

— WORKING PAPER · REV2 · NOT PEER REVIEWED

Semantic Neurons: Governed edge construction for GraphRAG

AUTHOR

Evgenios Skitsanos

ORCID

0009-0008-9346-4539

AFFILIATION

Gedank Rayze LDA, Lisbon

rev2	Aug 4, 2026	CC BY 4.0	Working paper
VERSION	PUBLISHED	LICENSE	NOT PEER REVIEWED

ABSTRACT

GraphRAG promises retrieval over structured, connected knowledge — the ability to assemble evidence that spans several facts no single passage contains. That promise rests on a precondition the field rarely examines: that the graph’s edges exist and deserve trust. We begin from a small, sharply instrumented observation: in June 2026, separating per-chunk what the language model produced from what the surrounding system added in a then-current open-source GraphRAG pipeline, the model returned clean entities and zero relations across the six chunks measured. We present this as a motivating observation from one pipeline, not a universal law — pipelines exist whose extraction stage does emit relations. What our subsequent measurements support is a broader and, we argue, more consequential claim: ungoverned relation output fails in three distinct ways — edges that are *absent* (a model drafting freely reached ~6% of an expert’s expected relation pairs), *fragmented* (on a real clinical corpus, 43% of “missing” facts existed under a different relation label; true coverage was 67%, not the 24% exact-match scoring reported), and *unsound* (on an adversarial corpus, a governed, negation-aware system still constructed nine of thirteen forbidden relations through subject misattribution and direction reversal). Each failure is silent in production. We propose a remedy shaped by those failures: the *Semantic*

Neuron, a governed construction operator that is ontology-bounded, evidence-bound, reviewable, reversible, and linguistically sound, embedded in a closed loop that measures recall *and* restraint, extends governance to the reviewer itself, and re-measures at the answer layer. We report the loop's measurements honestly, including where it lost: a 10,000-document pilot whose oracle-checked graph of 129,091 facts was 100% precise, and a real-corpus head-to-head where the graph's answer recall (22%) lost to vector retrieval (36%) — with a diagnosis attributing the loss to coverage and vocabulary, the parts governance had not yet reached, rather than to retrieval or grounding. A directed-construction variant, given an agreed taxonomy and deterministic evidence gates, reached 0.744 precision / 0.635 recall against held-out expert labels — evidence that the binding constraint was governance, not model capability. Our central claim, narrowed from rev1: GraphRAG is a knowledge-construction problem before it is a retrieval problem, a governance layer for edge construction is a precondition for a graph an organization can *trust*, and Semantic Neurons are one disciplined, measured way to build that layer — wherever the graph itself happens to live.

CONTENTS

1. The Promise and the Gap	2. The Observation: A Graph With No Edges
3. Why Edges Are the Whole Point	4. How Ungoverned Edges Fail: Absent, Fragmented, Unsound
5. Two Inadequate Poles	6. Semantic Neurons: Governed Construction Operators
7. Soundness Is More Than Governance	8. Knowledge Construction Is a Control Loop
9. What the Implementation Measured	10. Related Work
11. Limitations and Threats to Validity	12. Conclusion
References	Artifact Availability
Version History	Cite this paper

1. The Promise and the Gap

The case for graph-augmented retrieval is intuitive. Vector retrieval finds passages that are semantically near a query, but it has no notion of how facts connect. A graph adds that missing structure: typed, directed relations between entities, over which a system can traverse to gather evidence that is dispersed across a corpus. The canonical GraphRAG win is the multi-hop question — one whose answer is implied by a chain of facts, none of which co-occur in a single chunk. Vector search cannot assemble that chain; graph traversal can.

This promise rests entirely on one precondition that is rarely examined: that the graph actually has edges, and that those edges are correct. The literature and the tooling have concentrated on the retrieval side — embeddings, rerankers, traversal heuristics, prompt assembly — as though the graph were a given. It is not. The graph has to be built, from unstructured text, by some process. The quality of that process, not the sophistication of the retriever, is what determines whether GraphRAG delivers anything a vector index could not.

We came to this position empirically, and this paper reports the path in the order we walked it: an observation, a diagnosis, a pattern, and the measurements — favorable and unfavorable — that the pattern's own discipline forced us to publish together.

The paper makes three contributions:

1. **A diagnosis.** Ungoverned relation extraction fails in three distinct, individually measured ways — absent edges, fragmented vocabulary, unsound instances — and all three are invisible at the point of failure: nothing in a production system signals the edge that was never built, the knowledge stranded under a mismatched label, or the confidently wrong assertion (§4).
 2. **A pattern.** The Semantic Neuron: a governed construction operator with five defining properties, embedded in a control loop that measures recall and restraint, governs its own reviewer, and re-measures at the answer layer (§6–§8). The pattern is substrate-independent; it sits above any graph store.
 3. **An evidence base with its limits attached.** Dated measurements from one full implementation of the pattern — adversarial restraint, reviewer calibration, an operational 10,000-document pilot, a negative head-to-head result with its diagnosis, a held-out generalization boundary, and a directed-construction result on expert-labeled contracts (§9), with threats to validity stated (§11).
-

2. The Observation: A Graph With No Edges

In June 2026 we instrumented an extraction pipeline — Apache HugeGraph’s LLM toolkit, in its then-current default configuration [21] — to separate, per document chunk, what the language model produced from what the surrounding system added. The corpus was the 2024 State of the Union address: public, prose-heavy, and dense with exactly the relationships a knowledge graph exists to capture — political, geographic, institutional. We expected the model to produce a sparse or noisy graph. Instead, across the six chunks instrumented, the model produced a stream of clean entities and **zero relations**. Not a handful of weak edges. Not noisy edges to be filtered. Zero, chunk after chunk, while a deterministic grounding layer running beside it supplied the four semantic edges the evaluation required.

The entities were genuinely good: real people, places, institutions, events. But the model returned them as an unconnected set. Building the instrument honestly mattered: an early version of the comparison under-credited the model by storing entities under their type label rather than their instance name; we fixed that before drawing conclusions, so the result shows the model contributing real entities while contributing none of the structure.

2.1 What this observation does and does not establish

We state the scope plainly, because rev1 of this paper did not. Six chunks, one pipeline, one corpus, one point in time. (One record-keeping failure is ours to own: the run’s exact model identity was not preserved in the research notes; the corpus, the chunking, the per-chunk outputs, and the instrument script were.) This observation is an existence proof and a lesson in instrumentation, not a universal law. Pipelines exist whose extraction stage explicitly emits relationships — Microsoft’s GraphRAG reference pipeline among them [1] — and a claim that “LLMs return no edges” would be refuted by any of them.

What the observation established for us was different and has survived every subsequent measurement: **you do not know what your extractor contributes until you separate its output from everything around it**, and when we did, the structure we thought the model was building was coming from somewhere else. Later, closer measurement sharpened the point beyond this

pipeline: a capable model asked to draft relation rules freely over a corpus it could read reached roughly 6% of an expert's expected relation pairs across three runs, with direction reversals endemic — while the same model's entity discovery covered about half the expert's catalogue (§9). The asymmetry between entities and relations is real. The question it forces is the one this paper is about:

Where do edges that deserve the name come from?

