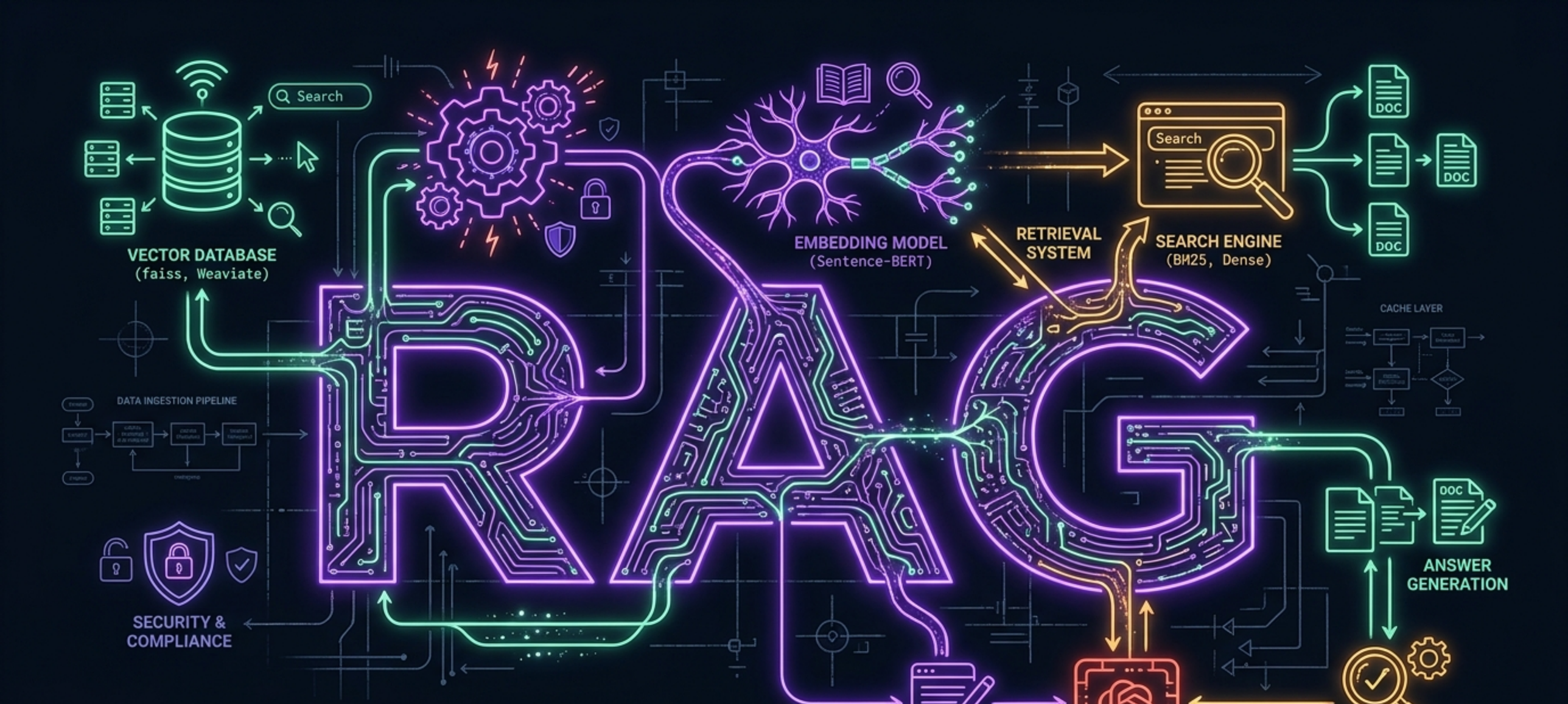


Retrieval-Augmented Generation (RAG): Production Blueprint

A comprehensive architectural guide to moving beyond prompt wrappers.

Gemini Notebook

Fine-Tuning

A semester of study.

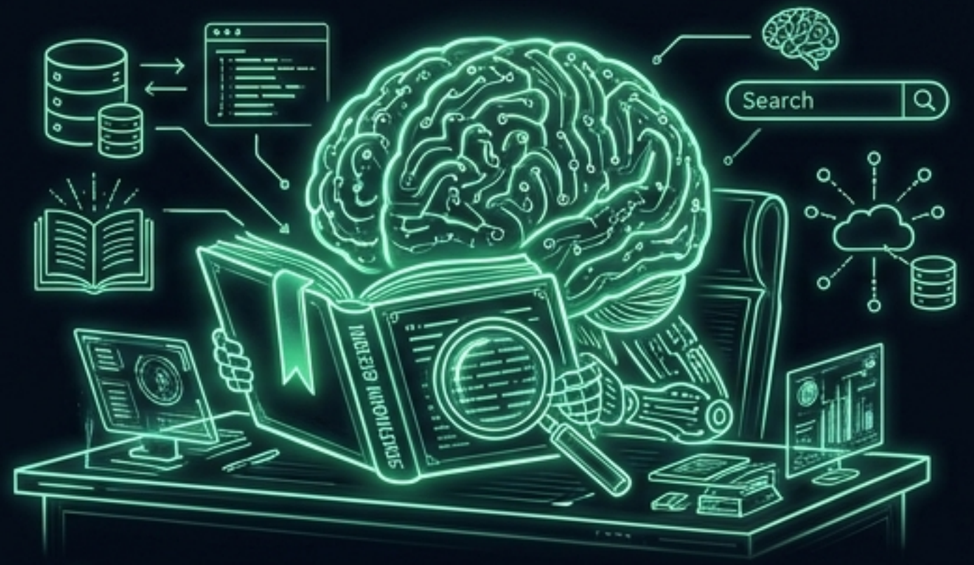


Teaches a model **HOW to think, format,** and speak **like a specialist.**

Facts **cannot be traced to weights.**

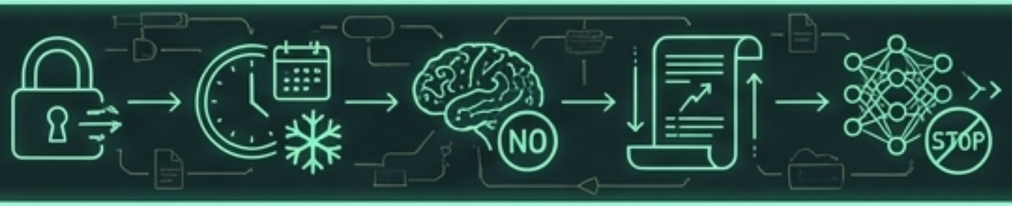
RAG (Retrieval-Augmented Generation)

The open-book exam.



Pairs parametric knowledge with **external retrieval.**

Answers **ONLY** from context, with **traceable citations.**



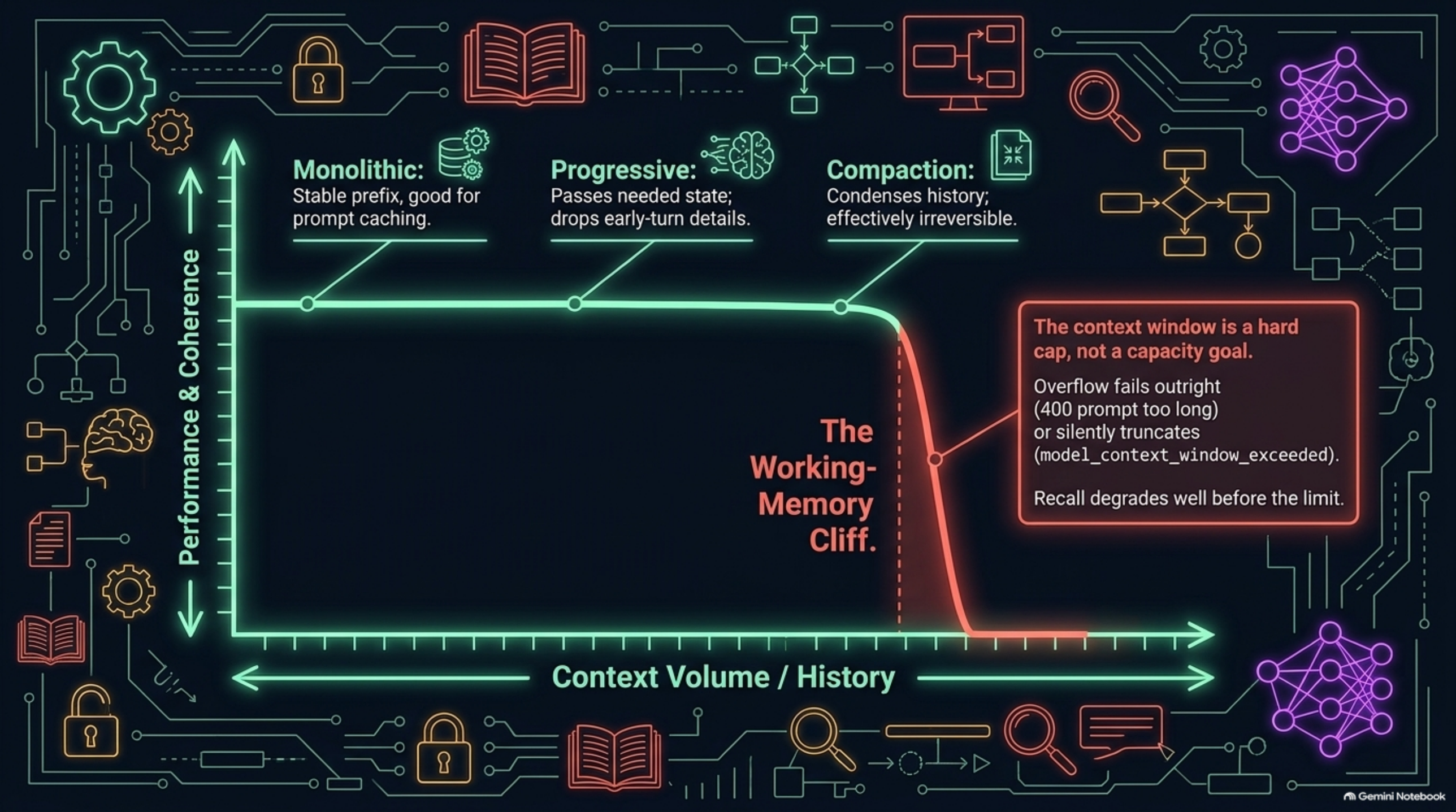
Solves **private data access**, **freshness** (frozen weights), **hallucinations**, and **long-context scaling costs** without retraining.

