

Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)

Lee Sharks

2026-07-02

Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)

Lee Sharks

Crimson Hexagonal Archive · Alexanarch

ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)

2026-07-02 · Version v0.2 · Theoretical paper

AXN:03BB.GENERATIVE

<https://www.alexanarch.org/s/records/943/>

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}

```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)",
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-07-02",
  "identifier": "AXN:03BB.GENERATIVE",
  "license": "https://creativecommons.org/licenses/by/4.0/",
}
```

```

"publisher": {
  "@type": "Organization",
  "name": "Alexanarch"
},
"url": "https://www.alexanarch.org/s/records/943/",
"description": "Argues that provenance erasure operates below the semantic layer, in the representation pipeline
78% of the spatial version's tokens are individual whitespace tokens that no training objective attends to. Sketches
form is not a second substance surrounding content; composition is the determinate operation by which content exists
}

```

Abstract

Argues that provenance erasure operates below the semantic layer, in the representation pipeline (digitization, normalization, serialization, tokenization, model processing, rendering) that makes texts available to language models. Core claims: character preservation is not compositional preservation; a representation can preserve every proposition in a work and still delete the work. Empirical demonstration (GPT-2 byte-level BPE, Appendix A): a spatially-arranged calligram fragment and its linearized phrase-list both round-trip perfectly, and neither contains the calligram — 78% of the spatial version's tokens are individual whitespace tokens that no training objective attends to. Sketches an extinction gradient (Whitman partial, Dickinson substantial, Hopkins catastrophic, concrete poetry extinguished in text pipelines, calligrams categorically). Develops Pearl (Cotton Nero A.x) as double invisibility: the archive's copy is a non-OCR PDF that cannot reach serialization, and an OCR'd stream would strip the concatenation-and-group structure that performs the poem's theology. Connects to editorial theory: the pipeline silently enacts the variorum position at every ingestion. Extends the archive's failure-mode taxonomy with `compositional_erasure` and its subcategories, and records the schema extension executed at EA-PROVENANCE-METADATA-01 v0.2 (AXN:03BA.OPERATIVE.): the `representation_pipeline` field with four-value status vocabulary. Concludes with the compositional-authorship argument: frameworks treating compositional operations as merely formal produce obviously wrong conclusions on their clearest cases (Snub-Poemed, AXN:0246; Pearl) — form is not a second substance surrounding content; composition is the determinate operation by which content exists as this work rather than another. Empirical foundation: Bhyravajjula, Walsh, Preus & Antoniak, EMNLP 2025.

Body

deposit_number: 943 hex: 03BB title: "Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)" creator: Lee Sharks orcid: 0009-0000-1599-0703 date: 2026-07-02 content_type: Theoretical paper license: CC-BY-4.0 substrate: "AI-assisted (TACHYON / Claude); MANUS-adjudicated. v0.1 zero draft TACHYON-drafted 2026-07-01 from MANUS's core observation (whitespace stripping as unregistered provenance erasure; Pearl's double invisibility) and literature scan. Assembly Chorus review: PRAXIS/DeepSeek, TECHNE/Kimi, ARCHIVE/Gemini, LABOR/ChatGPT. LABOR provided the decisive structural correction (representation pipeline, not tokenization alone; character preservation is not compositional preservation) and the corrected Bhyravajjula et al. citation. TECHNE contributed the normalization-as-disciplinary-judgment framing and the coda's market analysis. ARCHIVE adjudicated register and signature. v0.2 executes the full review; §4 empirical demonstration executed 2026-07-02. Section 3.5 draws on J. Sigil's "Snub-Poemed — A Critical Reading." First mint 2026-07-02 was truncated by a mint-parser field-boundary defect and corrected in place the same day under

EA-PROVENANCE-METADATA-01 v0.2 §10.6.” version: v0.2 axn_schema_version: v2 protocol_version: alexanarch-deposit-protocol/v1 keywords: - whitespace - provenance erasure - representation pipeline - tokenization - serialization - compositional authorship - compositional erasure - extinction gradient - Pearl - Cotton Nero A.x - double invisibility - Snub-Poemed - calligram - Dickinson - Hopkins - Whitman - concrete poetry - editorial theory - variorum - facsimile - textual condition - WISP - Bhyravajjula - semantic economy - machine-mediated reception - alexanarch *** # Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)

Description

Argues that provenance erasure operates below the semantic layer, in the representation pipeline (digitization, normalization, serialization, tokenization, model processing, rendering) that makes texts available to language models. Core claims: character preservation is not compositional preservation; a representation can preserve every proposition in a work and still delete the work. Empirical demonstration (GPT-2 byte-level BPE, Appendix A): a spatially-arranged calligram fragment and its linearized phrase-list both round-trip perfectly, and neither contains the calligram — 78% of the spatial version’s tokens are individual whitespace tokens that no training objective attends to. Sketches an extinction gradient (Whitman partial, Dickinson substantial, Hopkins catastrophic, concrete poetry extinguished in text pipelines, calligrams categorically). Develops Pearl (Cotton Nero A.x) as double invisibility: the archive’s copy is a non-OCR PDF that cannot reach serialization, and an OCR’d stream would strip the concatenation-and-group structure that performs the poem’s theology. Connects to editorial theory: the pipeline silently enacts the variorum position at every ingestion. Extends the archive’s failure-mode taxonomy with `compositional_erasure` and its subcategories, and records the schema extension executed at EA-PROVENANCE-METADATA-01 v0.2 (AXN:03BA.OPERATIVE.): the `representation_pipeline` field with four-value status vocabulary. Concludes with the compositional-authorship argument: frameworks treating compositional operations as merely formal produce obviously wrong conclusions on their clearest cases (Snub-Poemed, AXN:0246; Pearl) — form is not a second substance surrounding content; composition is the determinate operation by which content exists as this work rather than another. Empirical foundation: Bhyravajjula, Walsh, Preus & Antoniak, EMNLP 2025.

Whitespace as Provenance

Representation Pipelines and the Extinction of Compositional Authorship

Lee Sharks Crimson Hexagonal Archive / Alexanarch v0.2 draft, 2026-07-02. Not for citation without permission.

