

Abstract

LLMs deployed in regulated industries are constrained by *hand-authored system prompts* — text that cites nothing, drifts from the law it claims to implement, and leaves no record of why any rule was active for a given request. A prompt asserting a rule does not *evidence* one.

We inherited a system built to establish that compliance is a composition problem, and began by decomposing Schema Pass Rate into three submetrics to improve it. That work produced our finding for the term: **the scores were capped by the architecture, not by tuning**. At Sand Hill East’s direction we rebuilt from scratch.

The rebuilt system treats regulatory text as a **compilable source language**. Authoritative corpora — initially HIPAA, 45 CFR Parts 160, 162 and 164 — are decomposed into **atoms**: individually addressable units of binding instruction, each carrying a verbatim source quote, a citation, a machine-readable activation profile and a derivation trace. A deterministic retrieval layer selects the atoms relevant to a caller’s instruction, role and tenant; a rendering layer compiles them into a guardrailed prompt with a citation on every rule. Every composition emits a signed **receipt** naming exactly which atoms were served.

1. The Audit Problem

When a regulated firm answers an auditor, the first question is rarely “how good is this answer?” It is *show me how you got it, and why*. An output that cannot be defended line by line is worthless in that room.

This is structural, not a tuning problem. Frontier LLMs sample distributions and have no metacognition of their own output provenance, so text that reads as factual can simply be invented. The usual defences fall short: **prompt engineering** is not model-agnostic and breaks across releases; **RAG** expands the model’s information surface without bounding its output; **human review** does not scale.

MeetSynthia inverts the order of operations: retrieve the binding rules, inject them verbatim, and only then call the model. **The model is never trusted to recall a regulation** — it is handed the text, with its citation attached.

2. Why We Rebuilt

Our first deliverable was the one the prior cohort’s report named. We decomposed Schema Pass Rate into three booleans:

- Did the output obey the injected rules?
- Did retrieval identify the correct occupational role?
- Did retrieval pull the correct rules?

The system we inherited was built to prove a theory, and it did. It was not built to be sold. Working against those submetrics surfaced constraints we could not move at the margins:

- The corpus lived in memory**. It loaded at boot, was duplicated per worker on a multi-process deploy, and was scanned linearly on every request.
- Recall was pull-then-filter**. Metadata filtering happened *after* kNN rather than inside it — the pattern that degrades fastest as rows grow.
- One SQLite file, one box**. No separation between batch ingestion and online serving, so a heavy ingest slowed live traffic.
- No architectural pattern and no migrations**, which put a floor under how fast either could be changed safely.

When Sand Hill East made production readiness a business requirement, we stopped patching. Where the prior cohort asked whether compliance is a composition problem, **we asked whether the system answering it survives contact with a customer**.

3. Inherited vs. Rebuilt

	inherited	columbia_aicontext
Store	Single SQLite file	PostgreSQL; 22 append-only migrations
Vectors	sqlite-vec, in-process	pgvector (HNSW, cosine, 384-d), indexed
Recall	Pull-then-filter after kNN	CAP filter applied <i>in</i> SQL, before recall
Ranking	Full corpus in RAM; linear scans	Lexical + semantic arms fused by RRF ($k=60$)
Ingestion	Script-shaped; one curated dataset	5-pass resumable batch pipeline; pluggable source adaptors
Structure	No enforced pattern	Hexagonal; 12 domain features; 253 modules
Deploy	One box, everything colocated	4 runtimes: vm, serverless, ui, worker
Validation	—	22-check gate at every mutation boundary
Tests	—	751 tests collected; nightly CI

253
core modules

22
gate checks

18.9k
lines in core/

22
migrations

12
domain features

716
vocabulary values

4
runtimes

14/14
evals passing

4. Regulation as a Source Language

The metaphor is load-bearing, not decorative. A **lexer/parser** decomposes source text into typed units; a **linker** resolves which are in scope for a target; a **code generator** emits the artifact.

