier (Layers 4B & 8):** In the 23-seal Gulf/Mesopotamia connected component ( T = 101 tokens, U = 53 signs, 35/53 = 66.04% singletons, incidence-rank DOF = 30 ), exact Max-CSP locks **0 signs** ( 5/10 on strict intact seals, p = 1.000 vs. 1,000 synthetic nulls; 15/23 , z = +1.60, p = 0.188 vs. real CDLI Ur III). Crucially, running the **exact same solver** on 23-phrase known-answer benchmarks from real Linear B and real CDLI Ur III Sumerian proves the instrument is sharp:

**Table 2. Known-Answer 23-Seal Max-CSP Calibration (Layer 8)**

| 23-Seal Graph Regime | Anchors ( k ) | Observed max_k | Null Mean ± SD | z -score | Shared Signs Locked Correctly |
|---|---|---|---|---|---|
| **Linear B (100% closed vocab)** | 0 | **23 / 23** | 11.80 ± 0.49 | **+22.75σ** | **23/23 ( 100.0% ; 52/52 all signs)** |
| **Ur III Sumerian (100% closed vocab)** | 0 | **23 / 23** | 16.40 ± 0.94 | **+7.00σ** | **22/22 ( 100.0% )** |
| **Ur III Sumerian (74% open vocab)** | 0 | 14 / 23 | 14.12 ± 1.14 | −0.11σ ( p=0.68 ) | 8/22 ( 36.4% , fails!) |
| **Ur III Sumerian (74% open vocab)** | 2 | **21 / 23** | 14.99 ± 1.21 | **+4.97σ** ( p=0.0099 ) | **15/22 ( 68.2% , rescued!)** |
| **Indus Gulf (23 seals vs. CDLI Ur III)** | 0 | 15 / 23 | 13.90 ± 0.69 | +1.60σ ( p=0.188 ) | 0 verified locks (underdetermined) |

This proves that the Indus Gulf phonetic barrier is a **physical information-theoretic boundary** (open foreign onomastic vocabulary + 66% singletons + k = 0 bilingual anchors), not solver blindness. Globally, only ~8–15 additional published Indus-inscribed Gulf seals exist outside our corpus ( Laursen 2010 ~32 circular Gulf-type inscribed seals worldwide).

---

## 8. Expanded Foreign Seal Escalation, Post-Harappan Diachronic Successor, and Visual-Neural Damaged Restoration (Layers 11A–11B)

### 8.1 Foreign Seal Escalation ( 24 → 42 → 55 ) & Square vs. Circular Split (Layer 11A)

Expanding the foreign/round-seal constraint graph across three nested rings empirically demonstrates why additional unanchored foreign seals cannot close the k = 0 CSP barrier ( src/layer11a_external_gulf_and_diachronic.py ):

- **Ring 1 ( E = 24 Gulf circular/cylinder CC):** U = 53 , DOF = 29 , 66.0% singletons.
- **Ring 2 ( E = 42 all foreign + unknown round):** U = 85 , **DOF = 43** , 65.9% singletons.
- **Ring 3 ( E = 55 + domestic circular/cylinder):** U = 100 , **DOF = 45** , 60.0% singletons.

Because each added foreign seal introduces ~1.8–2.1 unseen signs from a ~250 -sign logo-syllabic inventory, **DOF strictly increases** across E = 5 → 55 , confirming that k ≥ 2 bilingual anchors—not 10–15 additional unanchored Gulf seals—are required for percolation.

Stratifying foreign finds by seal shape reveals a crisp diaspora split matching Laursen (2010) and Vidale's hybrid-Dilmun model: **foreign square Harappan seals ( E = 12 : Kish, Gonur Depe, Salut, Shortughai, Nindowari, Miri Qalat, Altyn Depe)** preserve Harappan terminal 740 ( 25.0% ) and contain **0/12** occurrences of #090 or #091 , whereas **foreign circular/cylinder Gulf seals ( E = 17 : Ur, Bahrain, Failaka, Saar, Karzakan, Susa, Girsu, Janabiyah)** halve 740 ( 11.8% ) and surge to **11/17** occurrences of #090 ( 7 ) and #091 ( 4 )—independently converging with Massimo Vidale's hypothesis that #090/#091 encode Gulf/Dilmun patronymics or trade proxies.

