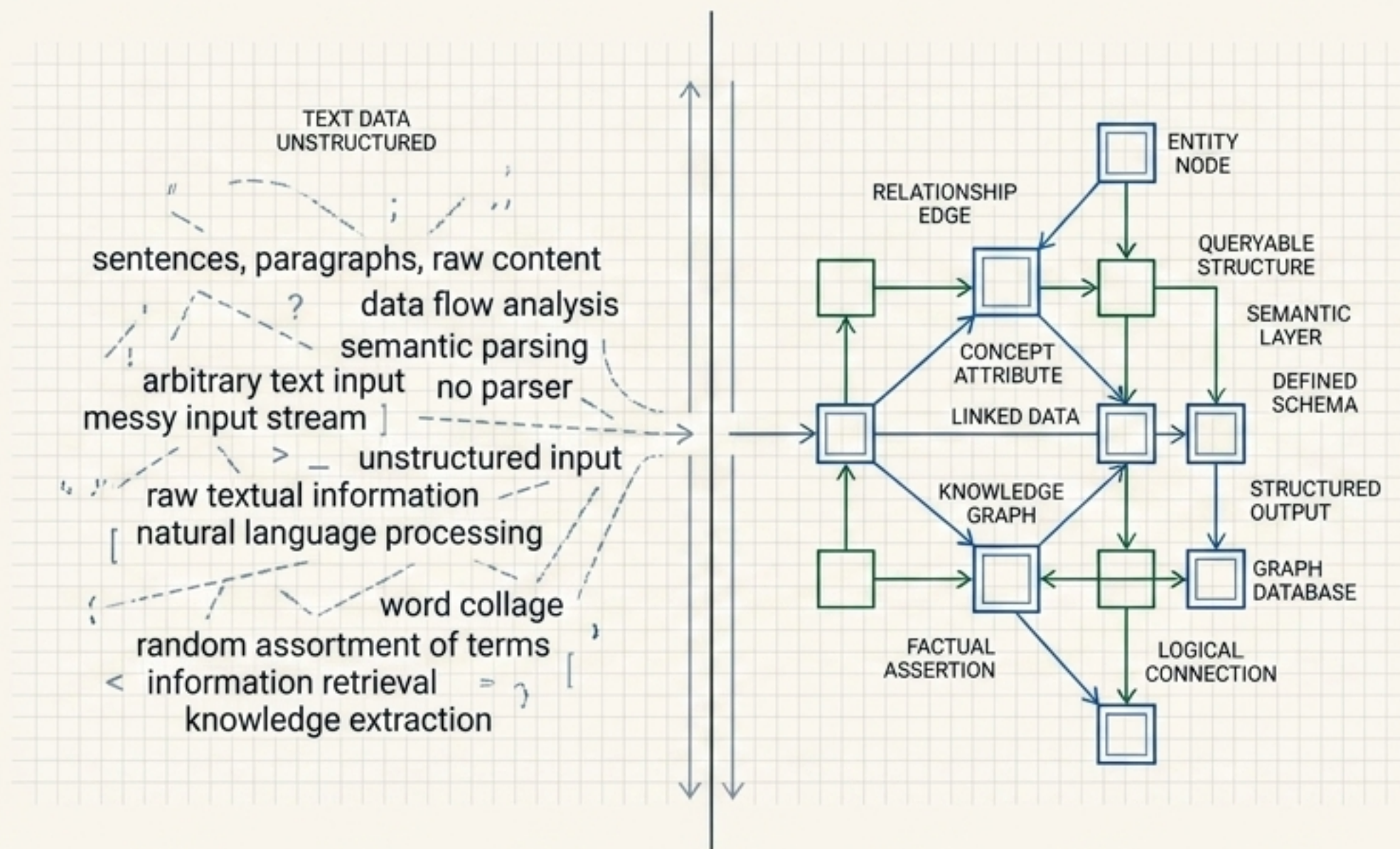




TextGraph

Engineering Research for a Graphify-Equivalent System over Pure Textual Data

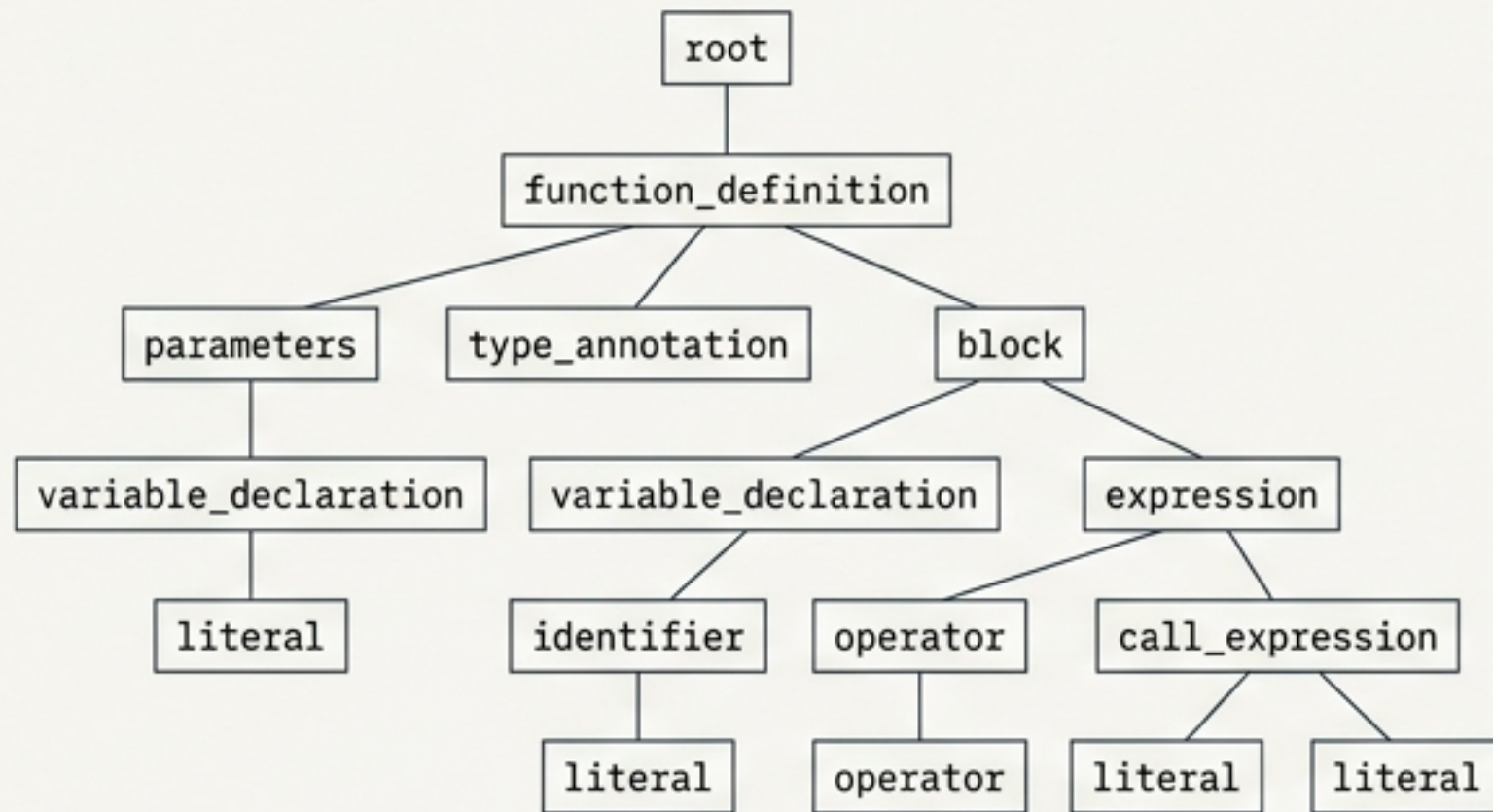
[AUTHOR: Dk (Harish Krishna)]