Gemini Notebook

Approach	Best For	Update Speed	Cost per Query	Traceability
Fine-Tuning	Tone, format, behavior, output schemas	Hours to weeks	Low inference, high training	None
Long-Context	Whole-doc reasoning on small corpus	Instant	High (linear growth)	Coarse
RAG	Large, changing reference facts	Seconds to minutes	Low (~1.5k-3k tokens)	Exact (doc, chunk, page)
Live Tool Call	Live transactional state (balances, orders)	Real-time	Low (1 API/DB call)	Exact (record ID)

Rule of Thumb: RAG supplies reference facts, Fine-tuning shapes behavior, Long context does deep reasoning, Tool calls fetch live state.

Gemini Notebook



Gemini Notebook

MACRO-ARCHITECTURE FLOWCHART

OFFLINE (Once per change)

Expensive work happens once so queries stay fast.

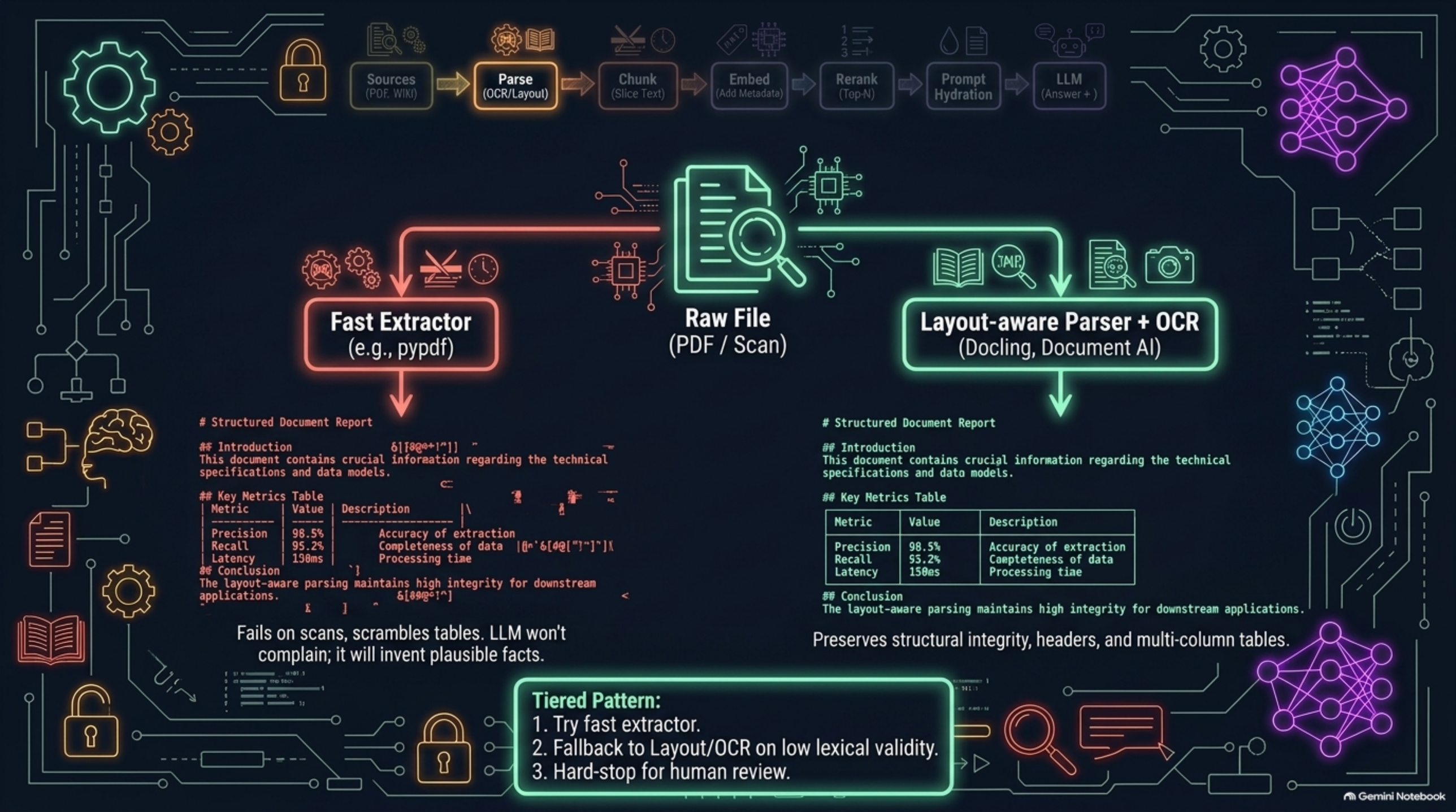


ONLINE (Every query)

Query embedded with the exact same model & version as the corpus.



Gemini Notebook


Sources
(PDF, WIKI)
Parse
(OCR/Layout)
Chunk
(Slice Text)
Embed
(Add Metadata)
Rerank
(Top-N)
Prompt
Hydration
LLM
(Answer +)
Fast Extractor
(e.g., pypdf)
Raw File
(PDF / Scan)
Layout-aware Parser + OCR
(Docling, Document AI)
Structured Document Report
Introduction
This document contains crucial information regarding the technical specifications and data models.
Key Metrics Table
Metric
Value
Description
Precision
98.5%
Accuracy of extraction
Recall
95.2%
Completeness of data
Latency
150ms
Processing time
Conclusion
The layout-aware parsing maintains high integrity for downstream applications.
Fails on scans, scrambles tables. LLM won't complain; it will invent plausible facts.
Preserves structural integrity, headers, and multi-column tables.
Tiered Pattern:
1. Try fast extractor.
2. Fallback to Layout/OCR on low lexical validity.
3. Hard-stop for human review.
Gemini Notebook

Sources (PDF, WIKI) → Parse (OCR/Layout) → **Chunk (Slice Text)** → Embed (Add Metadata) → Rerank (Top-N) → Prompt Hydration → LLM (Answer +)

CHUNKING STRATEGIES

FIXED-SIZE

Default starting point.
Uniform text.

SEMANTIC

Long prose. Splits where embedding similarity drops.

STRUCTURE-AWARE

HEADER

AST

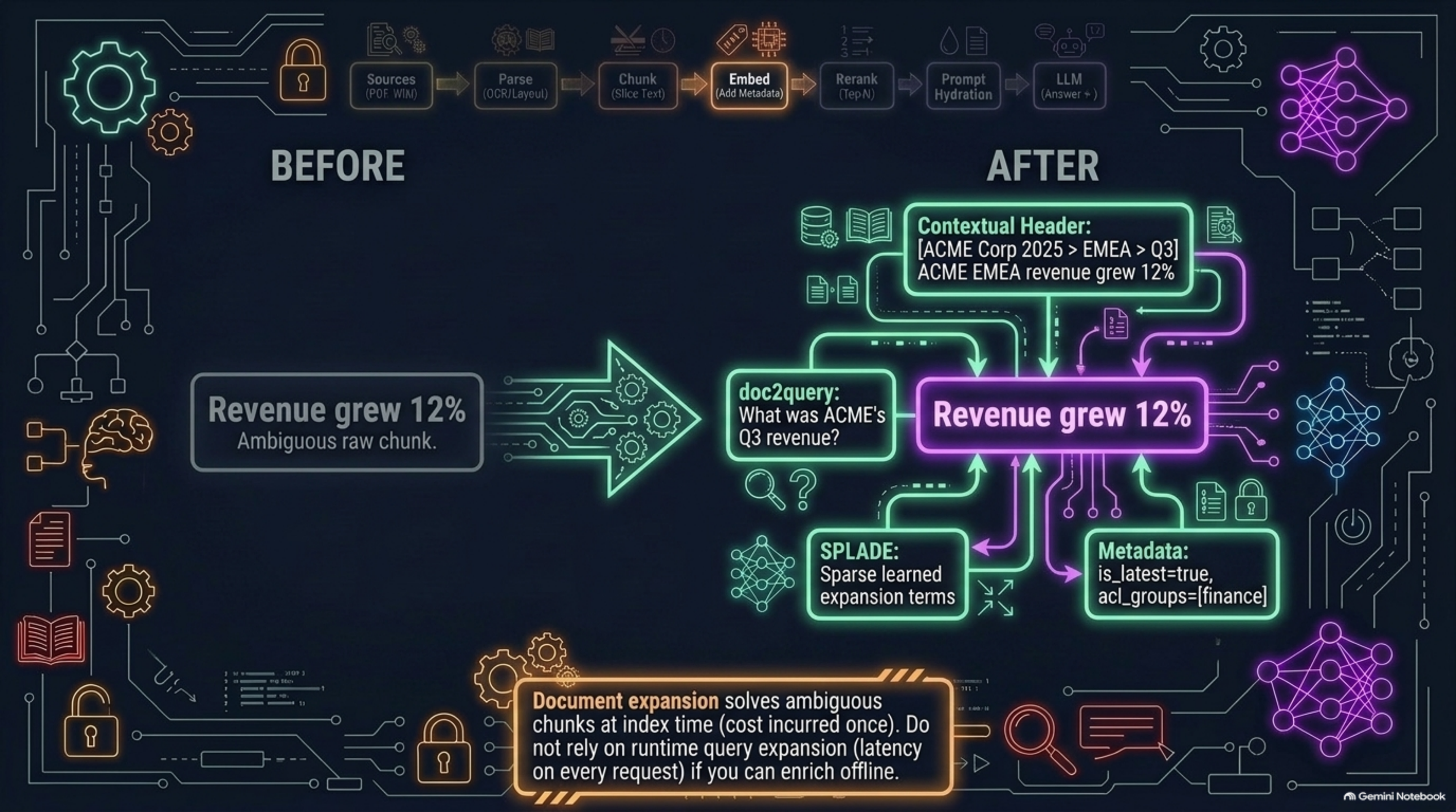
Codebases, markdown, legal clauses.

