

AI KNOWLEDGE BASE PLAYBOOK

Powering faster, smarter decisions with Retrieval Augmented Generation (RAG)

YOUR IP

1 Day

The average time employees
waste every week **searching** for
the right information



Source: McKinsey and IDC's Information Worker Survey

That's the invisible drag on
innovation, compliance, and growth.

“

“Where’s the latest version?”

“Where’s the process again?”

“Hasn’t someone done this already?”

”

THE SOLUTION = RAG

Custom AI Powered Knowledge Base - aka Your Company "Brain"

What is RAG?

Retrieval-Augmented Generation uses your documents and internal knowledge to answer questions accurately.

No guessing.

No making stuff up.

RAG = Retrieval + Generation

Why It Matters

Answers grounded in your knowledge
Works across formats (PDFs, Sheets, Docs)
Understands images
Delivers context-aware responses

Step 1

Search your curated and indexed content.



Step 2

Generate an answer using only what was retrieved.



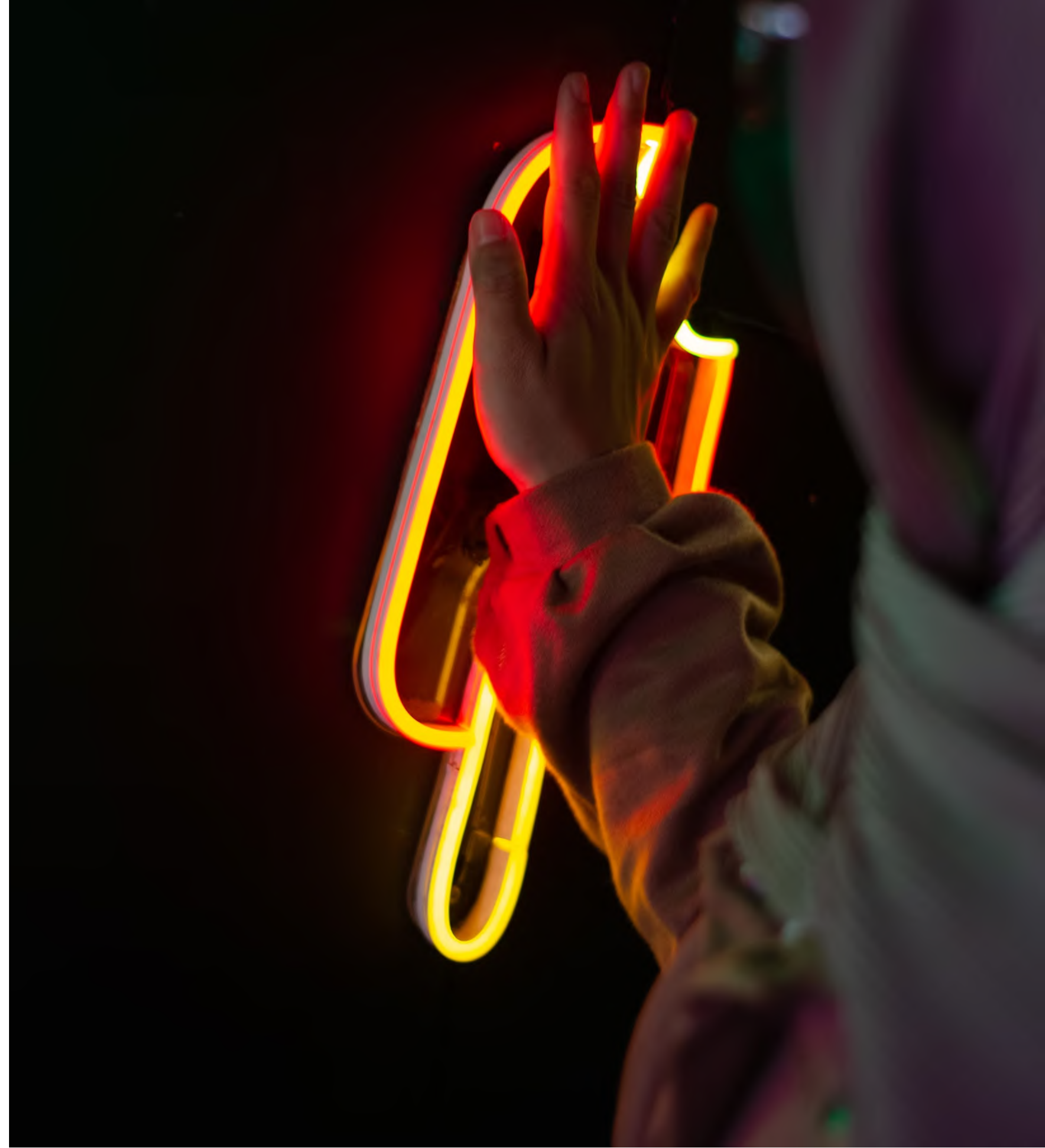
Step 3

Format the answer based on users role, experience and needs.

