Enforcing an Editorial Style Law Across a Million-Word Catalog with Tiered Language Models

A field report on a model-agnostic detect / judge / gate / apply pipeline

July 2026

Abstract

We describe a working pipeline that enforced a human editor's prose-style rules across a catalog of 47 book-length manuscripts (1.13 million words) in a single working session, using three capability tiers of commercial large language models in different roles. The pipeline found 801 candidate violations, made 324 verified edits, and deliberately preserved 477 flagged instances judged legitimate — with a final human-gate correction rate of roughly 1%. Three findings may generalize. First, **the system's intelligence should live in artifacts, not in any model**: rules as scripts, taste as exemplar pairs, judgments as an accumulating written brief — making every model an interchangeable part. Second, an editorial task decomposes into stages of very different sensitivity to model quality — **detection and application should be deterministic code; only judgment needs a capable model**, and a mid-tier model with a well-calibrated brief matches a frontier model on most of it. Third, when models disagree, the disagreements are *systematic*, not random — an aggressive model fits a kill-biased rule and a conservative model fits a nuance-biased rule — and adjudicated disagreements, written back into the brief, permanently improve the cheaper model's future performance without any model getting smarter.

1. The Problem

A small fiction publishing operation drafts genre novels with LLM assistance under a strong house style. The human operator — the sole editor and final quality gate — identified a class of prose defects that survived normal editing passes: **overused reflex vocabulary**. Three examples from the live rule set:

- **"weather" as metaphor.** Emotion-as-weather, presence-as-weather, "his own weather," rooms and moods with weather in them. The operator's rule: *unless the text is about actual sky, the word dies.* Literal meteorology, "weathered" as an adjective, and weather-as-verb are exempt.
- **breath as filler.** Rooms, towns, and mornings "holding their breath"; *"A breath."* as beat punctuation; "breathed" as a speech verb; breath as a time unit ("for one breath"); relief clichés ("let out a breath she hadn't known she was holding"). The rule: *breath must be earned* — literal lungs in a true story beat, a plot mechanic, an idiom, or a motif the book deliberately architects. Dialogue is fully exempt: characters may speak however people speak.

- **A specific word ban** (a term the operator never wants in narration), with an exemption for titles.

Two constraints made this hard at catalog scale:

1. **The judgment is genuinely contextual.** The same word is a defect in one sentence and load-bearing in the next. One manuscript's "breathing house" is a tic; another's is the central horror mechanic. A search-and-replace would vandalize the books.
2. **The best available model was temporary.** The operation had short-term access to a frontier-tier model but would fall back to mid- and small-tier models within days. Any process that only worked with the frontier model in the loop was a process about to break.

2. Design Principles

P1 — The system lives on disk, not in the model. Every rule, exemption, precedent, and process step is a file: a detection script (plain code, no model), a judge brief (the law plus its case history), a taste file (pairs of *killed* vs. *lived* sentences with the editor's reasoning), and apply/verify scripts. Any model — or a different model tomorrow — reads the same corpus of artifacts cold and executes the same process. The model is the interchangeable part; the files are the factory.

P2 — Decompose by sensitivity to model quality. The task splits into five stages, and only one of them needs intelligence:

STAGE	NATURE	WHO RUNS IT
1. DETECT	Deterministic regex extraction of candidate lines	A script (a small model can run it and return the report)
2. JUDGE	Contextual verdicts: defect or legitimate? If defect, render a replacement	Mid-tier models in parallel, guided by the brief
3. QA GATE	Review of *every* proposed edit before anything touches disk	The most capable model available
4. APPLY	Exact-string replacement, `assert count == 1` per edit per file	A script — never a model's memory of the text
5. VERIFY	Residual scan; every surviving instance must map to a recorded "keep" verdict	A script

P3 — Teach taste by exemplar pairs, not rules. Rules produce avoidance; pairs produce style. The judge brief includes killed→lived sentence pairs with the editor's stated reason. A mid-tier model does not need to derive the aesthetic — it imitates the pattern.

P4 — Edits are adversarially verified, never trusted. A judging model must copy its target string byte-exactly from the file (verified by search, count exactly 1) and the apply step re-verifies before writing. A replacement may not contain any banned word, and may not be a *dodge* — swapping the banned word for an equally vague synonym. Replacements must render a concrete image sourced from the book's own world.

P5 — Scoped laws. Every rule carries its exemptions in writing (dialogue, idiom, literal use, architected motif, titles). An earlier failure in this system came from applying a narration rule inside quoted dialogue; the fix was not a better model but a better-scoped written law.

3. The Pipeline as Run

Stage 1 — Detection. A ~20-line script extracted every line matching the banned-word patterns from the 47 manuscripts: **801 flags**. Four small-tier model agents ran the script across the catalog partitions and returned per-book reports. All four outputs were graded against a direct run of the same script: **byte-identical**. (Two agents initially failed on a malformed command involving shell-expansion over unusual file paths; notably, two of the four independently debugged and fixed the invocation, while two reported the failure and stopped. The codified rule: small-tier runners get zero-expansion commands — scripts read their own input files.)