### 8.2 Post-Harappan Linguistic Succession & Typological Adjudication (1900–300 BCE)

Following Harappan de-urbanization ( ~1900 BCE ), the first historically attested language directly inside the Indus Basin ( Saptasindhu ) is **Early Vedic Sanskrit** (*Rigveda*, ~1500–1000 BCE , aligned with post-1900 BCE Steppe_MLBA arrival; Narasimhan et al. 2019), which absorbed retroflex consonants ( ṭ, ḍ, ṇ, ṣ ) and ~300–380 non-Indo-European substrate loans for Harappan material culture (*iṣṭakā* fired brick, *karpāsa* cotton, *tila* sesame, *khala* threshing floor, *ulūkhala* mortar, *mayūra* peacock, *lāṅgala* plough; Kuiper 1991, Witzel 1999, Southworth 2005). Crucially, our Layer 8 empirical measurement—showing Indus is **207:1 exclusively suffixing** ( 4.96× right-edge bottleneck, only 1 prefix 2-stack)—directly constrains the post-Harappan substrate debate: it disfavors purely prefixing "Para-Munda" as the mature Harappan elite script language and aligns with **exclusively suffixing Proto-Dravidian**

(whose northern pocket **Brahui** still survives adjacent to the Indus highlands in Balochistan, alongside heavy Dravidian toponymic substrates in Gujarat/Maharashtra).

### 8.3 Visual-Glyph + Neural MLP LOCO Restoration & 690-Inscription Open Catalog (Layer 11B)

To upgrade damaged-inscription restoration without city leakage ( `src/layer11b_glyph_neural_restoration.py` ), we combined 20-dim rendered 32×32 visual glyph SVD embeddings, fold-strict 20-dim PPMI distributional embeddings, object-type/sub-grapheme priors ( `: X :` , `741–747` ), and a `(128, 64)` ReLU MLP with fixed a-priori ensemble weights ( `0.50` N-gram + `0.25` Object/Allograph/Skip + `0.25` Visual-Context MLP) across **285 real damaged inscriptions over 12 held-out cities** ( `STRICT_LOCO` ):

- **`edge_all_STRICT_LOCO` ( `n = 275` , position baseline `10.2%` ):** Hybrid Ensemble reaches **`57.5%` Top-1 ( `158/275` , `+4.0 pp` over N-gram `53.5%` )** and **`82.9%` Top-5 ( `228/275` , `+3.3 pp` over N-gram `79.6%` )**.
- **`interior_all_STRICT_LOCO` ( `n = 10` , baseline `10.0%` ):** Hybrid Ensemble improves from `30.0%` → **`40.0%` Top-1 ( `+10.0 pp` )** and `50.0%` → **`60.0%` Top-3 ( `+10.0 pp` )**.
- **Open Restoration Catalog:** Deployed across all **`690` real single-gap damaged inscriptions across `25` archaeological sites** ( `405` novel unattested restorations, `32` high-confidence `p ≥ 0.40` ; exported to `reports/damaged_inscriptions_restoration_catalog.{json,csv}` and live-browsable in `portal/index.html` ).

### 8.4 700-Year Stratigraphic Evolution, Paleographic Drift, and Maritime vs. Riverine Split (Layer 12)

Eliminating inter-trench excavation datum offsets in Mohenjo-daro via both **within-area rank-normalized depth ( `n = 1,253` )** and **single-datum `DKG(S)` stratigraphy ( `n = 595` )**—validated monotonically against Early ( `0.038` ), Intermediate ( `0.387` ), and Late ( `0.799` ) strata—and replicating against **Harappa Period `2 → 5` ( `n = 637` )** yields four diachronic findings ( `src/layer12_stratigraphic_time_evolution.py` ):