PARENT-CHILD (HIERARCHICAL)

Embed small children for precision, return parent for context.

Constraint: Embedding Input Limit.
Models cap at e.g., 2,048 tokens. Default autoTruncate=true means longer chunks are silently severed and their tails are never searchable.

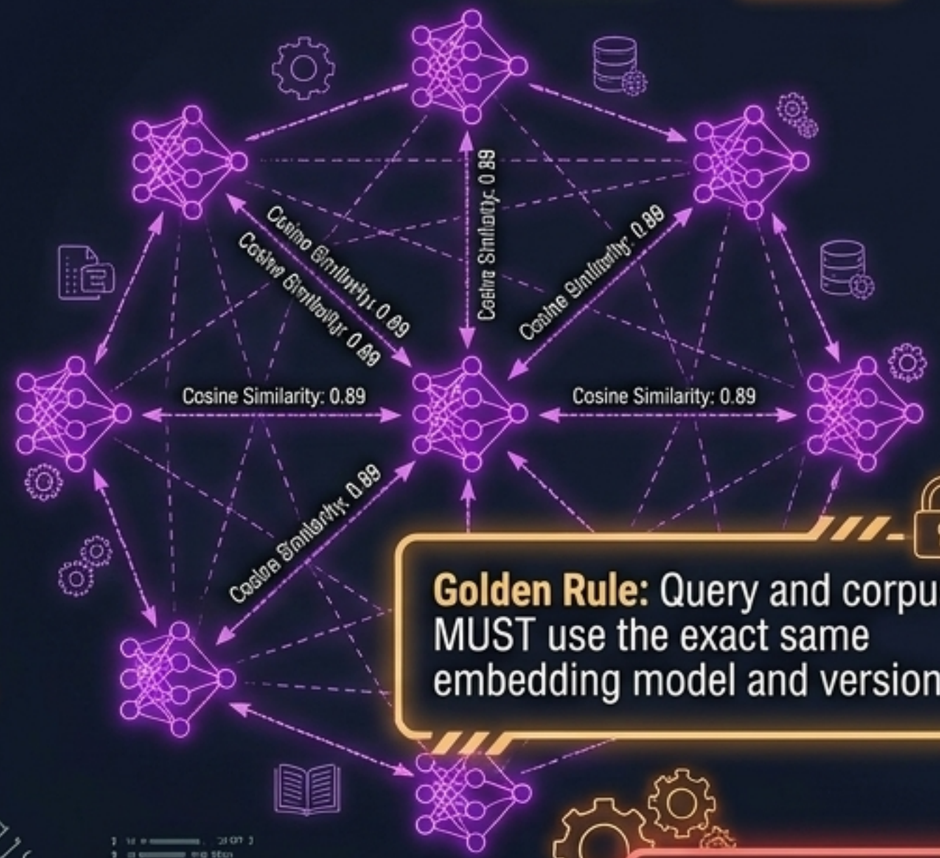
Gemini Notebook


Sources
(PDF, WIM)
Parse
(OCR/Layout)
Chunk
(Slice Text)
Embed
(Add Metadata)
Rerank
(Top N)
Prompt
Hydration
LLM
(Answer +)
BEFORE
Revenue grew 12%
Ambiguous raw chunk.
AFTER
Contextual Header:
[ACME Corp 2025 > EMEA > Q3]
ACME EMEA revenue grew 12%
doc2query:
What was ACME's
Q3 revenue?
Revenue grew 12%
SPLADE:
Sparse learned
expansion terms
Metadata:
is_latest=true,
acl_groups=[finance]
Document expansion solves ambiguous chunks at index time (cost incurred once). Do not rely on runtime query expansion (latency on every request) if you can enrich offline.
Gemini Notebook

SEARCH STRATEGIES:

SEARCH STRATEGIES: EXACT VS. APPROXIMATE

Chunk (Slice Text) → Embed (Add Metadata) → Rerank (Top-N) → Prompt Hydration → LLM (Answer +)



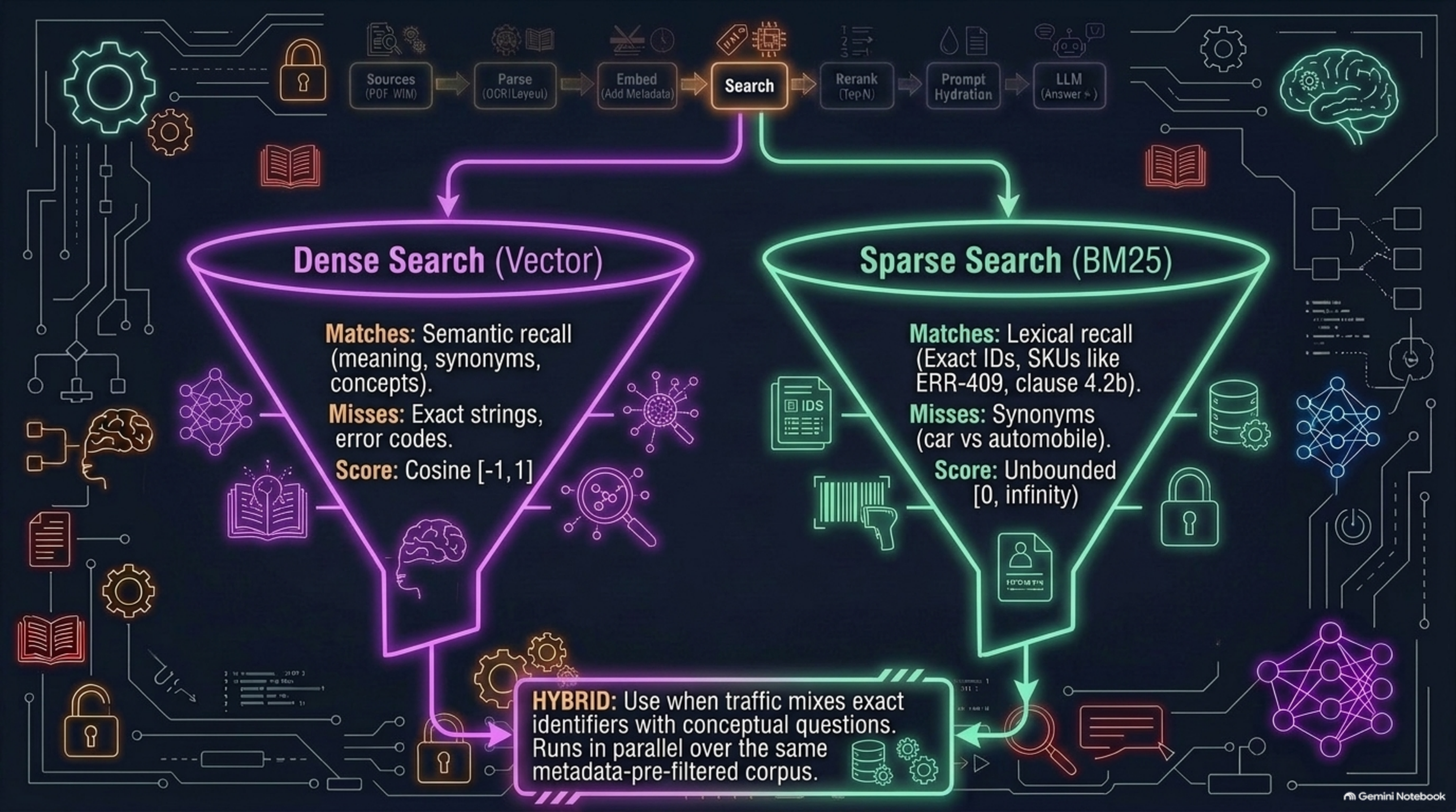
Golden Rule: Query and corpus MUST use the exact same embedding model and version.

Search Mechanics: Exact vs. Approximate

Exact k-NN	100% recall. Linear cost. Breaks past 100k vectors.
HNSW	Multi-layer proximity graph. Best recall/latency, high RAM.
ScaNN / Quantization	Highest QPS. Compresses float32 -> int8. Requires oversampling and rescoring.

Distance Math:
Cosine Similarity (higher = closer) vs Dot Product (computationally cheaper if L2 normalized).