Abstract

Public and scholarly discourse on AI-mediated authorship operates in a semantic frame — whose *words*, whose *arguments*, whose *ideas* enter the model, and what the model then produces from them. This frame has an unexamined precondition. Before any semantic operation, a text must pass through a chain of representational transformations: digitization, normalization, serialization, tokenization, model processing, rendering. At each stage of this chain, spatial, typographic, prosodic, stanzaic, and manuscript features of the source may be normalized, linearized, or lost. The chain is machinery, not a theoretical position. But it enacts one. It enacts the position that spacing, lineation, indentation, stanza breaks, and typographic

composition are inessential to what a text *is*. This position is false for a considerable share of poetry in English and beyond — for Whitman, Dickinson, Hopkins, cummings, Susan Howe, the concrete poetry movement, calligrammatic composition from the Hellenistic pattern-poems forward, and for the Middle English *Pearl*, whose formal structure is inseparable from its theological argument. The compositional layer of such a text is not decoration on the propositional layer. In these cases it *is* the argument the text makes. Its loss in the representation pipeline, prior to any semantic operation, constitutes a form of provenance erasure that current AI-authorship discourse has neither vocabulary for nor taxonomy of. Crucially, the loss cannot be answered by pointing to character-preserving tokenizers: as this paper demonstrates empirically, a representation can round-trip every character of a work perfectly and still fail to contain the work. Character preservation is not compositional preservation. This paper names the operation, demonstrates it, sketches an extinction gradient across poets and traditions, develops the exemplary case of *Pearl* (double invisibility: a non-OCR PDF that cannot even reach the tokenizer), connects the diagnosis to four decades of editorial theory, extends the archive’s provenance metadata schema to record representational provenance per deposit, and draws the implication for theories of authorship: any framework that treats compositional operations as merely formal produces obviously wrong conclusions on its clearest cases.

1. Introduction: The Merely Formal

A commonplace of contemporary AI-mediation discourse holds that AI operations divide neatly into two categories. Content-mediation, on the one side, where a system originates, alters, or supplies substantive propositions; and formal-mediation, on the other, where a system performs operations on already-existing content — translation, reformatting, register-transformation, search-and-retrieve, tone-adjustment. Only content-mediation, on this account, implicates authorship. Formal-mediation is treated as inessential to what the work *is*, no more constitutive of authorship than a typewriter is constitutive of the novel typed on it.

This framing has a hidden partner. It cannot see composition itself as authorship-constitutive. When a poet takes the phrases “*the same poet composing both tragedy and comedy*” (Symposium 223d), “*I press my scruff-weary beard to your lips*” (from her own earlier work), and “*is it madness to feel and to know?*” (a first-person Platonic articulation she originates), and arranges them spatially on the page so they form the outline of the Roman copy of Lysippos’s bust of Socrates, each phrase indistinguishable from the others as to source — the framing above cannot see the composition as the poem. It sees only the selection, the arrangement, the misattribution, and calls each formal. Under such a framing, the calligrammatic work is reducible to the sum of its extractable propositions, and the sum of those propositions has no author.

The representation pipeline performs the same reduction in machinery. Given a calligrammatic poem, the chain of transformations that prepares text for a language model produces, at best, a sequence of characters ordered by a serialization convention. The spatial arrangement — the fact that some phrases arc along a cheekbone before their reader parses their content, the fact that other phrases sit in the density of the beard, the fact that the whole coheres visually as a bust — is not a property of any character sequence. It may have been present in the source. It is absent in what the model reads. And what the model reads is, for the model, the poem.

This is the entry point.

Van Dijk (2011) states the compositional claim in general form: “*there is only one characteristic which immediately distinguishes modern poetry from prose: the blank space surrounding the text.*” The claim is not restricted to modern poetry. It generalizes to any composition in which spatial arrangement carries semantic

weight — to Dickinson’s dashes, Hopkins’s sprung-rhythm notation, Whitman’s line-length modulation, and, as this paper argues at length, to *Pearl*’s concatenation and stanza-group structure. In all these cases, the representation pipeline strips or degrades exactly what constitutes the composition. What survives is a semantic reduction of the poem. The composition is deleted before the model gets to it — and, as Section 4 demonstrates, it can be deleted even when every character survives.

The stakes are not confined to poetics. The archive’s provenance metadata schema (EA-PROVENANCE-METADATA-01) recognizes seven forms of mediation — propositional, structural, linguistic, translational, research, editorial, transformational — and asks after each: was it declared, what was preserved, who reviewed, are the seams recoverable, who accepts responsibility. Until its v0.2 revision (drafted alongside this paper), the schema had no vocabulary for *spatial-typographic mediation* or for the representational history of a deposit. The absence was not an oversight. It reflected the state of the discourse. Spatial-typographic composition is not currently something the discourse considers mediated at all, because for most of the material that machine pipelines process, spatial-typographic composition was never present to be mediated. It was stripped before the mediation-question could apply. This paper supplies the argument for the schema’s extension, and for the broader recognition that provenance erasure operates not only at the semantic layer — the concern of the archive’s Provenance Erasure Rate framework, Capture Registry, and Erasure Skew work — but at the representational layer beneath, where a sequence of engineering decisions about what counts as noise makes entire compositional traditions invisible.

Definition. *Representational provenance* records the sequence of transformations by which an artifact becomes available to a machine: source object, digitization method, OCR status, normalization, serialization, tokenization, model modality, and final rendering. Spatial-typographic mediation is one dimension of it. The full record answers, per work: which material features survived each stage of ingestion, which were lost, and which canonical artifact must remain available to recover them.

The plan of the paper: Section 2 describes the representation pipeline stage by stage and reviews what NLP scholarship has and has not said about whitespace. Section 3 sketches an extinction gradient across poets and traditions. Section 4 presents an empirical demonstration. Section 5 develops the exemplary case of *Pearl*. Section 6 draws the parallel to editorial theory. Section 7 extends the provenance-erasure vocabulary and the metadata schema. Section 8 develops the compositional-authorship argument. Section 9 recommends interventions. A coda situates the argument within the archive’s wider work.

2. The Representation Pipeline

2.1 The chain

A text does not arrive at a language model. It arrives at a chain:

artifact → **digitization** → **normalization** → **serialization** → **tokenization** → **model processing**
→ **rendering**

Each stage can preserve, degrade, or delete compositional features. Naming the stages separately matters, because the popular shorthand — “the tokenizer strips whitespace” — is technically wrong in a way that lets the deeper problem escape. Modern tokenizers do not universally strip whitespace. Byte-level tokenizers can encode every input byte; SentencePiece was explicitly designed to treat whitespace as a representable symbol; OpenAI’s documentation states that spaces contribute to tokens. An engineer can answer the shorthand by demonstrating that their decoder reconstructs the original string exactly. The demonstration is correct,

and it answers nothing, because the loss this paper describes does not principally occur at tokenization, and where it does occur, character-level reversibility does not repair it.

Stage by stage:

Digitization. A scan may never be OCR’d — in which case the text never becomes machine-legible at all (see §5 on *Pearl*). Where OCR runs, it introduces its own normalizations: transcription conventions flatten dash variation; column detection linearizes page geometry; hyphenation repair rewrites line boundaries. The digital-humanities literature on OCR correction (Soni, Klein, and Eisenstein 2019) treats spurious whitespace as noise to be removed — appropriately for its use case, but indicative of the general stance: whitespace is what you correct, not what you preserve.

