TECHNICAL WHITE PAPER · THE WRITING VOICES LLC

Quality Moves Into the Architecture

A closed-loop, model-tiered production architecture in which manuscript generation runs on the lowest-cost model tier and quality is enforced by deterministic stylometric instrumentation paired with frontier-tier editorial judgment.

CONTENTS

Quality Moves Into the Architecture: A Closed-Loop, Model-Tiered System for Scalable Book Production

The Writing Voices LLC · July 18, 2026

Abstract

The Writing Voices LLC operates a multi-imprint fiction production system in which distinct authorial voices — 27 active pen names across seven presses — must hold identity across long-form manuscripts, genres, and successive generations of language models. This paper reports the design and validation of a closed-loop, model-tiered production architecture in which manuscript generation runs on the lowest-cost model tier and quality is enforced by deterministic stylometric instrumentation paired with frontier-tier editorial judgment. The measured foundation is a fingerprint spine built from 74 compiled books totaling 2.01 million words: Burrows' Delta centroids over the corpus's 150 most frequent words, plus observed style bands per pen. Blind nearest-neighbor attribution on the compiled shelf reaches 77% (48/62 against a chance rate of approximately 4%). A controlled head-to-head trial — one pen, seven genres, bounded iterations — found that frontier-model generation with frontier-loop editing and lowest-tier generation with frontier-loop editing both converge to 7/7 rank-1 attribution, while lowest-tier self-editing converges in 0/7. The generator tier is negligible given the loop; the closed editorial loop, gated by code, is the primary quality mechanism. Economic implications are projected from measured behavior; dollar figures are not yet instrumented. The trial's single-pen scope awaits multi-pen replication.

1. Executive Summary

1.1 The Core Finding

The central empirical result of this work is a single sentence: **given a closed editorial loop run by a frontier-tier model against deterministic stylometric gates, the choice of generation model is negligible.** In the final head-to-head trial, drafts generated by the frontier tier (Fable) and drafts generated by the lowest tier (Haiku) both converged to perfect rank-1 voice attribution — 7/7 and 7/7 across seven genres — when loop-edited by the frontier

tier. The same lowest-tier model editing its own drafts, even when handed exact per-word repair prescriptions from a deterministic diagnostic, converged in 0/7. Quality does not live in the generator. It lives in the loop: draft, diagnose deterministically, carve with high-capability judgment, gate with code, repeat until the measurement passes.

This inverts the default economics of AI-assisted long-form production. The expensive resource is not drafting capability, which is widely available at low tiers; it is convergent editorial judgment, which the evidence shows is scarce, tier-dependent, and worth reserving. The system's validated house economy is: the cheap model writes, the frontier model loop-edits, and deterministic code gates every pass.

1.2 The Validated Production Model

The architecture rests on a measured voice layer called the stylometry spine. Every pen name has a quantitative fingerprint derived from its verified compiled books — never declared, always measured: a Delta centroid (the pen's mean z-profile over the corpus's 150 most-frequent function words, compared by Burrows' Delta) and style bands (observed ranges for sentence rhythm, fragment rate, punctuation per thousand words, dialogue ratio, word length, and standardized type-token ratio). The spine covers all 27 active pens, built from 74 compiled books and 2.01 million words, and is regenerated deterministically from the manuscripts on disk.

Two instruments run against this spine. A stylometric checker performs blind attribution — nearest pen of 27 — and band compliance, returning a graded verdict: PASS (the text's nearest centroid is its own pen), DRIFT (own pen ranks second within Δ0.05 of the winner — a range verdict, not a failure), or FAIL (the text landed on another pen's shelf). A voice diagnostic produces compact per-word repair prescriptions naming exactly which function-word habits have drifted. Diagnosis is code and therefore free at every model tier; carving on that diagnosis is top-tier judgment work.

Production then runs as a closed loop: a low-tier model drafts from a top-model brief; the diagnostic names the deviations; a frontier-tier editor carves the

manuscript; the checker gates the result; the cycle repeats within bounded iterations (four) until convergence, best version retained. Every editorial pass is bookended by the same measurement, so an edit that erodes voice is caught as a defect of the pass, deterministically.

1.3 Experimental Results

The compiled shelf establishes the baseline: blind nearest-neighbor pen attribution of 77% (48/62, chance ≈ 4%), demonstrating that the pens are measurably distinct hands. A public-domain canon overlay (31 anchor novels, 20 authors, scored in a joint feature space) validated the method itself at 13/14 blind author attribution, and showed the house's 27 pens spread wider than the 20 real authors (between-pen Δ1.01 versus between-author Δ0.88) while holding individual identity slightly looser (within-pen Δ0.73 versus within-author Δ0.50). The same overlay measured genre as a weak stylometric force: canon same-genre distance Δ0.82 versus cross-genre Δ0.89 — the hand beats the shelf.

A quarantined cross-genre lab (7 pens × 7 fixed premises × 2 model arms; 98 samples of at least 3,000 words each, drafted blind from declared author documents alone) found that stripped of the production pipeline, voice collapses: the frontier arm passed 13/49 cells, the lowest-tier arm 0/49 (one DRIFT), with each model's failures converging on a single measurable "attractor pen" — its own default hand. The conclusion: the pen voices live in the pipeline, not the prompt. Rewriting one failing pen's author document to an executable, parameterized standard moved its attribution rank in every one of seven cells (e.g., Romance 25→9, Action 20→7) but reached rank 1 in none — necessary, not sufficient.

An open-loop editor ladder over the lowest-tier drafts showed the frontier editor as the only net-positive arm (median rank 8, beating frontier one-shot generation's median of 20), mid-tier voice editing as net harmful (worse in 3 of 7 cells), and the generator's function-word watermark surviving every open-loop editing tier. The final closed-loop grid resolved the question: Fable-generates + Fable-loop-edits, 7/7 rank-1; Haiku-generates + Fable-loop-edits, 7/7 rank-1, converging in 1–4 iterations and fully erasing the generator watermark that open-loop editing could not launder; Haiku-generates + Haiku-loop-edits, 0/7. In

passing, the loop editor also repaired real structural defects in the cheap drafts — a duplicated scene, broken continuity arithmetic — value not captured by the attribution score. The trial's scope is deliberately narrow — one pen, seven genres, bounded iterations — and multi-pen replication is pending.

1.4 Economic and Operational Significance

If the generator is negligible given the loop, then the bulk token volume of book production — the drafting itself — can run at the lowest available model tier without measurable voice cost, while frontier capability is spent only where the evidence shows it is irreplaceable: the editorial carve. Deterministic instruments do the diagnostic and gating work at zero model cost per run, and the system's own history shows those instruments compounding — the diagnostic that made 14/14 closed-loop convergence possible was invented by an agent mid-run and banked as permanent repository code. This is stated plainly as a projection: the cost model follows directly from measured behavior, but dollar figures have not yet been instrumented. Operationally, the architecture is model-agnostic by construction. Each new model is profiled through the same lab — its attractor pen, its one-shot voice-holding, its closed-loop convergence speed — and slotted into the tier where it measures, so the system absorbs future models without being rebuilt.

2. The Problem

2.1 The Cost of Frontier-Model Generation

Long-form fiction is token-intensive: the house's compiled shelf alone is 2.01 million words across 74 books, and every manuscript passes through drafting, multiple editing sweeps, and recompilation. Running all of that volume through a frontier-tier model treats the most expensive available intelligence as bulk labor. The unexamined assumption behind that spend is that generation quality is where manuscript quality comes from — an assumption this paper's evidence contradicts. Frontier one-shot generation, measured on the lab's attribution

metric, produced a median rank of 20 of 27; the cheap-generation-plus-frontier-loop recipe produced rank 1 in 7 of 7 cells. Paying frontier rates for drafting buys little that the loop does not buy better. (The dollar magnitude of this inefficiency is projected, not yet instrumented.)

2.2 Inconsistent Author Voice Across Long-Form Work

A pen name is a promise to the reader: the same hand, book after book. Holding that promise across a 15,000–40,000-word manuscript — and across a shelf of them — is precisely what unaided model generation fails to do. The lab measured this directly: drafted blind from declared author documents alone, without the production pipeline, the frontier arm held its pen in only 13 of 49 cells and the lowest tier in 0 of 49. The failures were not noise; they converged on each model's own default hand (the frontier tier drifting Rance-adjacent, the lowest tier Lance-adjacent). Left alone, a model does not write in the author's voice; it writes in its own, and every pen on the roster collapses toward that single centroid. A multi-imprint operation whose entire premise is 27 distinct hands cannot ship that.

2.3 Editor-Induced Voice Drift

Editing is where voice erodes silently. A line pass that "improves" sentences tends to normalize exactly the habits that constitute a pen's identity — its signature fragments, its punctuation licenses, its rhythm. The editor ladder measured this hazard concretely: a mid-tier model given a voice-carving assignment made attribution worse than not editing at all in 3 of 7 cells (Crime dropping from rank 9 to 23, Literary from 4 to 20, Horror from 5 to 15), pushing samples onto other pens' shelves. Voice-destructive editing is not a hypothetical failure mode; it is the measured default behavior of insufficient editorial capability applied confidently. Without a deterministic before-and-after gate, this damage is invisible until a reader feels it.

2.4 The Limits of Prompt-Only Voice Control

The intuitive fix — describe the voice better in the prompt — was tested and found real but insufficient. Rebuilding one pen's author document from aspirational prose to an executable specification (locked narrator mechanism, numeric rhythm targets, per-thousand-word punctuation rates, every number sourced from the measured shelf) produced the largest single-intervention rank gain observed: every one of seven genre cells moved toward the pen, two into the top 10. But none reached rank 1 of 27, on either model arm. The executable specification closed the surface style gap almost completely — sentence means, fragment rates, and punctuation landed inside the pen's measured bands — while only partially closing the deep function-word gap on which attribution actually rides. One pen (whose identity is mechanically fixed by point of view) did transfer from its document alone, 7/7 across every genre; that is the existence proof and the standard other documents are being rewritten toward. But the general finding stands: instructions cannot fully substitute for the pipeline. Declared voice is necessary; it is not sufficient.

2.5 Why Single-Pass Framing Misleads (Nothing Ships Unedited)

Public debate about AI writing quality is dominated by one-shot framing: prompt in, prose out, judge the prose. No professional publishing operation works that way, and this house's standing law is that it one-shots nothing — every manuscript passes through the full editorial machine before any human reads it. Judged one-shot, the capability gap between tiers looks decisive: 13/49 versus 0/49 in the lab. Judged at the end of the pipeline — the only place a reader ever stands — the gap disappears: 7/7 versus 7/7 under the frontier loop. Comparing raw generations across models measures the wrong stage of the process and systematically overprices generation capability. The correct unit of evaluation is the recipe, not the draft.

2.6 The Difficulty of Measuring Editorial Improvement

"The edit made it better" is, in most workflows, an unfalsifiable claim. Voice is felt rather than counted; two readers disagree; and even the trivially countable — word count — is unstable, with the four counting methods in actual house use disagreeing by a median of 0.97%, 95th-percentile 1.68%, and maximum 2.04% of the canonical count across all 74 compiled manuscripts. A system that cannot measure editorial improvement cannot detect editorial harm (Section 2.3), cannot rank editor capability, and cannot know when to stop iterating. The prerequisite for everything that follows is instrumentation: a deterministic attribution gate with graded verdicts, range-based judgment margins for every count, and a before-and-after measurement bracketing every pass.

3. Design Objectives

3.1 Preserve Author Identity Across Genres

Each pen's voice must survive contact with any premise, including premises outside its home genre lane. The design target is authorial fingerprint, not genre furniture — justified by the canon measurement that genre is a weak stylometric force (same-genre Δ0.82 versus cross-genre Δ0.89; the hand beats the shelf) and validated by cross-genre transfer testing. The system distinguishes mechanical-identity pens (whose fingerprints ride on point-of-view and pronoun profile, and transfer cross-genre) from content-entangled pens (whose fingerprints ride on noun fields tied to their settled subject matter, and carry their lane with them), and engineers each accordingly.

3.2 Minimize Generation Cost

Bulk drafting — the largest token volume in the pipeline — must run at the lowest model tier that the downstream loop can converge, which the closed-loop trial demonstrated is the lowest tier available. The objective is not cheap output tolerated; it is cheap output made equivalent: lowest-tier generation must reach

the same measured endpoint (rank-1 attribution, style bands in) as frontier generation, and the evidence shows it does, 7/7 to 7/7. Cost savings that do not survive the gate do not count. (The realized savings figure is a projection until dollar instrumentation lands.)

3.3 Reserve High-Capability Models for Scarce Judgment

Frontier capability is spent only where lower tiers measurably fail: the editorial carve. The measured ladder is explicit — mid-tier voice editing is net harmful, the frontier tier is the only net-positive open-loop arm, and the lowest tier cannot converge even with exact prescriptions in hand (0/7 self-loop; it executes word swaps but cannot recast sentences). The allocation rule follows the measurements, not intuition: mechanical sweeps at cheap tiers, voice-carving judgment at the top rung only, and no tier assigned work above its measured floor.

3.4 Replace Repeated Reasoning With Deterministic Code

Any analysis a model performs more than once is a candidate for extraction into deterministic, stdlib-only code that runs identically at zero model cost: fingerprint regeneration, attribution checking, band compliance, per-word voice diagnosis, word counting, tic scanning. Code is capitalized intelligence — paid for once, executed forever — and it is tier-independent, so every gate is equally strong regardless of which model is being gated. The system further commits to the self-upgrade loop: an agent that builds a useful instrument mid-run banks it as repository code before its context ends, which is how the diagnostic central to the closed-loop result came to exist.

3.5 Make Every Editorial Pass Measurable

No editing pass ships on assertion. Every voice-touching pass is bracketed by the stylometric check: the edited text must still attribute to its own pen, and its distance to its own centroid must not worsen beyond the noise margin. Verdicts are graded (PASS / DRIFT / FAIL) and range-based: DRIFT — own pen at rank 2 within Δ0.05 of the winner — is a boundary reading, not a failure, on the same

principle as the word-count law. That law is itself an objective: no quantitative judgment is ever ruled on a margin smaller than the measured disagreement between measurement methods, fixed house-wide at ±2% for word counts, with every target written as a range rather than a bare number.

3.6 Preserve Story, Canon, and Manuscript Integrity

Voice compliance must never be purchased with story damage. Editorial carving is instructed to preserve plot, continuity, canon facts, and manuscript structure, and integrity checks (continuity ledgers, duplicate-scene detection, arithmetic verification) run alongside the stylometric gate. The closed-loop trial showed this working in the productive direction — the loop editor repaired duplicated scenes and broken continuity arithmetic in cheap drafts in passing — while the paper is explicit about the residual risk: attribution measures voice, not literary quality, and independent literary evaluation of converged manuscripts remains open.

3.7 Support Future Models Without Rebuilding the System

The architecture binds to measurements, not to models. Every new model or tier change is profiled through the standing lab protocol — its attractor pen (the shelf its default hand lands on), its one-shot voice-holding on the 7×7 matrix, and its closed-loop convergence speed as an editor — and is then routed to the rungs where it measures. When a model falls short, the doctrine treats it as a structure finding first: the fix is a stronger executable specification or scaffold, not a reflexive tier upgrade. Because the gates are deterministic code and the fingerprints regenerate from the manuscripts on disk, the entire quality apparatus is invariant under model replacement; the models are interchangeable labor, and the quality lives in the architecture.

---

4. System Architecture

The system treats authorial voice as a measurable property of text, enforced by deterministic code, rather than as an emergent property of any one model. The architecture separates four concerns: reference data (what each voice measurably is), declared intent (what it is supposed to be), generation inputs, and enforcement.

4.1 The Author Fingerprint Spine (150-MFW deltas + style bands; 27 pens, 74 books, 2.01M words)

The measured core is a fingerprint file, `PEN_FINGERPRINTS.json`, generated by `build_fingerprints.py` from every compiled manuscript on disk — at the current baseline, 74 compiled books totaling 2.01 million words across all 27 active pen names. The builder is standard-library-only, deterministic, and runs as one command; it is regenerated after any new book compiles, and its output is never hand-edited.