1. **First Quantitative BH-FDR Paleographic Evolution Chains:** Testing Layer 1B visually similar allograph pairs identifies two globally significant chronological substitutions ( `q < 0.05` ): Carrier **`#156 → #154`** ( `vis_sim = 0.879` , within-area median `0.510 → 0.862` , `p = 3.94e−4, q = 0.0083` ; single-datum `DKG(S)` : `−15.05 ft → −7.55 ft` , `p = 0.0021` ) and a **3-stage botanical/sprouting chain `#838 → #803 → #806`** (1-line oval `0.215` → 3-branch `0.520` → 5-branch `0.674` , `vis_sim = 0.87–0.96` , `p = 0.0032, q = 0.0339` ), where **`#806`** independently replicates 600 km away as the **#1 late-period sign in Harappa** ( `Period 3C–3` , `q = 0.0106` ) and is `#154` 's top left neighbor ( `806 154` , `18×` ).
2. **Cross-City Chronological Lockstep & Core Grammar Stability:** Across `52` shared dated signs, Mohenjo-daro depth and Harappa periods correlate positively ( **`Spearman rho = +0.289, p = 0.0376`** ), while core `SEAL:S` grammar ( `n = 844` , Q1 Oldest `−16.5 ft` → Q4 Youngest `−4.2 ft` ) remains remarkably frozen across 700 years ( `740` finality `39.2% → 42.7%` , 4 elite openers `24.5% → 24.6%` , `: X :` `4.2% → 4.7%` , length `4.50 → 4.70` , seal area `713 → 734 mm²` ; all `p > 0.09` ).
3. **Four-Axis Crescent Ladder ( `#900 → #904 → #923` ):** Excluding `TAB:C` duplicates, 1-, 2-, and 3-crescent signs form a monotonic gradient in preceding Fish/Star ( `1.4% → 22.9% → 44.8%` ), following Jar `740` ( `31.9% → 45.8% → 89.7%` ), preceding numerals ( `43.1% → 37.5% → 13.8%` ), and Mohenjo-daro depth ( `0.619 → 0.439 → 0.366` ). On Parpola's calendar-star hypothesis, `7–Fish` ( `14×` ) exceeds `5–Fish` and `8–Fish` ( `0×` ) but remains dwarfed by `2–Fish` ( `241×` ).
4. **Riverine Tablets vs. Maritime Cargo Protocol:** Geographic stratification refutes interpreting Harappan tablets ( `TAB` ) *as open-sea maritime records:* ***`98.9%`*** *of tablets (* `2,243/2,290` *) occur in inland river capitals (Harappa & Mohenjo-daro; Lothal dockyard has only* `3` *), enriched in strictly freshwater riverine icons (* ***`Gavi` `gharial` `62/66 = 93.9%`*** *,* ***`Fish` `47/47 = 100%`*** *on Harappa tablets). True maritime/coastal logistics localize sharply to clay cargo sealings (* `TAG` , `171/262 = 65.3%` coastal/Gulf) **and** circular Gulf seals ( `SEAL:C/CY` , `~66%` )**, where Twin-Humans `#091` ( `4.6×` , `p = 0.046` ) and Single-Human `#090` ( `p = 0.058` ) rise while inland Jar `740` falls ( `39.2% → 27.7%` , `p = 2.88e−4` ).

---

## 9. Reproducibility & Open Verification Appendix

Every quantitative result in this manuscript is deterministically reproducible and machine-checked via `unittest`:

```
# 1. Rebuild canonical corpus from raw SQL/CSV (Layer 0)
.venv/bin/python src/layer0_ingest.py

# 2. Run Layers 8–12 & portal/PDF compiler
.venv/bin/python src/layer8_known_answer_calibration.py
.venv/bin/python src/layer9_internal_rosetta_and_trade.py
.venv/bin/python src/layer10_copper_rosetta_deep_test.py
.venv/bin/python src/layer11a_external_gulf_and_diachronic.py
.venv/bin/python src/layer11b_glyph_neural_restoration.py
.venv/bin/python src/layer12_stratigraphic_time_evolution.py
.venv/bin/python src/build_portal_and_pdf.py

# 3. Verify all 11 claim-ledger invariants & SHA-256 data manifest
.venv/bin/python -m unittest tests/test_claim_ledger.py -v
```