[DATE: August 2026]

[STATUS: Architecture Specification]

[TARGET: Claude Code, Cursor, Codex, Gemini CLI]

The Core Translation Problem



Source Code -> tree-sitter -> Lossless AST

Zero LLM calls, zero non-determinism, near-linear cost.



Natural Language -> The AST Void

No ground-truth parser.

The answer is not “use an LLM instead.” The answer is a stratified extraction stack where each layer has a different determinism/cost/recall profile, and every edge is stamped with its **provenance**.

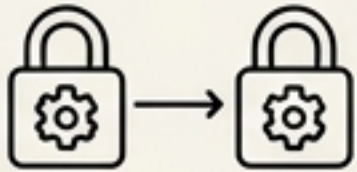
The Core Bet: Encoders Over LLMs

Property	Autoregressive LLM extraction	Encoder IE (GLiNER family)
Determinism	Poor even at T=0 (batching/kernel nondeterminism)	Effectively deterministic
Throughput	~1 chunk / 1-3 s / call	100s of chunks/s batched on GPU; workable on CPU
Cost model	Linear in tokens x price	One-time model download
Schema control	Prompt-dependent, drifts	Label set is an explicit runtime argument

Use GLiNER-class encoders (~200-500M params, local) as the default semantic layer. Reserve LLMs for narrow, cacheable rationale passes.

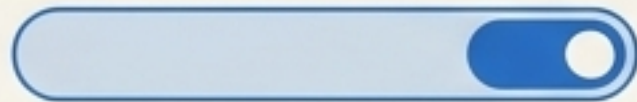
The 7 Non-Negotiable System Constraints

G1: Determinism



Same corpus + same version = byte-identical graph.json. Pinned weights, fixed seeds.

[STRICT]

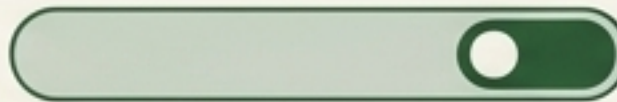


G2: Local-First



Zero network calls by default. LLM pass is an explicit opt-in flag.

[NETWORK: OFF]



G3: Provenance

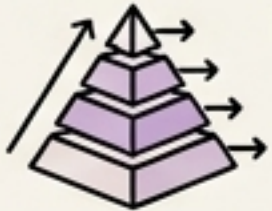


Never emit an assertion without a byte-range citation (source_spans).

[CITED]

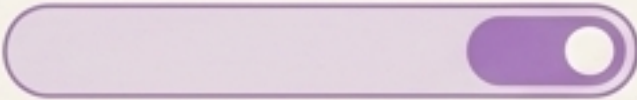


G4: Confidence Stratification

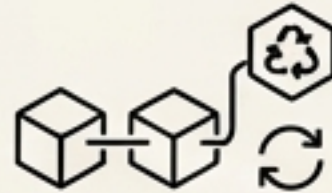


4-tier provenance tags (Structural to Generated).

[TIERED]



G5: Incrementality

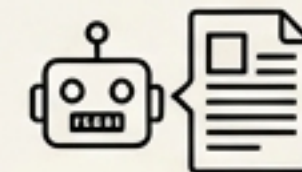


Content-addressed chunk cache. One changed file must not rebuild the world.

[INCREMENTAL]



G6: Agent-Legible Output

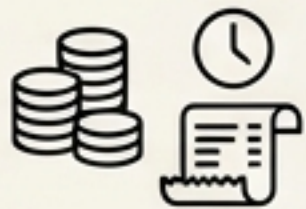


Serve bounded, ranked, cited context to LLM token budgets.

[LEGIBLE]