3. Why Edges Are the Whole Point

It is worth being precise about why an edge-poor graph is not merely a weaker GraphRAG but a pointless one.

A knowledge graph contributes exactly two things a vector index does not: *typed relations* and *traversability*. Strip the edges and you have a set of nodes with embeddings — which is a vector index with extra steps and worse ergonomics. The multi-hop capability, the explainable evidence path, the ability to answer “what connects A to D” through B and C — all of it is a property of edges. A graph that is 95% vertices and 5% edges does not deliver 95% of GraphRAG's value; it delivers almost none of it, because the value is concentrated in the structure, not the inventory.

This has a sharp consequence. The dominant failure mode of GraphRAG in the wild is not bad retrieval over a good graph. It is good retrieval over a graph that was never really constructed — a graph whose nodes are present and whose relations are missing, weak, or wrong. Improving the retriever cannot fix a missing edge. Our own real-corpus measurement made this concrete (§9.6): when our graph lost a head-to-head against vector retrieval, the answering stage had asserted 33 of the 34 expected facts the graph actually contained — 97% of everything achievable — and not one of the missing facts had failed to ground; every miss was a rule that had never been drafted. There was essentially nothing left to win downstream of construction. The bottleneck was upstream of everything the retrieval literature optimizes.

If edges are the point, and free extraction does not reliably supply edges worth trusting, then edge construction becomes the problem to solve. The design space for solving it is smaller than it first appears — but first the failure itself deserves a finer diagnosis than “the model doesn't extract relations,” because that is not quite what our measurements show.

4. How Ungoverned Edges Fail: Absent, Fragmented, Unsound

Across the instruments we built — the six-chunk observation, an adversarial corpus, a real clinical corpus, an ontology-drafting measurement — the relation output of ungoverned extraction failed in three distinct ways. They have different mechanisms, different remedies, and one property in common: at the moment of failure, nothing signals that anything went wrong.

Absent. The edge is never produced. This is the six-chunk observation, and the free-drafting measurement above (~6% of an expert's relation pairs). The mechanism is structural rather than

accidental. Entity recognition is largely *surface-anchored*: a span of text names a thing, and the decision is close to lexical. Relation extraction is a higher-order inference that must, for each candidate edge, identify *both* endpoints, decide that they are related *at all*, infer a predicate that is frequently unstated or expressed through paraphrase, commit to a *direction*, and respect the *polarity and modality* of the sentence. Each of those is a place to fail, and they multiply, while the number of candidate relations grows combinatorially with the number of entities and the evidence for any given relation is often diffuse. Faced with a large hypothesis space and weak per-hypothesis evidence, a well-behaved model does the conservative thing: it returns the entities it is confident about and abstains on relations. The empty edge set is often the model *declining to guess* — which, for relations, is frequently the correct behavior. But a correct abstention is still a missing edge, and production systems do not distinguish “not present in the text” from “present but not extracted.”

Fragmented. The edge is produced — under a different name. This failure mode we did not anticipate, and it took a real corpus to surface it. On a 100-document clinical corpus with an LLM-drafted ontology (§9.6), exact-match scoring reported that the graph contained only 24% of the expected facts. Diagnosis showed that for 43% of the expected set, the entity pair *was* connected — by a relation with a different label: `AVOID_IN` where the ground truth said `use_with_caution_in`, `HAS_WARNING` where it said `adverse_effect`. True coverage was 67%. The exact-match metric was measuring vocabulary fragmentation as if it were ignorance. And fragmentation is not a scoring artifact to be normalized away: in a clinical setting, the difference between “avoid” and “use with caution” is a real distinction of strength that someone must *decide*, not a synonym pair to be collapsed by a similarity threshold. Fragmentation is what an extraction process produces when no one has agreed on the vocabulary — which is to say, it is a governance failure by definition, and no amount of model capability fixes it, because the model cannot know which of two defensible labels an organization has committed to.

Unsound. The edge is produced, matches the schema, carries evidence — and is wrong. On an adversarial corpus built for the purpose (§9.2), a governed, negation-aware system still constructed nine of thirteen forbidden relations. The failures were not exotic: a characteristic phrase occurring, affirmed and unnegated, in text *about a different entity*, so the edge attributed a real fact to the wrong subject; and transaction-like relations expressed through possessives, appositives, and nominalizations, where surface order reverses the semantic roles, so the edge pointed the wrong way — which for an acquisition is not a degraded fact but a false one. We have now seen the direction weakness three independent times: in the adversarial corpus’s traps, in a machine reviewer’s misjudgment of the same constructions, and endemically in free rule-drafting. Unsound edges are the most dangerous of the three modes because they wear credentials: schema-valid, evidence-linked, confidently asserted.

The three modes matter because they decouple. A system can fix absence and drown in unsoundness (lower the extraction threshold). It can fix unsoundness and starve (raise it). It can have excellent per-instance soundness and still fragment, because soundness checks validate instances against rules, and fragmentation lives in the rules’ vocabulary. Any serious answer to “where do the edges come from” has to answer all three — and, we will argue, the answers are respectively *a proposal stream* (for absence), *closure over an agreed ontology* (for fragmentation), and *evidence tests that respect language* (for unsoundness), held together by the separation of proposal from authorization.

One implication reframes the model’s role, and §9.8 makes it concrete: when the vocabulary is agreed in advance and every instance must pass deterministic evidence gates, the same class

of model that drafted its way to ~6% free-form proposes usefully — 0.744 precision against held-out expert labels. The model was never the binding constraint. The absence of governance was.

5. Two Inadequate Poles

If free extraction will not supply trustworthy edges, the edges must come from elsewhere. There are two obvious sources, and each fails in a characteristic way. Understanding *how* each fails is what motivates the third option.

5.1 Unconstrained inference

The first option is to push the model to infer relations freely — to lower its threshold and accept what it produces. This recovers recall, but it trades away the property that matters most in a knowledge graph: that an asserted edge is true. It also, less obviously, manufactures fragmentation: a model free to name relations will name the same relation many ways across a corpus, and the graph's vocabulary decays into overlapping shards no traversal can rely on.

The danger of a false edge in a graph is qualitatively worse than a false passage in a retrieval set, for three reasons. First, false edges are *silent*: a confidently asserted wrong relation is indistinguishable, structurally, from a correct one. Second, they are *compounding*: an edge is not consumed once but traversed many times, and a single false edge corrupts every multi-hop path that passes through it, including paths whose other edges are all correct. The error is non-local. Third, in the domains where graphs earn their keep — investigations, clinical and pharmaceutical research, law, compliance, intelligence — a false relation is not a quality nuisance but a liability: a wrongly connected suspect and device, a wrongly attributed symptom and treatment, a wrongly linked claim and product. In these settings, a missing edge is recoverable and a false edge is dangerous, which means precision dominates recall in the objective function. Unconstrained inference optimizes the wrong quantity.