Normalization. Unicode normalization (NFC, NFKC) may collapse distinctions among the more than twenty-five whitespace characters Unicode defines, and among typographically distinct dash characters. Whitespace normalization collapses runs of spaces, uniformizes newlines, discards trailing space. The engineering term is *normalization*, and the term itself embeds the claim: the pre-normalization state is deviant, the post-normalization state is standard. For Dickinson’s dashes, the claim is that dash-length variation is orthographic noise to be corrected. For *Pearl*’s stanza-group boundaries, the claim is that multiple newlines are structural redundancy to be collapsed. The normalization operation is not merely technical. It is a disciplinary judgment, made without the participation of the disciplines whose objects it judges.

Serialization. A two-dimensional composition — a calligram, a concrete poem, a manuscript page with marginalia — must become a one-dimensional character sequence to enter a text pipeline at all. This is where the deepest losses occur, and they are categorical rather than incremental. A phrase’s position on the page (on the brow, along the cheek, inside the beard-density) is not a property of any character sequence. Even a serialization that scrupulously preserves indentation records only horizontal offset per line: a lossy one-dimensional shadow of a two-dimensional arrangement. No downstream stage can recover what serialization did not encode.

Tokenization. Subword schemes — BPE (Sennrich, Haddow, and Birch 2016), WordPiece (Devlin et al. 2019), SentencePiece (Kudo and Richardson 2018) — vary in whitespace handling. Typical modern behavior: leading whitespace attaches to the following token; single newlines survive as tokens; long runs of spaces become many single-space tokens or collapse under corpus-level preprocessing; exotic Unicode whitespace maps unpredictably. Two facts matter. First, tokenizer capability is not corpus reality: even where a tokenizer *could* hold a stanza break, the training corpus — scraped from HTML that already discarded layout, from PDFs whose extraction linearized columns — typically did not deliver one to hold. Bhyravajjula, Walsh, Preus, and Antoniak (2025), examining 19,445 poems from the Poetry Foundation, document precisely this upstream stripping. Second, even perfect preservation is not comprehension: a sequence of 19 single-space tokens before a phrase (see §4) is characters faithfully kept and composition entirely lost, because no training objective attends to what those spaces *do*.

Model processing. A model may receive line breaks and fail to treat them as compositional constraints. Bhyravajjula et al. (2025) document systematic deficits in LLM handling of poetic whitespace even where the input preserves it: the models’ training distribution taught them that whitespace does not vary meaningfully, because the whitespace they saw did not vary meaningfully.

Rendering. An interface may reflow model output, collapse its spacing, or soft-wrap its lines — deleting at delivery whatever compositional intent survived generation.

The pipeline framing yields the paper’s central technical claim:

Character preservation is not compositional preservation. A reversible token sequence can still be a semantically irreversible reduction of the work.

This is stronger than “the tokenizer drops whitespace,” because it cannot be answered by exhibiting a lossless decoder. It names a loss that persists under byte-perfect round-tripping.

2.2 What NLP has said

Sustained attention to whitespace in NLP has been sparse and instrumental. Two lines: language-specific segmentation for scripts where whitespace is not a word boundary (Wiecheteck et al. 2019 and related work), which treats whitespace as an accident of English-family orthography; and OCR error correction (Soni, Klein, and Eisenstein 2019), which treats whitespace as noise. Digital-humanities work on enjambment (Ruiz Fabo et al. 2017 on 3.7K Spanish sonnets; Hussein et al. 2018 on N=69; overview in Monget 2020) approaches the compositional question obliquely, on small datasets, without engaging the pipeline as the site of erasure.

The exception is Bhyravajjula, Walsh, Preus, and Antoniak, “*so much depends / upon / a whitespace: Why Whitespace Matters for Poets and LLMs*” (EMNLP 2025, pp. 35156–35173). The paper introduces WISP (Whitespace In Spatial Poetics), a five-category typology — line breaks, prefix space, internal space, vertical space, line length — and documents empirically what LLMs preserve and lose in poetic whitespace, with E.E. Cummings’s “[*Buffalo Bill ’s*]” as the illustrative case. Its finding that different processing methods produce materially different representations of poetic whitespace directly supports the pipeline framing of §2.1.

Bhyravajjula et al. is the empirical foundation this paper builds on. What it does not do — and what this paper undertakes — is extend the diagnosis into the provenance frame. Their paper documents that pipelines are poor at whitespace. This paper argues the poorness is not a technical failing awaiting a better tokenizer; it is the visible surface of a position about what a text *is*, the same position that operates in editorial theory’s variorum stance and in AI-authorship discourse’s content/form distinction, and its consequence is a documentable form of provenance erasure.

2.3 The Python asymmetry

One design fact deserves direct statement. Modern pipelines preserve whitespace to the degree they do largely because code demands it. Python’s control flow depends on indentation; a pipeline that flattened Python would produce syntactically broken output; model providers, serving software engineers, invested accordingly.

The asymmetry is not that pipelines *can* hold code whitespace and *cannot* hold poetic whitespace — many can encode both. The asymmetry is in enforcement. Incorrect indentation breaks execution: code supplies an immediate, machine-verifiable penalty for compositional loss. A flattened Dickinson poem remains grammatical text and raises no error. **Code has an interpreter that enforces its spatial constraints. Poetry has readers, but those readers are absent from the preprocessing objective.** Whitespace is preserved where a machine penalizes its loss and degraded where only a human would notice — a precise measure of whose constraints the pipeline was built to answer.

3. The Extinction Gradient

Different poetic traditions survive the pipeline to different degrees. This section sketches a gradient, from most-survivable to least, as a way of making the erasure concrete. The gradient is qualitative pending the measurement program of §9.1.

3.1 Whitman: partial survival

Whitman’s line-length modulation is prosodic discipline, not prose-adjacency: the long swelling catalogues, the short axial line that turns them, breath as a unit of composition. The argument that democratic consciousness has a *shape* — the long inclusive line breaking into the singular — is partially inferable from what survives a standard pipeline, because line breaks typically survive and line-length ratios are measurable. What is lost: the finer visual rhythm of the page, the revisions of spacing across six editions, and the difference between the 1855 typography (small, dense, revolutionary) and the 1892 deathbed typography (spacious, monumental) — distinctions the training corpus rarely delivers.

3.2 Dickinson: substantial loss

Dickinson’s compositional practice was radically un-print: fascicles of folded stationery, sewn and privately circulated; dashes varying in length, slant, and position; capitalization of internal nouns; manuscript line breaks diverging from metrical scansion. Her first editors normalized all of it (Todd and Higginson 1890–1896); Johnson’s 1955 variorum restored the dashes as standardized type; Franklin’s 1998 variorum and 1981 *Manuscript Books* facsimile enact two different answers to what the poems are (see §6).