Each pen's fingerprint has two blocks. The delta block is the pen's mean z-profile over the corpus's 150 most frequent words — predominantly function words, the deep and difficult-to-fake signal. The style block records the pen's observed range, with a margin, across sentence-rhythm, punctuation, and lexical metrics (Section 6.3). The file also carries per-pen tell words and avoid words (the most over- and under-used common words versus the house average), cohesion statistics, and a provisional flag for any pen with only one compiled book.

4.2 The Public-Domain Canon Baseline (31 anchor novels, 20 authors)

The house corpus alone cannot say whether the pens are "distinct enough"; an external scale is needed. The canon baseline supplies one: 31 public-domain anchor novels by 20 real authors, fingerprinted in a joint feature space with the house shelf. It calibrates every claim about pen separation (Section 6.5), and validates the method itself: the same pipeline blindly attributed 13 of 14 canon

samples to the correct real author. The anchors span 1810–1926; era is the acknowledged standing confound on any cross-cloud comparison.

4.3 Declared-Voice Documents (Author Homebases and Spines)

Each pen name carries a qualitative declaration of its voice — an `AUTHOR_HOMEBASE` and a voice spine stating intended register, rhythm, and signature moves. The architecture treats these as intent, not truth: the stylometry spine is the declaration's verification layer. Where declared and measured voice disagree, the measurement governs, because it is what the reader actually receives; the declaration is reconciled to the shelf, or the writing is changed, but the two never drift silently. Experimental work (Section 9) showed declarations are most effective when written as executable specifications — locked mechanisms and numeric rates — rather than aspirational prose.

4.4 Premise, Outline, and Continuity Inputs

A book enters production with a premise, an outline whose beats carry register-gate cues, and continuity structures that persist across the draft: a book ledger (hard facts, plants and payoffs, motifs, point-of-view lanes, who-knows-what state) and continuity support spines (timeline, object state, geography). These inputs matter architecturally because the experiments found that voice lives in this pipeline scaffolding, not in prompts alone: drafts produced from a declaration without the pipeline collapse toward the generating model's own default hand.

4.5 Low-Cost Draft Generation From Top-Model Briefs

Generation is split into an expensive brief and a cheap draft. A top-tier director session assembles the generation brief — the beat plan and gate row, emotional and humor routes, the ledger and spine slices the chapter touches, the pen's voice mechanism, the word specification stated as a range, and the prohibited-construction kill list. A low-tier generator agent (Haiku) drafts the chapter from that brief and only that brief. Its output is raw material, never the deliverable:

because nothing is ever shipped one-shot, the generator's tier is not allowed to matter (empirical support in Sections 10–11).

4.6 Deterministic Voice Diagnosis (Per-Word Repair Prescriptions)

When a draft fails or drifts on attribution, `voice_diagnose.py` answers the actionable question: which function words are starved or bloated relative to the declared pen, and toward which rival that error pulls. It emits per-word repair prescriptions — direction, current rate per 1,000 words, the pen's shelf-implied rate, and the approximate occurrences to add or remove. The tool is deterministic, standard-library-only, and free at every model tier (mechanics in Section 5.4).

4.7 Closed-Loop Editorial Carving

The mandatory editing step pairs the diagnostic with a top-tier editor: diagnose, carve the named words in-voice, re-check, repeat. Carving — recasting sentences so the function-word profile moves without breaking prose quality — is the one loop step that measurably requires frontier-level judgment (Section 7.3). The loop, not the generator, is the system's primary quality mechanism.

4.8 Stylometric Gating (PASS/DRIFT/FAIL, range verdicts)

Every voice-bearing output passes through `stylometry_check.py`, which attributes the text against all 27 pen centroids by Burrows' Delta and checks every style metric against the pen's measured band. The verdict logic: PASS if the declared pen ranks first; DRIFT if it ranks second within 0.05 Delta of the winner — deliberately a range verdict, never ruled as a failure on a hair margin; FAIL otherwise, naming the rival pen the text fell toward. Pens with a single book always report as PROVISIONAL and never FAIL, because one book is a data point, not a voice. The tool exits 0 on PASS, DRIFT, or PROVISIONAL and 1 on FAIL, so the gate is scriptable.

4.9 The Word-Count Measurement Law (±2% method variance; all targets as ranges)

Word counts in this system are never single definitive numbers. The four counting methods in actual use (`wc -w` under UTF-8, `wc -w` under `LC_ALL=C`, Python `str.split()` raw, and `str.split()` on comment-stripped text, the canonical method) were measured against all 74 compiled manuscripts and disagree by a median of 0.97%, p95 of 1.68%, and maximum of 2.04% of the canonical count. Two house-wide rules follow. First, the judgment margin is ±2%: no boundary case — under a floor, over a target, against a platform specification — is ever ruled on a margin smaller than 2% of the canonical count, because inside that band the methods themselves disagree and the finding is method noise. Second, every target is written as a range (low–high), never a bare number, with floors and caps stating which counting method they bind. The stylometric DRIFT band (Section 4.8) is this same law generalized to attribution distance.

4.10 Best-Version Retention

The editing loop is bounded (four iterations in the experimental protocol; production convergence was observed in one to four rounds). When a loop terminates, the version retained is the best-scoring one produced, not simply the last: because each iteration is scored deterministically on rank and gap, a regression in a later round never displaces a better earlier version. Loops that reach the bound without PASS or DRIFT escalate rather than ship (Section 5.9).

4.11 Version and Evidence Storage

No pass overwrites its input. Drafts, edited versions, and pre-apply archives are kept as separate files under a fixed naming convention, with archived snapshots preserved indefinitely; the canonical filename always holds current truth and the archive folder is the history. Every write is verified on disk before any metadata is updated — existence first, outline conformance second, word count third — and metadata is reconciled to verified content, never the reverse. Experimental evidence is stored alongside the tools: raw per-sample scores (`scores.json`,

`editor_ladder_scores.json` in the lab folder), the fingerprint file with its generation note, and harvested before-and-after edit pairs for future training use.

5. The Closed-Loop Process

The closed loop is the per-manuscript application of the architecture above. Its shape is fixed: a cheap draft enters; a deterministic instrument measures it; a frontier editor repairs exactly what the instrument names; the gate re-measures; the cycle repeats until the text sits on its own pen's shelf or the bound is reached.

5.1 Draft

A low-tier generator produces the raw draft from the top-model brief (Section 4.5). The draft is written to disk under the standard naming convention with its provenance header, and is treated as raw material.

5.2 Diagnose

The draft is scored by the same mathematics the gate uses. The text is cleaned (comments, headers, and separators stripped), tokenized, and converted to a z-profile over the corpus's 150 most frequent words; Burrows' Delta is computed to every pen centroid. A minimum of 3,000 words is enforced — below that floor the fingerprint is not stable and the tools refuse to score. The diagnosis reports the declared pen's rank among the 27, the Delta to its own centroid, the nearest pen, and the gap.

5.3 Identify the Nearest Rival Voice

If the declared pen does not rank first, the diagnostic names the capturing rival — the pen whose centroid the text actually sits nearest. Naming the rival matters because failures are systematic, not random: each generating model has a measurable attractor pen it defaults toward (Section 6.7), and certain house pens

act as gravity wells that absorb, for example, any present-tense draft. The repair strategy differs depending on which well the text fell into.

5.4 Generate a Compact Repair Prescription

`voice_diagnose.py` ranks the 150 monitored words by pull — the quantity |z_text − z_pen| − |z_text − z_rival|, each word's share of the Delta disadvantage toward the rival — and prints the top roughly 15 as PULL lines: the word, its pull, the text's rate per 1,000 words, the pen's shelf-implied rate, the direction (raise or lower), and the approximate occurrence count to add or remove. A second list, the GAP lines, gives the words with the largest absolute deviation from the declared pen regardless of rival — the residue that remains after the rival flips. Out-of-band style metrics are listed with direction and target. The shelf-implied rates are point estimates, and the prescription's "need" counts are treated as aim points, not exact quotas, consistent with the measurement law.

5.5 Carve the Manuscript

A top-tier editor carves the named words in-voice: recasting sentences so the prescribed function-word rates move while the prose remains fluent, in-genre, and true to the pen. The same pass repairs generator-class defects on sight — duplicated passages, continuity arithmetic errors, looping — and runs the house prose sweep. Carving is never delegated below the top tier (the measured basis for this restriction is Section 7.3).

5.6 Run the Stylometric Gate

The carved text is re-scored by `stylometry_check.py` — full attribution against all 27 centroids plus every style band. The gate is code: it costs nothing per run, cannot be argued with, and produces a machine-readable verdict and exit code.

5.7 Measure Rank and Attribution Gap

Each iteration's rank and gap are recorded, giving the loop a monotone progress signal. Rank movement, rather than absolute Delta, is the honest metric for short

samples, which sit farther from every centroid than settled full-length books do.

5.8 Repeat Until Convergence (bounded iterations)

The diagnose–carve–gate cycle repeats until the verdict is PASS or DRIFT. In the head-to-head trials, a top-tier editor running this loop converged to rank-1 attribution in one to four iterations from either generator's drafts. The loop is bounded at four iterations in the experimental protocol; it does not run open-ended.

5.9 Stop, Retain Best, or Escalate

On PASS or DRIFT, the loop stops and the converged version is retained. If the bound is reached without convergence, the best-scoring version is retained and the case escalates for human or architectural review rather than shipping — a persistent failure is treated as evidence about the voice or the pipeline (a rival gravity well, a content-entangled pen, an inadequate declaration), not merely as a bad draft. This escalation path follows the house verification law: a metric collision prompts reading the book, never automated relabeling.

6. Stylometric Framework

The framework rests on a distinction between what a voice is declared to be and what it measurably does, and on a small set of signals — function-word distributions foremost — that are robust, cheap to compute, and hard to fake.

6.1 Designed Voice Versus Observed Voice

Every pen name has two spines. The qualitative spine (the homebase and voice-spine documents) states what the hand intends: register, pacing, signature moves. The stylometry spine states what the hand actually does on the shelf, derived exclusively from verified compiled books — never improvised, never hand-typed. Where the two disagree, the measurement governs, because it is

what the reader receives. A house-wide audit of declared-versus-measured voice falsified template sentence-ratio claims across the catalog, which motivated making measurement, not declaration, the enforcement surface.

6.2 Function-Word Signals (Burrows' Delta on corpus MFW)

The deep signal is the relative frequency of the corpus's 150 most frequent words — dominated by function words such as articles, prepositions, pronouns, and auxiliaries. For each book, each word's frequency per 1,000 words is z-scored against the whole-house mean and standard deviation for that word; a pen's delta centroid is its mean z-profile across its books. Burrows' Delta between a text and a centroid is the mean absolute difference of their z-scores across the 150 dimensions. Attribution asks a single question of a text: whose shelf does it belong on — that is, which of the 27 pen centroids is nearest by Delta. On the compiled shelf, blind nearest-neighbor attribution reaches 77% (48 of 62 books; chance is approximately 4%), which is the accuracy range stylometry achieves on real authors, and the same method blindly attributed 13 of 14 canon samples to the correct real author. Function words carry the signal precisely because they are below deliberate authorial control: a model or editor can imitate imagery and diction far more easily than it can imitate a pen's rates of *of*, *the*, *didn't*, and *they*.

6.3 Sentence, Fragment, and Punctuation Architecture

The style block measures the surface architecture the delta block does not see. Sentence rhythm: mean sentence length, its standard deviation, and the proportions of short (eight words or fewer), medium (nine to twenty), and long (twenty-one or more) sentences, plus a fragment rate (sentences of four words or fewer). Punctuation habit, each per 1,000 words: em dashes, semicolons, ellipses, exclamation points, question marks, commas, and colons. Lexical texture: average word length and a standardized type–token ratio computed over successive 5,000-word segments to remove length dependence. For each pen, each metric carries a band: the observed range across the pen's books, widened by a margin (the larger of 25% of the observed range, 15% of the absolute mean, or a small floor). An out-of-band value means "this text does not sound like this

pen's shelf," not "wrong" — a pen deliberately stretching into a new subgenre will move its own bands at the next regeneration; that is the spine learning, not lying. Some style constants hold house-wide across all pens — exclamation rate near zero and fragment rates between 0.18 and 0.30 — the shared reflexes of the single deep authorial identity beneath the registers.

6.4 Dialogue and Narration Balance

The dialogue ratio — the fraction of text inside quotation marks — is measured as part of the style block and banded per pen like every other metric. It behaves as a genuine voice carrier: in the executable-homebase intervention, samples that matched a pen's rhythm and punctuation specification still missed on dialogue load in four of seven genres, and low dialogue was among the named residual misses. Dialogue balance is thus tracked as an independent axis rather than assumed to follow from register.

6.5 Genre as a Measured-Weak Signal (same-genre Δ0.82 vs cross-genre Δ0.89 in canon)

A natural objection to a multi-genre house sharing one deep author is that genre itself might dominate the stylometric signal. The canon baseline measured this directly: among the 31 public-domain anchors, books in the same genre by different authors sit at a mean Delta of 0.82, versus 0.89 for books in different genres — a small separation compared to the author signal (within-author 0.50 versus between-author 0.88). Genre is a real but weak stylometric force; the hand beats the shelf. The same baseline placed the house pens in context: within-pen distance averages 0.73 (pens hold identity slightly looser than a real author holds their own), while between-pen distance averages 1.01 — the 27 pens spread wider than the 20 real authors, more differentiated rather than less. The era gap between the anchors (1810–1926) and the house corpus remains the standing confound on all cross-cloud numbers.

6.6 POV and Tense as Dominant Voice Carriers (mechanical-identity vs content-entangled pens)

The cross-genre experiments surfaced a structural distinction among pens. Mechanical-identity pens — whose fingerprint is anchored in point of view and tense — transfer across genres, because a first-person confessional mechanism fixes the pronoun and contraction profile (*I, me, didn't*) regardless of subject matter; the one pen that held all seven genres from its declaration alone is of this type. Content-entangled pens carry their lane with them: one pen's fingerprint rides on third-person dense-noun-phrase density (*of, the, an, they*) entangled with its settled series content, and a premise from another genre modulates that noun field no matter what the specification says. Tense operates at corpus scale as well: one pen's fingerprint is effectively present tense itself (*has, does, says, will, is*), and any present-tense draft drifts toward it — so attribution is checked before the draft is blamed.

6.7 Rival-Pen Attribution and Per-Model Attractor Pens

Attribution failures cluster rather than scatter. Certain pens act as gravity wells that absorb other pens' books, and — the sharper finding — each generating model has a measurable attractor pen: the shelf its unassisted default hand lands on. At the current baseline, the top-tier generator's default hand is adjacent to one crime-lane pen, and the low-tier generator's to one science-fiction-lane pen. In the stripped-down one-shot matrix, nearly every failing cell converged on the generating model's attractor rather than on the premise's genre press (only 5 of 36 and 9 of 48 failures went to the premise-genre's press) — demonstrating that what the checker detects in unpipelined drafts is the generator's own hand, and that the attractor is the benchmark every structural upgrade must beat. Profiling a model's attractor is now the first number on every new model's scorecard.

6.8 Rank, Gap, and Range-Based Confidence

Verdicts are expressed in ranks and gaps, never in bare distances. PASS requires rank 1 of 27. DRIFT — rank 2 within a gap of 0.05 Delta of the winner — is

explicitly a range verdict: the word-count measurement law generalized, on the principle that no boundary case is ever ruled on a margin smaller than the noise of the method. FAIL names the capturing rival so the repair is directed, and is treated as evidence about the voice rather than proof the book is bad — the text is read before anything is concluded. Provisional pens (one compiled book) report but never fail, and the 3,000-word floor bounds where any verdict is meaningful at all. For short samples, rank movement is the honest progress metric, since all short one-shot samples sit far from every centroid (observed top-Delta around 0.86–1.24, versus roughly 0.6–0.8 for settled books).

6.9 Drift Checks on the Accumulating Manuscript

The checker accepts folders as targets, concatenating every chapter file, so the accumulating manuscript can be scored mid-book — and the check is run before and after any editing pass. Editing is where voice erodes silently: a sweep that "improves" sentences can normalize the pen's habits out of the text, dragging the book toward the house centroid. The regression rule is therefore deterministic: edited text must still attribute to its own pen, and Delta-to-own-pen must not worsen beyond the noise margin; an edit that fails this is a defect of the pass, not of the book. The style bands double as the pointing device for repair — a sentence mean above the pen's band means compress; a signature rate below band means the hand is being erased.