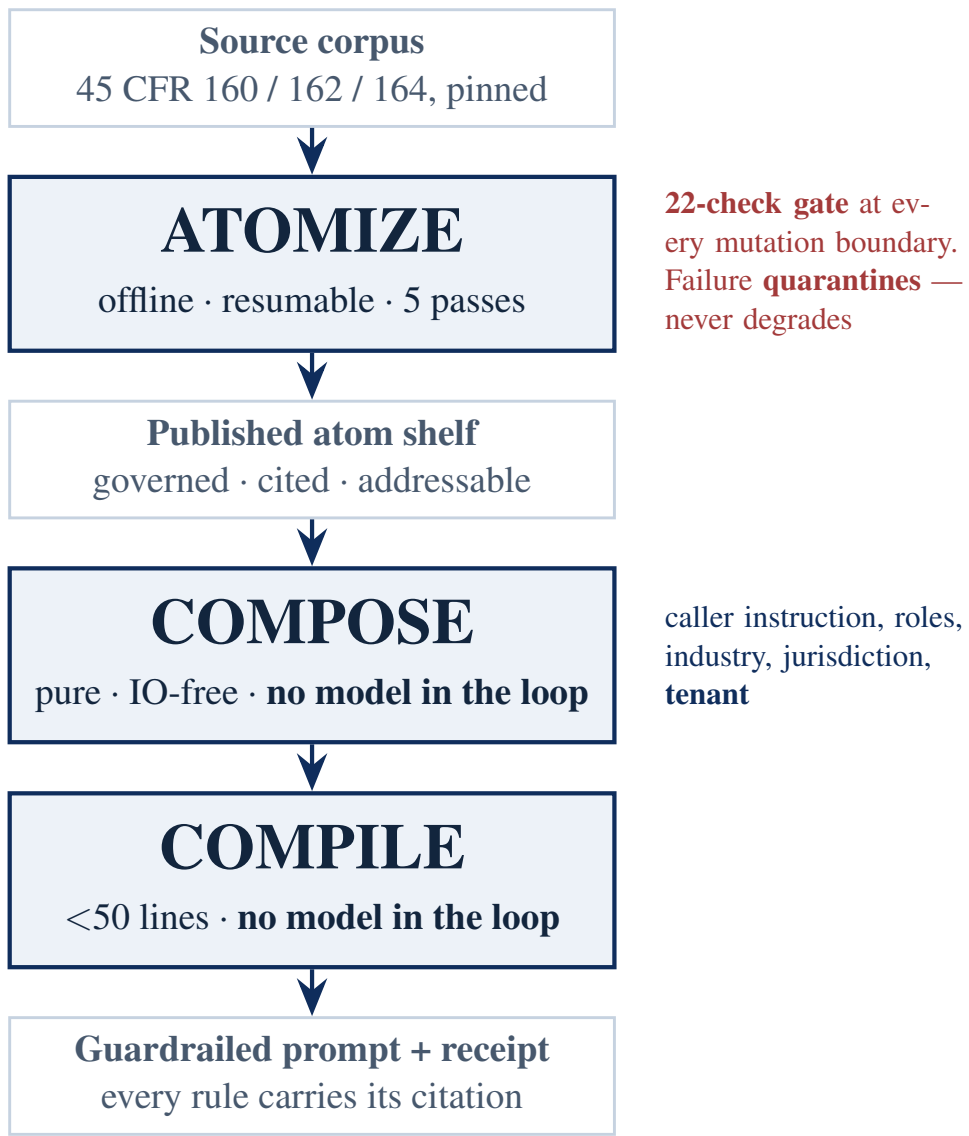


Figure 1. The model reads text and classifies. It never decides a ranking, a threshold, or a pass/fail.

5. Anatomy of an Atom

The smallest unit of instruction that can be independently *cited*, *activated* and *verified*. Identity is a UUIDv5 derived from content, so re-running the pipeline over unchanged source is idempotent and diffable.

human_text machine_text “A covered entity must encrypt electronic protected health information at rest.”
quote (verbatim, ≤25 words) + citation 45 CFR 164.312(a)(2)(iv)
CAP — machine-readable activation predicate applies_to industries jurisdictions events excludes topic_signals task_intents + temporal / exception predicates
normative_force binding advisory polarity obligation prohibition routing override_class normal floor hard_block weights + derivation trace {method, version, rationale} hashes over text and CAP provenance: model, prompt version, run
The check that carries the claim. QUOTE_IN_SOURCE verifies at write time that the recorded quote appears in the pinned canonical source under bounded normalisation — NFKC, quote and dash folding, whitespace collapse, digits compared exactly. An atom whose quote cannot be located is quarantined. This one check is what lets us claim an atom is <i>evidenced</i> rather than <i>generated</i> .

6. Stage 1 · The Atomizer

Offline, asynchronous, resumable. Adaptors normalise arbitrary sources — eCFR XML, Mark-down, structured JSON — into `SourceUnits` carrying a stable ID, a citation anchor such as §164.312, and a breadcrumb of ancestor headings, so binding language in a parent heading is not lost. Canonical documents are hashed: re-ingesting an unchanged source writes nothing.