Same question. Different roles. Different answers.

Advanced AI Knowledge Bases don't just retrieve information. They **frame the answer** based on who's asking - whether they're a junior exec, a team lead, or C-suite.

They deliver **role-specific**, decision-ready insight - not just documents.



66

Q: “How do I report a data breach?”



**Junior
Legal**

Answer: Step-by-step SOP, forms
required, reporting deadline
(Plain English, with tooltips and links)



**Data
Protection**

*Answer: Full compliance pack, audit
history, contact escalation matrix*
(Risk-first framing)



COO

*Answer: Exec-level summary, risk
impact, dept exposure, mitigation*
(Focus: Business cont. + stakeholder)

THE TECH

What powers RAG systems and how do they work?

RAG OVERVIEW

Retrieval-Augmented Generation (RAG) gives AI access to *your* knowledge - safely and scalably.



- Documents, sites, videos → split into chunks
- Chunks converted to *embeddings* (mathematical representations of meaning)
- Embeddings stored in a *vector database*
- When a user asks something, the system retrieves relevant chunks and the LLM answers based on them

WHY CHUNKING & EMBEDDINGS MATTER

Before an AI can answer smartly, it needs to *understand* your content. That starts by breaking your documents down into chunks - and turning those chunks into mathematical meaning.

The better this is done, the better your Knowledge Base performs.

Let's break it down.

CHUNKING

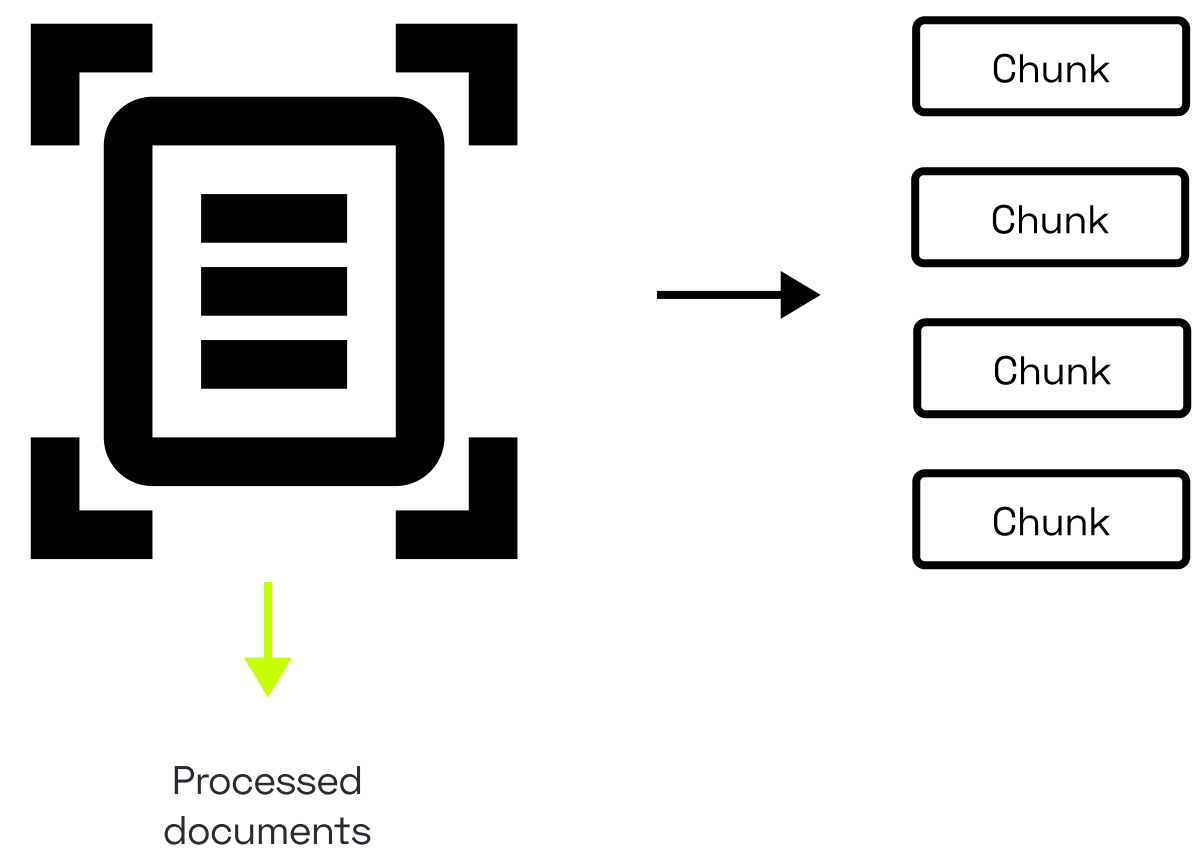
How your content is chunked directly affects RAG performance.

