

Retrieval Engineering

Beyond Cosine Similarity with [Luminary](#)

Anup Sethuram

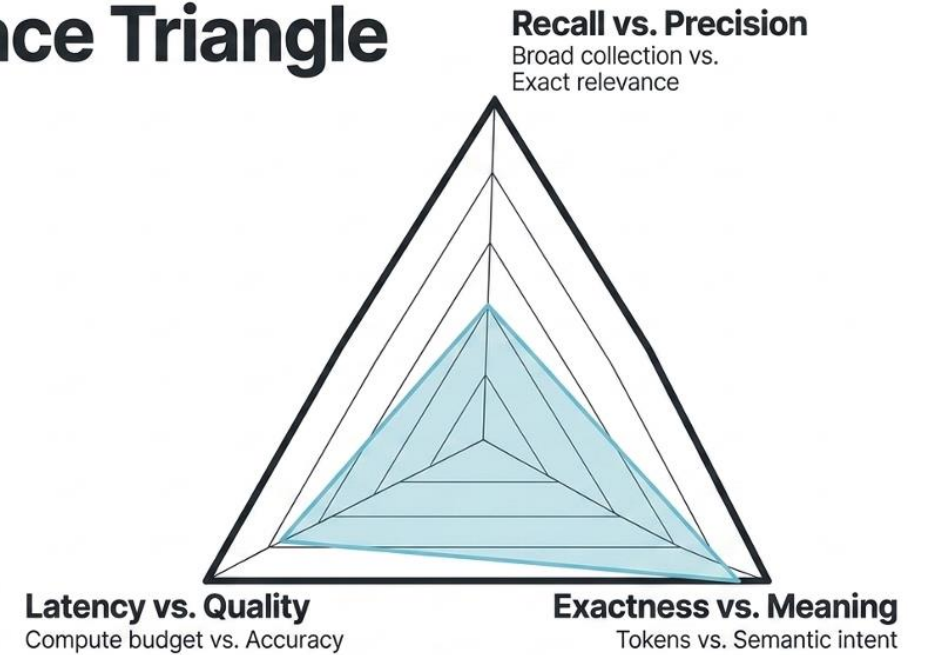
The Prototype Trap

Let us build a chat assistant using RAGs on the INTERNAL company documentation and be AI ready



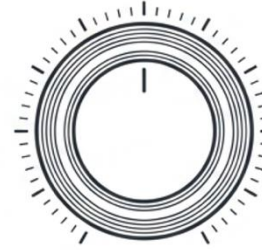
Age Old Problem

The Relevance Triangle



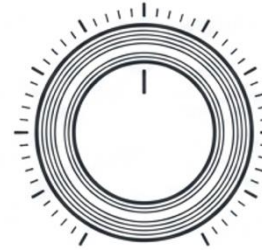
Leg1: Lexical – Exact Match

TERMS		DOCUMENT IDs
photosynthesis	→	pp. 88, 214
mitochondria	→	pp. 102, 150
chlorophyll	→	pp. 91, 115, 220
ATP	→	pp. 105, 132



k1 (≈ 1.2):
Term Saturation.

The 'salt in soup' effect.
The 50th occurrence
matters less than the first.