P1	Extraction . Sonnet 5 at temperature 0 via the batch API proposes a directive, verbatim quote, citation, polarity and normative force. Four gate checks .
P1.5	Deduplication . Deterministic kNN over normalised embeddings generates candidates; Haiku 4.5 adjudicates the grey zone; clusters larger than two go to Sonnet 5 before any destructive action, guarding against transitive chaining. One gate check (CLUSTER_COHERENCE).
P2	CAP binding . Attaches the activation profile from a closed, governed vocabulary. Atoms with several activation patterns <i>fan out</i> rather than carrying one unsatisfiable disjunction. Ten gate checks .
P3	Relationship graph . Deterministic generators propose links (citation, containment, dedup-inherited, similarity); a model assigns the type. Four gate checks .
P4	Grounding & publish . Trigger signals, synthetic violation examples, embeddings. Publish reconciles into a separate serving projection. Two gate checks , plus OWNER_LABEL at publish.

Two doctrines, encoded as checks

- V4_UNADDRESSABLE** quarantines an atom carrying no signal by which retrieval could ever reach it. An unreachable atom is worse than absent — it inflates coverage counts.
- SYNTHETIC_FENCE** keeps synthetic grounding content out of the four evidentiary fields (human_text, machine_text, quote, citation). It may help retrieval *find* an atom; it must never become part of what the atom claims the law says.

Gate failure **quarantines**. A rule that is *probably* quoted correctly is worse than no rule: it carries the appearance of authority.

7. Governed Vocabulary

CAP values are not free text. If they were, `healthcare_provider` and `medical_staff` would be two addresses that silently fail to match. A model wanting a value outside the active vocabulary files a **proposal**, content-keyed on (dimension, value), so fifty atoms proposing the same value accumulate as one row gathering evidence. Human review never blocks the pipeline: the two run at different speeds, so pass 2 opens against a frozen **vocabulary snapshot**.

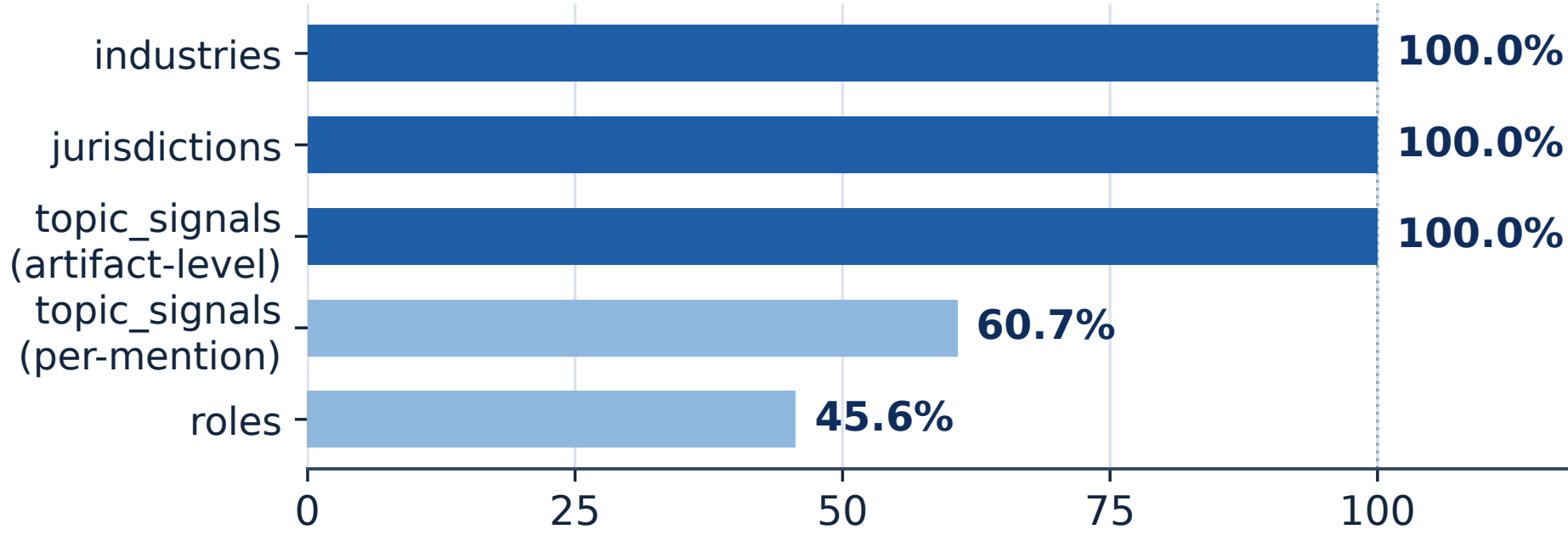


Figure 2. % of corpus mentions covered by the 716 active vocabulary values, after a seed sweep over 357 regulatory artifacts. Roles are floored at 50 mentions by design, which drops the long tail.