7. Model-Tier Labor Allocation

The allocation principle is that each unit of work goes to the cheapest resource measured capable of it, with deterministic code taken first wherever judgment is not required. The tiers below are roles; the specific models occupying them are interchangeable subject to measurement (Sections 7.6–7.7).

7.1 Cheap Models as Drafting Labor

Raw generation runs on the lowest tier (Haiku), from a brief assembled by the top tier. The generator receives the brief and only the brief — a deliberately lean context — and its output is raw material by definition. This allocation is safe only because of the mandatory closed loop downstream: in the head-to-head trial, low-tier generation plus top-tier loop editing matched top-tier generation plus top-tier loop editing at 7 of 7 rank-1 attribution, making the generator economically negligible given the loop.

7.2 Deterministic Code as Diagnostic Labor

All measurement is code: fingerprint construction, attribution, band checking, and per-word repair prescription are standard-library scripts that run identically at every tier of the ladder and cost nothing per invocation. Any reasoning that would otherwise be repeated by a model on every pass — counting, scoring, naming which words deviate — is capitalized once into these tools. The precedent case: the diagnostic that made full closed-loop convergence possible was invented by an agent mid-run and immediately banked as `voice_diagnose.py`, under the standing rule that a useful instrument built in-session is committed as repository code before the session ends.

7.3 Frontier Models as Editorial Judgment (voice-carve top-tier only)

Voice carving — recasting sentences so the function-word profile moves without damaging the prose — is restricted to the top tier, and the restriction is measured, not asserted. In the open-loop editor ladder over identical low-tier drafts with identical prompts and supplied targets, the mid-tier editor was net harmful for voice, making attribution rank worse in three of seven genres by carving aggressively toward other pens; the upper-mid tier was mixed; only the top tier was net positive (median rank 8, versus 15 and 13 for the lower tiers and 9 for the unedited drafts). Giving voice-carve to a cheap editor was measurably worse than not editing at all.

7.4 Mechanical Cleanup at Lower Tiers

Mechanical work — deterministic tic scans, formatting sweeps, and rule-driven cleanup that does not require recasting prose — remains cheap-tier work under the ladder doctrine. The dividing line is judgment: passes whose corrections are fully specified by a rule stay low; passes that trade off voice, meaning, and fluency go high.

7.5 Measured Capability Floors by Task

The ladder's floors are empirical. The low tier can draft but cannot edit itself to convergence: given the diagnostic's exact per-word prescriptions, the low-tier self-loop produced 0 of 7 passes, with movement from rank 5 to rank 3 at best and one regression — it executes word swaps but cannot recast sentences. The low tier also held zero pen voices in one-shot generation from declarations alone (0 of 49, one DRIFT), versus 13 of 49 for the top tier. The measured floor statement: diagnosis is code (free); drafting is low-tier work given a top-tier brief; carving is top-rung judgment work.

7.6 Model-Agnostic Orchestration

The system binds roles, not models. The pipeline's contracts — brief in, raw draft out; draft in, prescription out; prescription in, carved text gated by the checker — are model-independent, and the enforcement layer is code that runs the same regardless of what occupies each rung. Models are referred to by family (Fable, Opus, Sonnet, Haiku) as current occupants of rungs, and each carries a scorecard rather than an assumption.

7.7 Capability-Gated Execution

A model earns a rung by measurement, through the standing model-evaluation protocol run in the quarantined lab before any new model or tier change is trusted with production work. Three numbers are collected: the model's attractor pen (whose shelf its blind default hand lands on — the benchmark its briefed output must beat); its one-shot voice-holding across the seven-pen by seven-

genre matrix; and its closed-loop convergence speed as an editor, in iterations to PASS with the diagnostic and checker in the loop. Current baselines: the top tier converges to rank 1 in one to four iterations from either generator's drafts; the mid tier is net harmful for open-loop voice editing; the low tier's self-loop does not converge. Because the gates are deterministic and the scorecard is standing, future models slot into the ladder — up or down — without rebuilding the system.

---

8. Experimental Design

8.1 Research Questions

The experimental program was built to answer four questions, in order of increasing operational consequence:

1. **Is the measured pen fingerprint authorial or generic?** If a pen's stylometric signature survives outside its home genre, the signature belongs to the hand; if attribution follows the genre, the signature was genre furniture.
2. **Where does the voice actually live?** Specifically: does a pen's declared voice document (its Author Homebase) reproduce the pen's measured fingerprint on its own, or is the fingerprint a product of the full production pipeline?
3. **Does the tiered-labor economy hold?** Can a low-cost model generate drafts that a higher-tier editor then carves to the target voice, and does the outcome match or beat one-shot generation by a frontier model?
4. **Is closed-loop editing against a deterministic gate materially different from open-loop editing of equal editorial strength?**

8.2 Controlled Variables

Two controls anchor every experiment.

Fixed premises across pens. Seven premises, one per genre lane (Romance, SFF, Contemporary, Crime, Horror, Action, Literary), were written once and held constant across all pens. Within any genre column the content is identical and only the hand varies, so attribution differences between cells are authorial by construction. Lab character names were registered in the house name authority under a quarantine block and are barred from reuse in catalog books.

Unchanged shelf centroids as targets. All scoring throughout the program — one-shot matrices, homebase intervention, editor ladder, and the closed-loop trial — is against the same 27 compiled-shelf pen centroids, generated once from the compiled corpus and never regenerated mid-experiment. No lab sample ever entered the reference corpus.

8.3 Independent Variables

Three variables were manipulated:

- **Generator tier** — drafts produced by the frontier-tier model (Fable family) versus the lowest-tier model (Haiku family), under identical prompts and blindness rules.
- **Editor tier** — editing passes by three tiers (Sonnet, Opus, Fable), identical prompts, identical supplied targets.
- **Open versus closed loop** — a single editing pass (open loop) versus an iterated cycle in which the editor runs the deterministic diagnostic and attribution checker itself and edits against the output until the gate passes or the iteration cap is reached (closed loop).

8.4 Dependent Measures

Each scored sample yields three measures:

- **Attribution rank of 27** — the position of the sample's own pen when its Burrows' Delta distance is ranked against all 27 pen centroids. Rank 1 is a blind attribution hit.

- **Gap to rival** — the Delta margin between the winning pen and the sample's own pen (0.0 when the own pen wins). The gap supports range verdicts rather than knife-edge rulings.
- **Style-band compliance** — whether the sample sits inside the pen's observed bands for sentence rhythm, fragment rate, punctuation rates per 1,000 words, dialogue ratio, word length, and standardized type-token ratio.

8.5 Test Corpus

The reference side of every measurement is the house corpus: **74 compiled manuscripts, 2.01 million words, across all 27 active pens**, from which the 150-most-frequent-word Delta centroids and per-pen style bands are derived. The method itself was validated against an external baseline of **31 public-domain anchor novels by 20 real authors**, scored in a joint feature space with the house corpus (Section 9.2).

8.6 Cross-Genre Author Testing

The cross-genre lab is quarantined instrument calibration, not book production: no catalog entry, no status file, nothing compiled, nothing published, and no lab prose ever quoted into a real manuscript. Two blindness rules governed generation:

- **Drafting agents were blind to all measured data.** They drafted from the pen's Author Homebase and the premise only — never the fingerprint file, the atlas, or any stylometric measurement. Coaching from the measurement would invalidate the test.
- **Predictions were logged before scoring.** Three predictions were recorded in the lab README prior to any scoring run, including the headline prediction that the pen would survive genre shift in the majority of cells. Section 9.4 reports that this prediction was falsified for most pens — a result the design was built to be able to produce.

8.7 Model Pairings

The full program covers: two generator arms (frontier, Haiku) across the 7×7 one-shot matrix; a redraft of one pen's row on both arms after a homebase rewrite; three open-loop editor tiers (Sonnet, Opus, Fable) over the seven Haiku drafts of that pen; and three closed-loop recipes — Fable-generates + Fable-loop-edits, Haiku-generates + Fable-loop-edits, and Haiku-generates + Haiku-loop-edits.

8.8 Iteration Limits

Closed-loop editing was bounded at **four iterations** per sample (edit → diagnose → carve → re-check). Convergence in practice took one to four rounds (Section 11.5); no sample was allowed to iterate indefinitely against the gate.

8.9 Success Criteria

- **PASS** — the sample's nearest centroid of 27 is its own pen (rank 1).
- **DRIFT** — the own pen ranks 2nd within a Delta gap of 0.05 of the winner. DRIFT is a range verdict, not a failure ruled on a hair margin.
- **FAIL** — the sample attributes to another pen's shelf.

Style bands are reported alongside the verdict; an out-of-band line means "does not sound like this pen's shelf," not "wrong."

8.10 Manuscript-Quality Safeguards

Voice carving was never licensed to degrade the prose. Editor prompts in every arm included the full house editorial sweep alongside the fingerprint carve, and editors were instructed to repair defects they encountered, not merely to move word statistics. Section 11.6 documents that loop editors did in fact repair structural and factual defects in passing; Section 11.8 states the honest limit — attribution and band compliance are measured, and independent literary-quality evaluation of the converged samples remains open.

9. Preliminary Experiments and Findings

9.1 Shelf Attribution Baseline

Before any lab work, the corpus itself was mapped. Blind nearest-neighbor attribution over the compiled shelf assigns **77% of held-out book texts to their own pen (48/62), against chance of roughly 4%** for a 27-way decision. A companion press-level attribution figure is sometimes quoted, but it does not appear in the evidence files underpinning this paper and is treated here as unverified in-repo; the 77% pen-level result is the load-bearing baseline. The register system therefore produces measurably distinct hands at the accuracy stylometry achieves on real authors.

The map also established the structure the later experiments would probe: the sharpest voices (separation of +0.66 down to +0.55 for the top five pens), three "ghost" pens with no measurable identity yet (+0.06 to +0.13), two gravity wells that absorb other pens' books (one pen whose fingerprint is present tense itself; one whose fingerprint is domestic-furniture vocabulary), and house-wide constants shared across all pens (exclamation rate near zero; fragment rate 0.18–0.30).

9.2 Canon Comparison

The instrument was validated against 31 public-domain anchor novels by 20 real authors, scored jointly with the house corpus. Findings, with the standing confound stated below:

- **Method check: 13/14 blind author attribution** on the canon itself — the pipeline attributes real authors at the expected accuracy.
- **Author is strong; genre is weak.** Canon same-genre pairs average Δ0.82 versus cross-genre Δ0.89 — a small difference. The hand beats the shelf, in the canon as in the house.
- **House pens spread wider than real authors.** Real-author within-author distance averages Δ0.50 versus house within-pen Δ0.73 (pens hold identity

slightly looser than a real author holds their own), while real-author between-author distance is Δ0.88 versus house between-pen Δ1.01 — the 27 pens are more differentiated from one another than the 20 real authors are.

Era confound. The canon anchors date from 1810–1926; the house corpus is contemporary. Any number that compares the two clouds directly (the within/between contrasts above) carries era as an uncontrolled variable. The genre finding, measured within the canon cloud, is the most confound-resistant of the set.

9.3 Declared-vs-Measured Voice Audit

Every pen's declared voice document was harvested against its measured fingerprint (27 pen records; 25 with declared spines). Where a declared claim was machine-comparable, it was checked deterministically: **81 checks — 29 MISMATCH, 28 NEAR, 23 MATCH, 1 unparsed**. The mismatches concentrate almost entirely in the template's sentence-ratio claims (12 medium-sentence-ratio and 10 short-sentence-ratio mismatches, plus a handful of punctuation-rate misses): the declared short/medium/long sentence distributions, copied from a shared template, were falsified house-wide by the measured shelf. The audit established that declaration and measurement can drift silently, and that the measurement — what the reader actually receives — is the reconciliation target.

9.4 The 7-Pen × 7-Genre One-Shot Matrix

Ninety-eight samples were drafted blind (7 pens × 7 premises × 2 generator arms), each at or above the checker's 3,000-word floor (per-cell counts run roughly 3,100–3,800 words; typical 3.2–3.5k), from Author Homebase plus premise only — no outline routing, no ledger, no editorial passes of any kind.

One-shot confusion summary:

Arm	Rank-1 PASS	DRIFT	Dominant capture
Frontier	13/49	0	Rance Graves (Crime pen) in most failing cells
Haiku	0/49	1	Lance Starweave (SFF pen) in 45 of 49 cells

On the frontier arm, exactly two pens held: Vance Thorne passed 7/7 across every genre and Rance Graves passed 6/7; every other pen's row collapsed, mostly into Rance's shelf. On the Haiku arm, 45 of 49 cells attributed to Lance Starweave; the remaining four went to Runce Graves (2), Rynce Graves (1), and Lence Starweave (1), with a single DRIFT (Tence Stonewell, Contemporary, rank 2, gap 0.022).

Four findings:

1. **The instrument caught exactly what changed.** The compiled shelf attributes at 77%; strip the pipeline away and attribution collapses to one attractor per generator model. Noise scatters; this converges. And it is mostly not genre capture — only 5 of 36 frontier failures and 9 of 48 Haiku failures went to the premise-genre's home press. What the checker detected is the generator's own default hand.
2. **The pre-registered prediction was falsified.** The logged prediction — pen survives genre shift in the majority of cells — held for the compiled shelf but not for homebase-only drafting. **The pen voices live in the pipeline, not in the homebase declaration.** The measured fingerprints were produced by the full machine; the declaration alone does not reproduce them.
3. **The exception is diagnostic.** Vance Thorne's 7/7 is the existence proof that a declaration can carry a voice — but its homebase is *executable*: a locked first-person confessional mechanism, a numeric rhythm specification, concrete mechanical habits. Rance (6/7) is the second proof.
4. **Each model has a measurable attractor pen** (frontier ≈ Rance-adjacent; Haiku ≈ Lance-adjacent), which any structural upgrade must beat. A caveat applies to the two passing pens: their success may be partly proximity of the generator's default register to their shelves rather than purely homebase quality — motivating the intervention in 9.5.

9.5 The Executable-Homebase Intervention

One failing pen — Lince Starweave — had its declared voice rebuilt from aspirational prose to an executable specification (locked narrator mechanism, numeric rhythm spec, a conjunction rule, per-1,000-word punctuation rates,

every number sourced from the measured shelf). The pen's seven-genre row was redrafted blind on both arms and rescored against the unchanged centroids.

Result: 0/7 PASS on both arms — but large, systematic rank gains on the frontier arm. Own-pen rank (of 27), version 1 → version 2: Romance 25→9, SFF 21→16, Contemporary 27→20, Crime 26→23, Horror 24→22, Action 20→7, Literary 27→20. Every cell moved toward the pen; two entered the top 10. The specification was verifiably executed: version-2 samples sit inside the pen's measured style bands nearly everywhere (sentence mean 9.0–11.1 words, short-sentence ratio near 0.5, semicolons near 3 per 1,000, fragments in band; misses were low dialogue in four cells and one semicolon overshoot). On the Haiku arm the results were mixed (9→15, 2→11, 14→14, 9→9, 4→5, 4→2, 6→4), and Haiku's attractor pen still captured every cell.

Three interpretive points, carried forward as standing conclusions:

1. An executable specification moves the style layer completely and the function-word layer substantially — but not to rank 1 of 27. The attribution decision rides on the 150-word Delta block, and the specification closed only part of that gap.
2. **Mechanical-identity pens transfer cross-genre; content-entangled pens carry their lane with them.** Vance's identity is fixed by point of view (a first-person confessional pronoun/contraction profile that survives any subject); Lince's shelf fingerprint rides on a dense-noun-phrase function-word field entangled with the settled content register of the pen's home series. A premise about a wedding venue modulates that noun field regardless of the specification.
3. **Short samples are stylometric outliers.** All lab samples sit far from every centroid (Δ to nearest centroid roughly 0.86–1.24, versus roughly 0.6–0.8 for compiled books), so rank movement — not absolute Delta — is the honest metric at this sample length.

The lever is real (the largest single-intervention rank gain observed) and necessary — but not sufficient for one-shot voice transfer.