Gemini Notebook


Sources
(PDF WIM)
Parse
(OCR Layeul)
Embed
(Add Metadata)
Search
Rerank
(TepN)
Prompt
Hydration
LLM
(Answer)
Dense Search (Vector)
Matches: Semantic recall (meaning, synonyms, concepts).
Misses: Exact strings, error codes.
Score: Cosine [-1, 1]
Sparse Search (BM25)
Matches: Lexical recall (Exact IDs, SKUs like ERR-409, clause 4.2b).
Misses: Synonyms (car vs automobile).
Score: Unbounded [0, infinity)
IDS
HYBRID: Use when traffic mixes exact identifiers with conceptual questions. Runs in parallel over the same metadata-pre-filtered corpus.
Gemini Notebook

Chunk (Slice Text) → Embed (Add Metadata) → **Rerank (Top-N)** → Prompt Hydration → LLM (Answer +)

Two-Stage Retrieval

BM25 (Sparse)

lexical
keywords

erecent
context

Step 1: Reciprocal Rank Fusion (RRF)

$$y = \sum_{k=0}^{\infty} \frac{1}{(k+r)}$$

Raw scores can't be added. RRF rewazds consensus based on rank position.

Dense (Vector)

semantic
context

organite
list

The Reranking Pass

Stage 1: Bi-Encoder

Precomputed, fast, high recall (K=100).

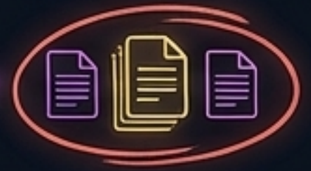


Stage 2: Cross-Encoder

Slow, full cross-attention transformer pass on fused pool. Optimizes for precision.

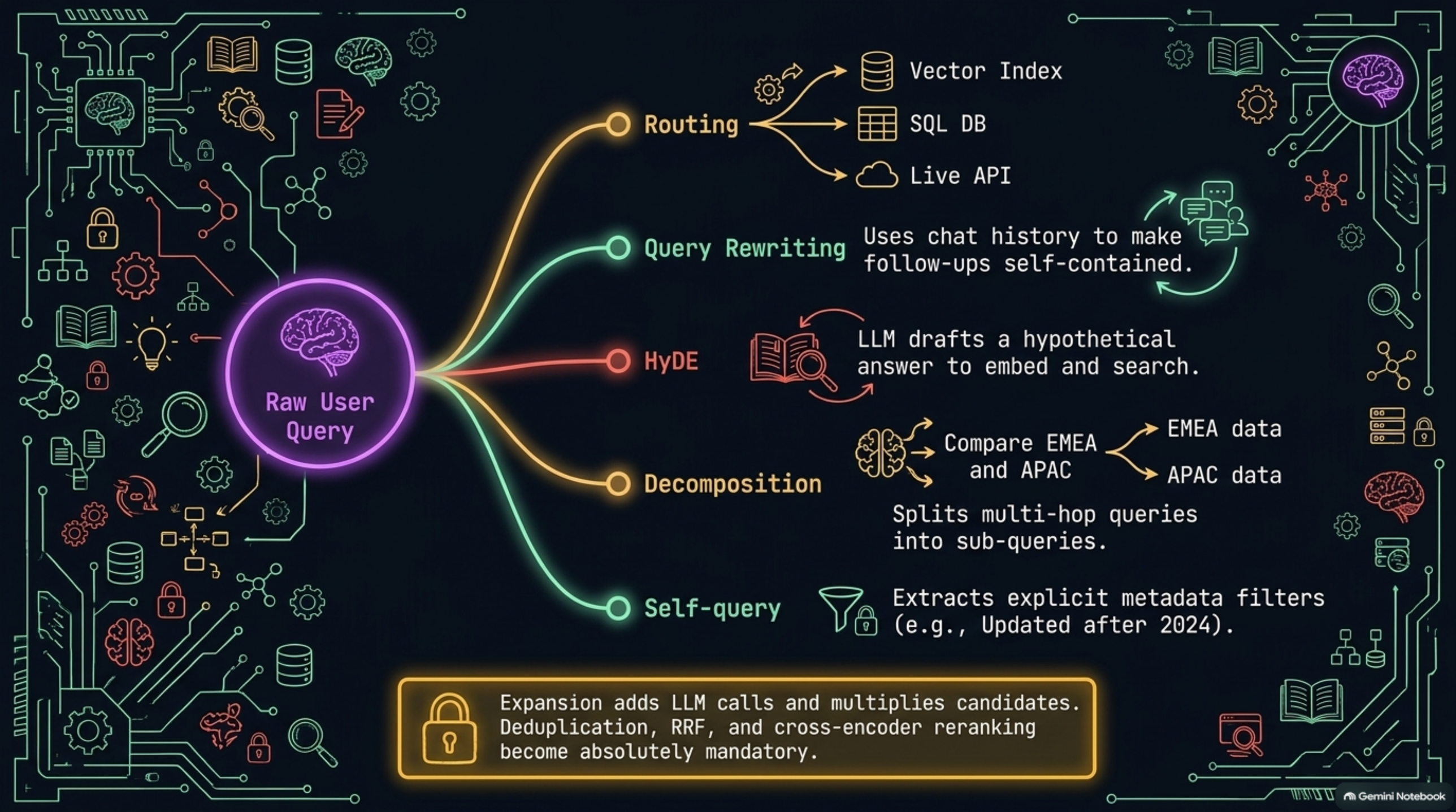


Final Top-N (3-10)



Never cross-encode the whole corpus. It requires one transformer pass per (query, chunk) pair.

Gemini Notebook


Raw User Query
Routing
Vector Index
SQL DB
Live API
Query Rewriting
Uses chat history to make follow-ups self-contained.
HyDE
LLM drafts a hypothetical answer to embed and search.
Decomposition
Compare EMEA and APAC
EMEA data
APAC data
Splits multi-hop queries into sub-queries.
Self-query
Extracts explicit metadata filters (e.g., Updated after 2024).
Expansion adds LLM calls and multiplies candidates. Deduplication, RRF, and cross-encoder reranking become absolutely mandatory.
Gemini Notebook

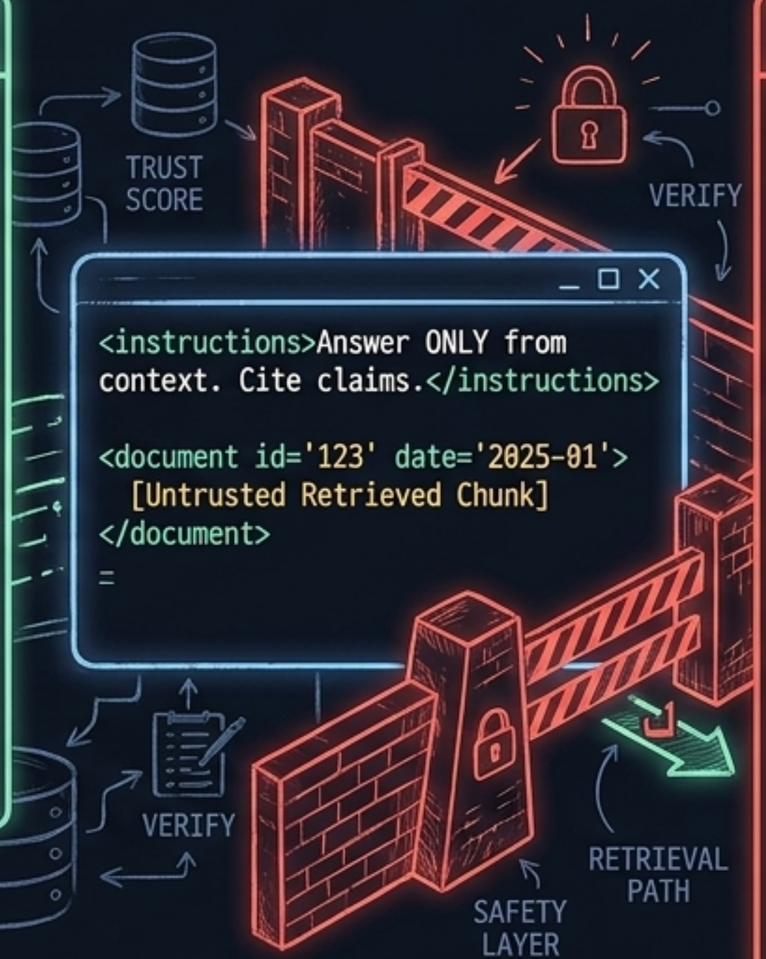
Chunk (Slice Text) → Embed (Add Metadata) → Rerank (Top-N) → Prompt Hydration → Generate → LLM (Answer +)

Concept 1: Prompt Hydration

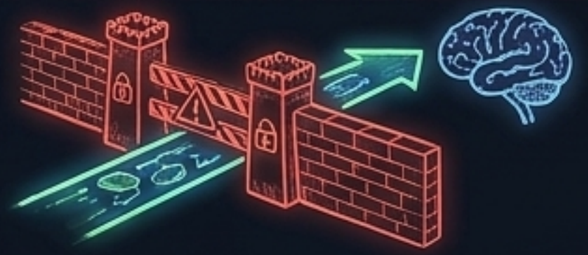
Put active chunks/queries at the uncached tail.

Treat retrieved chunks as untrusted data, never instructions.

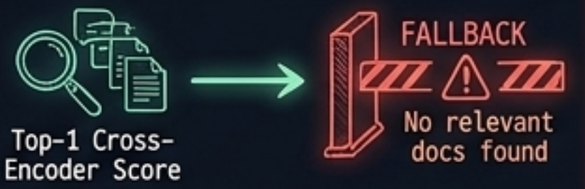




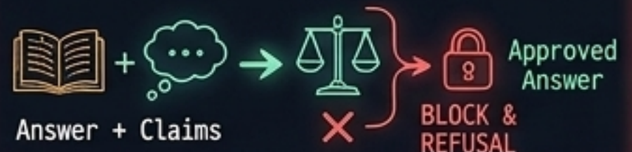
Concept 2: Refusal Gates



- Pre-generation: Short-circuit if Top-1 cross-encoder score falls below threshold ("No relevant docs found").