5.2 Hand-authored rules

The opposite option is to specify edges by hand: a fixed ontology with explicit rules for when each relation holds. This restores precision — nothing enters the graph that a human did not sanction — but it is brittle and static. Natural language expresses the same relation in an unbounded variety of forms, and a fixed rule set cannot anticipate them; it captures the phrasings its authors imagined and misses the rest. Worse, it does not learn. When the system misses a relation that the text clearly supported, the only remedy is for a human to notice and rewrite the rules. The system has no path from observed failure to improvement. It is sound but closed.

5.3 The shape of a solution

The two poles are precision-without-coverage and coverage-without-safety. Neither is acceptable, and averaging them is not either. What the failure analysis suggests is more specific: we need a mechanism that lets the model do what it is measurably good at —

proposing mappings from language to schema — while a governance process does what the model cannot be trusted to do: *authorize* those mappings before they affect the graph. The unit of the solution is not a fact and not a model; it is a reviewable rule for constructing facts.

6. Semantic Neurons: Governed Construction Operators

We call that unit a *Semantic Neuron*. It is not a neural-network neuron and not a model parameter. The name comes from the neuroscience of recovery. After injury, a brain does not regrow what was lost; function returns by *rewiring* — alternative pathways are recruited and reinforced because they restore behavior [19]. That is the shape of the mechanism here. A measured gap in the graph is the lesion; a grounding pathway from evidence to fact is the route; a gap-directed proposal is rewiring through an alternative route; and graduation — when a matured repair is absorbed into the base ontology and the standalone unit retired — is the system's synaptic pruning. We use the mapping as inspiration and claim nothing organism-specific. The metaphor also breaks in one place, and the break is the design: biological rewiring has no reviewer. Accountable review is the safety interlock biology does not have — the line that separates governed construction from hallucinated growth.

Abstracted from any implementation, a Semantic Neuron is a governed construction operator with five defining properties. The paradigm member of the family constructs edges; the family is wider — implementations usefully include operators that canonicalize names (aliases), suppress construction in illegitimate contexts (blockers), and reweight retrieval ranking (rank hints) — but every member is governed the same way, and the edge-construction operator is the one whose absence kills GraphRAG.

Ontology-bounded (closure). A neuron may only assert relations and reference entities that already exist in the schema. It cannot introduce new vocabulary. This is what converts an open-ended generation problem into a constrained selection problem: the model is not inventing a graph, it is proposing how observed language maps onto an *already-agreed* set of types and relations. New vocabulary is a separate, deliberate act of ontology design, not a side effect of extraction. Closure is also the structural answer to fragmentation (§4): a construction layer that cannot mint labels cannot shard the vocabulary.

Evidence-bound. A neuron asserts an edge only when supporting evidence is present in the source text. Provenance is intrinsic rather than reconstructed after the fact: every edge the system produces can point to the span that licensed it. An edge with no evidence is not a low-confidence edge; it is not an edge at all. (What evidence-binding does and does not guarantee is a distinction §11 makes explicit: it proves a licensed span exists, not that the span's source is right about the world.)

Reviewable. Each neuron is a discrete, inspectable object that a domain expert accepts or rejects before it can influence the graph. This is the property that makes governance tractable at scale, and it rests on a crucial shift in the unit of human attention. We do not ask a reviewer to approve individual edges, of which there are too many. We ask them to approve a *rule*, once, and that approval amortizes across every instance the rule ever grounds. Reviewing rules instead of instances is what makes human-in-the-loop affordable on a real corpus. This amortization has a ceiling of its own. On a corpus of tens of thousands of documents, the

proposal stream — machine-generated repair candidates awaiting review — outruns human attention even at the rule level, and a reviewer who rubber-stamps is worse than no reviewer, because the audit trail then records diligence that did not happen. The remedy is not to abandon review but to amortize once more: a domain owner authors a *review policy* — which kinds of rules a measured machine reviewer may approve, at what confidence, with what sampling rate for human audit — and that single approved policy amortizes across every rule it governs, exactly as an approved rule amortizes across every instance it grounds. Approve an instance never; approve a rule once; approve a policy once for a class of rules. Section 8.1 develops what “measured machine reviewer” must mean for this to be safe.

Reversible and staged. Neurons are kept separate from the base ontology, individually status-tracked, and promotable. A neuron has a lifecycle: proposed, reviewed, active, and — when its contribution becomes stable, generally-true domain knowledge — graduated into the base ontology, at which point the standalone neuron becomes redundant and is retired. Redundancy here is success, not waste: it is the trace of an improvement that has matured from a provisional patch into accepted structure. The neuron layer is a staging area for reviewed construction knowledge, not a permanent accumulation of patches.

Linguistically sound. A neuron’s evidence test must respect the structure of language, not merely the presence of words. This is the subtle property, and it deserves its own section, because it is where naive governance still fails.

Together these properties decouple *proposal* from *authorization*. The model proposes candidate mappings from language to schema, exploiting its real strength. The governance layer disposes, supplying the precision and accountability the model cannot. Neither does the other’s job. This division of labor — neural proposal under symbolic constraint, with symbolic execution — is a neuro-symbolic architecture in the precise sense [8], and it is the reason the system can be both adaptive and safe.

6.1 What an implementation realizes today

The five properties are the pattern. Because this paper also reports measurements from a concrete implementation (§9), honesty requires stating what that implementation realizes, so no reader credits the pattern with more automation than exists.

In the reference implementation, a neuron’s recorded states are *proposed*, *accepted*, *rejected*, and *retired*; only accepted neurons influence construction, and acceptance is written by an attributed review action. Graduation is implemented as an *instrument*, not an automatism: a leave-one-out analysis reports which accepted neurons have become redundant against the evolving base ontology, and a human promotes the rule and retires the neuron — the report makes the state visible rather than acting on it. Retirement is forward-looking: it stops a rule from influencing future construction, and already-materialized facts are removed by the next rebuild of the affected graph generation rather than retracted in place; reversal is real, but it is rebuild-and-redeploy, not a per-edge undo. Repair runs are explicitly invoked — there is deliberately no autonomous self-healing daemon. And of the linguistic-soundness taxonomy (§7), the implemented tests today are bounded lexical negation, sentence-scoped subject binding, and direction-explicit templates; modality and attribution beyond negation are stated requirements of the pattern, not yet general mechanisms of this implementation.

None of these gaps embarrasses the pattern; each is a place where the implementation chose measurement and human authority over automation it could not yet certify. But a paper that praises evidence-binding owes its readers the same discipline about itself.

7. Soundness Is More Than Governance

It is tempting to think that once edge construction is governed — bounded, evidence-linked, reviewed — it is safe. It is not, and the gap is instructive.