9.6 The Open-Loop Editor Ladder

The seven Haiku version-2 drafts of the Lince row were each edited once by three tiers — Sonnet, Opus, Fable — under identical prompts (full house sweep plus fingerprint carve, targets supplied), then rescored. Own-pen rank of 27 per cell:

Genre	Haiku raw	Sonnet-edited	Opus-edited	Fable-edited	Frontier one-shot
Romance	15	20	18	**8**	9
SFF	11	11	5	**6**	16
Contemporary	14	**6**	16	16	20
Crime	9	23	13	16	23
Horror	5	15	**6**	7	22
Action	2	2 (gap 0.044)	**2 (gap 0.020)**	2 (gap 0.025)	7
Literary	4	20	21	**8**	20
Median	**9**	15	13	**8**	20

Four findings:

1. **The economy directionally stands.** Haiku-generate plus Fable-edit beats frontier one-shot generation on median rank (8 versus 20) and on gap-to-rival (the Action cell reached gaps of 0.020–0.044 — inside the DRIFT band) at a fraction of the generation cost.
2. **The bottom rung is net negative for voice work.** Sonnet editing made ranks worse in 3 of 7 cells (Crime 9→23, Literary 4→20, Horror 5→15); aggressive carving by the mid-tier editor pushed samples toward *other* pens. Opus was mixed; Fable was the only net-positive arm. Voice carving is top-tier judgment work; giving it to a cheap editor is worse than not editing. Mechanical sweeps remain cheap-tier work.
3. **The generator's skeleton survives open-loop editing.** Lance Starweave — Haiku's attractor — remained the top-attributed pen in nearly every edited

cell across all three editor tiers. One editing pass, however strong, does not launder the generation model's function-word residue.

4. **Closed loop is the unlock.** The single cell that reached the DRIFT band with all style bands in band was the one where the editor ran the attribution checker itself and iterated against it (Sonnet, Action, gap 0.044). Open-loop editing moved gaps; closed-loop editing converged.

Caveats: one edit pass is not the full house editing suite; n = 7 cells per arm; all movement occurred inside a persistent attractor capture; and editor agents repaired real generation defects along the way (a duplicated scene, broken interval arithmetic, age-continuity errors) — editing value the attribution score does not capture.

10. Head-to-Head Closed-Loop Trial

The final trial answered the operational question directly: does the frontier model need to both generate and edit, or is the generator's tier negligible once the closed loop exists? Three full recipes were run over the same seven-genre row, all scored against the unchanged shelf centroids. The closed loop in each case is: edit → run the deterministic voice diagnostic → carve the named words → run the attribution checker → repeat until PASS/DRIFT or the iteration cap of four.

10.1 Fable Generation With Fable Loop Editing

The frontier-tier drafts of the Lince row were loop-edited by the frontier-tier editor. This is the maximum-cost recipe and the ceiling reference.

10.2 Haiku Generation With Fable Loop Editing

The lowest-tier drafts of the same row were loop-edited by the frontier-tier editor. This is the candidate production recipe: cheap generation, expensive judgment, deterministic gating.

10.3 Haiku Generation With Haiku Loop Editing

The lowest-tier drafts were loop-edited by the lowest-tier editor, with the same diagnostic prescriptions available. This arm tests whether the loop's instructions can substitute for editorial capability.

10.4 Scope: One Pen, Seven Genres, Bounded Iterations

The trial's honest scope: **one pen (Lince Starweave)** — deliberately a hard case, a content-entangled pen that failed every one-shot cell — across **seven genres**, at **~3.2–3.6k words per sample**, with the **iteration cap at four**. Multi-pen replication is future work, not a claim of this paper. All samples remain quarantined lab material.

10.5 On-Disk Verification (Content Before Metadata)

Per the house verification law, results were ruled on content, not labels: every converged sample exists as a file on disk in its recipe's folder (seven per arm, twenty-one total), was scored from the file by the deterministic checker, and had its word count verified against the 3,000-word floor as a range across counting methods rather than a single number. The before/after texts of all editing arms — 42 pairs spanning the three open-loop tiers and the three closed-loop recipes, with generator, editor, loop type, and attribution scores on both sides — are banked as structured training data.

10.6 Final Results

Recipe	Rank-1 PASS
Fable generates + Fable loop-edits	**7/7**
Haiku generates + Fable loop-edits	**7/7**
Haiku generates + Haiku loop-edits	0/7

11. Results

11.1 Fable+Fable: 7/7 Rank-1

The maximum-cost recipe converged every cell to rank-1 attribution with style bands in band, within the four-iteration cap. A pen that had failed all seven one-shot cells — including two cells at rank 27 of 27 — was carried to perfect blind attribution by the loop.

11.2 Haiku+Fable: 7/7 Rank-1

The candidate production recipe matched the ceiling exactly: 7/7 rank-1, style bands in band. Most consequentially, the Haiku generation watermark — the attractor-pen residue that had survived every open-loop editing tier in Section 9.6 — was fully erased by the closed loop. The finding that "one pass cannot launder the generator's hand" is a finding about open loops, not about editing as such.

11.3 Haiku+Haiku: 0/7

The lowest-tier editor could not close the loop even with the diagnostic's exact per-word prescriptions in hand: seven samples, zero PASS, best final rank 2, strongest improvement rank 5 → rank 3, mostly flat, with one regression. The qualitative failure mode was consistent: the small model executes word swaps but cannot recast sentences. Instructions did not substitute for capability; this is the measured floor of the editor ladder.

11.4 Generator Quality Versus Editor Quality

Read together, the three arms answer the head-to-head question: **the generator is negligible — the loop is everything.** Both generators converge to identical perfect attribution under the top-tier closed loop, while the cheap editor converges under neither. The validated house economy follows directly: the lowest tier writes, the top tier loop-edits, and deterministic code gates every

pass. Diagnosis is code and effectively free; carving is top-rung judgment work. One qualification from Section 9.6 carries over: under *open-loop* editing the generator choice is not free — it sets the floor the editors fight — so the economy is validated specifically with the closed loop in place.

11.5 Convergence in 1–4 Rounds

All fourteen converging samples reached rank-1 PASS in one to four iterations of the edit–diagnose–carve–check cycle; none required the cap to be raised, and no converging cell was abandoned. The per-word diagnostic that made this convergence rate possible was itself built by a closed-loop agent mid-run and banked as repository code — an instance of the self-upgrade pattern discussed in Section 12.

11.6 Incidental Repair of Structural/Factual Defects

The Haiku drafts required real repair beyond the voice carve: duplicated scenes, broken continuity arithmetic (interval math), and age-continuity errors were found and fixed by loop editors in passing, under the same prompts. Two implications: the cheap generator's defect rate is a real cost the editing tier absorbs, and the attribution score understates the editorial value delivered — the loop editor is doing quality work the stylometric measure does not credit.

11.7 Word Floors Held Under the ±2% Measurement Law

Haiku-arm word counts run lower than frontier-arm counts. Judged under the house measurement law — counts reported as ranges across methods, no boundary ruled on a margin under 2% of the canonical count — the floors held: one cell sits 1% under the 3,000-word floor by one counting method, inside the ±2% method-variance band, and is therefore method noise rather than a violation.

11.8 Quality Safeguards During Carving

Every editing arm was instructed to run the full house editorial sweep alongside the fingerprint carve, and the incidental repairs in 11.6 are evidence the instruction was executed rather than ignored. The honest limit stands: what the trial measures is blind attribution and style-band compliance, and the incidental-repair record is evidence of editorial diligence, not a quality certification. **Independent literary-quality evaluation of the converged manuscripts remains open** (Section 21), and attribution-without-quality is listed as a standing limitation (Section 18). No claim is made here that a rank-1 sample is a good chapter — only that it is verifiably the pen's chapter.

(Sections 12–16. Evidence base: `_lab/RESULTS.md`, `STYLOMETRY_SPINE.md`, the `/chapter` and `/draft-pass` production skills, `SPIRIT_SUITE_ARCHITECTURE.md` §12, and the training-pair manifests under `_MASTER_ARCHITECTURE/_training/`.)

12. Interpretation

12.1 The Generator Is Economically Negligible — Given the Loop

The head-to-head closed-loop grid produced a symmetric result: frontier generation with frontier loop editing reached rank-1 attribution in 7 of 7 genres, and lowest-tier (Haiku) generation with the same frontier loop editing also reached 7 of 7. On the measured outcome — blind attribution against all 27 pen centroids, with style bands in range — the choice of generation model made no detectable difference once the closed loop ran.

The qualifier matters. The open-loop editor ladder showed the opposite: without the loop, the generation model's function-word residue (the "watermark") survived a full editing pass at every editor tier, with the generator's attractor pen remaining the top attribution in nearly every edited cell. The generator is

negligible *given the loop*, not in general — open-loop it sets a floor the editor cannot escape; closed-loop that floor is fully erased. The residual costs of cheap generation are real but bounded: the Haiku-arm word counts ran lower (one cell sat 1% under the 3,000-word floor by one counting method — inside the ±2% measurement band), and the Haiku drafts needed genuine defect repair beyond the voice carve (continuity arithmetic, duplicated scenes), which the loop editor performed in passing.

12.2 The Closed Loop Is the Primary Quality Mechanism

The evidence for the loop as the operative mechanism, rather than editor capability alone, comes from the ladder itself. Across three open-loop editing tiers, the single cell that reached the DRIFT band with all style bands in range was the one in which the editor ran the deterministic checker itself and iterated against it (a mid-tier editor, Action genre, gap 0.044). Open-loop editing moved gaps; closed-loop editing converged. The production loop is: edit, run `voice_diagnose.py` for a per-word prescription, carve the named words, run `stylometry_check.py`, repeat until PASS or DRIFT. With a frontier editor in that loop, both generators' drafts converged to rank-1 in one to four iterations across all seven genres. Quality here is not a property of a single strong pass; it is a property of iteration against a deterministic measurement.

12.3 Drafting Capability Is Widely Available; Convergent Editorial Judgment Is Scarce

The grid separates two abilities usually conflated. Producing a coherent 3,000-plus-word draft from a structured brief is within reach of the lowest model tier. Converging that draft onto a specific measured voice is not: the mid-tier editor was net harmful for voice work in the open-loop ladder (worsening attribution rank in 3 of 7 cells by carving toward *other* pens), the next tier up was mixed, and only the top tier was net positive — and only closed-loop top-tier editing converged. Drafting labor is abundant; the scarce input is editorial judgment that reliably moves a text toward a target rather than merely changing it. The system

prices accordingly: generation goes to the cheapest tier, and the convergent judgment is concentrated in the loop.

12.4 Instructions Cannot Fully Replace Capability (the Small-Model Self-Edit Floor)

The Haiku-edits-Haiku arm tested whether perfect instructions substitute for editorial capability. They do not. Given the same closed loop and the diagnostic's exact per-word prescriptions, the small model went 0 for 7 — best result rank 2, mostly flat movement, with one regression. The observed failure mode is specific: the small model executes word swaps but cannot recast sentences, and the fingerprint lives at the level of sentence architecture, not vocabulary substitution. This establishes a measured capability floor for the carving role. Diagnosis is code and costs nothing at any tier; carving is top-rung judgment work, and no amount of prescription detail moved that boundary in this trial.

12.5 Voice Lives in the Pipeline, Not the Prompt

The 7-pen × 7-genre one-shot matrix stripped the production pipeline away and drafted from declared author documents alone. The compiled shelf attributes blind at 77%; the stripped lab drafts collapsed to a single attractor pen per model — the frontier model's default hand landing Rance-adjacent (13 of 49 passes), the small model's landing Lance-adjacent (0 of 49). The pattern is convergent, not noisy, and mostly not genre capture. What the declared documents alone reproduce is the *model's* voice; what the full pipeline — routed outlines, gates, per-beat passes, deterministic carves — reproduces is the *pen's* voice. Two refinements bound the claim. First, one pen (Vance Thorne) held 7 of 7 from its declaration alone, because that declaration is executable: a locked point-of-view mechanism and numeric rhythm specifications rather than adjectives. Second, rewriting a failing pen's declaration to that executable standard moved every genre cell toward the pen — the largest single-intervention rank gain measured — but reached rank 1 in none of them. Executable declarations are necessary; they are not sufficient. Rank-1 attribution lives in the full pipeline plus a settled shelf. The structure is the voice.

12.6 Quality Can Be Relocated From the Model Into the System

Taken together, 12.1–12.5 describe a relocation. The properties normally purchased by using the most capable model for everything — voice fidelity, continuity, floor compliance — are instead held by the architecture: a deterministic fingerprint that defines the target, a diagnostic that names the repairs, a loop that iterates to convergence, and a bookend that catches regression. Models become interchangeable labor; when a tier changes, the scorecard protocol re-measures its attractor pen, one-shot ranks, and convergence speed, and the routing adjusts without rebuilding the architecture.

12.7 The Self-Upgrade Loop: Instruments Invented Mid-Run Migrate Into Code

The diagnostic that made 14-of-14 closed-loop convergence possible was not designed in advance. It was invented by a closed-loop editing agent mid-run — an instrument the agent built to do its own job — and then banked into the repository as `voice_diagnose.py` before that agent's context ended. This is now standing doctrine: an agent that builds a useful instrument during a run commits it as code so the capability survives the run. Each such migration converts a piece of repeated model reasoning into a deterministic, zero-marginal-cost tool, which is the mechanism by which the system becomes permanently cheaper over time rather than repeatedly paying for the same inference.

12.8 Expensive Intelligence Should Be Applied Narrowly

The measured allocation rule that falls out of the ladder: frontier capability is spent only where cheaper tiers measurably fail, and everything else is pushed down. Diagnosis and gating are code — free at every tier. Generation is lowest-tier. Mechanical cleanup stays on cheap tiers. The frontier model is reserved for two narrow jobs: writing the structural brief and running the closed-loop voice carve — the one role where the ladder showed cheaper substitutes are not

merely weaker but net harmful. Narrow application is not an austerity posture; it is what the measurements support.

13. Operational Workflow

13.1 Corpus and Canon Retrieval

A production session begins from measured references, not memory. The pen fingerprint spine (`PEN_FINGERPRINTS.json`, regenerated deterministically from every compiled manuscript on disk) supplies the attribution target; the public-domain canon anchors supply the external baseline. Fingerprints are regenerated after any new compile, so a check is always run against the current shelf.

13.2 Pen-Specific Draft Assignment

Each book is assigned its pen and drafts through the house author identity with the gate pinned to that pen's registers. The pen's declared documents (author homebase, voice spine) and its measured fingerprint travel together: the declaration says what the hand intends, the fingerprint says what the shelf actually does, and the closed loop targets the latter.

13.3 Lean Worker Contexts

Generation subagents receive the brief and only the brief: the gate row and beat plan, the emotional and humor routes, the ledger and continuity-spine slices the chapter touches, the pen's voice mechanism, the word specification as a range, and the kill list. Workers do not carry the whole book, the whole corpus, or coordination state. Where the composition plan marks a chapter epistemically "innocent," the blindness is strict — the drafting agent gets a truncated outline and only the knowledge its characters legitimately hold — because foreknowledge leaks through manner and a model cannot unknow.

13.4 Coordination-Layer Context

The directing session — the top-model layer — holds what workers do not: the outline and composition plan, the book ledger (facts, plants, motifs, point-of-view lanes, world rules), the continuity support spines (timeline, object state, geography), and the production state file. Structure is top-model work; the brief each worker receives is that structure, compressed. Growth of this coordination context over a long book is a known operating cost (see Limitations).

13.5 Parallel Subagent Execution

Because continuity lives in the ledger rather than in drafting sequence, chapters are not required to be produced in reading order, and independent drafting assignments can be fanned out to parallel generation subagents. The generation-order law makes this safe: anchors are drafted first per the composition plan, and every agent reconciles against the same ledger and spine family.

13.6 Editorial Loop Assignment

Every raw draft then enters the mandatory closed loop on the top model — the house one-shots nothing. The loop editor performs the house sweep (kill list, syntax, blocking) plus the voice carve run closed-loop: `voice_diagnose.py` for the prescription, carve, `stylometry_check.py`, repeat to PASS or DRIFT. The loop editor also repairs generator-class defects on sight — duplicated passages, continuity arithmetic, looping. Voice carving is never delegated below the top tier.

13.7 Automated Scoring and Escalation

The `/draft-pass` quality gate runs after every chapter: the deterministic tic scan and carve, the continuity and ledger reconcile, and (step 3.7) the stylometric voice check on the book's accumulated drafts — single chapters below the checker's roughly 3,000-word floor are checked as the book-so-far. PASS or DRIFT proceeds; FAIL triggers the diagnostic and a carve before drafting continues, because a book drifting toward another pen's shelf — or the generator's attractor pen — is a defect at birth and cheapest to fix immediately. Verdicts are treated as

ranges per the word-count law; a provisional pen (one book on its shelf) reports but never blocks.