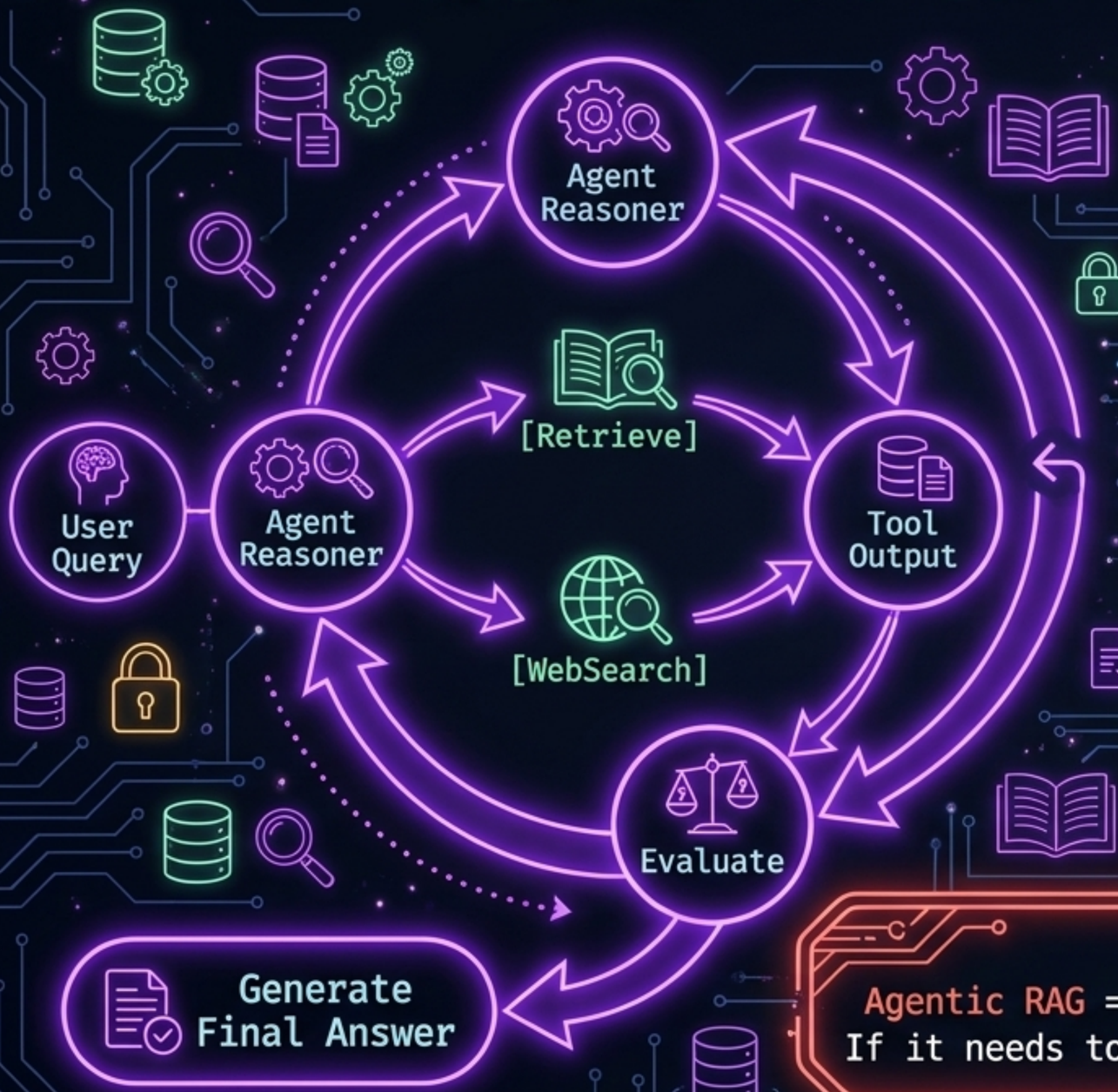
- Post-generation: Groundedness verification. If claim entailment fails, block answer and fallback to safe refusal.



Most hallucinations start in retrieval. Refusal is a first-class safety feature.

Gemini Notebook

AGENTIC RAG LOOP & RUNTIME JOBS



The Four Runtime Jobs of the Agent

Source Validation
Drop superseded hits via effective dates.

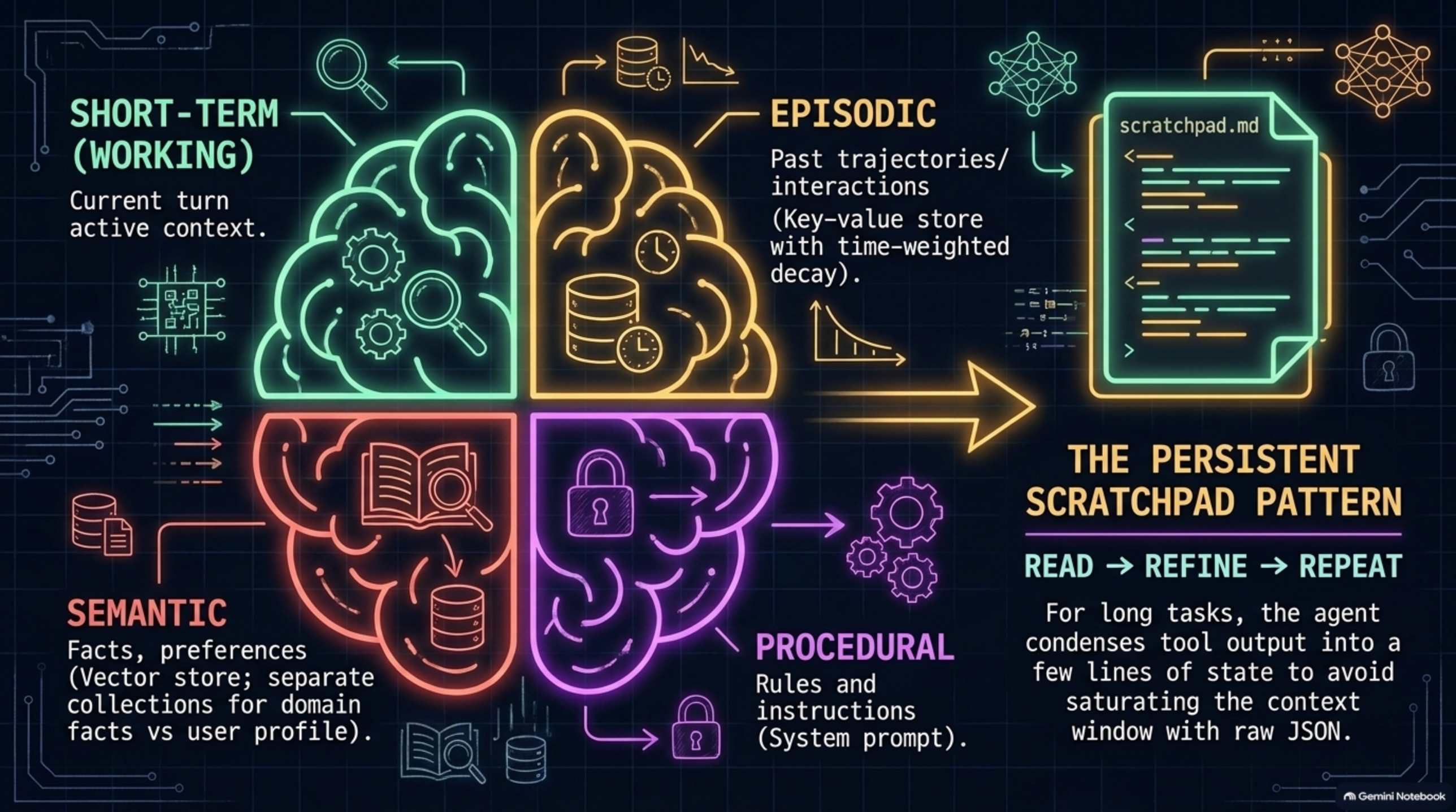
Conflict Reconciliation
Detect contradictory figures, ground in authority.

Decomposition & Synthesis
Multi-hop synthesis across tools.

Gap Detection
Fall back to live tools (CRAG) when local index misses.

Standard RAG = one-shot.
Agentic RAG = iterative retrieval inside the reasoning loop.
If it needs to KNOW, use RAG. If it needs to DO, use an agent.

Gemini Notebook



Gemini Notebook

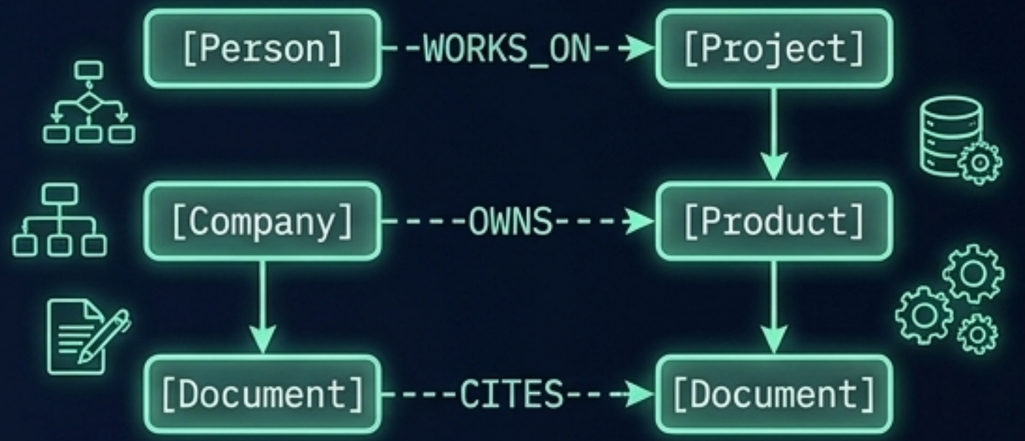
VECTOR RAG VS. GRAPHRAG & STRUCTURED DATA

Vector RAG



- **Use:** Conceptual search, local semantic match.
- **Fails at:** Math, sorting, explicit multi-hop entity chains.