G7: Auditable Cost



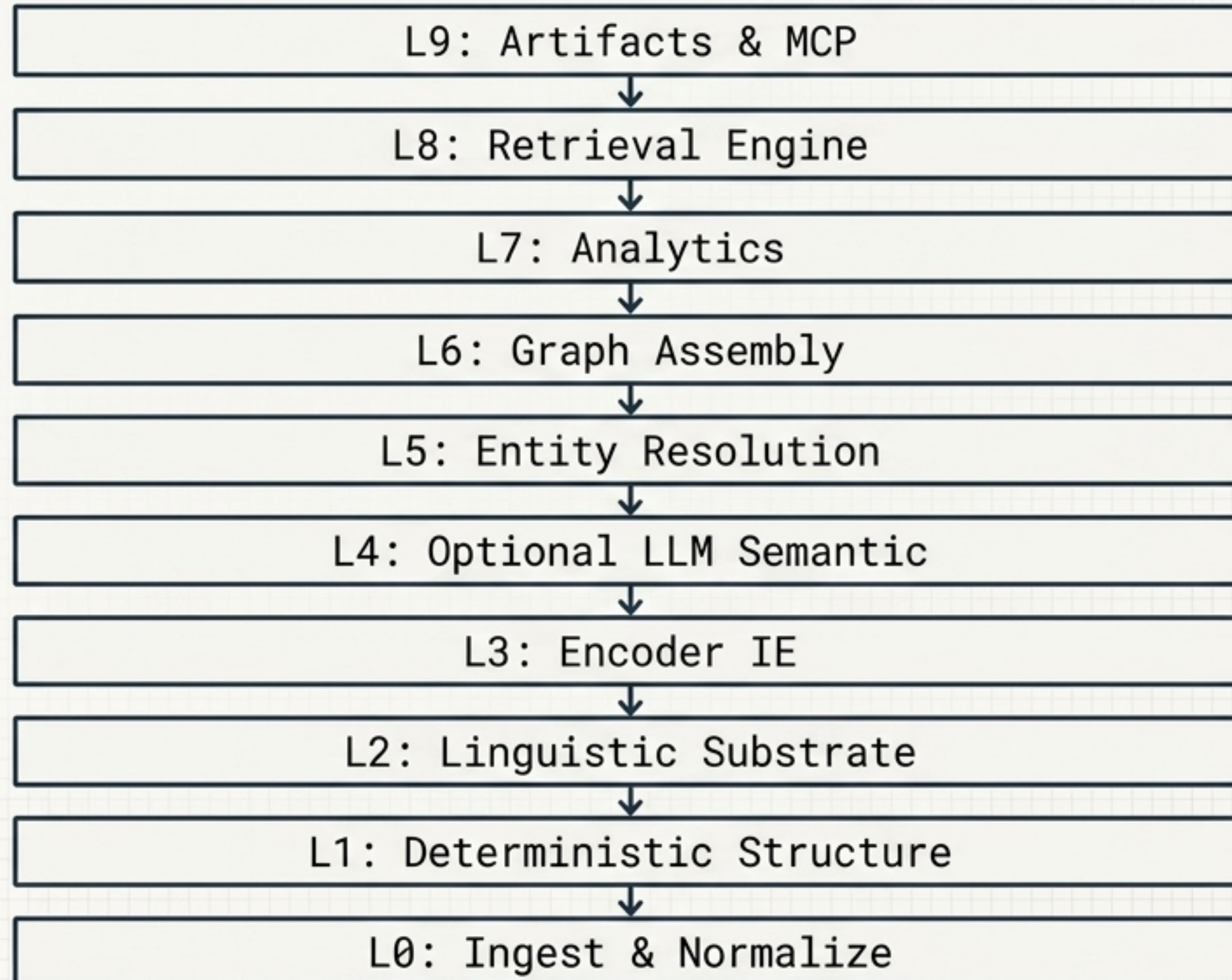
Per-stage token/time budgets emitted in the run manifest.

[AUDITABLE]

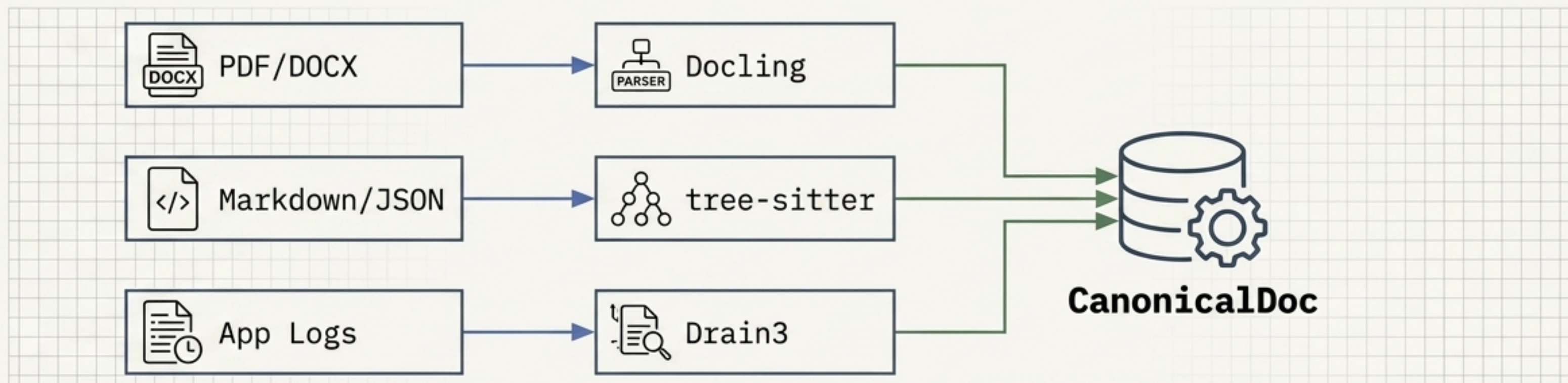


The Stratified Extraction Architecture

Every layer is a pure function of the layer below plus a pinned config hash.



Layer 0: Ingestion & Normalization



Key Engineering Invariants

- **Offset Fidelity:** Every character maps to (source_file, byte_offset). Run-length-encoded maps are mandatory for honest citations.
- **Content Addressing:** doc_id = blake3(raw_bytes). The foundation of incrementality.

Chunking Strategy

```
{
  "chunk_id": "b3:...",
  "doc_id": "b3:...",
  "breadcrumb": ["Doc", "$2 Architecture", "$2.3 Storage"],
  "span": [14203, 16891],
  "token_count": 612,
  "lang": "en",
  "layout_type": "paragraph|table|list|code|quote|transcript_turn",
  "prev_chunk": "b3:...",
  "next_chunk": "b3:..."
}
```

Hierarchical, layout-aware splitting. Never use fixed 512-token windows. Attach heading breadcrumbs to every chunk.

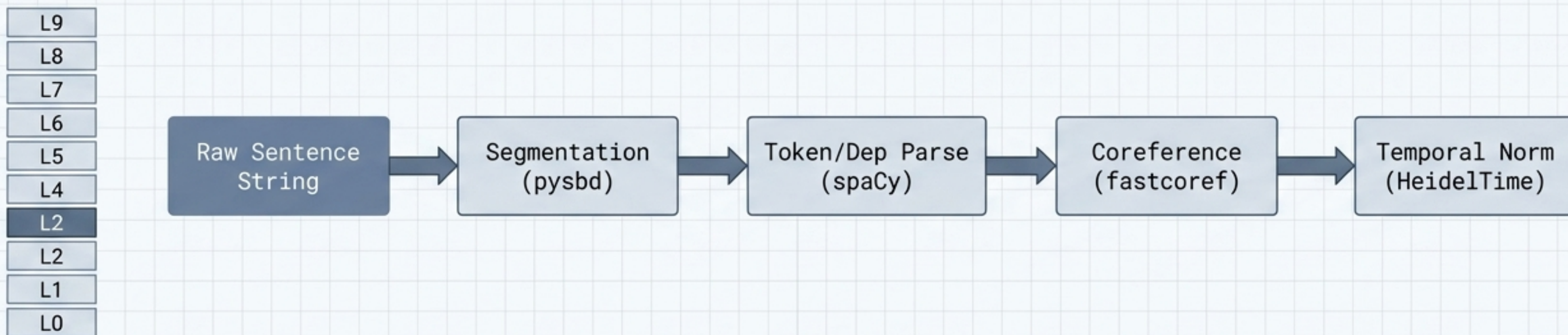
Layer 1: Deterministic Structure Parse (0 Models)

Typically 30-50% of final edge count, built in seconds.