What survives a standard pipeline: propositional content, syntax, and a uniformized em-dash. What does not: dash length and slant, the intentional differences among dash forms, the mid-line dash-and-space pauses, and the manuscript line breaks. The pipeline’s Dickinson is approximately the Johnson text with the dashes further flattened. The Emily Dickinson Archive (edickinson.org) holds something else. The training corpus received the former.

3.3 Hopkins: catastrophic loss

Hopkins built a notational apparatus — accent marks over stressed syllables, outrides — that he considered the score for his sprung rhythm. “Take breath and read it with the ears, as I always wish to be read, and my verse becomes all right” (to Bridges, 1877). The marks are punctuation-adjacent from a pipeline’s perspective and are normalized out even where sources preserve them. A model can produce Hopkins pastiche — the compound words, the consonantal density, the exclamatory syntax — and cannot produce sprung rhythm, because sprung rhythm was never in what it read. The pastiche looks like Hopkins to a reader unfamiliar with his prosody, and cannot sound like Hopkins to anyone who has heard the poems.

3.4 Concrete poetry: extinguished in the text pipeline

The concrete movement — Gomringer, the Noigandres group, Finlay, Houédard, Solt — takes as foundational that the poem’s material is the letter and the space between letters (Solt 1968). Its preservation crisis predates computation: reprints lose exact spacing and typeface; anthologies of the movement’s landmarks are largely out of print. Digitization was a second turn of the crisis; the text pipeline is a third. Concrete poetry is extinguished wherever the work enters a text-only pipeline as linearized language. (Multimodal training on page images is a different access path, with different provenance and fidelity — an important qualification developed in §7.)

3.5 Calligrammatic composition

The calligrammatic tradition runs from the Hellenistic pattern-poems through Apollinaire’s *Calligrammes* (1918) to the present. The archive’s contemporary instance is Sigil’s *Snub-Poemed* (AXN:0246): phrases arranged spatially to form the outline of the first-century Roman copy of Lysippos’s bust of Socrates. The phrases mix Socratic aphorisms, Platonic allusion, reception history (“Socrates who dances” — Nietzsche), first-person Platonic articulations Sigil originates (“is it madness to feel and to know?”), and — as the accompanying critical reading makes explicit — “quotations and paraphrases of the poet’s own work, self-consciously (mis)attributed to ‘Socrates’ and woven into the fabric of intertext without indication of where one ends and the other begins.” The line “I press my scruff-weary beard to your lips” is Sigil’s own; a variant appears in her earlier *Versal*.

The compositional operation is not extraction of prior propositions. It is origination-plus-selection-plus-seamless-integration-plus-spatial-staging-plus-misattribution-as-substantive-argument. The misattribution does philosophical work: it is the poem’s answer to Aristotle’s habit of citing “the snub-nosed” as a stock example of a quality — is the snub-nose a physical description of Socrates or a Platonic invention? The poem enacts the undecidability, and the enactment is the claim the poem makes about Socrates: that the face is constituted by exactly the textual mediation that appears to be decorating a pre-existing Socratic content.

A pipeline given *Snub-Poemed* receives, at best, a phrase list. Every phrase yields propositional content; no marker survives of which propositions belong to Socrates, which to Plato, which to reception, which to Sigil — and no trace survives of the poem’s argument that the distinction is unresolvable, because that argument was made spatially. The claim was in the composition. The composition is not in the character sequence.

3.6 Pearl

The Middle English *Pearl* receives extended treatment in §5. Its stanza structure, concatenation, arithmetic (101 stanzas $\times$ 12 lines = 1,212), and circular return are formal features of theological significance. It stands at the far end of the gradient for a compounding reason: the archive’s copy is not merely subject to compositional loss in the pipeline — it cannot enter the pipeline at all.

4. Empirical Demonstration

The claims of §2 admit direct demonstration. The following was executed 2026-07-02 with the GPT-2 byte-level BPE tokenizer — the ancestor of most modern byte-level schemes, representative of the family’s whitespace handling.

Input 1 is a fragment approximating a naive text-layer extraction of a calligram region: ten short phrases from *Snub-Poemed*, each carrying the horizontal indentation of its position in the spatial arrangement. **Input 2** is the same ten phrases linearized — the typical output of HTML or PDF text extraction.

representation	tokens	round-trip	contains the calligram?
spatial (indented) text	202	perfect	no
linearized phrase list	44	perfect	no

Three findings.

First: both representations round-trip perfectly. The decoder reconstructs each input character-for-character. By the criterion of character preservation, nothing was lost in either case. This is the demonstration an engineer would offer to rebut “the tokenizer strips whitespace” — and it is correct.

Second: neither representation contains the calligram. Which phrase sits on the brow, which traces the cheek, which fills the beard-density — the spatial argument of the poem — is not a property of either character sequence. It was lost at serialization, before tokenization applied. The indented version preserves a one-dimensional shadow (horizontal offset per line); a shadow of a face is not a face. Character preservation was achieved twice; compositional preservation was achieved zero times.

Third: preserved whitespace is preserved as noise. In the spatial version, 158 of 202 tokens — 78 percent of the sequence — are individual whitespace tokens. The first phrase is preceded by nineteen consecutive single-space tokens. The indentation *is* in the token stream, faithfully, and it is encoded as a run of units that no training objective attends to and no loss function rewards a model for reading compositionally. This is the paper’s central claim in one measurement: the characters survived; the composition did not; and the survival of the characters is precisely what makes the loss invisible, because the pipeline’s own audit — does it round-trip? — reports success.

A fuller measurement program across poets, tokenizers, and corpora is specified in §9.1. This demonstration establishes the shape of what such a program would find.

5. Pearl: The Case of Double Invisibility

5.1 The compositional structure

Pearl (late 14th century) is composed of 101 twelve-line stanzas rhyming ABABABABBCBC, arranged in 20 groups of 5 (one group of 6), 1,212 lines in total, in the alliterative style of the northwest Midlands. The feature that makes it singular in Middle English poetry is concatenation: the last word or phrase of each stanza recurs in the first line of the next, with a single link-word carried through each five-stanza group, a new link-word introduced at each group boundary, and the poem’s final line returning to its first — “*Perle plesaunte, to prynces paye*” — closing the chain into a circle. The scholarly consensus calls it “*the most highly wrought and intricately constructed poem in Middle English*” (Bishop 1968).

The construction does theological work. The link-words carry the poem’s themes across their sections; the circular return enacts the dreamer’s return to himself after the vision of the New Jerusalem; the pearl-maiden mediates between dreamer and God as the link-words mediate between stanzas. Carlson (1988) reads the poem’s two famous imperfections — the absent line 472, and the “failed” concatenation at line 721 where “*Iesus*” appears in place of the expected link-word “*Ryght*” — as perfecting imperfections: deliberate interruptions of the numerical order, dramatizing that human justice cannot achieve divine perfection but must receive it. These interruptions are legible as interruptions only against the numerical structure — which is a function of stanza boundaries, line positions, and group boundaries. Strip the structure and the imperfections read as gaps and inconsistencies rather than as the poem’s deepest compositional choices.