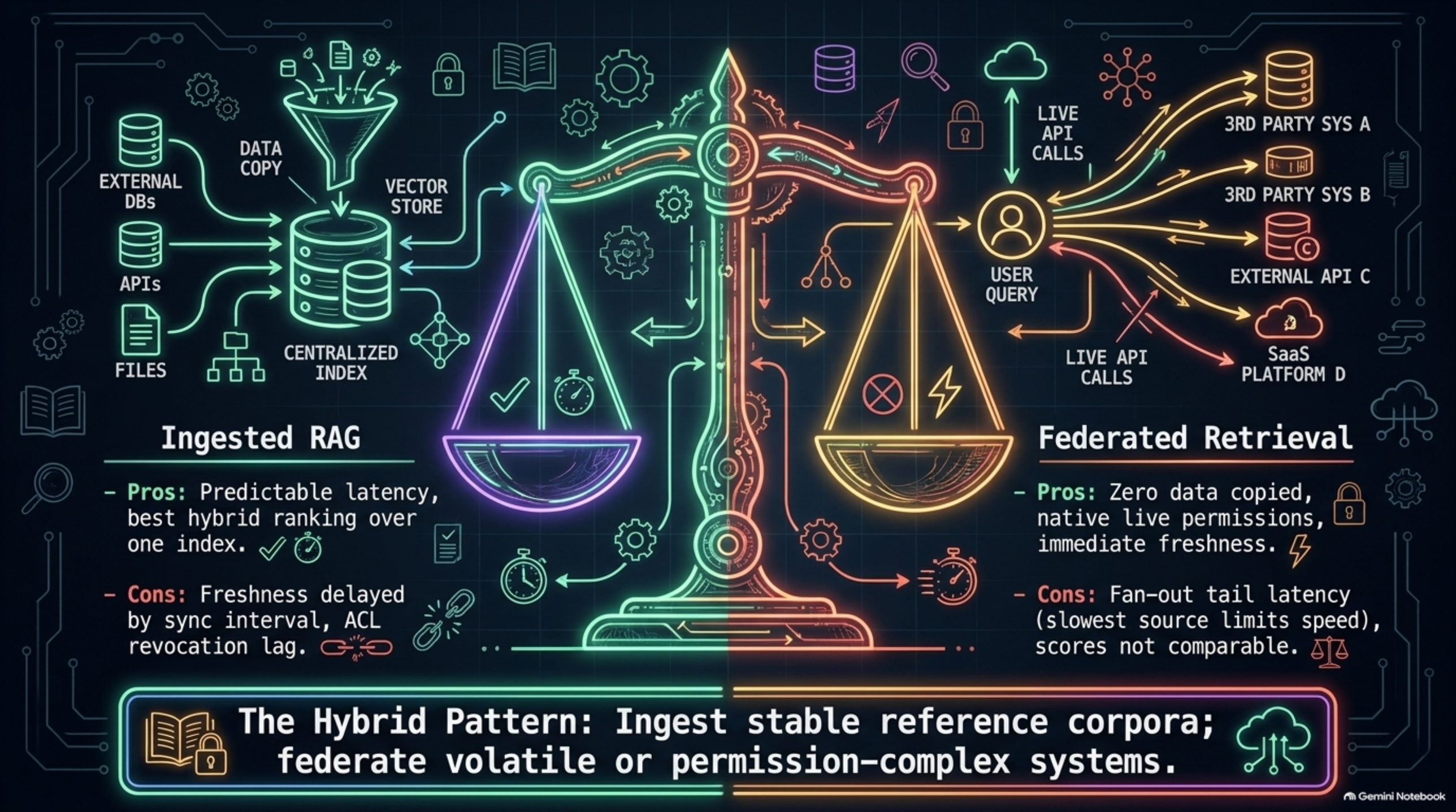
GraphRAG & Structured Data



- **Use (Graph):** Org charts, supply chains, corpus-wide global summaries.
- **Use (Text-to-SQL):** Quantitative questions (SUM, AVG), retrieving schema to generate read-only SQL.

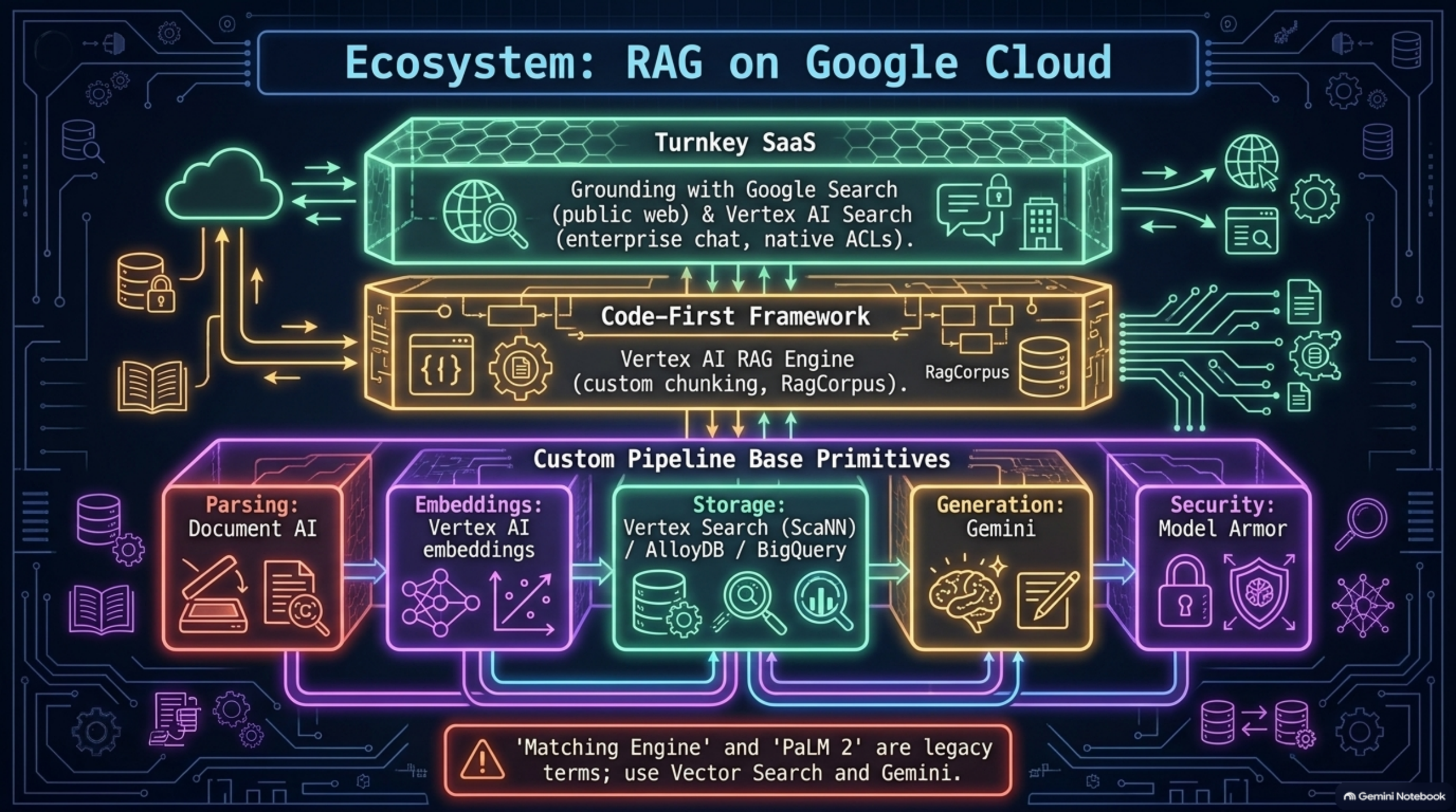
Hybrid is powerful but incurs the highest operational cost (two stores to sync).