Stage 2a — Calibration head-to-head. Before judging the full catalog, one mid-tier and one frontier-tier model independently judged the *same* 88 flags from two test manuscripts (a romance and a horror novel — chosen because each contained both genuine defects and legitimate uses of the same vocabulary), using the same written brief. Results:

	MID-TIER	FRONTIER-TIER
Verdicts	26 kill / 62 keep	18 kill / 70 keep
Agreement	74 of 88 flags (84%)	
Disputed	14 flags	
Adjudicated correct (per top-tier + written law)	10	4

The disagreements were **systematic, not noise**:

- The mid-tier model's *aggression* matched the kill-biased rule (the weather law is "kill on sight"); the frontier model's conservatism left real metaphors standing.
- The frontier model's *nuance* was better on the earned-vs-tic rule: it caught a relief cliché the mid-tier model kept, and correctly preserved an earned physical beat the mid-tier model wanted to kill.
- Neither model produced a single vague-synonym dodge in ~40 rendered replacements.
- One frontier-model edit failed byte-exactness on a curly-vs-straight apostrophe — producing a new written rule (copy target strings from search output, never retype).

Every adjudicated dispute was written into the brief as a **calibration addendum** — e.g., "uncertainty defaults to keep *for breath only*; for weather, uncertainty kills," "repetition un-earns a simile: if the same simile family recurs, it is habit, not architecture," "a characterizing exhale that does real work is earned."

Stage 2b — Main pass. Seven mid-tier agents, each with the *amended* brief, judged the remaining 713 flags in parallel (~100 flags each, bin-packed by manuscript). Output: **292 proposed kills, 415 keeps, 6 skips**, each kill carrying an exact target string, a rendered replacement, and a one-line reason.

Stage 3 — QA gate. The most capable model reviewed all 292 proposed edits as a compact old→new list, with special scrutiny on (a) any edit adjacent to quoted dialogue, and (b) the manuscript with the highest kill density. Findings: **1 miscopied target string and 2 style touch-ups — a correction rate of ~1%.** Notably, the mid-tier fleet had correctly preserved every architected motif flagged across the catalog (a breath-theft horror mechanic, a breathing-machine plot device, a deliberate house-breath motif chain), while killing the same vocabulary as tics in the same books.

Stages 4–5 — Apply and verify. A script applied all 292 edits with per-file `count == 1` assertions, propagating each edit into sibling draft files within the book's directory tree (excluding archived originals). A final scan confirmed residual counts fell from 234 / 698 / 46 (per banned pattern) to 119 / 496 / 29 — and that every residual was a recorded keep: literal use, dialogue, idiom, motif, or title.

4. Findings

1. **Tiered assignment works, and the boundaries are stage boundaries.** Small models run scripts and report faithfully (but should not be relied on to debug the harness). Mid models carry contextual judgment at volume *when the judgment criteria are written down*. The top model earns its cost at exactly two points: writing/amending the brief, and gating the output.
2. **The brief is the asset that compounds.** The measurable quality gain between the calibration round and the main pass came entirely from ~15 lines of adjudicated precedent appended to a text file. No model changed. This inverts the usual anxiety about model deprecation: if judgments are systematically banked, the *process* gets smarter while the models stay fungible.
3. **Model disagreement is diagnostic.** Aggressive-vs-conservative is not one model being better; it is a posture, and different rules want different postures. Writing the correct posture *into the rule* ("this law is kill-biased; that law defaults to keep") transfers the frontier model's judgment to the cheaper model permanently.
4. **The conservative failure mode belongs in the driver's seat.** For the orchestrating role — writing briefs, adjudicating, gating — a model that keeps-when-unsure fails safe; a model that kills-when-unsure fails at scale. Assign accordingly, independent of raw capability rankings.
5. **Determinism at the edges is what makes model-fallibility survivable.** Because detection and application are code, a wrong judgment is a *contained* wrong sentence — never a corrupted file, a phantom edit, or an

unverified claim of work done. Every model claim of "I edited X" is structurally impossible to fake: the apply step re-verifies the exact string on disk.

6. **Orchestration is a property of the harness, not the model.** The entire fan-out — parallel agents, model assignment per task, monitoring, resumption — is tooling available to whichever model is driving. Losing access to a frontier model does not mean losing the multi-model factory.

5. Limitations

The corpus is one house's fiction under one editor's taste; the ~1% gate-correction rate is measured against that editor's standards, not an external benchmark. Detection is lexical (regex), so it cannot flag a defect that avoids the banned vocabulary — the pattern list grows only as the editor names new offenders. The adjudication standard was a single frontier model applying a written law; a human spot-audit of the applied edits (the operator's normal listen-through pass) remains the final backstop and has not yet been completed for this pass at time of writing. Costs were not rigorously metered; the observed shape (expensive model on ~15% of tokens, cheap models on the volume) is reported as a design outcome, not a measured benchmark.

6. Conclusion

A million-word editorial enforcement pass — 801 contextual judgment calls, 324 verified edits, zero unverified writes — ran in one session on mostly mid- and small-tier models, because the intelligence was in the artifacts: a scoped written law, an exemplar-pair taste file, an accumulating adjudication brief, and deterministic scripts on both edges of the one stage that actually required a mind. The practical lesson for anyone building editorial (or any judgment-at-scale) systems on rented models: **write the judgment down, verify the edges with code, spend the expensive tokens on the brief and the gate — and the models become plumbing.**

Appendix A — Verdict schema

```
{
  "book": "path/to/manuscript.md",
  "line": 1424,
  "verdict": "KILL | KEEP | SKIP",
  "reason": "one line",
  "old_string": "exact text copied from file, unique in file",
  "new_string": "concrete replacement, no banned words, no vague-synonym dodges"
}
```

Appendix B — Judge brief skeleton

1. **The law** — each banned/earned-only pattern, in the editor's own words.
2. **Scope rules** — dialogue exempt; comments/headers skipped; per-law uncertainty posture (kill-biased vs. keep-biased).
3. **Replacement rules** — concrete image from the book's own world; whole sentences; smallest span that kills the defect; byte-exact target verified unique.
4. **Calibration addendum** — append-only adjudicated precedents from every test round.