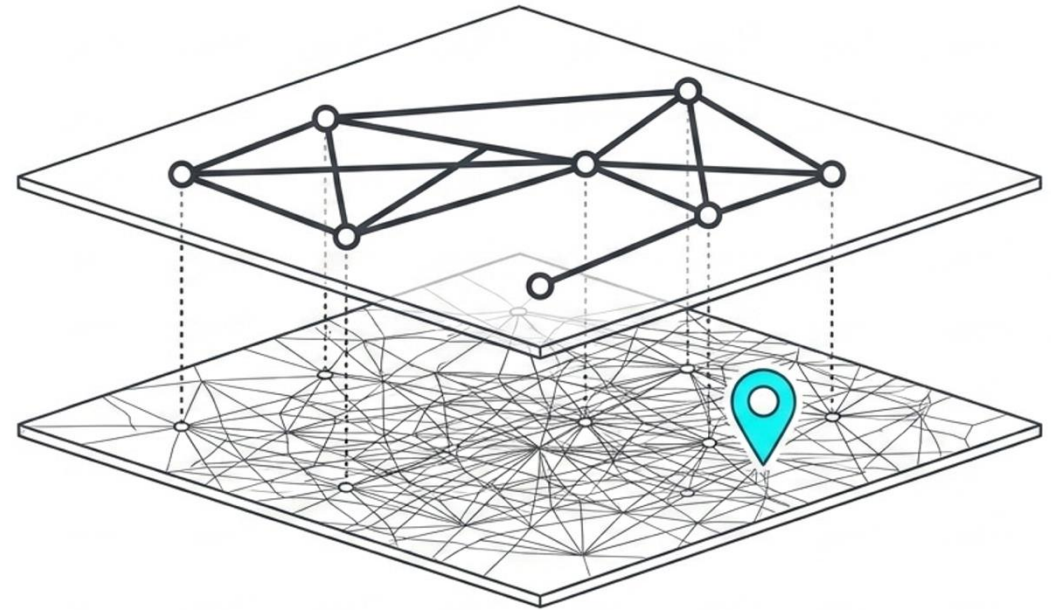
b (≈ 0.75): **Length**
Normalisation.

Grades documents fairly
regardless of word count.