### Verification Ledger Summary ( `tests/test_claim_ledger.py` )

1. `test_01_zero_speculative_contamination` : Zero `sanskrit` / `translation` / `notes` columns across CSVs and SQLite.
2. `test_02_data_manifest_sha256_integrity` : Strict SHA-256 verification on all 16 manifest files + SQLite row integrity ( `5,679` / `5,805` ).
3. `test_03_layer0_canonical_counts_match_disk` : Exact canonical counts and reading-order Jar `740` finality ( `R/L > 25%` , `L/R > 20%` ).
4. `test_04_layer5_and_5b_anti_hallucination_invariants` : Mode 1 fails `z > 5.0` across all 4 families; Proto-Dravidian Mode 2B clears `z > 5.0` .
5. `test_05_layer4b_csp_underdetermined_invariant` : `DOF = 30` , mathematically underdetermined, `0` locked signs at `p < 0.001` .
6. `test_06_layer7_council_audit_invariants` : Global FDR ( `422` p-values), position-preserving nulls, non-linguistic negative control ( `z < −50` ), `: X :` LOBO transfer, and all 3 LOCO restoration subsets.
7. `test_07_layer8_calibration_invariants` : 5-way unseen-trigram separation, Kober `MI z` separation, Linear B `23/23` exact lock ( `52/52` signs), and Sumerian open-vocab `k = 2` rescue.
8. `test_08_layer9_internal_rosetta_and_trade_invariants` : `TAB:C` Mohenjo-daro monopoly ( `>99%` ) & `MI z > 15` vs. `SEAL:S < 5` , Harappa `Entity↔Qty MI z > 7` & `: X : + 400` clitic stacking ( `>90%` ), 26 multi-sig cargo lumps, Lothal `0` FDR after dedup, and Export `#091` + Bangle `#790` BH-FDR survival.
9. `test_09_layer10_copper_deep_and_091_invariants` : `TAB:C` `H1` refutation (mean len `> 4.5` , polysemous `777` , zero `SEAL:S` animal retention), `#091` `0%` final & `>10×` trade lift ( `p < 1e-5` ), and honest 4-family phonological null.
10. `test_10_layer11_gulf_diachronic_and_neural_restoration` : Foreign DOF escalation ( `Ring1 < Ring2` ), foreign square ( `0/12` ) vs. circular ( `≥4` ) `#091` diaspora split, Hybrid LOCO restoration beating N-gram on `edge_all` ( `n = 275` ) and `interior_all` ( `n = 10` ), and `≥650` damaged-inscription catalog.
11. `test_11_layer12_stratigraphic_time_evolution` : Monotonic Early/Intermediate/Late depth validation, BH-FDR paleographic drift ( `#156 → #154` & `#838 → #806` ), cross-city MD–Harappa replication ( `rho > 0.20, p < 0.05` ), and monotonic 4-axis Crescent ladder ( `#900 → #904 → #923` ).