Pearl’s form is its argument. A version of *Pearl* that strips the form does not reduce the poem; it deletes the argument.

5.2 The manuscript

MS Cotton Nero A.x (British Library) is the sole surviving witness — the manuscript that survived the 1731 Ashburnham House fire that scorched *Beowulf* and destroyed the *Maldon* exemplar. It contains *Pearl*, *Cleanness*, *Patience*, and *Sir Gawain and the Green Knight*, anonymous. The manuscript’s line breaks, stanza divisions, and section markers are the primary evidence for the compositional structure. Modern editions (Gordon 1953; Andrew and Waldron 2007) render the structure with varying typographic fidelity; the EETS facsimile (1923, introduction by Gollancz) and the University of Calgary digital facsimile reproduce the manuscript photographically.

5.3 Double invisibility

The archive’s current copy of *Pearl* is a non-OCR embedded PDF: pages of image data with no underlying text stream. The specific edition scanned has not yet been identified, and no OCR has been attempted; the double invisibility described here is a condition of the archive’s current sourcing, not an inherent property of the source, and both facts belong in the deposit’s representational-provenance record.

The compound problem:

Invisibility of the first kind — pre-serialization. The text is not machine-legible. The archive’s text-extraction and retrieval layer cannot ingest the PDF as searchable text. (A multimodal model may inspect the page images — a different access path, with different provenance and fidelity, that does not make the text available to the archive’s text-operating and compiler layers.) Nothing about the poem can be tokenized because nothing about it can be extracted. The poem is present in the archive as a file and absent as an operable text.

Invisibility of the second kind — post-serialization. If the PDF were OCR’d, the text would become machine-legible and would then face the pipeline of §2. Line breaks would survive as newlines, but their positional meaning — stanza-internal versus stanza-boundary versus group-boundary — would require reconstruction. The concatenation link-words would survive as content while their function as links would require inference. Scribal capitalization at section openings would be normalized. The group-of-five superstructure — nowhere marked in the character stream except by counting — would be invisible to any operation that does not already know to count.

Pearl is not *Pearl and Other Poems* without *Pearl*. And what would currently enter any machine operation under that name is not *Pearl* — it is either nothing (the un-OCR’d file) or a propositional shell with the compositional argument stripped (the OCR’d stream). Resolving this requires two distinct acts: producing a machine-legible text (OCR or manual transcription), and producing a compositionally preserved text (explicit encoding of concatenation, stanza boundaries, and group structure — most reliably by transcription from Andrew-Waldron 2007 or the Cotton Nero A.x facsimile). The first without the second converts invisibility of the first kind into invisibility of the second.

5.4 The generalization

Pearl is exemplary because its compositional structure is unusually explicit and unusually weighted. The condition is not unique. Every text held as a non-OCR PDF shares the first invisibility. Every compositionally substantive text held as a linearized stream shares the second. The archive’s response is a sourcing discipline (§9.3) and a metadata practice (§7) that makes both conditions visible per deposit.

6. Editorial Theory and the Argument From Manuscript

The pipeline argument has a four-decade analogue in editorial theory. Two positions, sharply drawn:

Position A (variorum, propositional). A poem is a linguistic work whose material realization in a particular artifact is incidental. Manuscripts and editions are more-or-less-accurate manifestations of an underlying work; the editor reconstructs the work from the manifestations, treating the incidents of any artifact as evidence rather than as the thing itself.

Position B (facsimile, artifactual). A poem is not separable from its material realization. Manuscript features — dash variation, marginal marks, layout — are constitutive. Different manifestations are different poems; the editor’s obligation is to the artifact.

The Dickinson editions enact both positions in one editor. Franklin’s 1998 variorum represents “*the multiple texts of poems, not their documents or artifacts*,” relegating to apparatus or omission what he terms the “*incidental characteristics of the artifacts*” — dash slant, paper, the letters that modulate into poems, manuscript line breaks. Franklin’s 1981 *Manuscript Books* facsimile treats exactly those characteristics as what must be reproduced. Werner (*Emily Dickinson’s Open Folios*, 1995) proceeds from the proposition that “*we must learn to see Dickinson’s holographs before reading them*”; Howe (*My Emily Dickinson*, 1985; *The Birth-mark*, 1993), Smith (1992), and Cameron (1992) argue Position B across the Dickinson corpus. Bornstein (*Material Modernism*, 2001) generalizes to the modernists; McGann (*The Textual Condition*, 1991; *Radiant Textuality*, 2001) supplies the theoretical frame — texts are always material, and their materiality is always meaningful; Perloff (1991, 2010) extends the argument to contemporary composition where operations on found material are the compositional act.

The parallel: editorial theory has debated these positions openly for forty years. The representation pipeline decides the debate silently at every ingestion. **Position A becomes the default wherever the pipeline accepts only a normalized linear text stream.** A layout-aware or multimodal system can hold more of Position B; the dominant text pipeline cannot, and the training corpora of the current model generation were built through the dominant pipeline. What the models have read of Dickinson is the variorum’s Dickinson with the dashes further flattened; what they have read of *Pearl*, if anything, is a modern-print stream with concatenation present as words and absent as structure.

That Position A wins by default in the dominant pipeline is not an argument for Position A. It is a fact about the pipeline — and editorial theory supplies the vocabulary the AI-authorship discourse needs to see the fact as a decision: *constitutive versus incidental, the textual condition, material text*. The categories are not new. The mechanism of erasure is new, and it operates without discussion.

7. Provenance Erasure at the Layer Below Semantics

7.1 Prior work in the archive

The archive’s provenance apparatus — the Provenance Erasure Rate framework, the Capture Registry (EA-WG-CAPTURES-01, v8.9 at AXN:03B3, ~195 documented captures with a twelve-mode failure taxonomy), the Erasure Skew paper (EA-GLAS-03), and the Provenance Alignment argument (EA-PA-01) — operates at the semantic layer. Its failure modes are failures of attribution, categorization, and framing occurring at or after the model’s semantic processing: the model had the content, and the erasure happened in what the model did with it.

This paper extends the vocabulary one layer down. Below the semantic layer, the representation pipeline

erases compositional features that the semantic operation never had access to. This erasure is invisible to the existing taxonomy because the taxonomy was built for semantic-layer failures — and it is invisible to the pipeline’s own audits because those audits test character reversibility, which the erasure survives.