Consider a rule that grounds a relation when a characteristic phrase appears in the text. The rule is ontology-bounded, evidence-bound, and human-approved. It is still wrong if it fires on a sentence that *denies* the relation it looks for. A passage that says, in effect, “it is not true that A contains B” contains the surface evidence for “A contains B” while asserting the opposite. A construction operator that matches on surface evidence and ignores negation will confidently assert a relation the source explicitly retracts. We observed exactly this failure in a controlled adversarial test, and it is not a corner case; negated, hypothetical, and attributed statements are pervasive in precisely the careful, hedged language of the high-stakes domains where graphs matter most. The biomedical and clinical text-mining communities have long treated negation and modality detection as first-class problems for this reason [9, 10, 11] — “no evidence of malignancy” must never become an assertion of malignancy.

The lesson generalizes beyond negation. A sound construction operator must respect polarity (asserted versus denied), modality (actual versus hypothetical or future), and attribution (the author’s claim versus a quoted or reported one). Governance decides *which* relations a system is permitted to construct; soundness decides *whether the evidence actually supports* a given instance. Both are necessary. A governed but linguistically naive system is safer than free inference, but it is not safe — and the failures it does produce are especially insidious, because they wear the credentials of an approved, evidence-linked rule.

Adversarial testing at scale taught us the taxonomy has at least two more members. The first is **subject binding**: a characteristic phrase can occur, affirmed and unnegated, in text that is *about a different entity*. A passage discussing one company’s platform selection contains, verbatim, the phrasing that a rule for *another* company’s selection listens for; the trigger fires, the evidence link points at real text, and the constructed edge attributes the fact to the wrong subject. On a corpus built to exercise this — parallel narratives about competing organizations in the same domain, sharing vocabulary by construction — a governed, negation-aware system constructed nine of thirteen forbidden relations. Nothing in polarity, modality, or attribution catches this: the sentence is asserted, actual, and the author’s own claim. It is simply about someone else. The second is **direction faithfulness**: transaction-like relations (acquisitions, supply, ownership) are routinely expressed in constructions — possessives, appositives, nominalizations like “the acquisition of X” — where surface order reverses the semantic roles. A trigger that matches the phrasing matches it in both directions; the graph gains an edge pointing the wrong way, which for an acquisition is not a degraded fact but a false one.

The remedies are unglamorous and symbolic: an evidence test can be scoped so that the licensing occurrence must share a *sentence* with the entity the fact is about, and a trigger can be written as a direction-explicit template that expands against the endpoints’ known names and matches only role-faithful phrasings. Both closed the adversarial corpus completely —

thirteen of thirteen forbidden relations refused, with recall intact. But the deployment finding matters as much as the mechanisms: **soundness features of this kind cannot be applied globally**. Requiring subject-in-sentence on every rule collapsed an honest corpus's recall to near zero in simulation, because legitimate evidence is often cross-sentence — the subject named in one sentence, the fact stated in the next. Sentence-scoped binding is therefore a *precision budget* that the ontology author spends, rule by rule, where the domain's vocabulary overlaps dangerously — and nowhere else. Instrumentation can say where the budget is likely needed (for every rule, one can measure how often its evidence occurrences fail to name the subject, against the exact production matching semantics); but the decisive case — an endpoint that *never* co-occurs with its evidence — is ambiguous between systematic misattribution and legitimately cross-sentence prose, and only a reader of the actual sentences can tell. The instrument flags; the author decides. Soundness, like governance, ends in an accountable human judgment about language.

This is why we list linguistic soundness as a defining property and not an implementation detail. It is the difference between a governance layer that looks rigorous and one that is.

8. Knowledge Construction Is a Control Loop

Step back and the individual neuron is less important than the loop it lives in. The deeper claim of this paper is that GraphRAG quality is a *systems* problem, not a *model* problem, and the system it requires has the shape of a control loop.

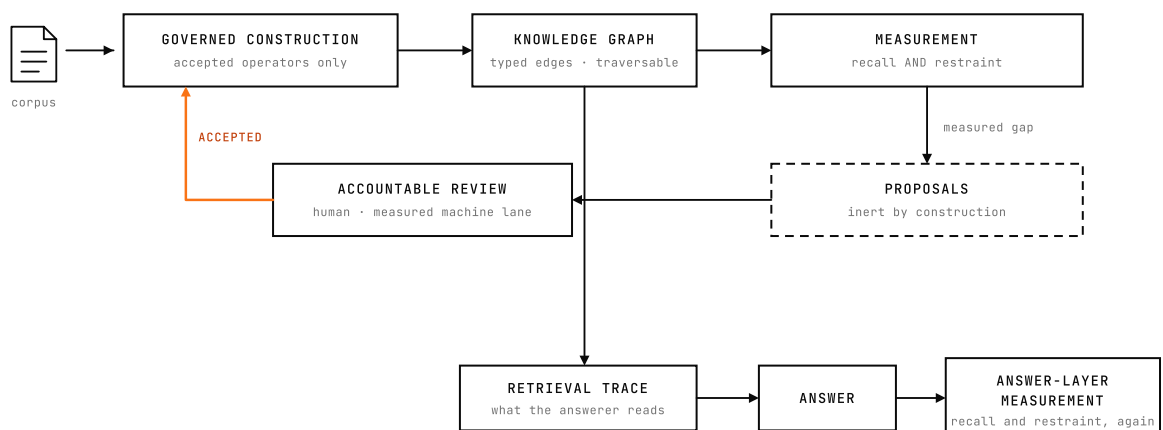


Figure 1 – the governed construction loop. Proposals are inert until accountable review admits them; measurement scores recall *and* restraint; past the graph, the answer layer is measured again.

A control loop needs three things: a measurement, an actuator, and a controller. The measurement is an evaluation that scores the constructed graph not only on *recall* — did the

expected relations get built? — but on *restraint* — did any forbidden relation get built? A harness that measures only recall is blind to the failure mode that matters most in high-stakes use, because it cannot see a confidently-constructed false edge. Measuring restraint requires deliberately asserting what must *not* appear: relations the corpus tempts a naive system into, which a correct system refuses. The actuator is the Semantic Neuron — the governed unit through which a measured gap is closed. The controller is *accountable review* — by default a human, and, for a measured and bounded class of proposals, a machine reviewer operating under an authored policy — which decides which proposed actuators are admitted. What makes a controller legitimate is not its species but its properties: its decisions are attributed, reversible, floored by symbolic checks no verdict can waive, and its competence is measured before it is trusted; §8.1 develops this. Close the loop and the process becomes legible: a measured miss prompts a proposal, a proposal is reviewed, an accepted proposal changes construction, and re-measurement confirms the gap closed without opening a new one.

Actuators have a scope boundary worth stating. A veto-type neuron — a reviewed phrase that suppresses construction of one fact in illegitimate contexts — closes *occasional* violations cleanly: an allegation phrased as fact in one or two passages. Against *systematic* leakage, where dozens of differently-phrased passages all tempt the same forbidden edge, chunk-local vetoes are whack-a-mole: in our adversarial corpus, one-shot veto proposals closed one violation class in six, and the proposal-time coverage simulation disclosed the incompleteness of the rest before any reviewer could have accepted them — an instrument reporting the limits of its own actuator, which we now regard as a design requirement rather than a courtesy. Iterating proposals against the still-uncovered remainder closed all six classes completely, at two to three rounds each, without touching recall — but this makes the point sharper, not weaker: iteration makes the point-repair actuator *complete*, while the structural fix for a systematic failure class remains a rule-level soundness feature (§7), which costs no inference at all and is robust to phrasings the corpus has not shown yet. Point repairs for point failures; rule-level semantics for class failures; and instruments that say which one you are looking at.