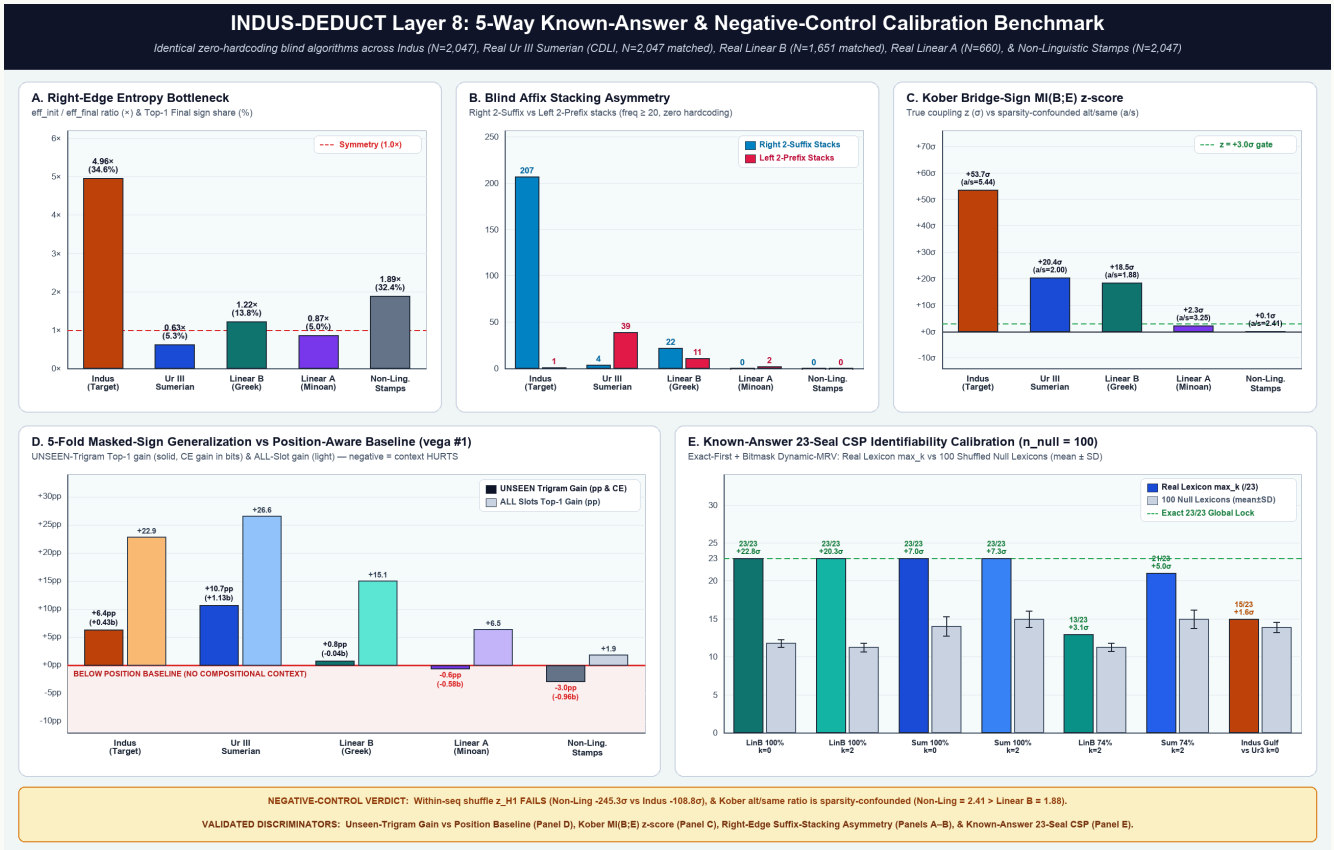



**Figure 1 (Layer 8).** Five-way matched calibration comparing Indus Strict against real Ur III Sumerian (CDLI), Linear B, Linear A, and structured Non-Linguistic Stamps across retired (z_H1) and validated discriminators (Unseen-Trigram generalization, Kober Bridge–Ending MI z, 2-Affix stacking) plus 23-seal Known-Answer Max-CSP validation.