Text Signal	Extracted Edge	Determinism
Heading hierarchy	CONTAINS(section_a, section_b)	100%
Table structure	Table -> Row -> Cell (via Docling)	100%
SQL DDL	TABLE -> COLUMN (via tree-sitter)	100%
Structured markers (TODO:, WHY:, ADR-003)	Rationale Nodes	100%

Steal this from Graphify: Elevating rationale/decision blocks to first-class nodes lets agents answer why, not just what.

Layer 2: The Linguistic Substrate



The Highest-ROI Component

Coreference Resolution
(fastcoref / maverick-coref).

Without resolving 'the company', 'it', and 'they' at the document scope, multi-hop recall is destroyed.

Temporal Normalization

HeidelTime resolves relative dates like 'last quarter' against a document's `t_ref`.

Mandatory for bi-temporal modeling.

Modality/Negation

NegEx rule packs prevent extracting 'X CAUSES Y' from 'X does not cause Y'.

Attached as edge attributes.

Layer 3: Encoder-Based IE (The tree-sitter Analog)

L9

L8

L7

L3

L4

L3

L2

L1

L0

The Stack: GLiNER (Zero-shot NER) + gliner-relex (Joint NER & Relation Extraction in one forward pass).

(A) Schema-Free (OpenIE)

Extract raw (subject, predicate, object).
Max recall, max mess.
Requires PPR at retrieval.

(B) Induced Schema (Recommended)

Cheap pass over corpus
sample -> cluster ->
propose schema -> freeze
to schema.yaml ->
constrain full extraction.

(C) User-Supplied Ontology

Accept SHACL/YAML,
reject type-violating triples.
Kills large classes of
hallucinated edges.

The Confidence Taxonomy

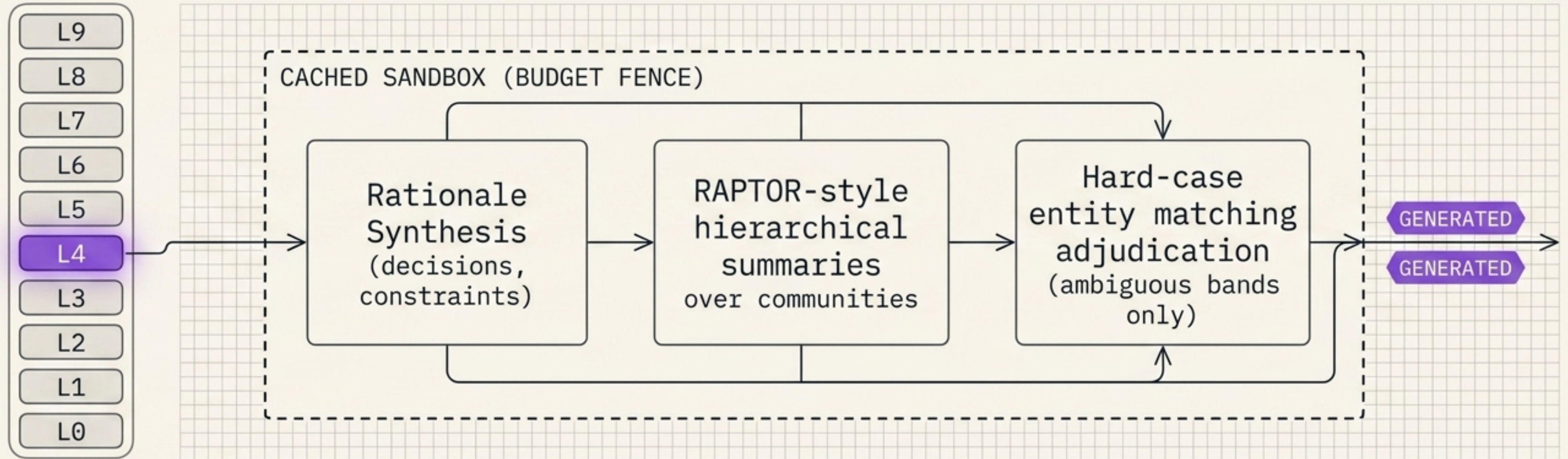
Every edge carries a confidence tag and evidence count.

STRUCTURAL L1 STRUCTURAL (L1) 6.Spt: sanet-esocitit-royact	Read directly from machine-readable structure. 6.Sptx: machine-readable layess
EXTRACTED L2/L3 EXTRACTED (L2/L3) 6.Spt: sonst-csocitln-byead	Stated explicitly in text; exact supporting span stored. 0.Sptlx: wapposting span stored
INFERRED L2/L5/L7 INFERRED (L2/L5/L7) 0.Spt: sead Antezhbnbest	Derived deterministically by the engine (coref, alias propagation). Not stated explicitly in text. Not expliitly in text
GENERATED L4 GENERATED (L4) Soft. Anethyst Purple	Produced by an LLM. Never treated as primary evidence. Never trested as pcinary evidence

Rule: Retrieval must filter by tag. An agent can request "STRUCTURAL & EXTRACTED only."