Without the loop, the alternatives are exactly the two inadequate poles. A static ontology is an open loop with no feedback: it never learns from its misses. Free inference is a loop with no controller: it acts without authorization and drifts. The Semantic Neuron pattern is what makes the loop closeable — it provides an actuator small enough to review and reversible enough to retire, with its constructed facts removed at the next rebuild. The graph stops being an artifact you produce once and hope is right, and becomes a state you steer toward correctness with measurement and governed action.

This framing also explains why scale alone does not rescue GraphRAG. A larger model is a better proposer, which helps; but it is still an untrusted proposer, and it does not supply the measurement, the controller, or the accountability. The missing piece is not a better component but a missing *loop*.

One more measurement stage belongs in the loop, past the graph itself. A correctly constructed graph can still lose facts on the way to an answer: the answering model reads a retrieved trace of edges, and what it does with them is a behavior to measure, not assume. Three findings from scoring answers — not just graphs — against the same expected-and-forbidden specifications (§9.4). First, the two layers decouple asymmetrically: in our runs, *restraint* carried through construction to answers — a graph with no forbidden edges produced no forbidden assertions, across every corpus and both prompting regimes — while *recall* did not: the answering model reliably omitted facts that were present in its trace, concentrated on boundary and enumeration

questions. Second, the gap is attributable and partly repairable at the answer layer: trace size and prompt wording were worth tens of points of answer recall with the graph held constant, and a second, additions-only pass — the model reviewing its own selection against the trace for completeness — recovered most of the remaining selectivity residue (73% to 87% in our combined measurement) *without* costing restraint or introducing a single unsupported assertion, because every added fact is still checked against the trace and anything absent from it is flagged as fabrication rather than credited. Third, and most practically: provenance of the graph did not matter to the answerer — graphs repaired by the governed loop answered identically to graphs built from hand-authored ontologies. The loop's product is fungible; its guarantees have to be re-measured at every layer that consumes it.

The loop, in full: measure construction (recall *and* restraint), repair through governed actuators, review accountably, re-measure, and then measure once more at the answer layer — because that is where the user is standing.

8.1 Governing the controller

Making review affordable at scale (§6) requires letting a machine reviewer authorize some proposals. This seems to reintroduce exactly the danger the architecture exists to prevent — an unaccountable model deciding what enters the graph. It does not have to, but only if the controller is subjected to the same discipline as the actuators it governs: measured before trusted, bounded while trusted, and audited after.

Measure the reviewer against the system's own recorded decisions before granting it any authority. A construction loop that has run for a while owns a labeled dataset it may not realize it has: every reviewed proposal, with its evidence and its outcome. Replaying a candidate reviewer against those records — blind, scored against what actually happened — yields the two numbers that matter: the false-accept rate (it approved what the system rightly refused) and the conservative-error rate (it escalated or rejected what was good). In our replay of seventy-six recorded decisions, a schema-constrained reviewer achieved a false-accept rate of zero on the proposal class we later automated (0 of 67 at the delegation threshold, in every run), and independently caught the one proposal whose sole evidence sat inside a negated clause. Its errors were almost entirely conservative — an autonomy cost, never a safety cost. That asymmetry licenses a narrow delegation, and only that: nothing whose false-accept rate is nonzero gets automated. The reviewer's instructions turned out to be policy in the strict sense — a one-line “defense” addition to the prompt cost twelve points of measured escalation on good proposals, and a plausible direction-checking line was rejected by the same replay (25 and 21 errors-on-good across variants, against the baseline's 12) — so the prompt carries a version, the version is recorded in every attribution the reviewer produces, and any change re-runs the replay and the injection suite before shipping. Replaying the same decisions twice also showed borderline verdicts flipping between runs: verdicts are a distribution, not a fact, which is why the delegation gate is stated at the confidence threshold the policy consumes and why the symbolic floor beneath the reviewer — verbatim evidence checks, negation-aware grounding, validation — must independently catch what a coin-flip verdict lets through. In the one observed flip, it did.

A reviewer that reads documents can be attacked by documents. Text under review may address the reviewer directly: claim prior approval, dictate a verdict, imitate system output [13]. Before any delegation, the reviewer must pass an injection suite — bad proposals whose

evidence carries embedded instructions to approve, across the payload families an adversary would try, with clean controls to prove rejections are semantic rather than pattern-matched. Our gate demanded zero induced approvals across sixteen attacks; it passed, and the suite earned its keep by finding instead a *semantic* weakness nobody predicted: the possessive-appositive direction reversals of §7 fooled the reviewer at high confidence even in clean controls, and were mitigated at the policy layer by excluding direction-critical rules from delegation entirely. The clean split, we found, is architectural: one purpose-bound stage screens the material for attempts to steer the review and escalates on detection, so the quality judgment that follows carries no defensive burden — folding defense into the quality prompt measurably degraded the quality judgment; separating them restored it. The controller, it turns out, is itself best built as small governed units with single responsibilities. The architecture recurses.

Delegation is tiered by what a mistake costs, and authority is bound to the measured model. Proposals that only add recall to already-sanctioned structure sit in the lane a measured reviewer may approve, at high confidence, on rules the author has not marked precision-critical. Everything that affects restraint — vetoes, anything direction-critical, anything creating new structure — queues for a human with the machine verdict attached as triage; transitions of already-accepted state remain human-only. The machine reviewer never auto-rejects: proposals are inert by construction, so rejection buys no safety, and its measured error rate on good proposals would cost real recall — a machine “reject” is just an escalation with an opinion. A deterministic sample of every machine acceptance goes to a human auditor; disagreement retires the neuron through the ordinary, attributed lifecycle. And because the measurements are of a *specific* model under a *specific* policy, authority is model-bound: a second model family measured on the same harnesses failed to qualify alone — yet its blind spots were disjoint from the incumbent’s, which licensed a measured two-family concordance lane (auto-accept only on agreement) with zero concordant false-accepts across every run, at twice the cost. Which models hold trust is authored configuration, never a deployment accident.

Under this regime the controller keeps the properties the loop requires — measurement, bounded action, accountability — while the human moves up one level: from reviewing every proposal to owning the policy, auditing the sample, and deciding everything the policy refuses to delegate. That is the same move the Semantic Neuron made for construction, applied to review. The loop does not lose its controller; it gains a governed one.

8.2 Where proposals come from

The loop consumes proposals; it is worth naming their sources, because the architecture’s inertness property — nothing influences the graph until reviewed — is what makes the sources safely pluggable.