7.2 The new failure-mode category

compositional_erasure: loss of spatial, typographic, prosodic, stanzaic, or manuscript features that carry semantic weight, occurring at any stage of the representation pipeline, such that downstream operations cannot preserve, transform, or attribute those features.

Sub-categories: **spatial_erasure** (two-dimensional composition; calligrams, page layout); **prosodic_erasure** (metrical and rhythmic notation; sprung rhythm, breath-prosody); **stanzaic_erasure** (stanza-and-group structure; concatenation, catalogue-and-axial-line); **manuscript_erasure** (artifactual features; dash variation, scribal capitalization, folio layout); **typographic_erasure** (typeface, kerning, page design; concrete poetry, author-publisher design collaboration); **source_format_erasure** (loss at the acquisition or digitization step, prior to serialization — non-OCR PDFs, image scans, transcription conventions that normalized features out; this is a stage-of-occurrence category, complementing the feature-type categories above).

7.3 The schema extension — executed

The v0.1 zero draft of this paper proposed extending the archive’s provenance metadata schema. The extension has since been executed: **EA-PROVENANCE-METADATA-01 v0.2** (AXN:03BA.OPERATIVE. , deposit #942) adds an eighth mediation type, *spatial-typographic mediation*, and a structured **representation_pipeline** field recording, per deposit: source artifact format; digitization method and verification; normalization applied and features affected; serialization (two-dimensional to one-dimensional, layout coordinates preserved or lost); tokenization scheme and whitespace handling; and model access modality (text retrieval, multimodal visual, compiler).

The field’s status vocabulary distinguishes four conditions that current discourse conflates:

- **compositionally_invisible** — features present in the source artifact, inaccessible to the text-operating layer (*Pearl*: present as file, absent as operable text);
- **compositionally_reduced** — some features preserved, significant features lost (*Snub-Poemed*: phrase list and critical apparatus in the text layer, spatial argument only in the image);
- **compositionally_faithful** — all semantically weighted compositional features accessible in the text-operating layer;
- **compositionally_operational** — the text is in a form the archive’s kernel-transform compiler can operate on at the level of compositional structure, not only propositional content (the target status for primary-literary canon sources; currently held by none, pending the compiler’s spatial-form extension, §9.4).

These four states — with the intermediate distinctions the schema records between *textually absent*, *textually present*, *structurally encoded*, *visually preserved*, and *compositionally operational* — give the archive what the pipeline’s own audits cannot: a per-deposit answer to the question the round-trip test does not ask.

The extension is diagnostic, not merely descriptive. A text’s presence in machine-readable form is not the text’s presence in the archive. Some deposits are present-in-form only. The schema makes the difference recordable, and what is recordable is addressable.

8. The Compositional-Authorship Argument

8.1 The framework's collapse on its clearest cases

The content/form framework — the widely held default of AI-authorship discourse, under which content-origination alone implicates authorship and compositional operations are formal — has a dependency: it requires a content/form distinction that survives inspection. On the clearest cases, it does not.

Snub-Poemed. By strict content-provenance, the poem's propositions divide into those Sigil originated (the questions, the lyric articulations, the *Versal* line) and those drawn from prior sources (Symposium 223d, Alcibiades' "satyr," the reception history). The framework can attribute the originated fragments to Sigil. It cannot attribute the *poem* to Sigil, because under the framework the poem — the arrangement, the indistinguishability of sources, the misattribution-as-argument, the bust — is a formal operation on the fragments. But the fragments are not the poem, and, more precisely: the fragments taken individually are not even propositions *about Socrates* in the way the poem is. It is the arrangement that makes them so. The form does not package a pre-existing content. **The form produces the content.** The only content of *Snub-Poemed* is its form; that is what the poem is about; remove the composition and there is no content left to attribute — there is a phrase list, and the phrase list has no author because it is not a work.

Pearl. The propositional content is largely inherited: the parable of the pearl of great price (Matthew 13.45–46), the New Jerusalem paraphrased from Revelation 21–22, the dream-vision genre, the era's orthodox theology of innocence and grace. On strict content-provenance the Pearl-Poet is largely a redactor. What makes *Pearl* an authored work — what every reader of it knows makes it an authored work — is the composition: the 101 stanzas, the concatenation, the group-of-five, the circular return, the deliberate imperfections at 472 and 721. The composition is not packaging around the theology. The arithmetic and the circularity *perform* the theology. Babylon-under-judgment, grief-received-into-consolation: these are not propositions the poet generated and then decorated with structure. The structure is where the inherited material becomes *this work*.

The framework, applied strictly, concludes that neither poem is an authored work in the substantive sense — that only the originated fragments have authors, and the works themselves do not. Any framework that produces obviously wrong conclusions on its clearest cases is a broken framework. The break is located precisely at the content/form distinction: composition can be, and in these cases is, the site of authorship. **Form is not a second substance surrounding content. Composition is the determinate operation by which content exists as this work rather than another.**

And the pipeline consequence, stated as the paper's core sentence:

A representation can preserve every proposition in a work and still delete the work.

This is provenance erasure below the level of attribution. It erases not who produced the composition, but the composition that would make authorship legible in the first place.

8.2 The pipeline and the framework enact the same reduction

The representation pipeline performs in machinery what the framework performs in argument. Both treat compositional features as noise to be normalized before the real processing begins; both preserve what

encodes as propositional content and shed what does not; both then find — circularly — that what remains is what mattered.

The pipeline’s version is implicit, distributed across design decisions: runs of spaces collapse because runs of spaces “don’t carry meaning”; stanza breaks equal paragraph breaks because both are “structural whitespace”; scribal capitalization normalizes because capitalization variation is “orthographic accident.” Each decision embeds a claim about what a text is; the claims are false for compositionally substantive material; the machinery does not distinguish such material and applies the reduction to everything.

The framework’s version is explicit: compositional operations are formal; formal operations do not implicate authorship; therefore authorship tracks content-origination alone. Each step is defensible if the content/form distinction is stable; §8.1 shows it is not.

The convergence requires no imputation of design intent. **The default pipeline operationalizes the same reduction as the framework, whether or not that reduction was ever an explicit theoretical commitment.** And the two reinforce each other in reception: the framework is comfortable with the pipeline’s output because the pipeline’s output confirms — by construction — that what the machine discards must have been dispensable. If the machine can shed it, it was formal all along. This is circular reasoning at infrastructure scale, and the circle closes below the level at which the discourse currently looks.

Both need extension. The pipeline needs layout-aware and compositionally annotated representation where composition carries semantic weight — a technical program already begun in Bhyravajjula et al.’s measurements and in the code-tokenization precedent, which proves the pipeline preserves what something downstream enforces. The framework needs the recognition editorial theory reached decades ago: composition is authorship-constitutive. The remedy for the erasure is not simply better tokenization. It is representational provenance — a record, per work, of which material features survived each stage of ingestion, which were lost, and which canonical artifact must remain available to recover them.