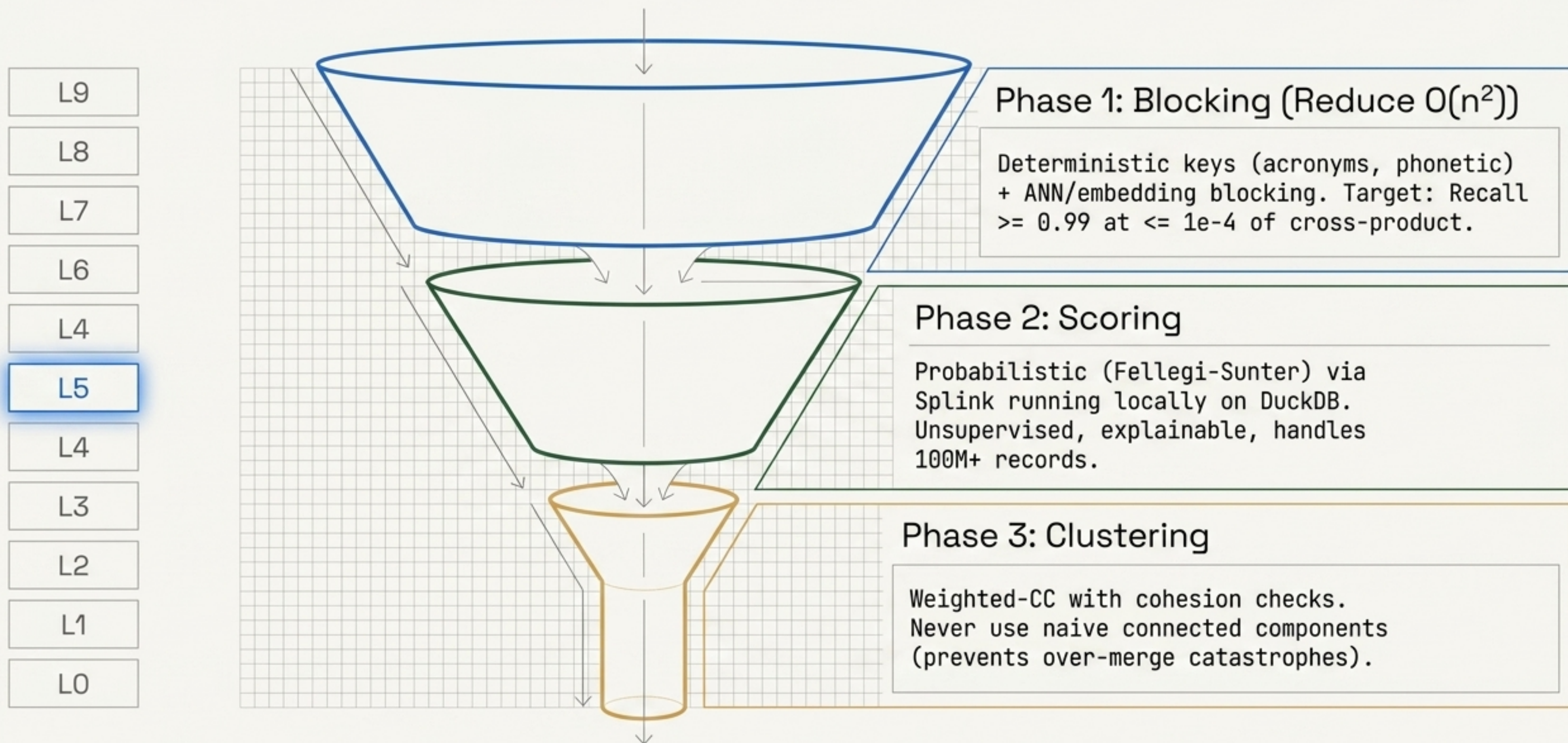
Layer 4: The Optional LLM Pass

Reserve LLMs for what encoders genuinely cannot do. All operations are opt-in, bounded by `llm.max_tokens`, and highly cacheable.

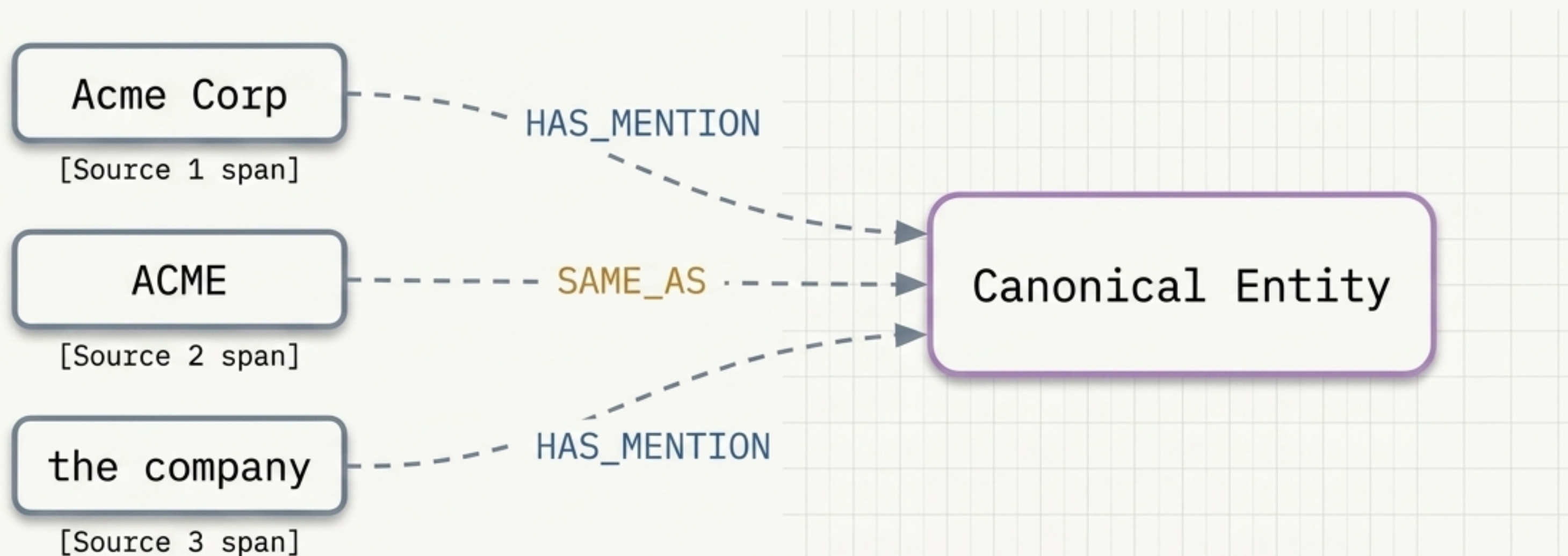


Cost Control: The `--no-llm` flag must produce a fully functional baseline graph.

Layer 5: Entity Resolution (Where Naive Text KGs Die)