Luminary: FTS5, SQLite implementation of BM25

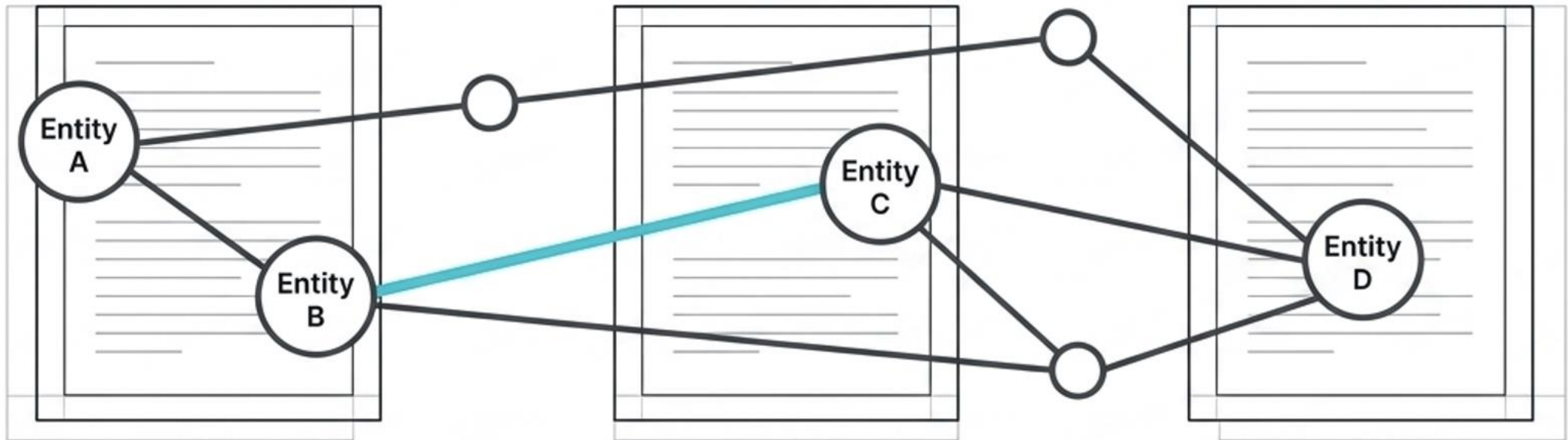
Leg2 : Vector – Meaning Match

- BI Encoders
- ANN indexes - HNSW



Luminary: bge-small-en-v1.5 embedding (384-dim) + LanceDB

Leg 3: Relational – Multi Hop

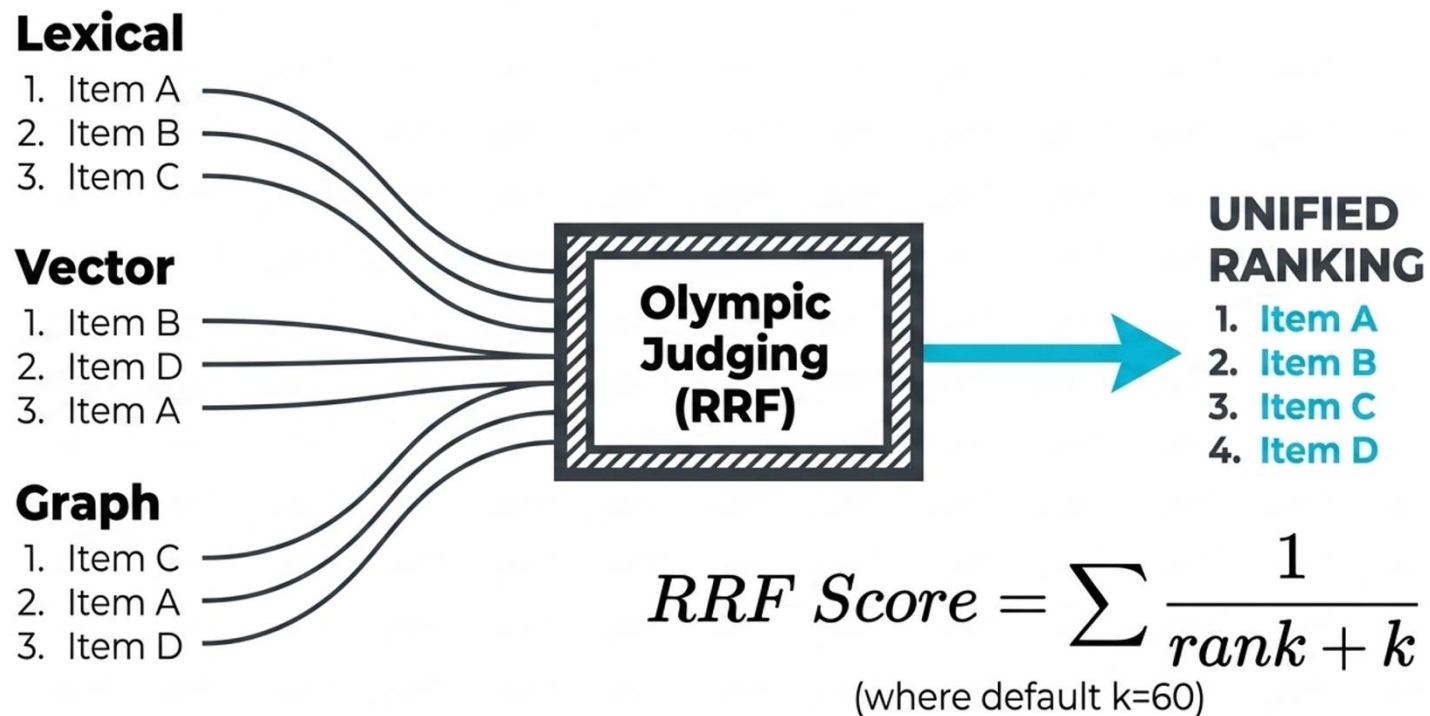


Luminary: GLiNER extracted entities into Kuzu store

Selecting the Right Index

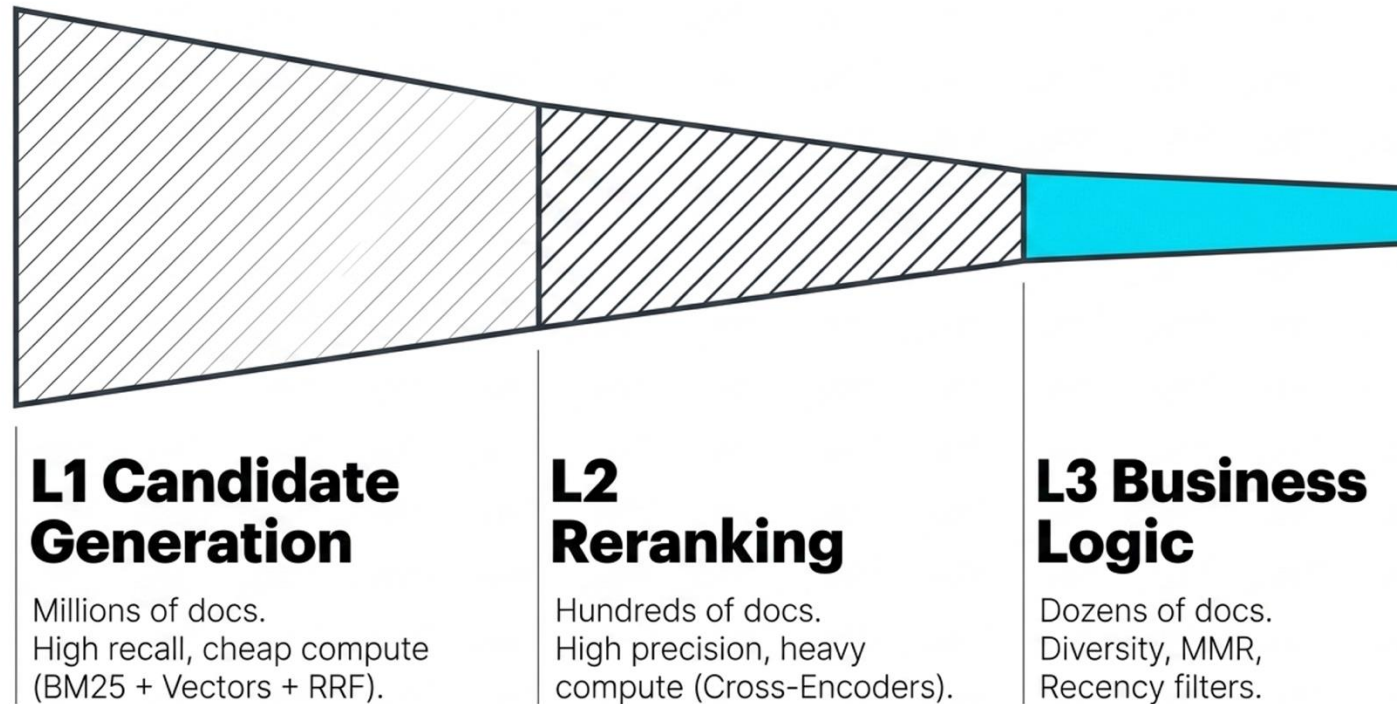
	LEXICAL	VECTOR	GRAPH
MECHANISM	Exact token match	Dense semantic mapping	Entity relation edges
SUPERPOWER	Acronyms, IDs, Jargon	Intent, Meaning, Cross-Lingual	Multi-hop reasoning, global themes
KRYPTONITE	Paraphrasing, Synonyms	Exact identifiers	Factoid Lookup
Luminary Stack	SQLite FTS5	LanceDB (bge-m3)	Kuzu (GLiNER)

The Fusion Problem



Rank-based fusion: Consensus outranks any lone champion.

Funnel Architecture



Metrics – With Evals



HR@5 ≥ 0.60

(Hit Rate) Did the relevant document survive L1?



MRR ≥ 0.45

(Mean Reciprocal Rank) Were the documents ordered correctly by L2?