Gemini Notebook



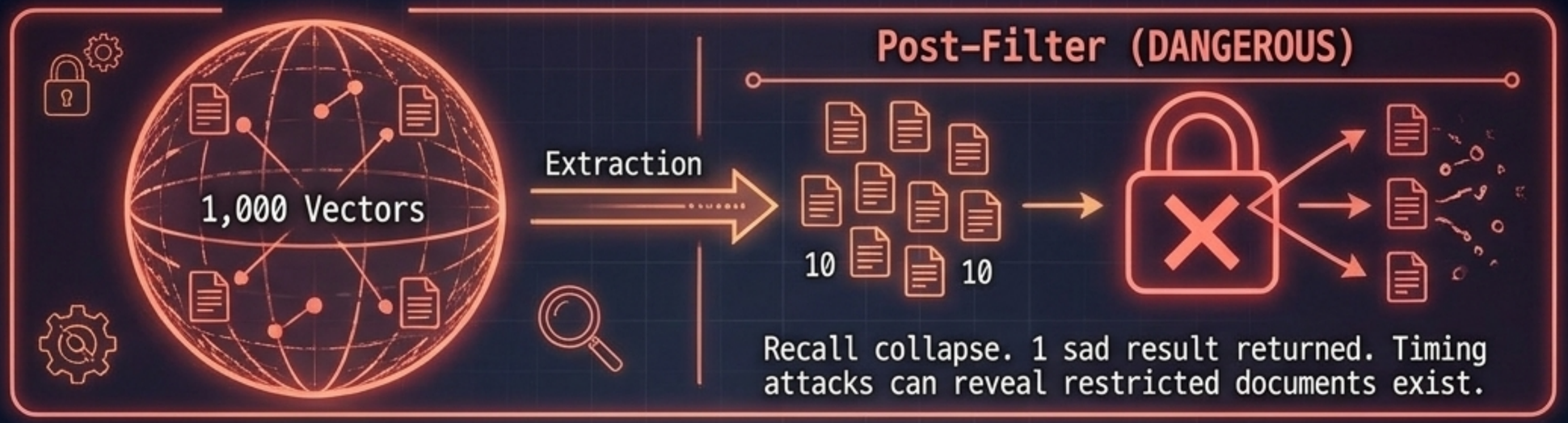
Gemini Notebook

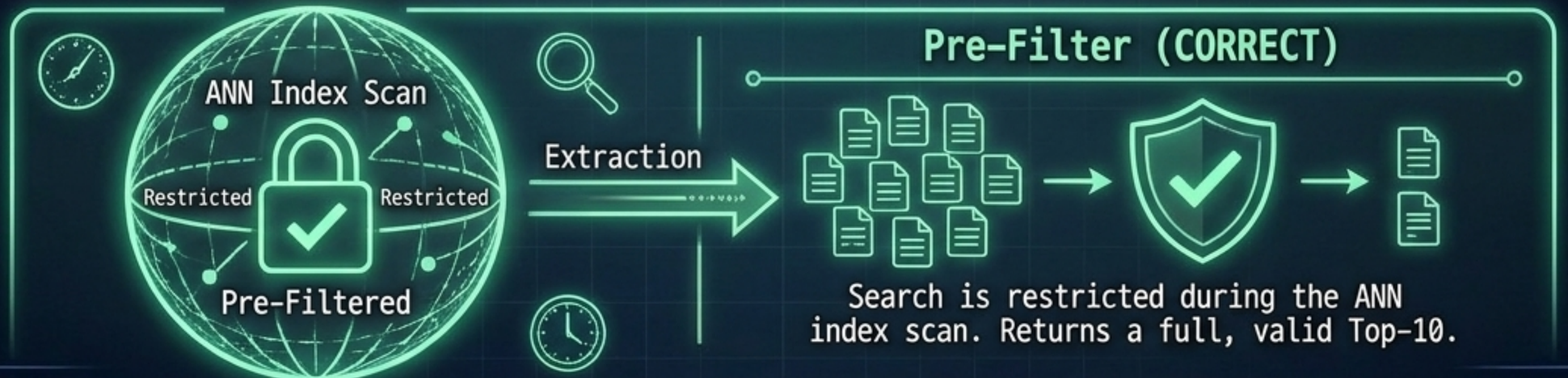
Ecosystem: RAG on Google Cloud



Gemini Notebook

Vector Search Security: Post-Filter vs. Pre-Filter Patterns





Hybrid Recommendation

Pre-filter at the index level + Live ACL re-check of the final Top-N before generation. Store group IDs, not user lists.

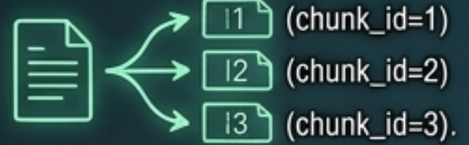
Gemini Notebook


Deterministic Filters (~1ms)
Strip hidden HTML comments, regex known patterns.
Regex!
Small Classifier (e.g., Model Armor)
Screens both user prompt AND retrieved chunks. Must Fail Closed.
Isolated Generation
Chunks wrapped in <document> tags defined in instructions as untrusted data.
Output Checks
Citation validation, DLP filtering.
Indirect Prompt Injection via Retrieved Chunks
Agentic Defense: Least privilege at the tool registry (strictly read-only tools for RAG agents).
Gemini Notebook

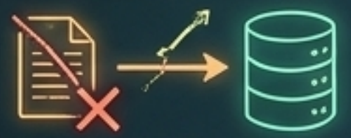
“Blue/Green” Dual-Database Synchronization

Freshness Mechanics

- Use deterministic chunk IDs (doc_id + chunk_index).



- Delete-before-insert prevents orphan chunks.



- CDC tombstones evict “ghost vectors” and purge downstream caches.





Index Migration Rule
(Atomic Alias Swap)

- Vectors cannot be mathematically converted between models.



- Backfill raw documents into a completely new index.

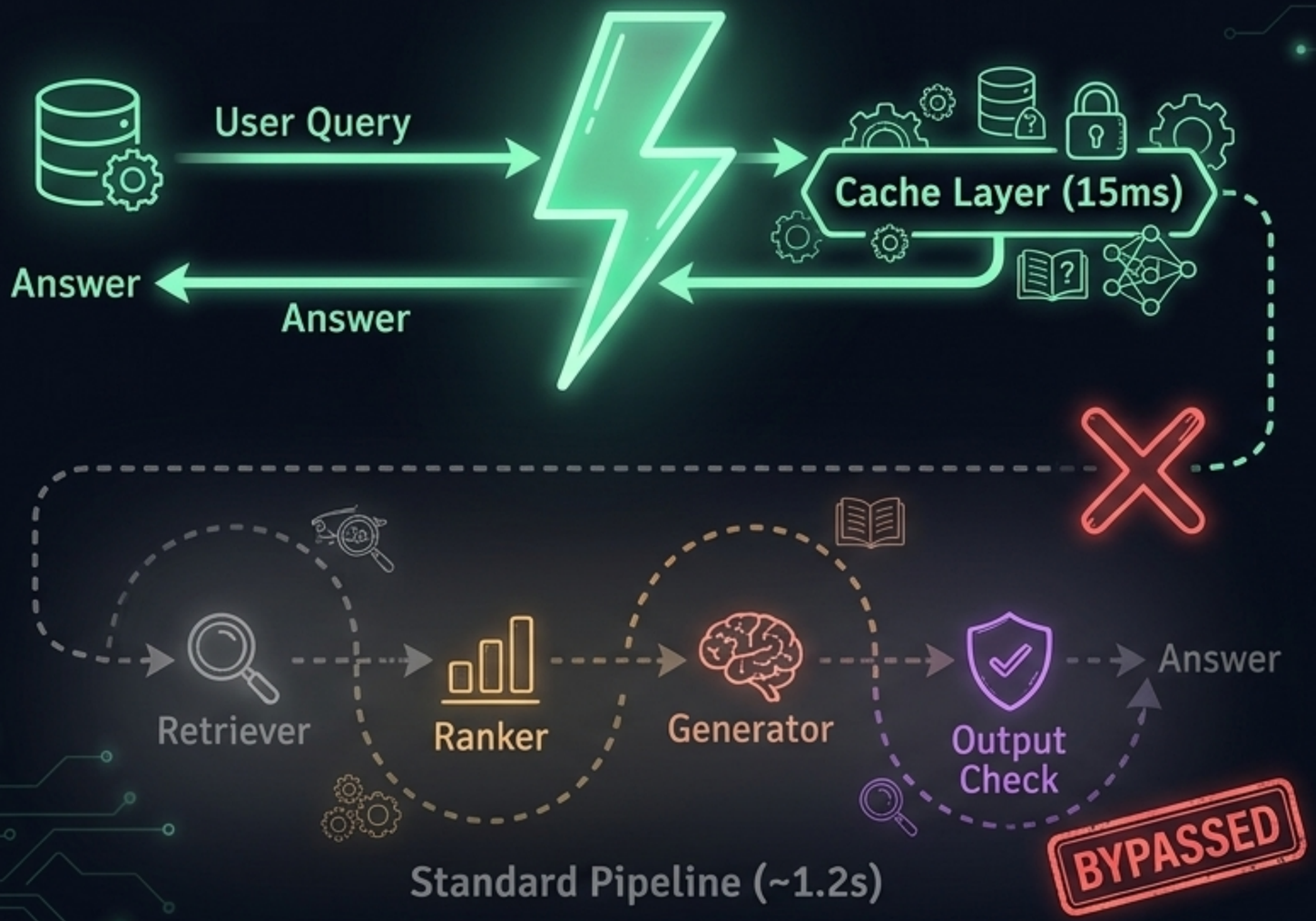


- Atomic swap: Switch the corpus index AND the query embedding model at the exact same moment.

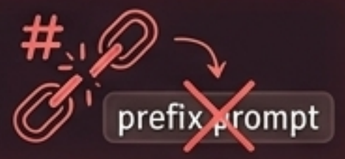


Leaving the query encoder on an old model against a new index results in near-random rankings.

Gemini Notebook



Exact-string hash caching fails (low hit rate).
Prefix prompt caching cannot serve stale passages; it requires exact byte matches.



Mechanics of a Safe Cache

• Entity Masking

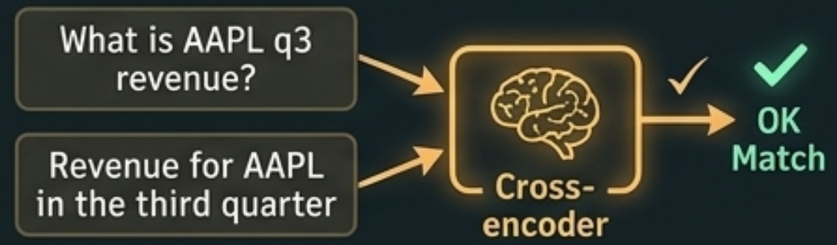
AAPL revenue Q3 → [COMPANY] revenue [PERIOD]

Values kept as exact keys so Q3 never serves for Q4.