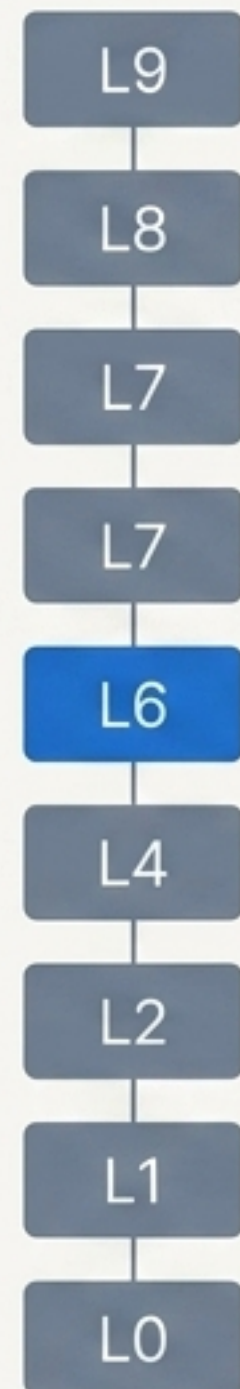


Reversible Entity Resolution

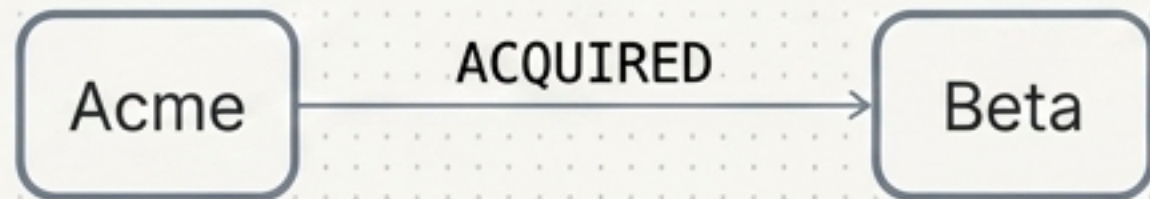


- The Rule: Emit SAME_AS edges rather than destructively merging nodes.
- Keep every original mention node tied to its exact byte-span.
- Result: Entity resolution becomes reversible, auditable, and tunable at query time. A massive operational advantage for provenance.

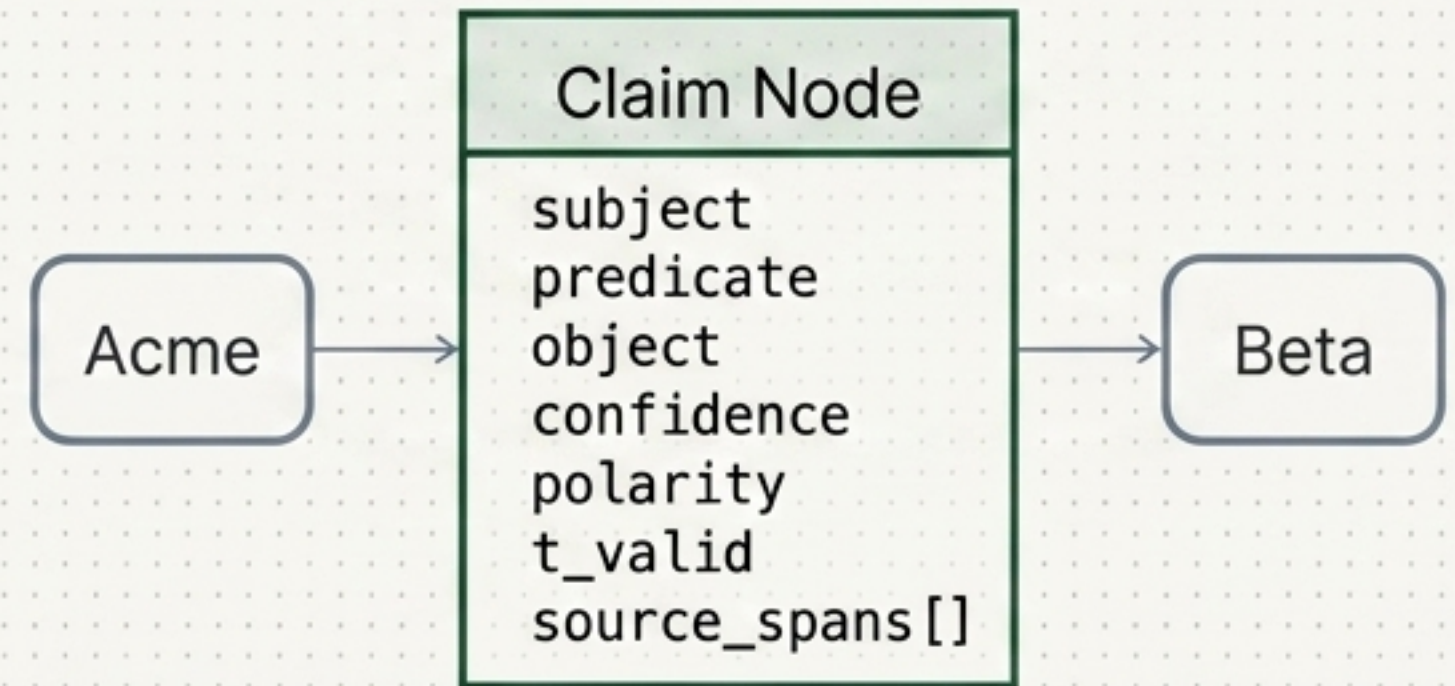
Layer 6: Graph Model & Reification



Before: Bare Edge (Anti-pattern)

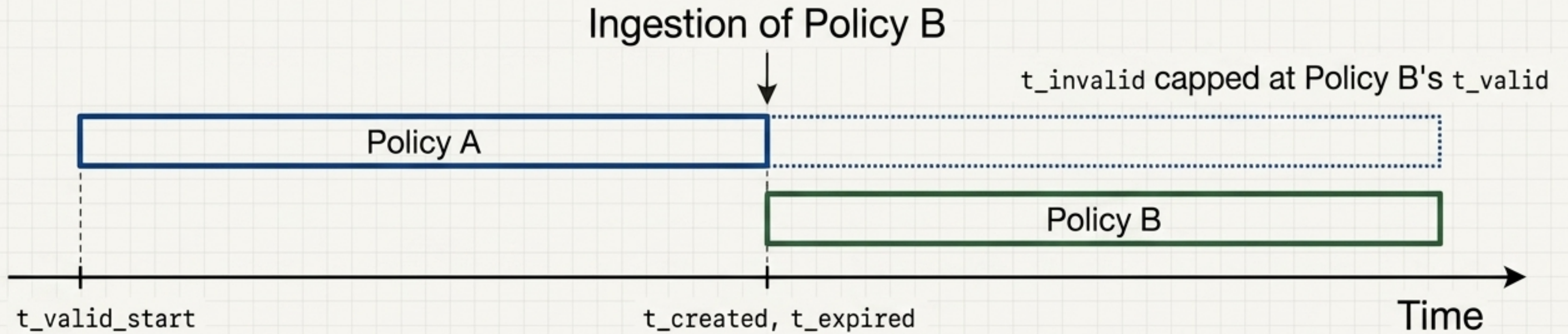


After: Reified Claim Node



- Why? Reification costs storage but buys **multi-source evidence aggregation**, contradiction handling, and honest **provenance**.