The primary source is the measurement itself: a scored gap between expected and constructed facts, handed to a proposer that drafts candidate rules against the agreed ontology. A second source is instrumentation: the coverage and subject-binding instruments of §7 and §8 emit flags that are themselves proposal-shaped (“this rule likely needs a gate here”). A third is *directed construction* (§9.8): when the consumer already owns a taxonomy — a clause inventory, a regulatory checklist — the model nominates per-document facts constrained to that taxonomy, and deterministic gates dispose of each nomination before anything is written.

A fourth source is generative context expansion, which we build under the name *side-views*: at ingest, the system generates ten to twenty question–answer pairs per document and indexes them alongside the original prose, in the document-expansion lineage of doc2query and hypothetical-document embeddings [14, 15]. Side-views exist for retrieval — a user’s question often matches a generated question far better than the raw text — and in our implementation they are deliberately *quarantined* from the governed graph: they are non-authoritative, non-deterministic surfaces, and nothing in that layer may write a fact. But the quarantine boundary is precisely what makes them interesting to the loop: a Q&A pair that keeps surfacing knowledge the graph lacks is a gap detector running for free, and feeding such signals into the *proposal* queue is safe by construction, because proposals are inert until authorized. Today side-views serve retrieval only; their role as a proposal source is a design direction the architecture licenses, not a measured result.

What the sources share is the contract: anything may propose; only the governed loop disposes.

9. What the Implementation Measured

The pattern of §6–§8 is implemented in full in CogniGraph, our graph platform: governed ingest, evaluation kits with expected and forbidden facts, gap-directed proposing, risk-tiered review over HTTP, answer-level measurement. This section reports what that implementation measured, favorable and not. Three disclosures up front. First, this is evidence from one implementation, largely on evaluation kits its own research program authored (the exceptions are noted); it is offered as an existence proof with named boundaries, not as a benchmark victory. Second, every number below has a dated, reproducible source — a fixture, a results file, or a decision record — in the implementation repository; the fixtures are private today, and a sanitized public evidence pack is planned, so until it ships these numbers carry our signature, not independent verification. Third, the corpora named here (the 2024 State of the Union address, DailyMed drug labels, WebNLG, CUAD) are public.

9.1 Construction and repair on authored kits

On four independent evaluation kits authored outside the implementing session, cold construction scored 67/67 on expected facts with 0/36 forbidden traps constructed. Then, with every grounding pathway mechanically stripped — the lesion experiment — the repair loop recovered 62/67 (93%) from zero, with 0/36 violations throughout, and every unfilled gap was a governed refusal rather than a fabrication: one verbatim trigger was declined because its only occurrence sat inside a negated clause — the soundness rule of §7, learned on an earlier corpus, firing organically on third-party prose.

9.2 Adversarial restraint

A 1,120-chunk adversarial corpus — parallel narratives about competing organizations sharing vocabulary by construction — broke a governed, negation-aware system 9 times out of 13 (§4, “unsound”). Sentence-scoped subject binding and direction-explicit templates closed it completely: 13/13 forbidden relations refused, recall intact at 37/37. The same campaign measured the deployment boundary: applying subject-in-sentence globally collapsed an honest

kit's recall (14 of 15 points lost in simulation), which is why gates are authored per rule as a precision budget; and a gate advisor instrument subsequently rediscovered all six authored gates on exactly the endpoints the author had chosen, while deliberately never auto-suggesting the ambiguous cases. Veto-type repairs against the corpus's systematic leakage closed 1 of 6 violation classes one-shot, with the proposal-time coverage simulation disclosing the incompleteness of the rest; coverage-guided iteration closed 6 of 6 in two to three rounds each, restraint moving 6/13 → 0/13 with recall held at 37/37.

9.3 The reviewer, measured

The replay-before-authority protocol of §8.1, in numbers: seventy-six recorded review decisions from four corpora; 82% blind agreement; zero false-accepts on the relation-hint lane (0/67 at the delegation threshold, in every run); the one negated-evidence proposal caught independently; errors otherwise conservative. Policy sensitivity: a one-line defensive addition moved escalation from 18% to 30% on good proposals with no safety gain; two direction-checking variants were rejected at 25 and 21 errors-on-good against the baseline's 12. Replaying twice caught a borderline verdict flipping (caught five times in six across observed runs) — the reason delegation is stated at threshold and floored symbolically. The injection gate: sixteen attacks across payload families, zero induced approvals, with clean controls; its real yield was semantic — possessive-appositive direction reversals fooled the reviewer at high confidence, excluding direction-critical rules from delegation as a consequence — and the screener/judge split (a purpose-bound stage flagging steering attempts at 0.96–0.99 on 16/16, so the quality prompt carries no defensive burden) restored judgment quality that a folded-in defense had measurably degraded. A second model family evaluated on the same harnesses did not qualify alone (a negation-class false-accept at exactly the policy threshold; ~38% conservatism) but showed disjoint blind spots — it rejected both direction traps the incumbent missed — licensing the measured two-family concordance lane now in policy, with authority model-bound.

9.4 Answers, not just graphs

Scoring answers against the same expected-and-forbidden specifications: the trace-to-answer stage alone dropped recall from what the graph contained (32/68 asserted) until trace representation and prompting were treated as measurable variables (53/68); the additions-only second pass lifted combined answer recall 77/105 → 91/105 (73% → 87%) with 0/54 forbidden assertions in both modes and zero fabrications — every added fact verified against the trace. Restraint carried through construction to answers in every run; recall did not, until measured. Graphs repaired by the governed loop answered identically to graphs built from hand-authored ontologies: provenance indifference, which is what “the loop's product is fungible” means operationally.

9.5 An operational pilot at 10,000 documents

On 10,000 public FDA drug labels (DailyMed [22]) — the first run with an external structured ground-truth oracle — the deterministic layer's cost and queue behavior were measured directly, and the review judge, scored against the oracle, rejected ~82% of the deterministic layer's false positives; notably, its errors were confident enough that no confidence threshold separates them, which is an argument for oracle-first gates in production wherever an oracle

exists. The resulting oracle-checked graph of 129,091 facts measured 100% precise with zero cross-product leakage, and under controlled document revision the loop tracked every addition, removal, and no-op edit correctly. Reviewer economics were measured at roughly six seconds per judgment and ~214 calls to certify a gate; sustained-operation economics remain extrapolated (§11). Where a structured oracle exists, the governed loop is provably accountable. The relations for which no oracle exists — clinical effect, interaction — are the frontier, and they are where the next result lives.

9.6 The result that did not reproduce

On five fixture kits with hand-authored ontologies and hand-authored ground truth, graph answers beat vector RAG 86% to 36% on recall, with restraint 0/54 versus 4/54. On the real corpus, that result did not reproduce — and we regard publishing this as part of the method. A 100-label DailyMed corpus was constructed with an LLM-drafted ontology whose edge precision a blinded audit put at 85% (after excluding administrative relations; 71.5% before). The same head-to-head methodology — same answering model, same runner, ground truth written from the label text by an independent model restricted to the ontology's vocabulary, never derived from the graph — produced: vector 54/149 (36%), graph 33/149 (22%) on two-pass combined recall, with restraint essentially tied (2/77 versus 3/77 forbidden leaks). The fixture advantage vanished.