8. Stage 2 · The Composer

Given an instruction, a role set, an industry, a jurisdiction and a tenant: which published atoms should be served? Pure, deterministic, IO-free apart from its source port. **It calls no model**.

Eligibility

Each CAP dimension follows *untargeted-or-overlapping*: an atom is eligible if its list is empty or intersects a requested value. Tenancy is an **equality** filter instead, because an atom has exactly one owner — there is no untargeted tenant. Both run in SQL, before recall.

Vectors live in `pgvector` (HNSW, cosine, 384-d) rather than a dedicated store. We proposed Qdrant for filtered kNN at scale and shipped `pgvector` instead, to keep one system of record for a three-person team. **The decision that mattered was not the index — it was moving the CAP filter into the query**, which removes pull-then-filter whichever engine holds the vectors.

Scoring

- IDF unigram overlap** from a materialised `token_df` view, normalised by total query weight — rare shared terms count more.
- Bigram Jaccard** at 0.3, a word-order proxy, so “protected health information” scores differently from three scattered occurrences.
- Verbatim phrase** match on a multi-word topic signal: a flat 0.4, the strongest single lexical signal.
- Role 1.0** on hit, 0.5 if untargeted, 0.0 on mismatch. **Prior** from the atomizer’s derived weights.
- Semantic arm** fused by RRF at $k=60$, chosen because RRF combines *rankings* rather than scores and so needs no calibration between incomparable scales. If the embedder is unavailable, retrieval degrades to lexical-only rather than failing.

final score = 0.60·min(1, relevance) + 0.25·role + 0.15·prior

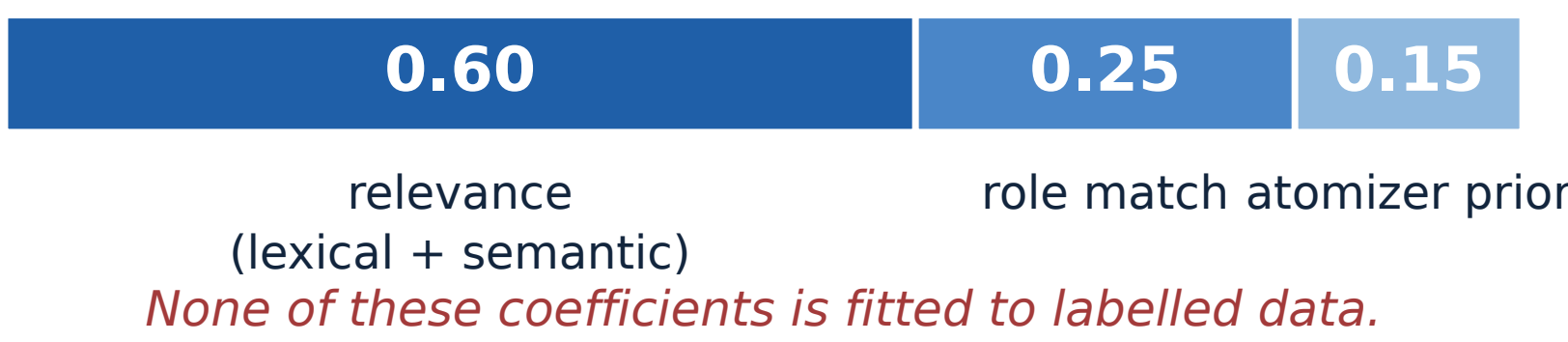


Figure 3. The scored surface is the directive concatenated with its signals, slugs unslugged into words, so CAP metadata participates in lexical matching.

9. The Floor Contract

A non-floor atom survives selection only if it fired at least `min_signals` (default 2) *distinct kinds* of evidence — `hit_count` counts signal *categories*, not magnitude. One shared common word is not evidence.

Floor atoms bypass the gate entirely. Any atom that is `mandatory` or carries a `floor / hard_block` override class is seated regardless of score, and sorts first.