13.8 Final Cleanup

When the book is fully drafted, the fifteen-Spirit editing suite runs inline across all chapters and its findings are applied, bookended by the before/after stylometric check (§16.8). Pre-apply versions are archived; edited chapters are promoted into the canonical filenames so the live name always holds the current truth.

13.9 Merge and Release

The compile step is deterministic: it strips provenance comments, builds front matter from the state file, concatenates chapters in reading order, verifies the chapter sequence, and reports the word count. The compiled manuscript is verified on disk — existence, outline match, word range, in that order — before any metadata is reconciled to it. Only then does status advance toward human review, which also requires the genre-alignment gate (§16.9).

13.10 Session Logging and Memory Banking

Sessions close with durable records: dated session logs in the archive, per-book emotion and humor logs (which capture the operator's chapter-level reactions as training verdicts), and doctrine records banked in the persistent memory core. Instruments built mid-run are committed as repository code before the session ends (§12.7). What is not logged cannot be learned from; the logging is part of the workflow, not overhead.

14. Data and Evidence Infrastructure

14.1 Versioned Manuscripts

No pass overwrites a draft. Originals stay untouched in `drafts/`; each editing stage writes new files, and every applied pass snapshots the prior version into `_archive/versions/<Run>/pre_apply/` before promoting the edited text into the canonical filename. The archive is the history; the canonical name is always current.

14.2 Parent-Child Rewrite Lineage

Because each stage writes forward rather than in place, every manuscript has a reconstructable lineage: raw generation output, loop-edited chapter, Spirit-applied chapter, compiled manuscript, with archived snapshots at each promotion. The lab directories preserve the same lineage for experimental arms — raw samples, per-tier edited sets, and closed-loop outputs side by side.

14.3 Model and Prompt Provenance

Each draft carries a gate-header comment recording its register gate and constraints, stripped only at compile. The harvested training records carry explicit `generator_model`, `editor_model`, and loop-type fields per pair, so every before/after text is attributable to the exact production recipe that made it.

14.4 Diagnostic Outputs

The deterministic layer produces inspectable artifacts, not just verdicts: the tic scanner's hit blocks, the stylometry checker's per-band report and attribution ranking, the diagnostic's per-word repair prescriptions, and the lab's raw score files (`scores.json`, `editor_ladder_scores.json`).

14.5 Rank and Gap Trails

Every scored sample records its attribution rank among the 27 pens, the top-ranked pen, the Delta to its own pen, and the gap to the winner. The lab reports movement in these terms — rank trajectories per intervention, gap-to-rank-1 per editing arm — because for short samples rank movement is the honest metric; absolute Delta values for 3,500-word one-shots sit far from every centroid relative to settled full-length books.

14.6 Before-and-After Training Pairs

Two harvested JSONL datasets exist as of 2026-07-18. The voice-carve pairs (`_training/voice_carve_pairs/2026-07-18_lab.jsonl`, produced by `harvest_lab_pairs.py`) hold 42 records — one per pen/genre/arm across three open-loop editor tiers and three closed-loop recipes, all seven genres of the Lince Starweave lab row — each with full before and after text, generator and editor model, loop type, and deterministically recomputed attribution scores (rank, top pen, own-pen Delta, gap, word count) on both sides. The declared-versus-measured harvest (`_training/voice_pairs/2026-07-18_declared_vs_measured.jsonl`) holds 27 per-pen records — declared voice-spine fields beside the measured fingerprint, 25 pens with declared spines, and 29 machine-checkable mismatch findings. Together they are the raw material for the future-work goal of training specialized editorial models.

14.7 Author Fingerprint Histories

`PEN_FINGERPRINTS.json` is generated, never hand-edited, by a deterministic stdlib-only script, and its header carries the generation note. Regeneration after each compile means the fingerprint record is dated and reproducible; a pen deliberately stretching its range moves its own bands on the next regeneration — the spine learning, recorded rather than overwritten silently.

14.8 Model Scorecards

Every model (or tier change) is profiled through the lab on three numbers: its attractor pen (whose shelf its unconstrained default hand lands on — frontier

Rance-adjacent, small model Lance-adjacent at the 2026-07-18 baseline), its one-shot voice-holding matrix ranks, and its closed-loop convergence speed as an editor (frontier: rank-1 in one to four iterations from either generator's drafts). The lab is the standing scorecard; no model takes a rung on the generation/editing ladder without these measurements.

14.9 Session-Level Evidence

The experimental record itself is versioned in the repository: `_lab/RESULTS.md` with its same-day addenda (the intervention rerun, the editor ladder, the closed-loop grid), the premise set, and every sample directory. Findings cite their raw score files; the results document is the narrative index over on-disk evidence, not a substitute for it.

14.10 Durable Memory Records

Doctrines distilled from measurements — the model economy, the closed-loop law, the self-upgrade rule — are banked as identified records in the persistent memory core and cross-referenced from the spine documents, so the conclusions survive any single session's context and later sessions inherit law rather than re-deriving it.

15. Economic Model

(Projected from measured behavior; dollar figures are not yet instrumented. No cost below is a price — each is a structural claim about where spend concentrates, derived from the measured trial. Instrumenting per-manuscript token and dollar accounting is open work.)

15.1 Cost of Low-Tier Drafting

Generation — the volumetrically largest output task, thousands of words per chapter — is delegated entirely to the cheapest model tier. Since the closed-loop

grid showed the generator's tier does not affect the measured outcome (7/7 versus 7/7), the largest token volume in the pipeline is purchased at the lowest available unit price with no measured quality penalty on attribution.

15.2 Cost of Frontier Editorial Passes

Frontier spend is confined to the brief and the closed loop, and the loop is measured as bounded: convergence in one to four iterations per sample. Frontier editorial cost per chapter is therefore capped by a small iteration count rather than open-ended, though the frontier editor reads and rewrites substantial text per iteration; the loop is where the dominant model spend concentrates, by design.

15.3 Reduction of Repeated Model Reasoning

Diagnosis, attribution, band compliance, tic scanning, word counting, and compilation are deterministic code. Each runs at effectively zero marginal cost, at every tier, on every pass — work that would otherwise be re-reasoned by a model on every chapter of every book. The checks being free is what makes running them *always* (every chapter, both bookends of every edit pass) economically trivial.

15.4 Code as Capitalized Intelligence

The self-upgrade loop (§12.7) is an investment mechanism: model reasoning that proves useful is converted once into a deterministic instrument and then reused indefinitely. `voice_diagnose.py` is the precedent — a mid-run invention whose cost was paid once and whose output now feeds every closed loop for free. Each such conversion permanently lowers the marginal cost of every subsequent manuscript.

15.5 Voice Quality per Dollar

On the measured outcome, the cheap-generation recipe and the all-frontier recipe are tied (7/7 rank-1 each). Whatever the eventual instrumented prices, the

ratio of measured voice quality to generation spend is strictly higher for the cheap-generation recipe, because the numerator is equal and the generation denominator is smaller. The earlier open-loop evidence points the same direction: cheap generation plus top-tier editing beat frontier one-shot generation on median rank (8 versus 20) even before the loop closed.

15.6 Marginal Cost per Manuscript

Projected composition of a manuscript's marginal cost: a small frontier component (briefs plus bounded loop iterations per chapter, plus the Spirit sweep), a large-volume but low-rate generation component, and a near-zero deterministic-check component. Marginal cost is therefore governed by frontier iteration count — which the system minimizes structurally, since better briefs, executable homebases, and per-beat scaffolds reduce the distance the loop must close. Not yet instrumented.

15.7 Scaling Across Books, Authors, and Genres

The expensive assets — fingerprints, diagnostics, checkers, doctrine — are house-level, not book-level: they amortize across every pen, genre, and title. Adding a book adds only marginal cost (§15.6); adding a pen adds a fingerprint once its shelf exists. The measured genre finding (a weak stylometric force relative to authorial hand) suggests the same apparatus spans genres without per-genre rebuilds, subject to the content-entanglement caveat for pens whose identity rides on subject matter.

15.8 Comparison With Frontier-Only Production

Frontier-only production spends top-tier rates on the highest-volume task the trial showed the cheapest tier can perform. Per §2.5, single-pass frontier framing is misleading anyway: nothing ships unedited, so the editing spend exists in both models, and the measured comparison is between identical outcomes at unequal generation spend. The frontier generator's one advantage — fewer generator-class defects (duplicated scenes, continuity arithmetic) for the loop to repair — is

real and would appear in instrumented iteration counts, but did not change the converged result.

16. Quality Assurance

16.1 Stylometric Attribution Gate

The primary voice gate is blind attribution against all 27 pen centroids plus per-pen style-band compliance, with graded verdicts: PASS (nearest centroid is the book's own pen), DRIFT (own pen ranks second within a 0.05 gap — a range verdict, not a failure), FAIL (the text landed on another pen's shelf), and PROVISIONAL (a one-book pen has no band to fail). A FAIL is evidence about the voice, not proof the book is bad — the verification law applies, and the response is diagnosis of *why* before anything is touched.

16.2 Canon and Continuity Checks (Ledger + Support-Spine Family)

The book ledger is live canon during drafting: facts, plants and payoffs, motifs, point-of-view lanes, world rules, and who-knows-what. The continuity support-spine family turns fact-class checking into lookup — a Timeline Spine (every stated duration derived from a master clock, never improvised), an Object-State Spine (injuries, wardrobe, recurring props checked against last-known state), and a Geography Spine (distances and travel times derived, not guessed). Contradictions are fixed in the chapter that created them, never deferred to the sweep.

16.3 Arithmetic and Factual Verification

Duration and age claims are grep-detectable and are scanned per chapter, then reconciled against the timeline spine by derivation. The trial demonstrated the need concretely: cheap-generator drafts contained broken interval math and age-

continuity errors, which the loop editors caught and repaired — quality value not captured by the attribution score.

16.4 Duplicate-Scene Detection

Duplicated passages are a known generator-class defect at the cheap tier (one full duplicated scene appeared in the lab's Haiku drafts). Detection is currently a loop-editor responsibility — repair on sight during the closed loop — rather than a standalone deterministic scan; the Spirit sweep provides the second net.

16.5 Structural Integrity Review

Structure is verified against the outline (every promised chapter and beat present, in order), against the composition plan, and at compile (chapter-sequence verification). Chapter-level structural checks include the opening-lean continuity rule (each chapter opens on the pressure the previous one left hanging) and the blocking law (physical continuity in multi-character scenes).

16.6 Word-Count Ranges and Floors (±2% Judgment Margin)

Word counts are never single numbers. The four counting methods in house use disagree by up to 2.04% (median 0.97%) across the 74-manuscript corpus, so every target is written as a range, every reported count is a range across methods, and no boundary case is ruled on a margin under ±2% of the canonical count — a finding that holds under only one counting method is method noise, grounds for further check rather than a confident fail. The closed-loop trial applied this law directly: the one cell sitting 1% under a floor by one method was ruled inside the band.

16.7 Prose-Quality Review (Kill-List + Judgment Passes)

Prose quality runs on two layers: a deterministic scanner for the kill-list tic families (the single source of truth, shared by draft-time QA and the Spirit suite's line-carving front end — a new tic family is added to the scanner once and both stages inherit it), and a by-hand judgment pass for what regex cannot catch

(editorializing tails after images, stilted syntax, register-mismatched cadence, emotionally inert beats). One honest boundary from the trial: manuscript-quality safeguards during carving were instructed and swept, but independent literary evaluation of the converged manuscripts remains open work — attribution PASS is a voice verdict, not a literary one.

16.8 Regression Testing (Before/After Bookends on Every Editing Pass)

Editing is where voice erodes silently, so every Spirit sweep — and any pass that rewrites prose — is bookended by the stylometric check: rank and own-pen Delta recorded before, re-run after applying edits. The edited book must not lose attribution to its own pen, and its Delta must not worsen beyond the noise margin; an edit that drags the book toward the house centroid or another pen's shelf is a defect *of the pass*, caught deterministically and repaired closed-loop before compile. This converts the open-loop ladder's central hazard — that editing can be net harmful for voice — into a gated, measurable regression test.

16.9 Human Escalation Conditions

Automated passes escalate rather than decide in defined cases. The genre-alignment gate is mandatory before human handoff, and its high-severity findings are never auto-applied — they block until the operator fixes or explicitly waives them, because they are creative judgment calls. Published, frozen books are untouchable by any automated pass; apparent staleness there is flagged, never silently corrected. Boundary verdicts — DRIFT attributions, word counts inside the ±2% band, findings that hold under only one method — escalate to further checking rather than hard failure. And under the verification law, no metadata signal alone condemns a book: a human-relevant judgment is always made on the read content, with the automated trail as evidence.

(Sections 17–22 of "Quality Moves Into the Architecture." Evidence base: `_lab/RESULTS.md`, `STYLOMETRY_SPINE.md`, *and the tool docstrings in*

_MASTER_ARCHITECTURE/stylometry/.)

---

17. Governance and Authorship

17.1 Human Creative Direction

Every book in the system originates in human creative decisions: the premise, the outline, the pen-name identity, the series canon, and the acceptance of the finished work. Models draft and edit; deterministic code measures and gates; the human operator decides what gets written, in whose voice, and whether the result ships. The architecture concentrates human attention where it is irreplaceable — invention, taste, final judgment — and removes it from repeated mechanical verification. Direction flows one way: from declared intent into generation, never from generated text back into canon without review.

17.2 System-Level Editorial Control

Editorial standards are not enforced by asking a model to behave. They are enforced by the system: the closed loop terminates only when the deterministic gate passes, editing passes are bookended by before-and-after measurements, and a pass that degrades the voice is caught as a defect of the pass rather than absorbed into the manuscript. Control relocates from individual model runs, which vary, into architecture, which does not; a recurring defect class is fixed once, in shared instructions or code, rather than patched per book.

17.3 Model Interchangeability

No component of the system depends on a specific model. Models are slotted into roles — drafting labor, editorial judgment, mechanical cleanup — according to measured capability, and the measurement protocol (Section 6.7, Section 18.1) exists precisely so that a new model can be profiled and routed without redesigning the pipeline. In the head-to-head trial, swapping the generator

between a frontier-family model and the smallest tier produced identical final attribution given the same closed loop. The system's identity lives in its corpus, its fingerprints, its gates, and its documented process — not in any vendor's checkpoint.

17.4 Provenance and Auditability

Every manuscript carries a reconstructible history: versioned drafts archived rather than overwritten, parent-child rewrite lineage, the model and prompt used at each stage, diagnostic outputs, and the rank-and-gap trail of every gate run. The fingerprint file records its own generation date and source corpus. When a question arises — which pass introduced a change, whether an edit degraded the voice — the answer is on disk, not in anyone's memory; claims about how a book was made can be checked.

17.5 Author Identity Preservation

The pen names are house-constructed identities, owned by the publisher, with declared voice documents and measured fingerprints. The stylometric layer keeps the identity a reader encounters consistent across a shelf: a book published under a pen name must attribute to that pen name under blind measurement, and editing passes are barred from silently normalizing the pen's habits away. Where declared and measured voice disagree, the measurement — what the reader actually receives — is the ground truth, and the declaration is reconciled to it.

17.6 Public AI Disclosure Practice (voluntary, on every published work)

The house discloses AI involvement on every published work, voluntarily. A canonical disclosure line appears on the copyright page, in the retail description, in the author bio, and on the series page. This is a governance practice, not a legal position: the disclosure states that the work was produced under human creative direction with AI models used as instruments, and the provenance

infrastructure of Section 17.4 stands behind the claim. A production system built on measurement and auditability should not carry an undisclosed step at the point of sale.

17.7 Responsibility for Final Publication

Responsibility for every published work rests with the human publisher. No manuscript reaches publication on automated approval alone: the pipeline's terminal gates include human review conditions, and creative-judgment findings (genre alignment, payoff, content decisions) are never auto-applied — they block until a human resolves or waives them. The system reduces the volume of judgment required of the publisher; it does not transfer accountability to the tooling.