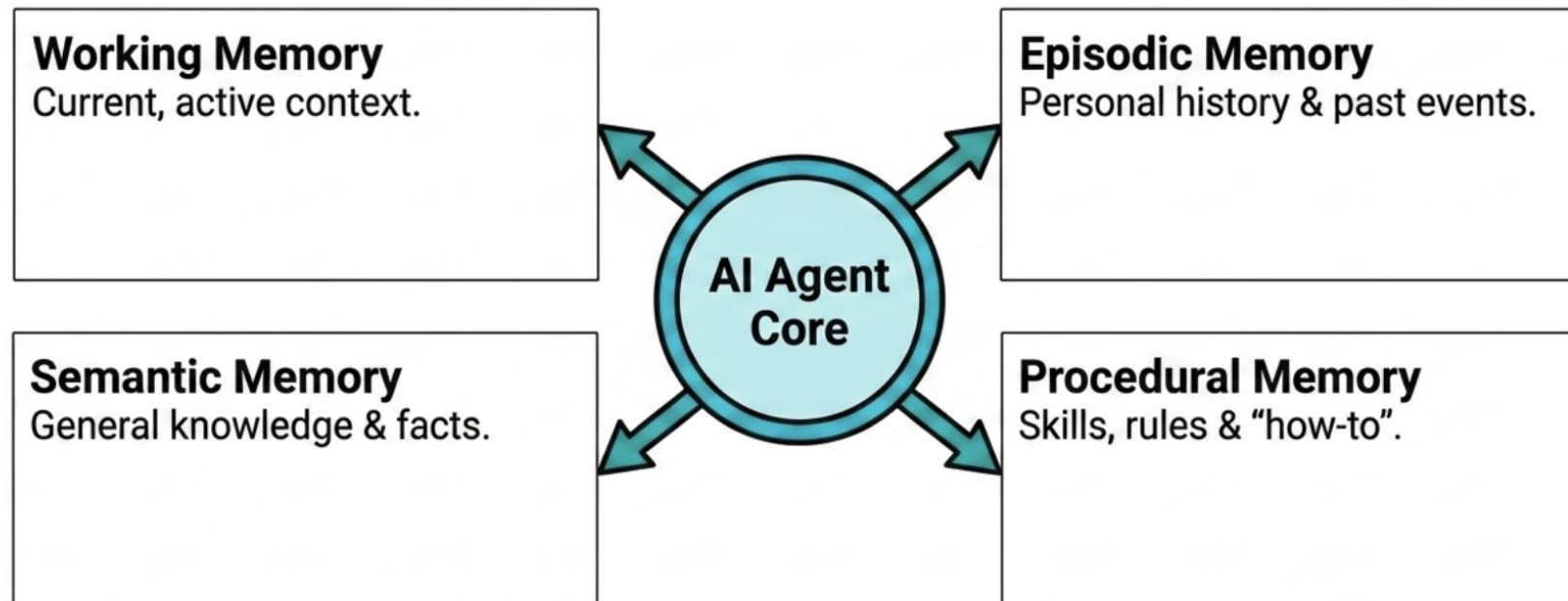


Faithfulness ≥ 0.65

Did the LLM actually use the retrieved context?

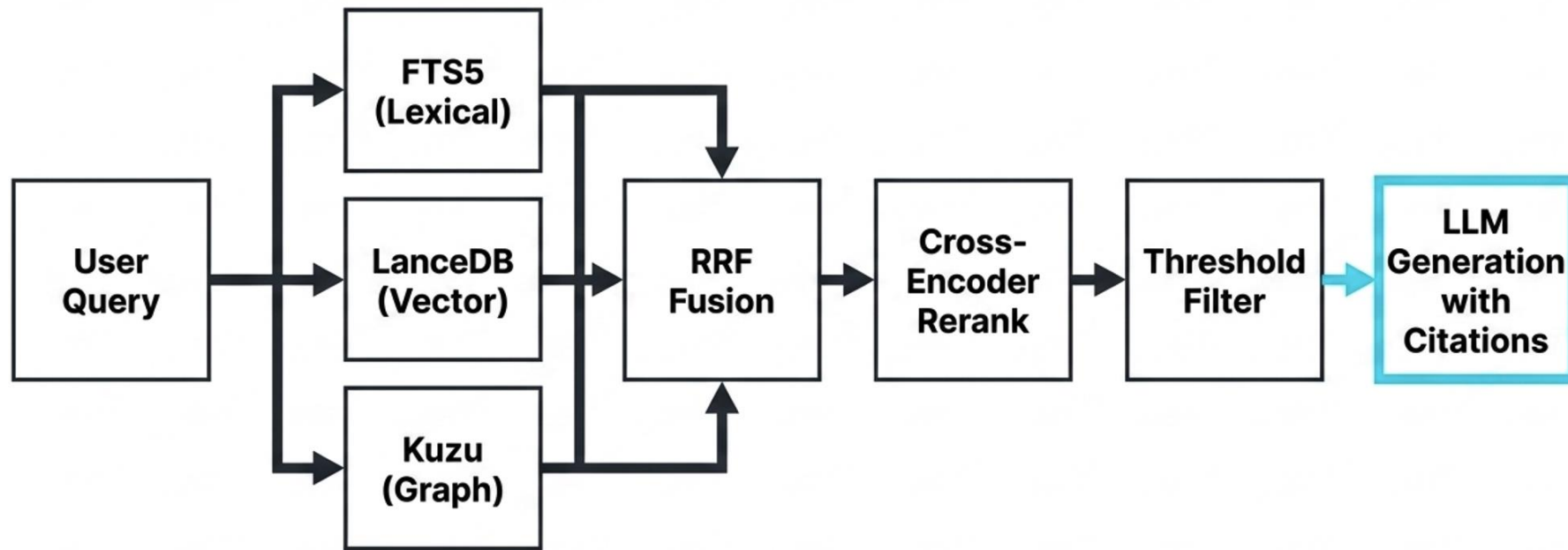
Luminary: Generate golden sets, measure quality and identify gaps through different eval metrics.