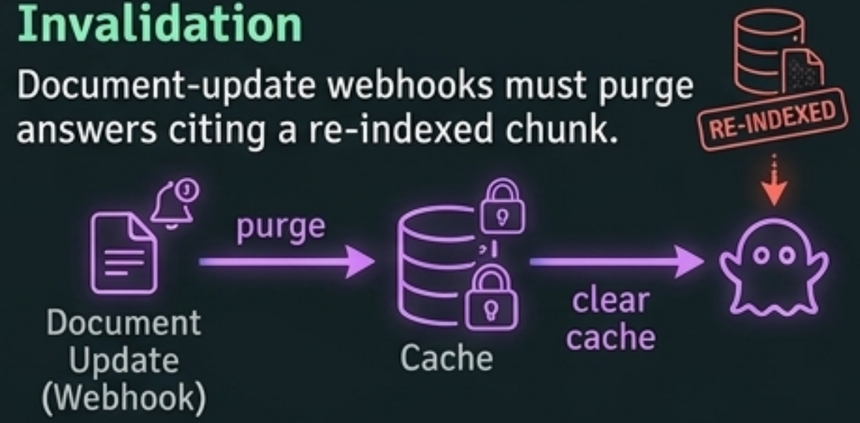
• Intent Check

Light cross-encoder confirms meaning match.



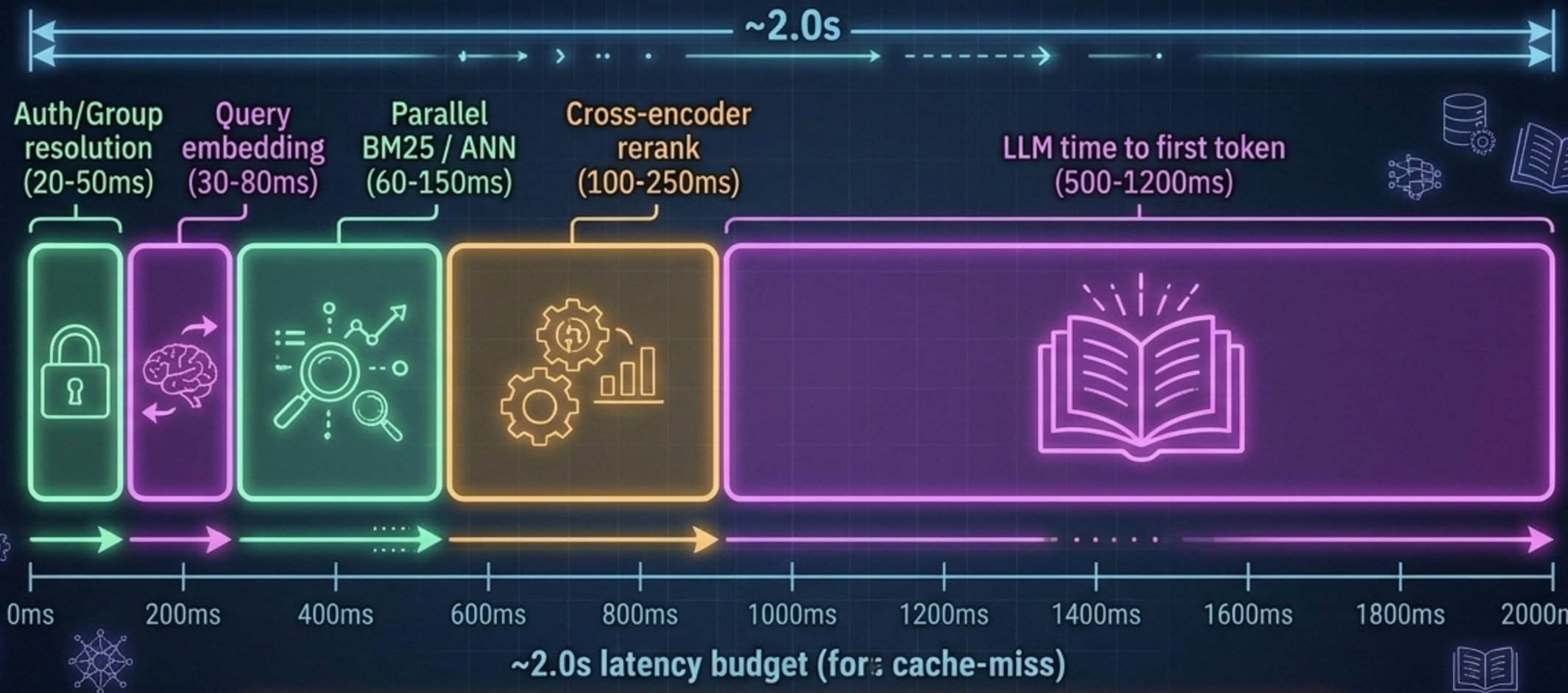
• Invalidation

Document-update webhooks must purge answers citing a re-indexed chunk.



Gemini Notebook

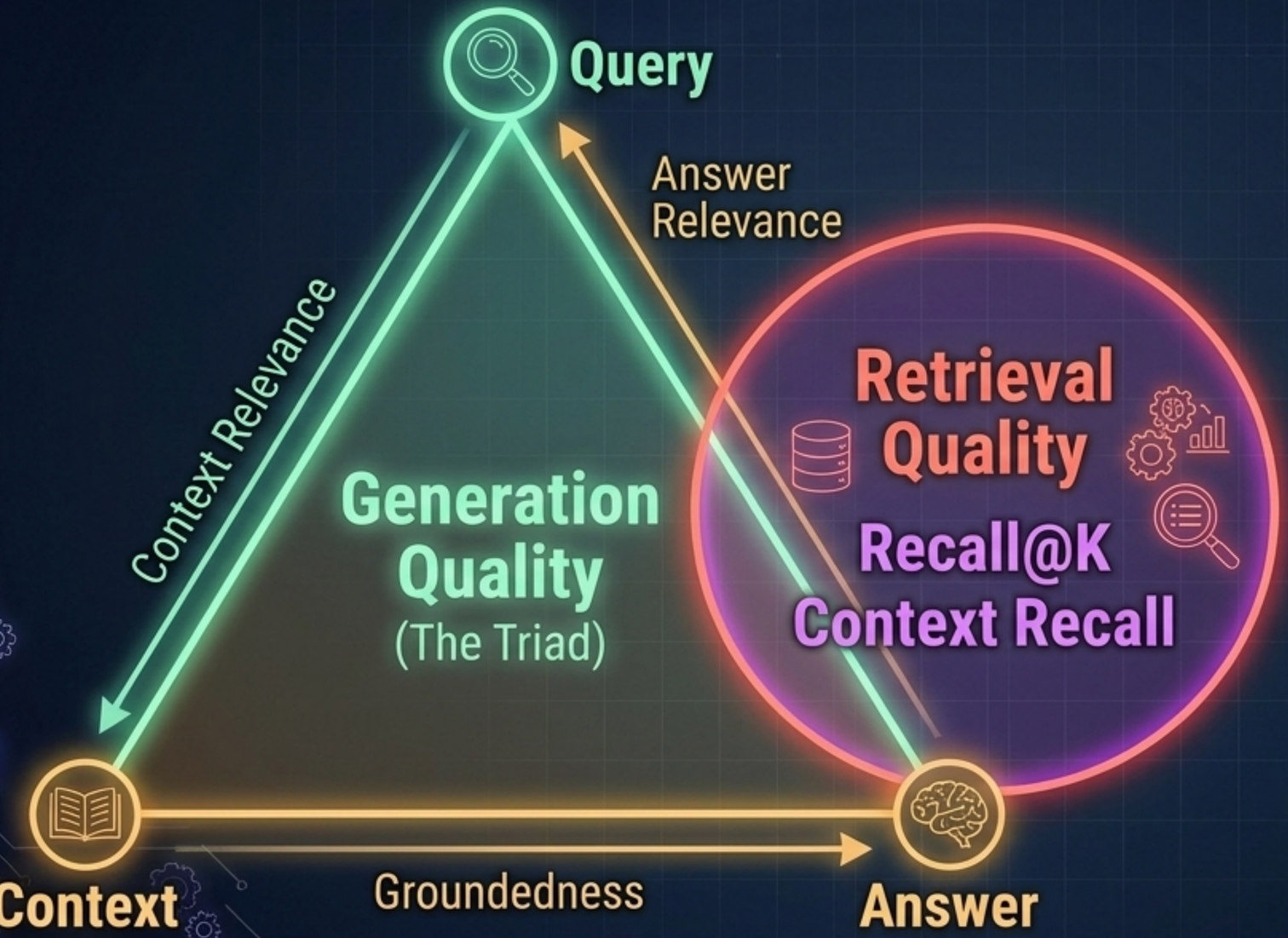
Performance: The Latency Budget



Key Takeaway: The LLM dominates the latency budget. Trimming the final N (fewer chunks) and caching the system prefix yield vastly better ROI than micro-optimizing the vector search.

Gemini Notebook

Blueprint: Evaluation Framework



Core Thesis:

The Triad evaluates generation, but the best LLM cannot answer from a chunk that was never retrieved.

Recall@K against a human-labeled **Offline Golden Set** (with unanswerable/restricted cases) is absolutely mandatory before evaluating generation.

Gemini Notebook

Synthesis: The Troubleshooting Grid

Symptom	Root Cause	Correct Fix	Tempting Wrong Answer
Fact buried in prompt ignored	Lost in the middle	Better chunks at start/end	Longer context window
Exact SKUs (ERR-742) not found	Dense-only search	Hybrid BM25+Dense	Larger embedding model
Confident answer beyond sources	Gap filled from memory	Grounding checks + abstain rule	Fine-tuning on docs
Outdated content returned	Ghost chunks	CDC tombstones, delete-before-insert	Post-filtering results
High Triad scores, answers incomplete	No recall metric	Measure Recall@K	Tuning the generator

Total structural mastery separates a production RAG system from a prompt wrapper.

Gemini Notebook