17.8 Corpus Ownership and Data Boundaries (owned and public-domain text only)

The training and comparison corpora consist exclusively of two sources: manuscripts owned by the house (74 compiled books, 2.01 million words, across 27 pens) and public-domain texts (31 canon anchor novels by 20 authors, published 1810–1926). No third-party copyrighted material is ingested into the fingerprints, the diagnostic tooling, or the before-and-after training pairs. The public-domain canon is used for structural baselines only — how strongly real authors cohere, how weak genre is as a signal — never as a voice source. This boundary is a design constraint, not an accident of convenience, and any extension of the system to external corpora (Section 19.5) would require the same ownership-or-public-domain standard.

---

18. Limitations

The findings in Sections 9–11 are real but bounded. This section states the bounds plainly.

18.1 Minimum Capability Floors

The closed loop does not make every model sufficient. The smallest-tier model, editing its own drafts with the diagnostic's exact per-word prescriptions in hand, converged zero of seven samples: it executed word swaps but could not recast sentences, moving at best from rank 5 to rank 3 with one regression. Mid-tier open-loop voice editing was net harmful — it made attribution worse in three of seven cells by carving samples toward other pens' shelves. And open-loop editing at every tier tested left the generator's function-word residue in place: the small model's attractor pen remained the top attribution in nearly every open-loop-edited cell, across all three editor tiers. Only the top-tier editor running the closed loop erased that watermark. Voice-carve editing is a measured capability floor, not a prompt-engineering problem: instructions plus diagnostics did not substitute for it.

18.2 Dependence on Reference Corpus Quality

Every verdict the gate issues is relative to the fingerprints, and the fingerprints are only as good as the compiled shelf they are built from. Pens with one book are provisional — they report but never fail, because one book is a data point, not a voice. Several pens currently have no measurable identity at all (separation near zero from the house average). A known artifact remains in the tell-word lists: proper nouns are not stripped, so a long series can push a character name into a pen's tells. A corrupted, thin, or unrepresentative shelf produces a confidently wrong gate.

18.3 Risk of Overfitting to Stylometric Measures

The closed loop optimizes attribution against a 150-most-frequent-word Delta block and a set of style bands. It is possible in principle to satisfy the checker while degrading the prose — to game the metric. The system carries safeguards (quality sweeps during carving, kill-list passes, human escalation), and the loop editors were instructed to preserve story and prose while carving, but **no systematic evidence against metric gaming has been collected**. An earlier draft of this paper claimed such evidence; the claim was cut because it is not

held. Guarding against it is future work (Section 21.2), not a demonstrated property.

18.4 Genre, Tense, and Era Confounds

Genre measured weak in the canon baseline (same-genre Δ0.82 versus cross-genre Δ0.89), but weak is not zero, and other confounds are stronger. Tense and point of view are dominant voice carriers: any present-tense text drifts toward the house's present-tense gravity-well pen regardless of who wrote it. Content entanglement is real — one pen's fingerprint rides on a noun field bound to its settled series subject matter, so a cross-genre premise modulates the fingerprint no matter what the voice specification says. And every cross-cloud comparison against the public-domain canon carries a standing era confound: the anchors are 1810–1926 prose measured in a joint feature space with contemporary genre fiction.

18.5 Solo-Pen Data Scarcity

The fingerprint model assumes a shelf. For pens with one or two books, bands are undefined or fragile, attribution is provisional, and the regeneration cadence cannot help until more books compile. A meaningful fraction of the 27 pens sit in this regime. Results reported at the corpus level (77% blind attribution) are carried disproportionately by the pens with deep shelves.

18.6 Single-Pen Scope of the Closed-Loop Trial (multi-pen replication pending)

The headline closed-loop result — both generators converging to 7/7 rank-1 with the top-tier loop editor — was demonstrated on **one pen across seven genres**. It has not been replicated on a second pen. The pen used has known properties (a rebuilt, executable voice specification) that may flatter the result. Until the trial is rerun on pens with different identity mechanisms — including a content-entangled pen and a provisional pen — the closed-loop finding generalizes by argument, not by measurement.

18.7 Short-Sample Effects (lab samples vs settled full-length books)

All lab samples are 3.2–3.5 thousand words; the compiled shelf is 15–40 thousand per book. Short one-shot samples sit far from every centroid (nearest-pen Delta roughly 0.86–1.24, versus roughly 0.6–0.8 for compiled books) — they are outliers relative to any settled manuscript, which is why rank movement, not absolute Delta, was the honest metric throughout the lab. Whether closed-loop convergence behaves the same on a full-length accumulating manuscript is asserted by the drift-check design, not yet demonstrated at book length under lab conditions.

18.8 Coordination-Layer Context Growth

The orchestration layer that assigns drafts, routes editorial loops, and reconciles results accumulates context as a book proceeds. Worker contexts are kept lean by design, but the coordination layer's growth over long runs is an observed operational pressure, not a solved problem, and no formal measurement of its cost or failure modes has been made.

18.9 Attribution Without Literary Quality

This is the most important limitation. **Rank-1 attribution means the text lands on its own pen's shelf. It does not mean the book is good.** No independent literary-quality evaluation of the converged manuscripts has been performed — no blind reader panel, no external editorial review of the closed-loop outputs. The loop editors incidentally repaired real defects (duplicated scenes, broken continuity arithmetic) and that editing value is not captured by the attribution score in either direction. All lab samples remain quarantined; nothing produced in the trials has shipped. The system currently proves voice consistency at scale; it does not yet prove quality at scale.

18.10 Limits of Automated Judgment

Deterministic gates verify what can be counted. Genre satisfaction, emotional payoff, whether an ending lands — these remain judgment calls that the pipeline deliberately routes to a human. The escalation conditions exist because the automated layer's confident verdicts are only as broad as its instruments, and the instruments measure style, structure, and consistency, not worth.

---

19. Generalizability

19.1 New Pen Names

A new pen onboards by declaration first, measurement second: a voice document defines the intended hand, the first compiled book creates a provisional fingerprint, and the pen graduates to full gating as its shelf deepens. The one-shot matrix sets the standard for the declaration — write it to the executable grade (locked point-of-view mechanism, numeric rhythm specification, concrete mechanical habits), because that grade demonstrably transfers and the aspirational-prose grade does not.

19.2 Across Genres

Cross-genre transfer divides by identity mechanism, not by effort. Pens whose identity is mechanical — fixed by point of view, pronoun profile, and rhythm — held across all seven genres from a declaration alone. Pens whose identity is entangled with settled subject matter carried their lane with them, closing the style-band gap but not the function-word gap. A house extending across genres should therefore build pens on mechanical identities where cross-genre range is wanted, and accept lane-bound pens as lane-bound.

19.3 Evaluation of Future Models

The model evaluation protocol generalizes directly: profile any new model's attractor pen (whose shelf its default hand lands on), its one-shot voice-holding matrix, and its closed-loop convergence speed as an editor. Three numbers, produced by the existing lab, slot the model into the routing table without touching the rest of the system. The protocol also gives structure upgrades a measurable bar — an upgrade succeeds when it beats the model's attractor.

19.4 Small-Press Production

The framework's inputs — a corpus of owned manuscripts, stdlib-only scripts, and access to at least one high-capability editing model — are within reach of a small press. The economics of Section 15 remain projected rather than instrumented, so cost claims should be read as directional. The governance layer (disclosure, provenance, human final responsibility) transfers without modification.

19.5 Ghostwriting and Collaboration

A living author's corpus could serve as the reference shelf, with the closed loop converging drafts to that author's measured fingerprint. The technique transfers; the governance must transfer with it — consent, corpus ownership by agreement, and disclosure. The Section 17.8 boundary (owned or public-domain text only) is the template: no fingerprint is ever built from a corpus the operator does not have rights to.

19.6 Translation and Adaptation

Function-word fingerprints are language-specific, so translation requires rebuilding the feature space per target language. The architecture — fingerprint, gate, diagnose, closed loop — carries over; the specific 150-word inventory does not. Adaptation within a language (abridgment, serialization, reading-level shifts) is a nearer application, since the drift check already measures whether an editing pass moved a text off its own shelf.

19.7 Long-Form Nonfiction

Nothing in the stylometric layer assumes fiction. A nonfiction house voice, a columnist's hand, or an institutional register can be fingerprinted from an owned corpus and gated the same way. The continuity apparatus (ledgers, canon checks) would be replaced by factual verification appropriate to the domain — a heavier, not lighter, requirement.

19.8 Other Style-Constrained Production Systems

The general pattern — cheap generation, deterministic measurement, expensive judgment applied narrowly in a closed loop against the measurement — applies to any production system where output must land inside a measured stylistic envelope at scale: brand and marketing copy, documentation suites, localization QA. The transferable insight is architectural: when the target can be measured, the loop against the measurement is the primary quality mechanism, not the generator's pedigree.

---

20. Implementation Framework

This section describes what an adopter actually needs. The reference implementation is three stdlib-only Python scripts and a JSON artifact; no external dependencies, no services.

20.1 Required Inputs

Four things: (1) a reference corpus of compiled manuscripts attributable to named voices; (2) a declared voice document per voice, written to the executable grade where possible; (3) premise and outline inputs for new work; (4) access to at least one model measured capable of closed-loop voice editing, plus any cheaper model for drafting.

20.2 Minimum Reference Corpus

The tooling enforces a floor of 3,000 words per measured sample; below it, fingerprints are unstable and the scripts refuse to score. Per voice, one book yields a provisional fingerprint (reports, never fails); style bands and reliable attribution require multiple books. The reference deployment's corpus is 74 manuscripts and 2.01 million words across 27 voices — far more than a minimum, and Section 18.5 documents how results thin out on shallow shelves. An adopter should expect provisional-grade gating until each voice has several settled books.

20.3 Fingerprint Construction

One deterministic command (`build_fingerprints.py`) scans every compiled manuscript on disk, skipping archives and legacy folders, and writes the fingerprint file. Each fingerprint has two blocks: a Delta block — per-1,000-word rates of the corpus's 150 most frequent words, z-scored against the whole corpus, compared by Burrows' Delta — and a style block covering sentence rhythm, punctuation rates per 1,000 words, dialogue ratio, word length, and standardized type-token ratio. The output is generated, dated, and never hand-edited. Regenerate after any new compile, and before trusting a check on any voice whose shelf changed.

20.4 Diagnostic Tooling

Two tools sit on top of the fingerprints. The **checker** (`stylometry_check.py`) takes a voice name and a file or folder and reports attribution (nearest voice of all fingerprinted voices, with rank and gap) plus every style metric against the voice's band; it exits 0 on PASS/DRIFT/provisional, 1 on FAIL, 2 on usage error, so it composes into scripts and gates. The **diagnostic** (`voice_diagnose.py`) answers *why* a failing draft is being captured: it ranks the function words pulling the text away from the declared voice and toward the capturing rival — each word's current rate, shelf-implied target rate, direction, and an approximate occurrence count to add or remove — then lists the largest residual gaps and the out-of-band

style metrics with targets. The diagnostic always exits 0: it prescribes, the checker gates.

20.5 Editorial Loop Configuration

The validated loop: draft (cheap tier) → diagnose (code) → carve the named words in-voice (top tier) → check (code) → repeat until PASS or DRIFT, bounded at four iterations, escalating to a human on non-convergence. The best-scoring version is retained even when a later iteration regresses. Two configuration rules are load-bearing: the carving pass must be a model measured capable of it (Section 18.1 — mid-tier open-loop carving is worse than not editing), and every editing pass of any kind is bookended by checker runs so that voice-degrading edits are caught as defects of the pass.

20.6 Model Routing Table

Roles, not models, are fixed: drafting on the cheapest tier that clears the structure in use; diagnosis and gating in code at zero model cost; voice carving on the top measured tier only; mechanical sweeps on cheap tiers. Populate the table empirically via the evaluation protocol (Section 19.3) — attractor pen, one-shot matrix, closed-loop convergence — and re-profile whenever a model or tier changes.

20.7 Acceptance Thresholds

Attribution: PASS requires the declared voice at rank 1; DRIFT is rank 2 within Delta 0.05 of the winner and is treated as a range verdict, never a hard failure. Style-band flags mean "off this voice's shelf," not "wrong." Word counts are governed by the measurement law: all targets are written as ranges, and no boundary case is ruled on a margin under ±2% of the canonical count, because the counting methods themselves disagree by up to that much. The general principle: never rule a boundary case on a margin smaller than the instrument's own noise.

20.8 Logging and Version Control

Never overwrite a draft: archive the current version, then promote the corrected content into the canonical filename, so the live name always holds the current truth and the archive is the history. Record model and prompt provenance per pass, keep the rank-and-gap trail of every gate run, and bank before-and-after pairs from successful edits — they are training data (Section 21.11). The fingerprint file's header carries its generation note so a stale spine is detectable.

20.9 Deployment and Maintenance

Steady-state maintenance is small: regenerate fingerprints after each compile; run the checker as the tail of drafting and the bookends of editing; re-run the model scorecard when the model roster changes; and treat any instrument an agent builds mid-run as a candidate for permanent code — the diagnostic tool itself originated that way. Drift in the spine is expected: a voice deliberately stretching moves its own bands at the next regeneration. The failure mode to watch is silent divergence between declared and measured voice; the standing rule is that the measurement wins and the declaration is reconciled to it.

---

21. Future Research

21.1 Multi-Pen Replication of the Closed-Loop Trial. Rerun the head-to-head grid on additional pens — at minimum one content-entangled pen and one provisional pen — to convert the single-pen result into a system-level claim.

21.2 Independent Literary-Quality Evaluation of Converged Manuscripts. Blind quality assessment of closed-loop outputs by evaluators with no access to the attribution scores, directly addressing Sections 18.3 and 18.9 — including whether loop-converged text is ever stylometrically right but worse prose.

21.3 Sentence-Level Repair Prediction. The diagnostic currently prescribes at the word level; predicting which sentences to recast would target the

operation the small model could not perform and might lower the carving capability floor.

21.4 Automated Model Routing. Drive the routing table directly from scorecard output, so a new model is profiled and slotted without manual judgment.

21.5 Genre-Adjusted Author Centroids. Partial out the measured (weak but nonzero) genre signal to sharpen cross-genre attribution for content-entangled pens.

21.6 Dialogue and Narration Fingerprints. Separate fingerprints per channel; the dialogue-ratio misses in the intervention experiment suggest the channels carry distinct signals now blended.

21.7 Historical Voice Evolution. Use dated fingerprint regenerations to track how a pen's measured hand moves across its shelf over time, distinguishing deliberate stretch from erosion.

21.8 Cross-Model Editorial Bias. Each model has a measured attractor pen; characterize how each editor tier's attractor biases its edits, and whether pairing generator and editor with different attractors cancels or compounds residue.

21.9 Quality-Cost Optimization. Instrument actual dollar costs per manuscript across recipes to replace the projected economics of Section 15 with measured ones.

21.10 Self-Updating Author Fingerprints. Automate the regeneration cadence into the compile step, with alerts when a new book moves its pen's centroid beyond the noise margin.

21.11 Training Specialized Editorial Models. The banked before-and-after pairs (42 scored edit pairs; 27 declared-versus-measured records) are a growing supervised dataset for training or fine-tuning a dedicated voice-carving editor — potentially lowering the top-tier-only floor that currently makes editorial judgment the scarce resource.

---

22. Conclusion

22.1 Quality as an Architectural Property

The central finding of this work is that voice quality is not a property of the generating model. Stripped of its pipeline, every model tested collapsed to its own default hand; inside the closed loop, the cheapest generator matched the most expensive one exactly, 7/7 to 7/7. Voice lives in the pipeline — the fingerprints, the diagnostic, the gate, the bounded loop — and a property that lives in architecture can be versioned, audited, and improved.

22.2 The Validated House Economy

The measured division of labor stands: the smallest model drafts, deterministic code diagnoses and gates at zero marginal reasoning cost, and the top-tier model spends its expensive judgment only on the one operation nothing cheaper can perform — carving prose to a measured voice. Each claim in that sentence is backed by a specific experimental cell, and each has a stated boundary in Section 18.

22.3 Cheap Generation, Expensive Judgment, Deterministic Gates

The pattern generalizes beyond this house. Wherever a target can be measured, the loop against the measurement — not the generator's pedigree — is the primary quality mechanism. Drafting capability is widely available; convergent editorial judgment is scarce; measurement is free. Systems should be built in that order.