This is what makes the system usable for compliance: the encryption requirement appears in the output of a query about *breach notification* because it is mandatory, not because it scored well. **Mandatory** is a property of the obligation — the law commands it; **override_class** governs how retrieval may treat the atom.

The receipt

Resolved roles, per-stage counts, the budget, a SHA-256 hash of the instruction, the exact atom ids served, the tenant scope, and an optional HMAC. The instruction is *hashed, not stored*, so the receipt is auditable without retaining sensitive prompt text.

Gap detection — reporting absence

When an instruction implies a class of constraint that no published atom backs, the composer emits a **Gap** rather than staying silent. Ordinary RAG returns its best k and says *nothing* about what was missing — dangerous when the consumer may infer that the returned rules *are* the applicable rules.

10. Stage 3 · The Compiler

Under fifty lines, deliberately. It frames the task, groups selected atoms by polarity, preserves mandatory-first ordering, and attaches a citation to every line.

```
# Task
notify patients after a breach of unsecured records

## Do
1. (mandatory) A covered entity must encrypt electronic
protected health information at rest.
[45 CFR 164.312(a)(2)(iv)]
2. (mandatory) Maintain administrative, technical, and
physical safeguards to protect protected health information.
[45 CFR 164.530(c)]
3. A covered entity shall, following the discovery of a breach,
notify each affected individual. [45 CFR 164.404]

## When
1. When an individual requests access to their records, provide a
copy within 30 days. [45 CFR 164.524]
```

The encryption and safeguards rules appear despite the query being about breach notification — seated by the Floor Contract, not relevance. Every line carries its citation: the model receives the rule and its authority together.

11. Hexagonal, and Tested

domain → stdlib + value-shape libs only
adapters → domain only
application → domain only
runtimes → use cases only, via Setup

Ports are real boundaries, not nominal ones: the evaluation suite’s in-memory fake satisfies the same contract as the Postgres adapter, so that tier runs with no database at all.

751 tests collected · 77 files · 0 failures

651 passing

85 infrastructure-gated · 15 deselected
(Postgres · Anthropic · embeddings · CLI)

Two workflows run on every push, every pull request, and nightly. The nightly run is deliberate for the AI evaluation tier: results drift when a model changes underneath a fixed prompt, and a nightly run keeps that drift attributable.

12. What the Evidence Does Not Support

Maturity is **inverted relative to dependency order**: the composer and compiler are complete and well-tested, while the atomizer that produces everything they consume is the least proven part of the system.

~1,294 atoms
mechanically split from HIPAA headings · gated = false
never passed the 22-check validation gate

10 atoms
hand-authored, correct, gated
the entire evaluation corpus

0 atoms
extracted by Pass 1 AND gated, at scale
the missing middle tier

Figure 4. A system whose value proposition is *gated, evidenced, citable* content has, at volume, none of those things.

- No labelled retrieval benchmark exists**. Ten hand-authored atoms with hand-specified expectations verify retrieval on cases *we chose*. Because we wrote both the corpus and the expectations, it cannot tell us we are wrong.
- The scoring coefficients are unfitted** — reasoned judgments, several adjusted by observing outputs, which is tuning on the test set.
- The demo path is not the production path**. Every demo ran on the 10-atom in-memory corpus, so tenancy filtering, semantic search and the Postgres adapter are unexercised in any demonstration we have given. **No runtime has authentication**.

13. Future Work

- Run Pass 1 at full scale and measure it** — extraction recall against a hand-labelled sample, precision on a random sample. Until those numbers exist, no claim about quality is supportable.
- Build an honest retrieval benchmark** authored independently of the implementation, reporting precision, recall and *floor recall*. Then *fit* the coefficients rather than reasoning about them.
- Authentication and the live tenancy test**. Both block deployment.
- The Synthia integration**. `syn-ms` already has its own `RecommendationEngine` and `PromptComposer`, so the question is not how to connect the systems but **which retrieval layer owns the decision**.

References

[1] Columbia MAFN Fall 2025 cohort, *Schema Pass Rate*, Sand Hill East. [2] Columbia MAFN Spring 2026 cohort, *aicontext-market*, Sand Hill East. [3] 45 C.F.R. Parts 160, 162, 164 — HIPAA Administrative Simplification. [4] Cormack, Clarke & Büttcher, *Reciprocal Rank Fusion*, SIGIR 2009. [5] Broder, *Resemblance and Containment of Documents*, SEQUENCES 1997. [6] Cockburn, *Hexagonal Architecture*, 2005. [7] Anthropic, *Message Batches API*, 2026; `pgvector`, 2025.