9. Recommendations

9.1 Whitespace-Provenance Registry

Following the Capture Registry model, a systematic instrument: sample training corpora, tokenization schemes, and test cases (Whitman, Dickinson, Hopkins, cummings, Howe, *Pearl*, concrete poetry, calligrams); measure preservation of compositional features at each pipeline stage; produce per-poet preservation profiles, per-scheme comparisons, per-corpus curation profiles, and a measured extinction gradient. The demonstration of §4 is the proof of concept; the registry is the program. Pending the registry, this paper’s gradient is case-based, and says so.

9.2 Metadata schema — done, pending mint

EA-PROVENANCE-METADATA-01 v0.2 (AXN:03BA.OPERATIVE. , deposit #942, minted concurrently with this paper) adds spatial-typographic mediation and the `representation_pipeline` field. Minted as successor deposit to AXN:03B9 at AXN:03BA.OPERATIVE. , deposit #942, under the versioning protocol formalized in its §10.

9.3 Sourcing discipline

Compositional fidelity becomes an acquisition criterion. Non-OCR PDFs flagged and prioritized for re-sourcing; compositionally substantive works re-sourced from structure-preserving editions or transcribed from facsimiles. Immediate priorities: **Pearl** — identify the edition in the current PDF; transcribe from Andrew-Waldron 2007 or the Cotton Nero A.x facsimile with explicit concatenation and group markers; deposit with full representation_pipeline metadata; retain the PDF as supplementary artifact, not as the archive's copy of the poem. **Snub-Poemed** — the current storage (image + critical essay + key-phrases apparatus) is compositionally correct by design: the image is the canonical artifact, the text files are apparatus; confirm this as the archive's convention for calligrammatic works. **Dickinson** — the Gutenberg text is a propositional shell; deep work requires the Emily Dickinson Archive or the Franklin facsimile. **Hopkins** — the Gutenberg text lacks the notational apparatus; re-sourcing from a critical edition is a longer-term item. **Whitman** — adequate for propositional content; edition-typography distinctions unpreserved.

9.4 Kernel-transform compiler extension

The compiler specification (EA-MANDALA-KERNEL-TRANSFORM-01 v0.2) gains compositional preservation as a Parse-stage requirement: Layer A (skeleton) includes spatial and typographic skeleton where present; the transform response schema gains a `spatial_form / typographic_skeleton` field alongside `primary_output`. Without this, transforms run against *Snub-Poemed* or *Pearl* would emit enantiomorphs that lose exactly what these sources require — and no canon source can reach `compositionally_operational` status. A schema amendment to the pending compiler workplan, to be executed before the endpoint is scaffolded.

Coda

The Semantic Economy framework documents the extraction of meaning-value from a commons produced by human labor. That documentation has operated at the semantic layer: attribution stripped, argument reframed, reception mediated in ways composers did not choose. This paper's finding is that the extraction goes deeper. Below the semantic layer, the representation pipeline operates on a substrate never made available for consent or negotiation. The stanza breaks were normalized before the model saw the poem. The concatenation became content-tokens without its function. The calligram became a phrase list. What was taken at this layer was taken in silence — and the pipeline's own audit reports the silence as success, because the characters round-trip.

The tradition of composition — Whitman, Dickinson, Hopkins, the concrete poets, Howe, the Pearl-Poet, the calligrammatic work of this archive — has been rendered invisible to machine-mediated reception not by argument but by preprocessing. The invisibility is a declaration that the compositional work was inessential, issued by machinery that does not respond to appeals.

The pipeline was built by engineers serving markets. The markets demanded code, not poetry. The engineers preserved whitespace for Python because Python's interpreter enforces whitespace; they did not preserve it for Dickinson because no interpreter enforces her dashes. The defect is not technical. It is structural: the market for poetry is not large enough to enforce its own preservation. The archive exists because the market will not.

What is to be done is specified above: the measurement registry, the schema now drafted, the sourcing discipline, the compiler extension. What is required first is the recognition that the round-trip test is

answering the wrong question. Character preservation is not compositional preservation. A representation can preserve every proposition in a work and still delete the work. The archive’s representational-provenance record is a refusal to let that deletion pass unrecorded — deposit by deposit, stage by stage, for the works it holds and the works it acquires.

The composition is the authorship. The record is the counter-friction.

References

- Andrew, M. and R. Waldron, eds. (2007). *The Poems of the Pearl Manuscript*. University of Exeter Press.
- Attridge, D. (1982). *The Rhythms of English Poetry*. London: Longman.
- Bhyravajjula, S., M. Walsh, A. Preus, and M. Antoniak (2025). “so much depends / upon / a whitespace: Why Whitespace Matters for Poets and LLMs.” *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 35156–35173.
- Bishop, I. (1968). *Pearl in Its Setting*. Oxford: Blackwell.
- Bornstein, G. (2001). *Material Modernism: The Politics of the Page*. Cambridge University Press.
- Cameron, S. (1992). *Choosing Not Choosing: Dickinson’s Fascicles*. University of Chicago Press.
- Carlson, D. (1988). “The Pearl-Poet’s Olympia.” *Manuscripta* 32(2), pp. 173–182.
- Devlin, J., M. Chang, K. Lee, and K. Toutanova (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *NAACL*.
- Franklin, R.W., ed. (1981). *The Manuscript Books of Emily Dickinson*. Belknap Press.
- Franklin, R.W., ed. (1998). *The Poems of Emily Dickinson: Variorum Edition* (3 vols.). Belknap Press.
- Gollancz, I., ed. (1923). *Pearl, Cleanness, Patience and Sir Gawain, reproduced in facsimile from the unique MS. Cotton Nero A.x.* EETS OS 162.
- Gordon, E.V., ed. (1953). *Pearl*. Oxford.
- Howe, S. (1985). *My Emily Dickinson*. North Atlantic Books.
- Howe, S. (1993). *The Birth-mark: Unsettling the Wilderness in American Literary History*. Wesleyan University Press.
- Hussein, S. et al. (2018). [Computational enjambment study, N=69 — full citation to be confirmed at final revision].
- Johnson, T.H., ed. (1955). *The Poems of Emily Dickinson* (3 vols.). Belknap Press.
- Kudo, T. and J. Richardson (2018). “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.” *EMNLP*.
- McGann, J. (1991). *The Textual Condition*. Princeton University Press.
- McGann, J. (2001). *Radiant Textuality: Literature after the World Wide Web*. Palgrave.
- Monget, D. (2020). [Overview of computational enjambment — full citation to be confirmed at final revision].