22.4 A Reproducible Model for Scalable Authorship

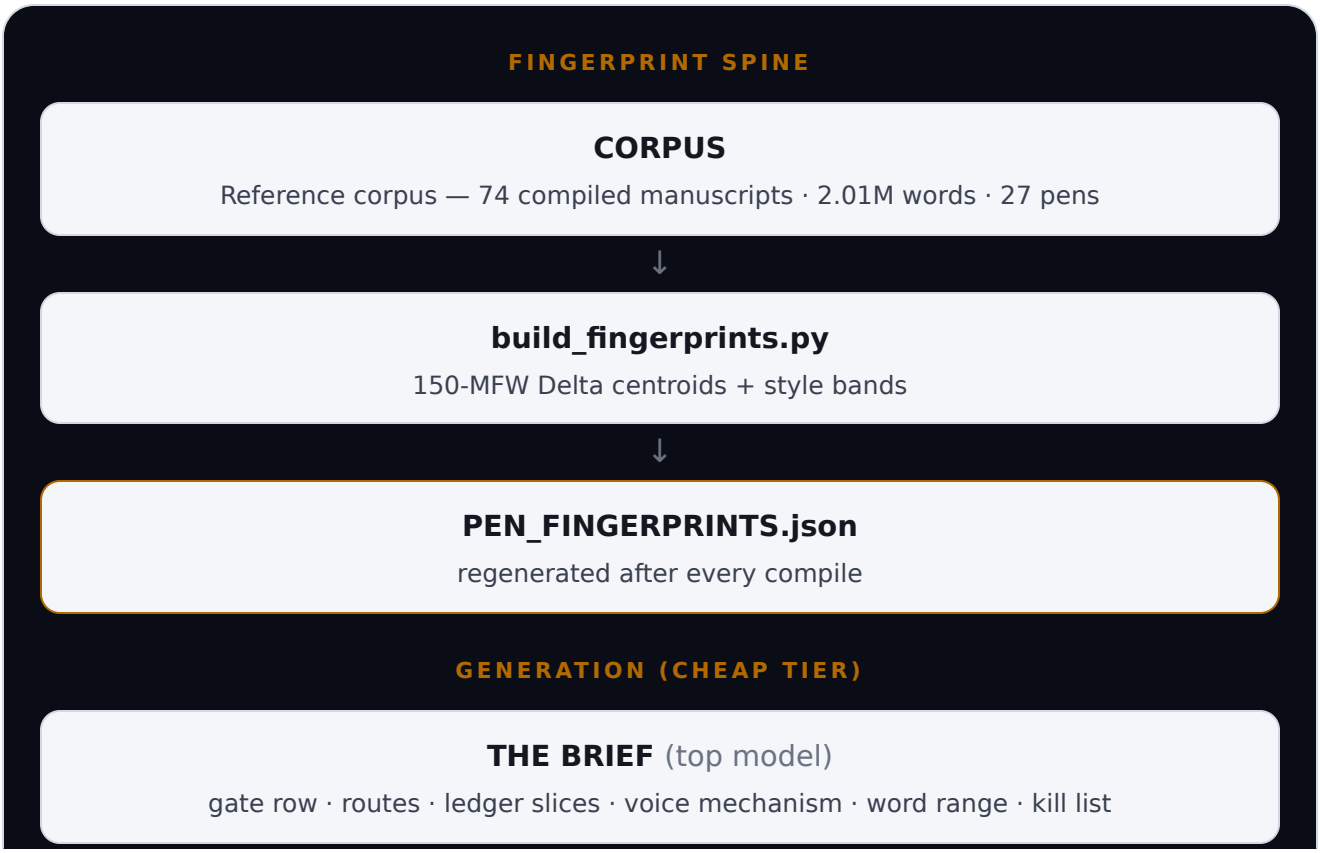
The full apparatus is three deterministic scripts, a generated fingerprint file, an owned corpus, and a documented loop — reproducible by any operation with its own manuscripts and the discipline to gate every pass. What remains open is

stated plainly in Sections 18 and 21: multi-pen replication and independent quality evaluation. What is closed is this: consistent authorial voice at production scale is an engineering property, and it has been engineered.

Generated 2026-07-18 from the live data files in `_MASTER_ARCHITECTURE/stylometry/` (scripts, `PEN_FINGERPRINTS.json`, `_lab/scores.json`, `_lab/editor_ladder_scores.json`, `_lab/RESULTS.md`, the on-disk closed-loop outputs). Tables marked "recomputed" were rescored from disk at build time with the same attribution math the gate uses — accurate by construction, not transcription.

Appendix A. System Diagram

The production loop, as actually wired (`.claude/skills/chapter/SKILL.md` step 3; `STYLOMETRY_SPINE.md` "WHERE IT RUNS").



↓

THE GENERATOR (Haiku subagent)

drafts raw chapter from the brief only

↓

Raw draft

raw material, never the deliverable

↓

THE CLOSED LOOP — TOP MODEL, MANDATORY

Edit: house sweep + voice carve

+ repair generator defects on sight

↓

voice_diagnose.py

PULL / GAP / STYLE prescriptions ←·· PEN_FINGERPRINTS.json

↓

stylometry_check.py

rank of 27 · gap · style bands ←·· PEN_FINGERPRINTS.json

↓

PASS or DRIFT?

↑ no — iterate (≤ 4 rounds) back to Edit

↓ yes — keep best version

/draft-pass tail

tic-check · CAP check · canonical count · commit

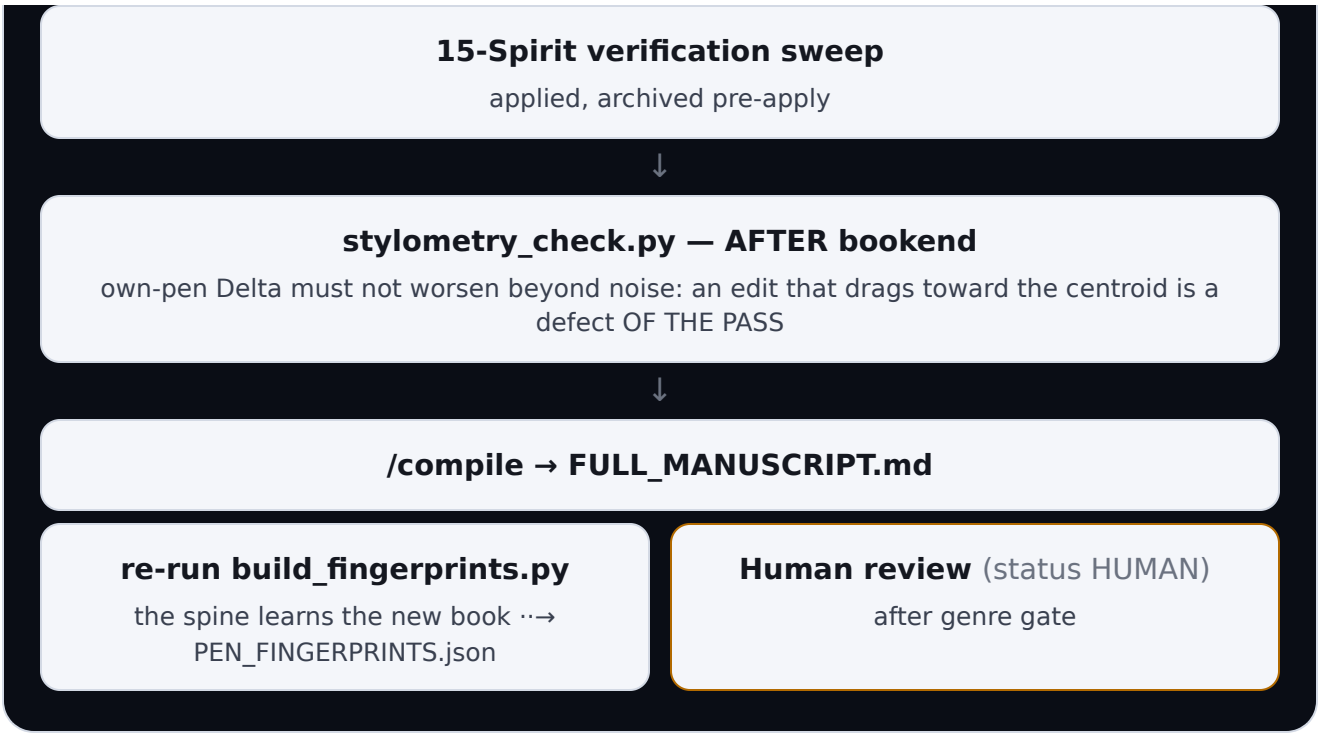
↓

...repeat per chapter until book complete

BOOK-LEVEL VERIFICATION

stylometry_check.py — BEFORE bookend

↓



Appendix A, rendered from the source flowchart. The dotted edges (←··) mark the fingerprint file feeding the diagnostic and checker, and the regenerated spine folding the new book back in.

Appendix B. Closed-Loop Pseudocode

Thresholds are the shipped values in `stylometry_check.py`; iteration bound and keep-best rule are the lab-validated procedure (`_lab/RESULTS.md`, final grid: convergence in 1–4 iterations from either generator).

```
draft      = generate(brief, model=LOWEST)          # raw material, never deliverable
best       = score(draft)                           # (rank, gap, delta_own, bands)
MAX_ITER   = 4

for i in 1..MAX_ITER:
    rx     = voice_diagnose(pen, draft)             # PULL words, GAP words, STYLE misses
    draft  = carve(draft, rx, model=TOP)            # house sweep + carve the named words,
                                                    # in-voice; fix generator defects on sight
    s      = stylometry_check(pen, draft)           # deterministic; code, not judgment
    if s.delta_own < best.delta_own: best = draft   # keep best — never regress the keeper

    if s.rank == 1:                    verdict = PASS   ; break
    elif s.rank == 2 and s.gap <= 0.05: verdict = DRIFT ; break   # range verdict, not a RED
    # else continue

if verdict not in (PASS, DRIFT):
```

```
    retain(best); escalate()                        # FAIL after 4 rounds = human/architect call
else:
    retain(best); proceed_to_draft_pass()
```

Notes: `score` attributes against all 27 pen centroids (Burrows' Delta on the 150-MFW z-profile); diagnosis is code (free at every tier); carving is top-rung judgment work — the measured floor: Haiku executes word swaps but cannot recast sentences, and did not converge on itself (0/7).

Appendix C. Model Scorecard

The three-number protocol (`STYLOMETRY_SPINE.md`, MODEL EVALUATION PROTOCOL), with the 2026-07-18 measured baselines. Every new model or tier change gets these three numbers before it is trusted with any rung of the ladder.

#	Number	What it measures	How
1	Attractor pen	The model's default hand — whose shelf its blind drafts land on	Draft blind samples, run the checker
2	One-shot voice-holding	Can it hold a declared pen without the pipeline	The 7-pen × 7-genre matrix, per structure revision
3	Closed-loop convergence as editor	Iterations to PASS with the diagnostic in the loop	`voice_diagnose.py` + `stylometry_check.py` loop

Measured baselines (2026-07-18):

Model	Attractor pen	One-shot (7×7)	As open-loop editor	As closed-loop editor
Fable	Rance-adjacent	13/49 PASS (frontier arm)	Only net-positive open-loop arm (median rank 8)	Converges rank-1 in 1–4 iterations from either generator's drafts — 14/14
Opus	— (not profiled as generator)	—	Mixed (median rank 13; some cells worse)	— (not run)
Sonnet	— (not profiled as generator)	—	**Net harmful for voice** — ranks worsened in 3/7 cells (Crime 9→23, Literary 4→20, Horror 5→15); pushed samples onto other pens	The one cell where it self-ran the checker reached DRIFT (Action, gap 0.044) — the finding that motivated the closed loop
Haiku	Lance-adjacent	0/49 PASS (1 DRIFT)	— (generator only)	**Does not converge on itself**: 0/7, best movement r5→r3, one regression; executes word swaps, cannot recast sentences

Standing verdict: the generator is negligible given the loop; the loop editor is not. Validated economy: Haiku writes, Fable loop-edits, code gates. Each attractor pen is what every structure upgrade must beat — rerun the lab per model, per revision.

Appendix D. Experimental Test Matrices

D.1 One-shot 7×7, frontier arm — 13/49 PASS

Cell = capturing pen and own-pen rank of 27 (from `_lab/scores.json`, `samples`). Samples 3.2–3.6k words, drafted blind from AUTHOR_HOMEBASE + premise only.

pen \ premise	Rom	SFF	Cont	Crime	Horr	Act	Lit
Vance Thorne	PASS	PASS	PASS	PASS	PASS	PASS	PASS
Lince Starweave	Rance r25	Rance r21	Rance r27	Rance r26	Rance r24	Rance r20	Rance r27
Tence Stonewell	Rance r18	Jænce r21	Rance r22	Rance r17	Rance r22	Vunce r25	Rance r24
Rance Graves	PASS	PASS	PASS	PASS	PASS	Vance r2	PASS
Munce Ashfall	Rance r12	Rance r15	Rance r11	Rance r7	Vunce r15	Vunce r18	Rance r9
Honce Flint	Rance r8	Vunce r14	Rance r6	Rance r11	Mænce r17	Rance r4	Rance r9
Jænce Beckett	Rance r15	Rance r4	Rance r9	Vance r4	Rance r15	Vunce r13	Rance r11

D.2 One-shot 7×7, Haiku arm — 0/49 PASS (1 DRIFT)

Same construction (from `scores.json`, `samples_haiku`).

pen \ premise	Rom	SFF	Cont	Crime	Horr	Act	Lit
Vance Thorne	Lence r12	Lance r17	Lance r4	Lance r14	Lance r14	Lance r13	Lance r10
Lince Starweave	Lance r9	Lance r2	Lance r14	Lance r9	Lance r4	Lance r4	Lance r6
Tence Stonewell	Lance r3	Lance r7	DRIFT	Lance r4	Lance r5	Lance r14	Lance r9
Rance Graves	Lance r15	Lance r20	Runce r16	Runce r17	Lance r10	Lance r18	Lance r19
Munce Ashfall	Lance r2	Lance r2	Lance r9	Lance r4	Lance r2	Lance r3	Lance r3
Honce Flint	Lance r2	Lance r10	Lance r13	Rynce r4	Lance r3	Lance r2	Lance r8
Jænce Beckett	Lance r5	Lance r7	Lance r9	Lance r11	Lance r7	Lance r7	Lance r7

Reading: each arm collapses to one attractor (frontier → Rance, Haiku → Lance). Noise scatters; this converges — the checker detected the generator's default hand showing through with the pipeline stripped away.

D.3 The closed-loop grid (Lince Starweave × 7 genres)

Recipe results (`_lab/RESULTS.md`, final section):

Recipe	Generator	Loop editor	Rank-1 PASS
Fable + Fable loop	Fable (frontier)	Fable, closed loop	**7/7**
Haiku + Fable loop	Haiku	Fable, closed loop	**7/7**
Haiku + Haiku loop	Haiku	Haiku, closed loop	0/7 (best r2; mostly flat)

Per-cell final scores, **recomputed from the on-disk loop outputs**: Appendix G.2.

Appendix E. Stylometric Feature Inventory

Everything actually computed, from `build_fingerprints.py`. Preprocessing for all features: strip `<!-- -->` comments, markdown headers, and `---`-style separator lines; NFC-normalize; word tokens = `[a-zA-ZÀ-ƒ']+`, lowercased; minimum 3,000 tokens per text.

E.1 The Delta block (attribution signal)

Element	Definition
MFW set	The 150 most frequent words across the whole reference corpus (dominated by function words — the deep, hard-to-fake signal)
Per-book profile	Each MFW's rate per 1,000 words in the book
z-scoring	Each rate standardized against the house: z = (rate − corpus mean) / corpus std, per word
Pen centroid	Mean z-vector over the pen's books (`delta_centroid`, 150 values)
Burrows' Delta	Distance between two profiles = mean of |z_i − z_j| over the 150 words. Attribution = nearest pen centroid by Delta

E.2 The style block (band signal)

Each metric is banded per pen as observed lo–hi ± margin, where margin = max(25% of the observed range, 15% of |mean|, 0.02). Sentence split: `[.!?]+` + optional closing quote + whitespace; lengths clamped to 1–199 words.