Bi-Temporal Modeling (Invalidation > Deletion)



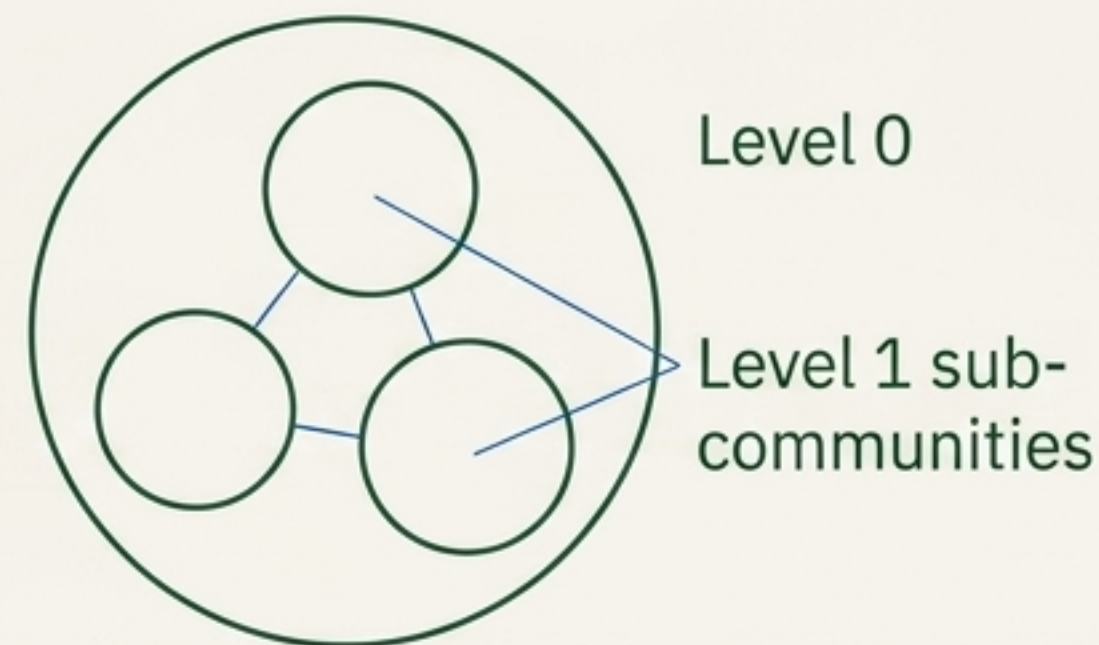
- Standard KGs treat facts as timeless. Temporal KGs track how facts change.
- Track 4 timestamps per fact: Event Time (t_{valid} , t_{invalid}) and Ingestion Time (t_{created} , t_{expired}).
- Enables agents to answer: "What did the contract say in March, and when did that change?"

Storage Engine Showdown

Option	Verdict for you
NetworkX + Parquet	Recommended for v0.1. Zero install friction (pipx install). Fits <5M edges in RAM.
DuckDB + Recursive CTEs	Underrated pragmatic choice. Single dependency, columnar, zero-server.
Kuzu / LadybugDB	Technically ideal embedded Cypher, but ecosystem stability dictates waiting.
Neo4j	Best tooling, but too heavyweight for a CLI default. Provide as a pluggable backend.

12
11
L9
L6
L8
L7
L6
L4
L1
12

Layer 7: Analytics & Leiden Communities



Community Detection

- Use **Leiden**, not **Louvain** (guarantees well-connected communities).
- Fix the random seed (G1 determinism).
- Apply edge weights: confidence * $\log(1 + \text{evidence_count})$.

Labeling without LLMs

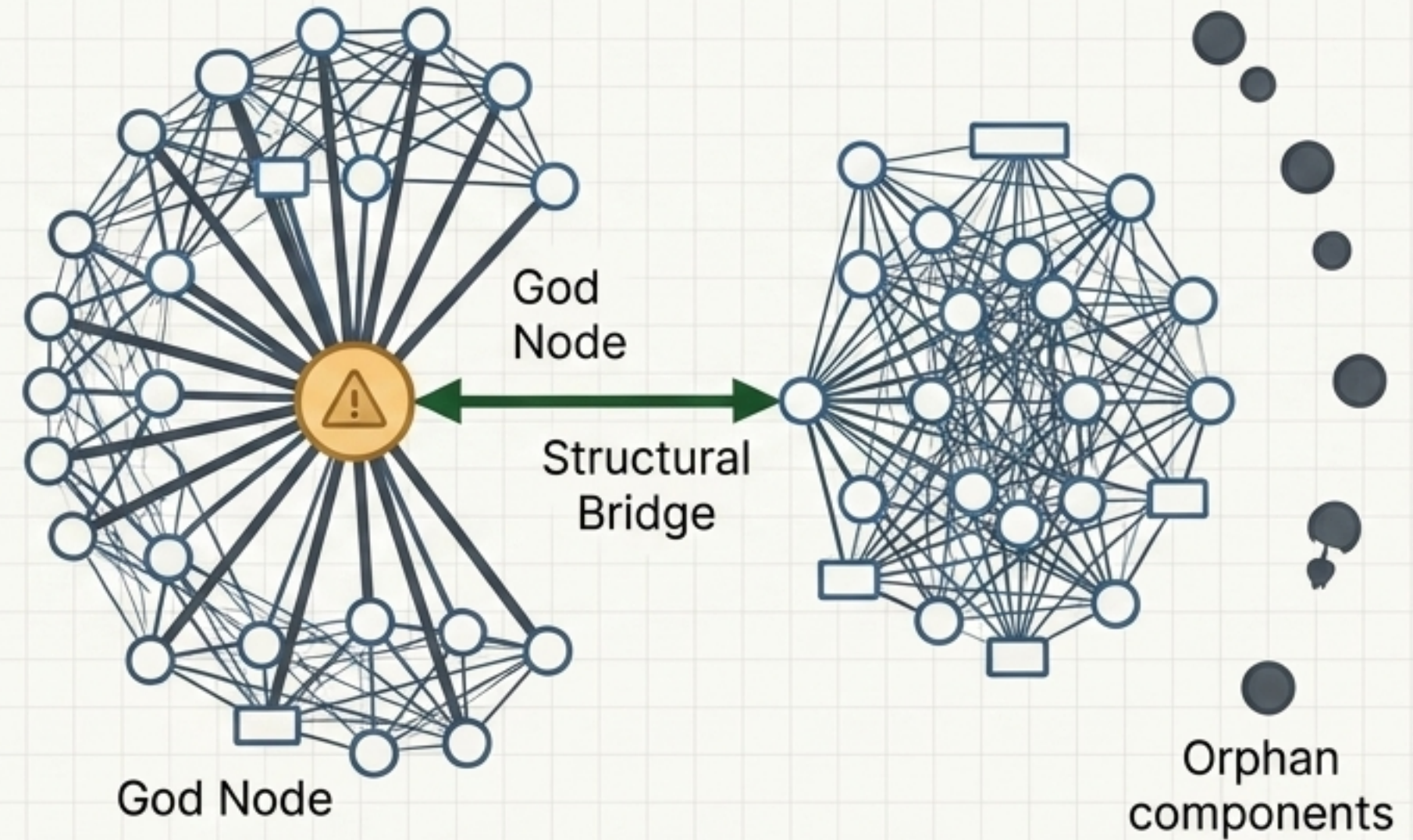
1. Collect mentions + section headings.
2. Compute c-TF-IDF (term freq within community vs across all).
3. MMR to reduce redundancy.
4. Label = Top terms + highest-centrality entity.