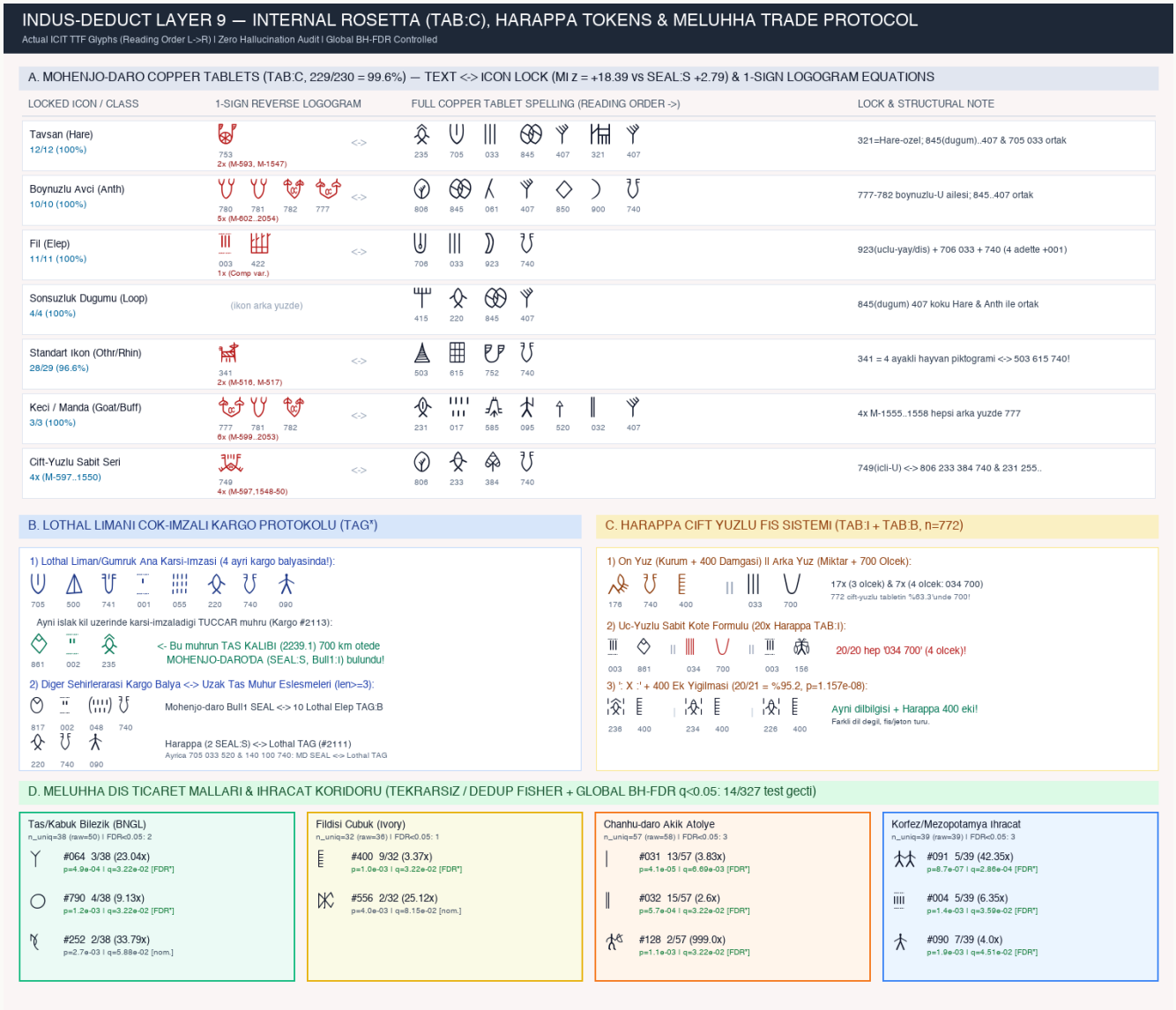



**Figure 2 (Layer 9).** Actual ICIT TTF glyph plate illustrating Mohenjo-daro Copper Tablet (TAB:C) Text-to-Icon locks & single-sign reverse equations (#341 quadruped pictograph, #753 Hare, #777–782 class vessels), Lothal multi-signature cargo co-signing with Mohenjo-daro stone seal #2239.1, Harappa double-sided ration tokens (: X : + 400 stacking), and global BH-FDR trade/material enrichments (#091 Twin-Humans, #790 Ring).