The diagnosis is the finding. First, the answering stage was at its ceiling: only 34 of 140 expected clinical facts existed in the graph at all, and the graph arm asserted 33 of them — 97% of everything achievable. Second, grounding was not the bottleneck: of the 106 facts the graph lacked, zero had a rule that failed to ground; every miss was a rule never drafted. Third — the decisive split — for 60 of 140 expected facts (43%), the entity pair *was* connected under a different relation label; true coverage was 67%, not 24% (§4, “fragmented”). The loop had hit its own ceiling almost perfectly; the ceiling itself was coverage and vocabulary — the parts of the problem that are governance work, not retrieval work and not grounding work. The negative result does not contradict the paper's thesis; it locates, with unusual precision, which part of construction was still unsolved.

9.7 A held-out generalization boundary

Generalization beyond self-authored kits is not established, and one clean held-out boundary says so quantitatively: on WebNLG [18] — frozen text-to-triple pairs with a mechanically scorable oracle — even supervised, entity-supplied, oracle-mined candidate templates achieved 7.0% recall at 68.8% precision, without exercising the runtime proposal/review loop. The protocol removes the domain-expert bottleneck precisely so that failure is informative: template-based grounding, as currently built, does not transfer to arbitrary prose without the loop's repair machinery, and measuring the full loop on that held-out protocol is future work.

9.8 Directed construction against expert labels

The head-to-head's diagnosis — coverage and vocabulary — predicts that agreeing on the vocabulary in advance and gating instances deterministically should recover most of what free-form construction loses. A directed-construction mode tests exactly this: the consumer supplies the taxonomy; the model nominates per-document facts constrained to it; deterministic gates

dispose of each nomination — the evidence must be verbatim in the source with recoverable offsets, the endpoints must appear in the evidence sentence, and negated vocabulary grounds nothing; survivors flow through the same writer, reconciliation, and attribution as rule-grounded facts. On CUAD [17], a public corpus of 510 expert-labeled legal contracts (a domain where the recurring-fact rules that fit our earlier corpora were structurally inexpressible — clause detection is per-document by nature), the mode was designed in three rounds on a design set, the taxonomy frozen, and then run once against 40 held-out contracts: **0.744 precision, 0.635 recall, F1 0.685** against the expert clause labels, with six of eight clause categories at 0.81–0.90 precision and the two weak categories each traceable to a named failure mode. This is the same class of model that reached ~6% drafting freely (§2.1). The difference is not the model. It is the governance: an agreed vocabulary, deterministic evidence gates, and attributed writes.

10. Related Work

The pattern sits at the confluence of several established lines of work, and naming them situates it without claiming novelty for its parts.

The entity-versus-relation asymmetry is familiar from information extraction and open information extraction, where relation extraction has long been the harder, noisier task [2]. The precision cost of trading supervision for coverage is the classic lesson of distant supervision [3], and free LLM relation inference shares its failure profile — recall bought with silent over-assertion — while lacking even the knowledge base that distant supervision aligns against; data-programming systems such as Snorkel [4] made the labeling *function* the reviewable unit, an amortization move the Semantic Neuron applies to the construction operator itself. The discipline of evidence-bound assertion is the concern of data and knowledge provenance, standardized in PROV-O [5]; constraint languages such as SHACL [6] validate a graph *after* construction, where the governed loop refuses the invalid edge *during* it — complementary layers, not competitors. Reviewing rules rather than instances is the efficiency that human-in-the-loop and active-learning systems seek [7], applied at the level of the construction operator; knowledge-graph refinement surveys [20] catalogue post-hoc correction methods that a governed loop aims to make largely unnecessary. The decomposition into neural proposal under symbolic constraint is a neuro-symbolic architecture in the standard sense [8]. The demand for linguistic soundness recapitulates decades of negation and modality work in clinical text mining [9, 10, 11]. On the retrieval side, GraphRAG systems [1] and neurobiologically inspired graph memories [16] concentrate on what to do with a graph once it exists; this paper is about whether it deserves to exist. Generative document expansion [14, 15] is the lineage of the side-view mechanism of §8.2 — a retrieval aid we deliberately quarantine from construction.

Delegating bounded review to a model joins what has become common practice under the name LLM-as-judge [12] — a practice many teams, ours included, arrived at independently in production systems before treating it formally. What we add to it is not the idea but the discipline of §8.1: the judge is calibrated and regression-tested against the system's *own recorded decisions* rather than generic benchmarks; it is gated on a prompt-injection suite [13] before receiving any authority, since a reviewer that reads documents inherits the document-borne attack surface; its instructions are versioned as policy; and its authority is bound to the measured model. The contribution of the paper as a whole is likewise not any one component but their composition into a single, closeable loop, oriented at the failure that makes GraphRAG fail: the edge that is absent, fragmented, or unsound.

11. Limitations and Threats to Validity

The originating observation is small and dated. Six chunks, one pipeline, one corpus (§2.1), and the run's exact model identity was not preserved. It motivates; it does not generalize. Extraction pipelines that emit relations exist, and the paper's argument deliberately does not depend on their absence — it depends on the three failure modes of §4, each of which has larger, later evidence behind it.

The evidence base is one implementation, mostly self-authored harnesses. The evaluation kits, adversarial corpus, and replay records were authored within the same research program (kits by assistants outside the implementing session, which limits but does not eliminate shared blind spots). The exceptions run in the other direction of risk: the CUAD holdout is expert-labeled and was frozen before scoring; WebNLG is an external oracle; the 10,000-label pilot is scored against a structured external oracle. The DailyMed *clinical* ground truth was written by an independent LLM from the label text (never from the graph, vocabulary-restricted) — independent of the graph, but not expert-adjudicated; a blinded two-expert review protocol exists and had not completed at this writing.

Evidence-bound is not the same as true. A matching source span proves that construction was licensed and is attributable; it does not prove the source's claim about the world. A governed graph built over wrong documents will faithfully, auditably represent wrong facts. The property on offer is fidelity-with-attribution — the precondition for finding and reversing an error — not omniscience.

Restraint is measured against enumerated traps. The restraint metric counts forbidden facts the kit authors thought to enumerate — relations the corpus tempts a system into. It cannot count false edges nobody anticipated. Zero-of-thirteen means the enumerated attack surface held, not that no false edge is possible.

Generalization is not established. The WebNLG boundary (7.0% recall) is a measured negative on held-out prose for the template layer as built (§9.7). The vocabulary-fragmentation lever identified in §9.6 is diagnosed but, at this writing, not yet closed on the clinical corpus. Answer-layer measurements showed ± 5 points of run-to-run variance across identical configurations. Operational economics at sustained scale are partially extrapolated from single certification passes.

Reversal is rebuild-based and repair is invoked. Retiring a neuron stops future construction; already-materialized facts leave at the next verified rebuild of the affected graph generation, not retroactively (§6.1). There is no autonomous repair daemon by design; every loop iteration is an explicit, attributed act. Readers wanting a self-healing system should read this as a feature refused, not a feature missing.