Extracting Structural Insight

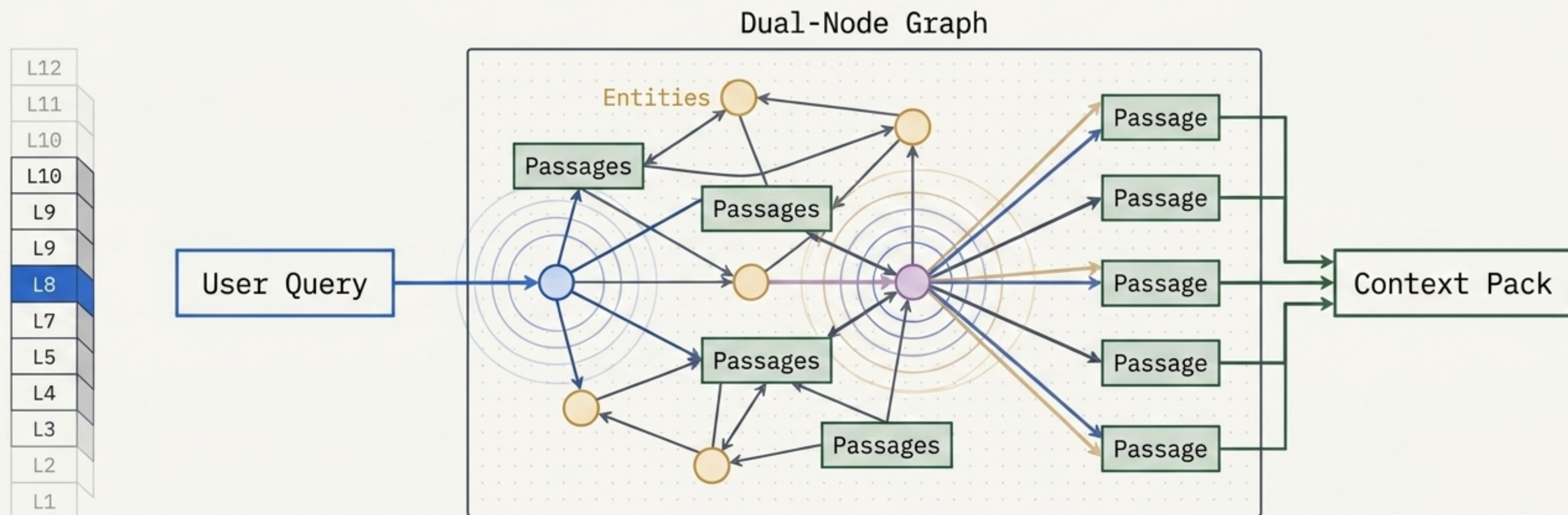
God Nodes: Top percentile by PageRank/betweenness. Usually signifies central concepts—or catastrophic ER over-merge bugs.

Bridges (Structural Holes): Edges whose removal disconnects communities. Generate the “surprising insights” for agent reports.

Contradictions: Claim nodes with identical subjects/predicates, incompatible objects, and overlapping validity intervals. Surfaced via CONTRADICTS edges.



Layer 8: The Retrieval Engine (HippoRAG Backbone)



Architecture:

A dual-node graph containing both phrase/entity nodes and passage nodes.

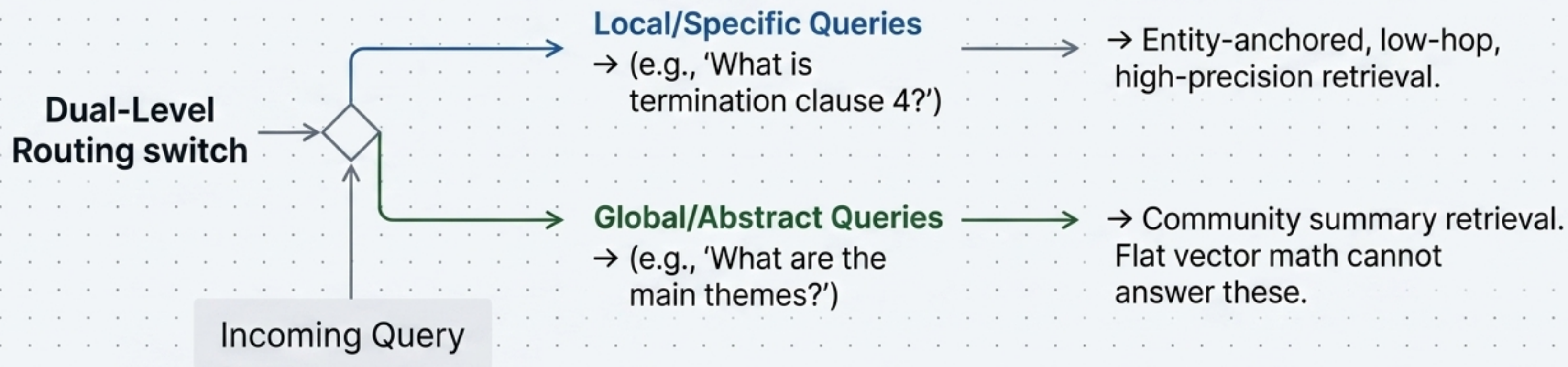
Why PPR over Cypher?

Soft, multi-hop traversal degrades gracefully when the graph is imperfect.
Unifies dense and sparse retrieval natively.

Hybrid Scoring & Query Routing

$$\text{score} = w1(\text{BM25}) + w2(\cos(\text{query}, \text{embed})) + w3(\text{PPR}) + w4(\text{structural_prior})$$

Fuse using Reciprocal Rank Fusion (RRF), then rerank the top 50 with a local cross-encoder (bge-reranker-v2-m3).



Synthesis: The Agent Query Lifecycle