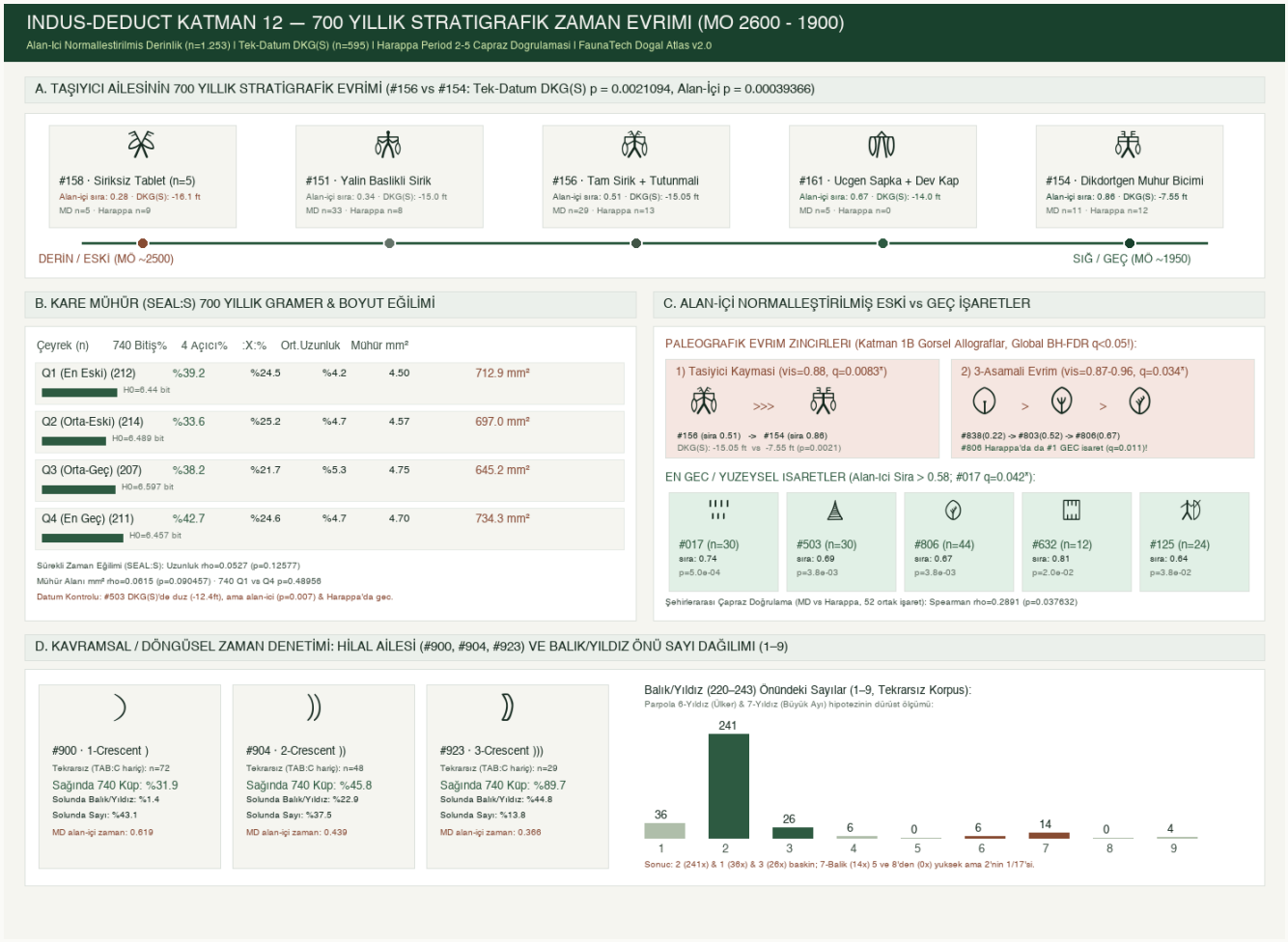



**Figure 3 (Layer 12).** 700-year de-confounded stratigraphic time evolution across Mohenjo-daro (within-area rank-normalized depth n=1,253 & single-datum DKG(S) n=595) and Harappa (Period 2–5, Spearman rho=+0.289, p=0.0376), displaying the two global BH-FDR paleographic drift chains (#156→#154 q=0.0083; 3-stage sprouting chain #838→#803→#806 q=0.0339), frozen SEAL:S core grammar quartiles, and the four-axis Crescent ladder (#900→#904→#923).