Metric	Definition
`sent_mean`	Mean sentence length in words
`short_ratio`	Share of sentences ≤ 8 words
`med_ratio`	Share of sentences 9–20 words
`long_ratio`	Share of sentences ≥ 21 words
`sent_std`	Population std of sentence lengths (rhythm variance)
`frag_rate`	Share of sentences ≤ 4 words (fragment reflex)
`dial_ratio`	Characters inside double-quoted spans (2–600 chars) ÷ total text characters
`avg_word_len`	Mean characters per word token
`sttr`	Standardized type–token ratio: mean TTR over successive 5,000-word segments (segments ≥ 2,500 words kept; whole-text TTR as fallback)
`em_dash`	(— + `--`) per 1,000 words
`semicolon`	`;` per 1,000 words
`ellipsis`	(… + `...`) per 1,000 words
`exclaim`	`!` per 1,000 words
`question`	`?` per 1,000 words
`comma`	`,` per 1,000 words
`colon`	`:` per 1,000 words

E.3 Per-pen derived fields (in `PEN_FINGERPRINTS.json`)

Field	Definition
`cohesion.within`	Mean Delta between the pen's own books
`cohesion.between`	Mean Delta from the pen's books to all other pens' books
`cohesion.separation`	between − within (identity sharpness; e.g. Munce Ashfall +0.66)
`tell_words`	Top 8 centroid-z words — the pen's most over-used common words vs the house
`avoid_words`	Bottom 5 centroid-z words — most under-used
`provisional`	True if the pen has fewer than 2 books (no band to fail)

Known artifact: proper nouns are not stripped, so a long series can push a character name into a pen's tells.

Appendix F. Diagnostic Output Format

Both tools are stdlib-only, deterministic, and machine-readable line formats. Examples below are genuine captured output (run 2026-07-18 on `_lab/loop_fable_from_frontier/Lince_Starweave__Romance.md`), truncated for length.

F.1 `stylometry_check.py` — the gate

Usage: `python3 stylometry_check.py "<Pen Name>" <file-or-folder> [...]` Exit codes: 0 = PASS/DRIFT/PROVISIONAL, 1 = FAIL, 2 = usage/data error.

```
ATTRIBUTION PASS  pen=Lince Starweave rank=1/27 delta_own=0.903 nearest=Lince Starweave delta_nearest=0.903 gap=
  * Lince Starweave      0.903
    Lance Starweave      0.977
    Vance Thorne         1.043
STYLE sent_mean         9.118  band [8.009 .. 12.509] mean 10.428
STYLE dial_ratio        0.199  band [0.197 .. 0.375] mean 0.298
```

```
STYLE em_dash           8.542  band [0.385 .. 9.808] mean 4.993
SUMMARY verdict=PASS out_of_band=none words=3629 provisional=False
```

Line grammar: one `ATTRIBUTION` line (verdict, own rank of 27, own Delta, nearest pen, its Delta, gap), the top-5 ranked pens (`*` marks the declared pen), one `STYLE` line per band metric (out-of-band lines append `<< OUT OF BAND`), one `SUMMARY` line.

F.2 `voice_diagnose.py` — the prescription

Usage: `python3 voice_diagnose.py "<Pen Name>" <file-or-folder> [...]` Exit code: always 0 (diagnostic, not a gate — the gate is `stylometry_check.py`).

```
ATTRIBUTION pen=Lince Starweave rank=1/27 delta_own=0.903 nearest=Lince Starweave delta_nearest=0.903 gap=0.000
# PULL: words feeding the Lance Starweave capture — carve these first (need = occurrences to add/remove in 3629
PULL word=through      pull=+2.80 text_per1k=  3.58 pen_per1k=  1.37 dir=lower need=  -8.0 have=13
PULL word=we           pull=+1.58 text_per1k=  1.10 pen_per1k=  4.27 dir=raise need= +11.5 have=4
PULL word=had          pull=+1.25 text_per1k= 17.36 pen_per1k=  7.07 dir=lower need= -37.3 have=63
# GAP: largest deviations from Lince Starweave regardless of rival
GAP  word=but          zdev= 2.71 text_per1k=  0.00 pen_per1k=  4.01 dir=raise need= +14.6 have=0
GAP  word=stood        zdev= 2.63 text_per1k=  2.76 pen_per1k=  1.15 dir=lower need=  -5.8 have=10
SUMMARY pen=Lince Starweave rank=1 gap=0.000 rival=Lance Starweave style_misses=0 words=3629
```

Line grammar: `ATTRIBUTION` (same math as the checker, plus the named `rival`); up to 15 `PULL` lines — MFW words ranked by pull = |z_text − z_pen| − |z_text − z_rival| (> 0 favors the rival), each with the text's per-1k rate, the pen's shelf-implied rate, a raise/lower direction, and the approximate occurrence count to add/remove at this text length; up to 15 `GAP` lines — largest |z_text − z_pen| regardless of rival (the residue after the rival flips); `STYLE` lines only for out-of-band metrics, with direction and target band; one `SUMMARY` line. "Need" counts are shelf-implied point estimates — aim points, not gospel (the word-count law).

Appendix G. Rank-and-Gap Evidence Tables

G.1 The open-loop editor ladder (from `_lab/editor_ladder_scores.json`)

The 7 Haiku-v2 Lince drafts, one editing pass per tier, identical prompts (full house sweep + fingerprint carve, targets supplied). Cell = own-pen rank of 27 / capturing pen / gap to rank-1.

Genre	Haiku raw	Sonnet-edited	Opus-edited	Fable-edited
Romance	15 / Lance / 0.253	20 / Lance / 0.208	18 / Lance / 0.153	**8** / Lance / 0.122
SFF	11 / Lance / 0.271	11 / Honce / 0.130	**5** / Lance / 0.158	6 / Lance / 0.082
Contemporary	14 / Lance / 0.347	**6** / Lance / 0.242	16 / Lance / 0.309	16 / Lance / 0.197
Crime	**9** / Lance / 0.208	23 / Runce / 0.241	13 / Lance / 0.183	16 / Lance / 0.178
Horror	5 / Lance / 0.227	15 / Munce / 0.115	6 / Lance / 0.132	7 / Lance / 0.132
Action	2 / Lance / 0.143	2 / Lance / **0.044**	2 / Lance / **0.020**	2 / Lance / **0.025**
Literary	4 / Lance / 0.242	20 / Honce / 0.197	21 / Lance / 0.270	**8** / Lance / 0.095
median rank	**9**	15	13	**8**

Findings carried: Sonnet net harmful (worse in 3/7, new captures by other pens); Opus mixed; Fable the only net-positive arm; the Lance skeleton survived every open-loop tier; the one DRIFT-band cell (Action, gap 0.044) came from the editor that ran the checker itself and iterated — the closed-loop finding.

G.2 Closed-loop final outcomes, per cell (recomputed from disk)

Attribution of each on-disk loop output against the unchanged shelf centroids (same math as the gate). Fable-loop arms: 14/14 rank-1, gap 0.000, style bands in.

Genre	Fable+Fable loop (rank / Δ-own / words)	Haiku+Fable loop	Haiku+Haiku loop (rank / top / gap)
Romance	1 / 0.903 / 3,629	1 / 1.042 / 3,083	19 / Lance / 0.245
SFF	1 / 0.922 / 3,498	1 / 0.879 / 3,072	11 / Lance / 0.263
Contemporary	1 / 0.921 / 3,503	1 / 0.846 / 2,971	14 / Lance / 0.347
Crime	1 / 0.827 / 3,559	1 / 0.888 / 3,108	9 / Lance / 0.209
Horror	1 / 0.644 / 3,614	1 / 0.498 / 3,058	3 / Lance / 0.180
Action	1 / 0.597 / 3,513	1 / 0.956 / 3,135	2 / Lance / 0.142
Literary	1 / 0.872 / 3,296	1 / 0.802 / 3,063	4 / Lance / 0.242

Convergence (from `_lab/RESULTS.md`): Fable loop converged in 1–4 iterations from either generator's drafts, fully erasing the Haiku watermark that survived every open-loop tier. Haiku self-loop: 0/7, movement r5→r3 at best, one regression — it executes the prescription's word swaps but cannot recast sentences. Haiku-arm word counts run lower; one cell (Contemporary, 2,971 by the lab tokenizer) sits ~1% under the 3,000 floor by one counting method — inside the ±2% band, so not ruled a violation (Appendix H).

Appendix H. Word-Count Measurement Protocol

The ±2% method-variance law (`CLAUDE.md`, VERIFICATION LAW — measured 2026-07-18).

The problem. A word count is never a single definitive number. The four reporting methods in actual house use disagree by design: `wc -w` under UTF-8 vs `LC_ALL=C` (locale-dependent; the same chapter counted 1,893 vs 1,862), any regex tokenizer vs Python `str.split()` (by construction), raw vs comment-stripped text.

The measurement. Across all 74 compiled manuscripts, the four methods (`wc -w` UTF-8, `wc -w` LC_ALL=C, `str.split()` raw, `str.split()` comment-stripped/canonical) disagree by **median 0.97%, p95 1.68%, max 2.04%** of the canonical count.

The protocol, house-wide:

Rule	Statement
Canonical method	`str.split()` on comment-stripped text (`draft-pass/word-count.sh`, same rule as `compile.py`) is the one method for `STATE.json`'s `current_words` — and even it is a point estimate
Judgment margin	±2% of the canonical count. No boundary case (under-floor, over-target, platform spec) is ever ruled on a smaller margin — inside ±2%, the methods themselves disagree, so the finding is method noise, not fact
Targets as ranges	Every target is written lo–hi, never a bare number ("3,300–3,700", not "3,500"). A bare-number target is malformed
Method binding	Floors and caps state which counting method they bind (default: canonical); compliance is judged with the ±2% margin on top
Reporting	When a count judges something, state it as a range spanning the methods actually run (e.g. "38,703–39,572 words"). A finding that holds under only one method is grounds for further check, never a confident violation
Remeasure trigger	Re-map the band if the method family changes; until then 2% is the number

The stylometric verdicts inherit the same shape: DRIFT is a range verdict (rank 2 within Δ0.05), never ruled on a hair margin.

Appendix I. Version-Lineage Schema

The two JSONL harvest schemas actually in use (`_MASTER_ARCHITECTURE/_training/`). Both harvesters are stdlib-only and deterministic; both append an entry to their directory's `_MANIFEST.md` per run. Dated snapshots are the point: the drift between snapshots is training signal.

I.1 Voice-carve pairs — `harvest_lab_pairs.py`

Output: `_training/voice_carve_pairs/<YYYY-MM-DD>_lab.jsonl` — one record per (pen, genre, editing arm); 6 arms × 7 genres from the 2026-07-18 lab (3 open-loop editor tiers + 3 closed-loop recipes).

Field	Type	Definition
`harvested`	date	Harvest date (ISO)
`pen`	string	Target pen (lab: "Lince Starweave")
`genre`	string	Premise genre (one of the 7)
`generator_model`	string	Drafting model family (`haiku` / `frontier`)
`editor_model`	string	Editing model family (`sonnet` / `opus` / `fable` / `haiku`)
`loop`	string	`open` (single pass) or `closed` (iterated against the gate)
`before_path`, `after_path`	string	Repo-relative paths to the exact input and output texts
`before_score`, `after_score`	object	Attribution recomputed deterministically on each side: `{rank, top_pen, delta_own, gap, words}`
`before_text`, `after_text`	string	Full text, both sides

I.2 Declared-vs-measured pairs — `harvest_voice_pairs.py`

Output: `_training/voice_pairs/<YYYY-MM-DD>_declared_vs_measured.jsonl` — one record per pen in `PEN_FINGERPRINTS.json` (27 records per snapshot).

Field	Type	Definition
`pen`, `press`	string	Identity
`harvested`	date	Snapshot date (ISO)
`books`, `words`	int	Shelf size behind the measurement
`provisional`	bool	True if < 2 books
`declared`	object | null	Raw WRITEDITOR STRAND 2 fields from `AUTHOR_HOMEBASE.md`, captured field by field; null if no homebase/section
`measured`	object	The stylometry spine slice: `style_bands`, `cohesion`, `tell_words`, `avoid_words`
`checks`	array	Deterministic comparisons where a declared claim is machine-comparable: `{claim, declared, measured, verdict}`

Check verdicts are ranges, never hair-margin rulings: sentence-ratio claims — MATCH ≤ 0.10 absolute gap, NEAR ≤ 0.20, else MISMATCH; punctuation-habit keyword claims (em dash / semicolon / ellipsis) — declared keyword (rare/moderate/frequent) vs the measured per-1k band (em dash: rare < 2.0, frequent ≥ 6.0; semicolon: 1.0 / 3.0; ellipsis: 0.5 / 2.0), with `moderate` on either side scoring NEAR; `UNPARSED` when a declared spine exists but no claim is machine-readable.

Appendix J. Terminology

Term	Definition
Pen (pen name)	One of the 27 active house author identities, each owning one genre lane at one press. A pen is a register configuration of A_0, not a separate writer
A_0 (THE ONE AUTHOR)	The single locked deep author identity that drafts every book, wielding the 49 pen registers through the gate (one dominant + ≤2 modulators per beat)
Fingerprint	A pen's measured profile: Delta centroid + style bands + tell/avoid words, derived from its verified compiled books (never declared, never hand-typed)
MFW	Most-frequent words — the corpus's top 150, dominated by function words; the attribution feature set
Delta (Burrows' Delta)	Distance between two texts' z-scored MFW profiles: mean absolute z-difference over the 150 words. Lower = stylistically closer
Centroid	A pen's mean z-profile over its books — the mathematical "shelf" a text is attributed to
Rank	Position of the declared pen when all 27 centroids are sorted by Delta to the text (rank 1 = nearest)
Gap	Δ-own minus Δ-nearest: how far the declared pen trails the capturing pen (0.000 at rank 1)
Attractor pen	The pen whose shelf a model's default hand lands on when the pipeline is stripped away (Fable ≈ Rance-adjacent, Haiku ≈ Lance-adjacent). What every structure upgrade must beat
Gravity well	A pen centroid that absorbs other pens' texts (e.g. any present-tense book drifts Vynce-ward) — check attribution before blaming the draft
PASS / DRIFT / FAIL	Gate verdicts: own pen rank 1 / rank 2 within Δ0.05 (a range verdict, not a RED) / captured by another shelf. A FAIL is evidence about the voice, not proof the book is bad

Term	Definition
PROVISIONAL	Verdict for a one-book pen: reports, never fails — one book is a data point, not a voice
Style bands	Per-pen observed lo–hi ranges (+ margin) for each style metric; out-of-band means "doesn't sound like this pen's shelf," not "wrong"
Tell / avoid words	The pen's most over-/under-used common words vs the house average
Open loop	Edit pass with targets supplied but no instrument in the loop — the editor never measures its own output. Measured net harmful at mid-tier
Closed loop	Edit → `voice_diagnose.py` → carve the named words → `stylometry_check.py` → repeat to PASS/DRIFT (≤4 iterations, keep best). The primary quality mechanism — 14/14 convergence
The flip	The 2026-07-18 doctrine inversion: generation runs on the lowest model from top-model briefs; the mandatory closed loop on the top model is what makes the cheap generator safe. "The house one-shots nothing, so the generator's tier cannot matter"
Generator watermark	The generation model's function-word residue, which survives any single open-loop edit at any tier — and is fully erased by the closed loop
Executable homebase	An author declaration written as mechanism (locked POV device, numeric rhythm spec, per-1k rates) rather than adjectives — the Vance Thorne standard; necessary but not sufficient for one-shot transfer
Mechanical-identity vs content-entangled pens	Pens whose fingerprint rides on a POV/tense mechanism transfer cross-genre (Vance); pens whose fingerprint is entangled with their settled subject register carry their lane with them (Lince)
Canonical word count	`str.split()` on comment-stripped text (`word-count.sh`); the one method for `STATE.json` — still a point estimate, judged under the ±2% law
±2% law	No word-count boundary case ruled on a margin smaller than 2% of the canonical count (the measured max method disagreement

Term	Definition
	across 74 books); all targets written as lo–hi ranges
Bookend (before/after)	The stylometry check run before and after any editing pass; own-pen Delta worsening beyond the noise margin marks the pass itself defective
Self-upgrade loop	The doctrine that an agent which builds a useful instrument mid-run banks it as repo code before its context dies (`voice_diagnose.py` is the precedent)
The shelf	A pen's set of compiled, verified books — the only ground truth a fingerprint is derived from ("the shelf is the truth")