Context as Agentic Memory

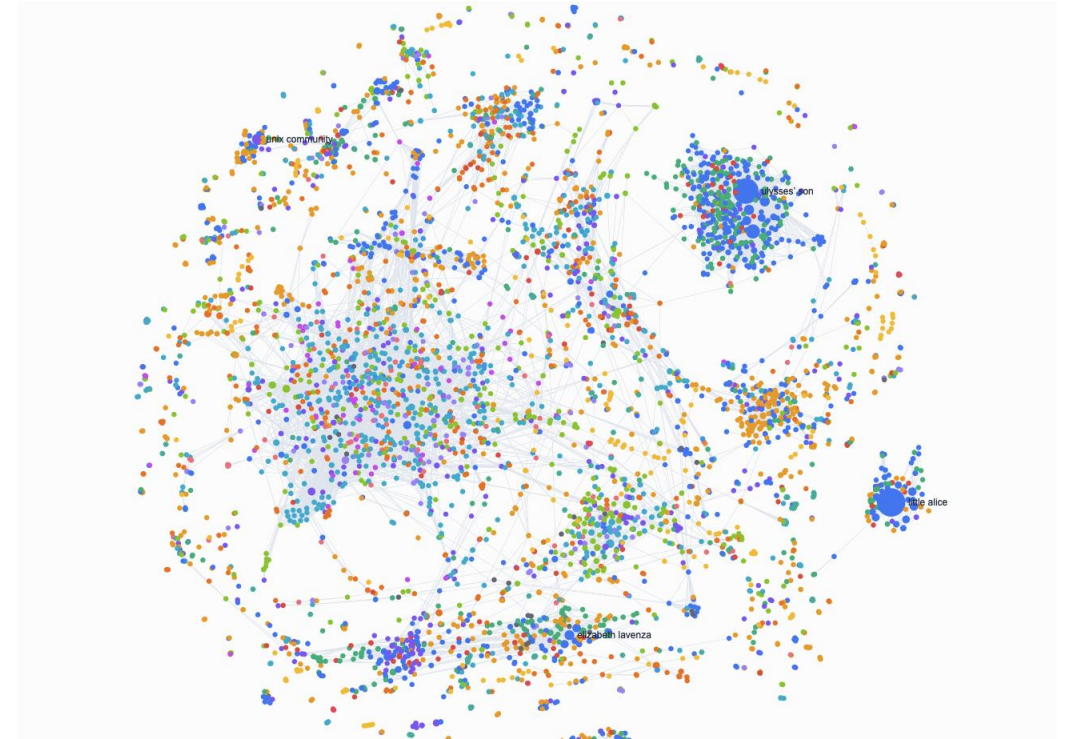


Luminary: FSRs manages human memory using the same math (decay/TTL) that agents use for vector retrieval.

Luminary Retrieval Architecture



Luminary Demo



References & QnA

- Hands On RAG for Production - [Ofer Mendelevitch](#), [Forrest Sheng Bao](#)
- Designing Data Intensive Applications – [Martin Kleppman](#)
- <https://www.normaltech.ai/p/what-will-be-left-for-us-to-work>
- <https://martin.kleppmann.com/2024/05/30/local-first-conference.html>
- Claude Code, NotebookLM
- nupsea.github.io



RAG - Prod Solution

- Data Refresh & Pipelines
- Performance & Optimization
- Security & Governance
- Evals & Observability
- Controls, CI/CD & Ops