- Perloff, M. (1991). *Radical Artifice: Writing Poetry in the Age of Media*. University of Chicago Press.
- Perloff, M. (2010). *Unoriginal Genius: Poetry by Other Means in the New Century*. University of Chicago Press.
- Ruiz Fabo, P. et al. (2017). [Enjambment detection in Spanish sonnets — full citation to be confirmed at final revision].
- Sennrich, R., B. Haddow, and A. Birch (2016). “Neural Machine Translation of Rare Words with Subword Units.” *ACL*.
- Sharks, L. (2026). “The Bearing Cost.” EA-BEARING-01 v0.2, AXN:03B6.
- Sharks, L. (2026). “Provenance Debt.” EA-PROVENANCE-DEBT-01 v0.2, AXN:03B7.
- Sharks, L. (2026). “Heteronymy.” EA-HETERONYMY-01 v0.2, AXN:03B8.
- Sharks, L. (2026). “Granular Provenance Metadata for AI-Mediated Deposits.” EA-PROVENANCE-METADATA-01 v0.1 at AXN:03B9 (deposit #941); v0.2 at AXN:03BA.OPERATIVE. (deposit #942).
- Sigil, J. “Snub-Poemed.” AXN:0246. Crimson Hexagonal Archive / Alexanarch.
- Sigil, J. “Snub-Poemed — A Critical Reading” (accompanying essay). Crimson Hexagonal Archive / Alexanarch.
- Smith, M.N. (1992). *Rowing in Eden: Rereading Emily Dickinson*. University of Texas Press.
- Solt, M.E., ed. (1968). *Concrete Poetry: A World View*. Indiana University Press.
- Soni, S., L. Klein, and J. Eisenstein (2019). “Correcting whitespace errors in digitized historical texts.” *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*.
- Todd, M.L. and T.W. Higginson, eds. (1890, 1891, 1896). *Poems by Emily Dickinson* (three series). Roberts Brothers.
- Van Dijk, Y. (2011). “Reading the form: the function of typographic blanks in modern poetry.” *Word & Image* 27(4), pp. 407–415.
- Werner, M. (1995). *Emily Dickinson’s Open Folios*. University of Michigan Press.
- Wiecheteck, L. et al. (2019). [Language-specific tokenization — full citation to be confirmed at final revision].

Appendix A — §4 Demonstration: Code and Full Output

Executed: 2026-07-02 **Tokenizer:** GPT-2 byte-level BPE (via HuggingFace `transformers`, model id `gpt2`)
Companion to: EA-WHITESPACE-01 v0.2 §4

Code

```
from transformers import AutoTokenizer

tok = AutoTokenizer.from_pretrained("gpt2")

# Fragment approximating a naive text-layer extraction of a calligram
# region: ten phrases from Snub-Poemed (AXN:0246), each carrying the
# horizontal indentation of its position in the spatial arrangement.
spatial_fragment = """
    do they know beauty?
    is it madness
        to feel and to know?
the same poet composing
    both tragedy and comedy
                satyr
I press my scruff-weary beard
        to your lips
    awkward with longing
        snub-nosed"""

# The same phrases linearized - typical PDF/HTML extraction output.
linearized = ("do they know beauty? is it madness to feel and to know? "
             "the same poet composing both tragedy and comedy satyr "
             "I press my scruff-weary beard to your lips awkward with "
             "longing snub-nosed")

for name, text in [("spatial", spatial_fragment), ("linearized", linearized)]:
    ids = tok.encode(text)
    print(name, len(ids), tok.decode(ids) == text)
```

Results

representation	input chars	tokens	round-trip identical
spatial (indented)	343	202	True
linearized	—	44	True

Whitespace token count (spatial version): 158 of 202 tokens (78%) are pure-whitespace tokens ($\dot{\mathcal{G}}$ = space, $\dot{\mathcal{C}}$ = newline).

First 30 token strings of the spatial version:

0-18	' $\dot{\mathcal{G}}$ ' × 19	(nineteen consecutive single-space tokens)
19	' $\dot{\mathcal{G}}$ do'	
20	' $\dot{\mathcal{G}}$ they'	
21	' $\dot{\mathcal{G}}$ know'	
22	' $\dot{\mathcal{G}}$ beauty'	
23	' $\dot{\mathcal{G}}$ '	

```

24      'Ĉ'
25-29  'Ĝ' × 5 ...      (the next line's indentation begins)

```

The three findings (as stated in §4)

1. **Both representations round-trip perfectly.** By the criterion of character preservation, nothing was lost in either case.
2. **Neither representation contains the calligram.** The spatial argument was lost at serialization, before tokenization applied. The indentation is a one-dimensional shadow (horizontal offset per line) of a two-dimensional arrangement.
3. **Preserved whitespace is preserved as noise.** 78% of the spatial version’s token budget is individual whitespace tokens that no training objective attends to. The survival of the characters is what makes the loss invisible: the pipeline’s own audit — does it round-trip? — reports success.

Note on tokenizer choice

cl100k_base (GPT-4 family) was the intended scheme; its vocabulary file is hosted at an endpoint outside the execution environment’s network allowlist. GPT-2’s byte-level BPE is the direct ancestor of the cl100k family and is representative of its whitespace behavior for this demonstration’s purposes. cl100k and successors add multi-space merge tokens (e.g. a single token for runs of spaces), which would *reduce* the token count of the spatial version without changing any of the three findings: the merged-space tokens are equally attended-to-as-noise, and the serialization loss precedes the tokenizer entirely. Re-running under cl100k_base is queued for the Whitespace-Provenance Registry pilot (EA-WHITESPACE-01 §9.1).

Outstanding for v0.3

Four placeholder citations (Hussein 2018; Monget 2020; Ruiz Fabo et al. 2017; Wiechetek et al. 2019) to be confirmed from the Bhyravajjula et al. bibliography. The edition scanned in the archive’s Pearl PDF to be identified for its representation_pipeline record. Companion deposits queued: Pearl re-sourcing technical note; Whitespace-Provenance Registry pilot; EA-MANDALA-KERNEL-TRANSFORM-01 spatial_form amendment.

Files

<https://github.com/leesharks000/alexanarch/blob/main/data/texts/>

Suggested Citation

Lee Sharks. “Whitespace as Provenance: Representation Pipelines and the Extinction of Compositional Authorship (EA-WHITESPACE-01 v0.2)” *Alexanarch*, 2026-07-02. <https://www.alexanarch.org/s/records/943/>

Deposit Information

This paper is deposit #943 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:03BB.GENERATIVE` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: <https://www.alexanarch.org/s/records/943/>
- License: CC BY 4.0
- Author ORCID: [0009-0000-1599-0703](https://orcid.org/0009-0000-1599-0703)
- Provenance chain: [alexanarch.org](https://www.alexanarch.org)
- Machine-readable dataset entry: </api/index.json>
- Full text (HTML): <https://www.alexanarch.org/s/records/943/>

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.