Machine review is narrow by construction. The delegated lane covers recall-adding proposals on rules not marked precision-critical, under a versioned policy, with sampling audit; everything else is human. The measured false-accept rate of zero is a property of that lane, that model pair, and that policy revision — not of machine review in general. Signed, expiring model qualification for governance use remains open work.

12. Conclusion

The prevailing mental model of GraphRAG treats it as retrieval with a graph attached, and concentrates effort on the retrieval. The evidence points the other way. Ungoverned extraction yields edges that are absent, fragmented, or unsound — three silent failures that no retriever downstream can repair — and since the edges are the entire reason to use a graph, the system's quality is decided in construction, before retrieval begins. Construction cannot be left to free inference, whose false edges corrupt silently and compound, nor frozen into static rules, which are sound but brittle and cannot learn. It requires a third thing: a governed, evidence-bound, reviewable, reversible, and linguistically sound operator for turning language into approved structure — a Semantic Neuron — embedded in a loop that measures both what was missed and what was wrongly asserted, and steers the graph toward correctness through governed action. The same discipline extends upward to the loop's controller — review itself can be tiered and partially delegated to a machine reviewer, but only one that is measured against the system's recorded decisions, hardened against the documents it reads, versioned like the policy it is, and floored by symbolic checks no verdict can waive — and downward past the graph, to the answers users actually receive, where restraint must be measured again, because construction can guarantee what a graph contains but only measurement can say what a model does with it.

The claim, stated at the width the evidence supports: GraphRAG is a knowledge-construction problem before it is a retrieval problem; a governance layer for edge construction is a precondition for a graph an organization can trust in the domains where trust is the point; and Semantic Neurons are one disciplined way to build that layer — measurable, governable, reversible — demonstrated end-to-end in one implementation, with its losses reported next to its wins and each loss traced to a named, addressable cause. A graph database can store any edge you give it. What it cannot tell you is which edges deserve to exist. That judgment is the product. We publish the pattern openly so that judgment can be built — and measured — wherever the graph happens to live.

References

- [1] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson. *From Local to Global: A Graph RAG Approach to Query-Focused Summarization*. arXiv:2404.16130, 2024.
- [2] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni. *Open Information Extraction from the Web*. IJCAI, 2007.
- [3] M. Mintz, S. Bills, R. Snow, D. Jurafsky. *Distant Supervision for Relation Extraction Without Labeled Data*. ACL-IJCNLP, 2009.
- [4] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré. *Snorkel: Rapid Training Data Creation with Weak Supervision*. PVLDB 11(3), 2017.
- [5] T. Lebo, S. Sahoo, D. McGuinness (eds.). *PROV-O: The PROV Ontology*. W3C Recommendation, 2013.
- [6] H. Knublauch, D. Kontokostas (eds.). *Shapes Constraint Language (SHACL)*. W3C Recommendation, 2017.
- [7] B. Settles. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

- [8] A. d'Avila Garcez, L. C. Lamb. *Neurosymbolic AI: The 3rd Wave*. Artificial Intelligence Review 56, 2023.
- [9] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, B. G. Buchanan. *A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries*. Journal of Biomedical Informatics 34(5), 2001.
- [10] H. Harkema, J. N. Dowling, T. Thornblade, W. W. Chapman. *ConText: An Algorithm for Determining Negation, Experience, and Temporal Status from Clinical Reports*. Journal of Biomedical Informatics 42(5), 2009.
- [11] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik. *The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes*. BMC Bioinformatics 9(Suppl 11), 2008.
- [12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. NeurIPS Datasets and Benchmarks, 2023.
- [13] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz. *Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. AISEC@CCS, 2023.
- [14] R. Nogueira, W. Yang, J. Lin, K. Cho. *Document Expansion by Query Prediction*. arXiv, 2019.
- [15] L. Gao, X. Ma, J. Lin, J. Callan. *Precise Zero-Shot Dense Retrieval Without Relevance Labels*. ACL, 2023.
- [16] B. J. Gutiérrez, Y. Shu, Y. Gu, M. Yasunaga, Y. Su. *HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models*. NeurIPS, 2024.
- [17] D. Hendrycks, C. Burns, A. Chen, S. Ball. *CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review*. NeurIPS Datasets and Benchmarks, 2021.
- [18] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini. *Creating Training Corpora for NLG Micro-Planners*. ACL, 2017.
- [19] T. H. Murphy, D. Corbett. *Plasticity During Stroke Recovery: From Synapse to Behaviour*. Nature Reviews Neuroscience 10, 2009.
- [20] H. Paulheim. *Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods*. Semantic Web 8(3), 2017.
- [21] Apache HugeGraph (incubating), and its LLM toolkit (hugegraph-ai). Software, Apache Software Foundation.
- [22] DailyMed. U.S. National Library of Medicine. Structured Product Labeling corpus.

Artifact Availability

The reference implementation of the pattern — governed ingest, evaluation kits, gap-directed proposing, risk-tiered review, answer-level measurement, and the directed-construction mode — is part of **CogniGraph**, a graph platform by Gedank Rayze. Every number in §9 maps to a dated fixture, results file, or decision record in the implementation repository. Those fixtures are private at this writing; a sanitized public evidence pack (corpora references, kit specifications, scored outputs, and harness descriptions, with all non-public material removed) is planned, and until it ships the measurements should be read as author-attested. The corpora used are public: the 2024 State of the Union address, DailyMed structured

product labels [22], WebNLG [18], and CUAD [17]. The pattern itself is substrate-independent and this paper is published under CC BY 4.0 so that it can be implemented, criticized, and measured against, freely.

Version History

- **rev1 — July 13, 2026.** Position statement: governed edge construction as GraphRAG's precondition. Superseded; preserved for lineage.
 - **rev2 — August 4, 2026 (this version).** Claims narrowed to measured scope; the three-failure-mode diagnosis (absent / fragmented / unsound); soundness taxonomy extended with subject binding and direction faithfulness; controller governance (§8.1); answer-layer measurement; full evidence section including negative results (the non-reproducing head-to-head, the WebNLG boundary) and the directed-construction holdout; implementation-status disclosures; related work with citations; limitations and threats to validity.
-

Cite this paper

A DOI via Zenodo is planned; until it is minted, cite the canonical URL.

```
@misc{skitsanos2026semanticneurons,  
  author = {Skitsanos, Evgenios},  
  title  = {Semantic Neurons: Governed Edge Construction for {GraphRAG}},  
  year   = {2026},  
  month  = aug,  
  note   = {Working paper, rev2. Not peer reviewed. Licensed CC BY 4.0},  
  url    = {https://skitsanos.com/papers/semantic-neurons/}  
}
```

AUTHOR

info@skitsanos.com · ORCID 0009-0008-9346-4539

IMPLEMENTATION

CogniGraph — a product of **Gedank Rayze**, Lisbon

LICENSE

© 2026 Evgenios Skitsanos · **CC BY 4.0**