- LLMs don't read like humans – they retrieve chunks
- Bad chunks = wrong answers, hallucinations, slow queries
- Good chunks = precise, fast, context-aware results
- This one step can make or break your Knowledge Base



CHUNKING

Here’s how to think about building the right chunking approach.



1. Content Type

----->

FAQ = short Q&A
Manual = long sections
2. Chunk Size

----->

Smaller = better precision
Larger = better context
3. Chunk Strategy

----->

Fixed = fast, naive
Semantic = slower, smarter
4. Metadata

----->

Great metadata = faster,
more accurate answers

EMBEDDING

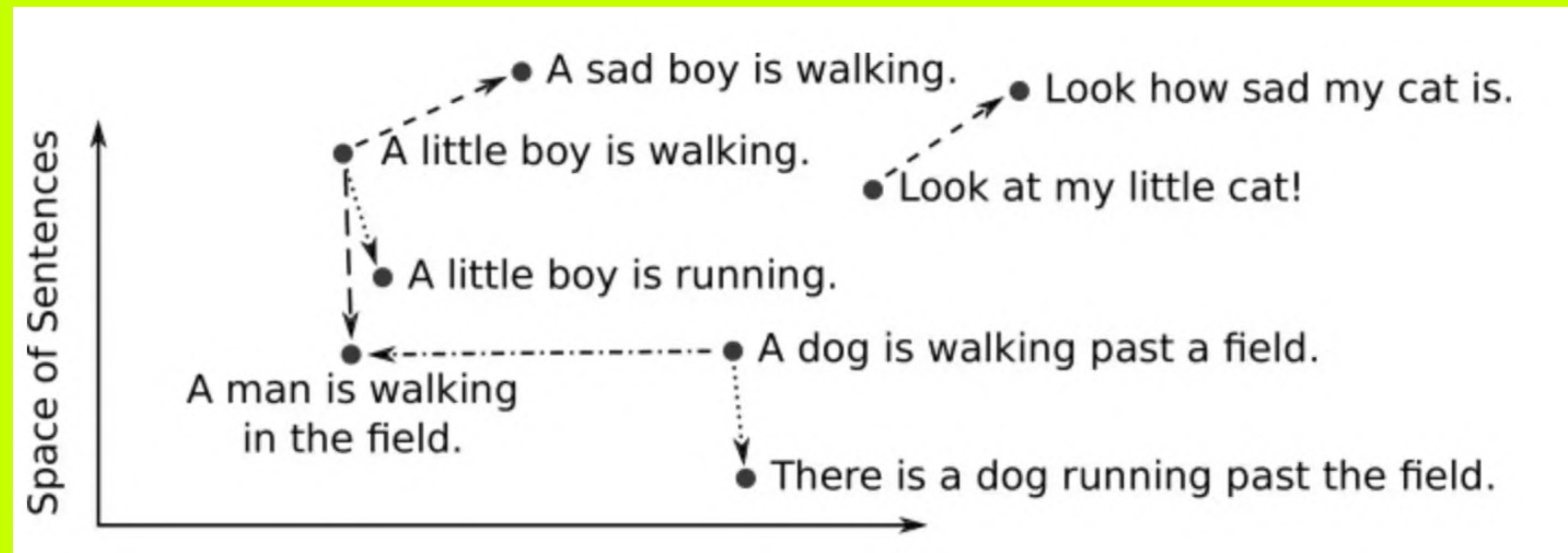
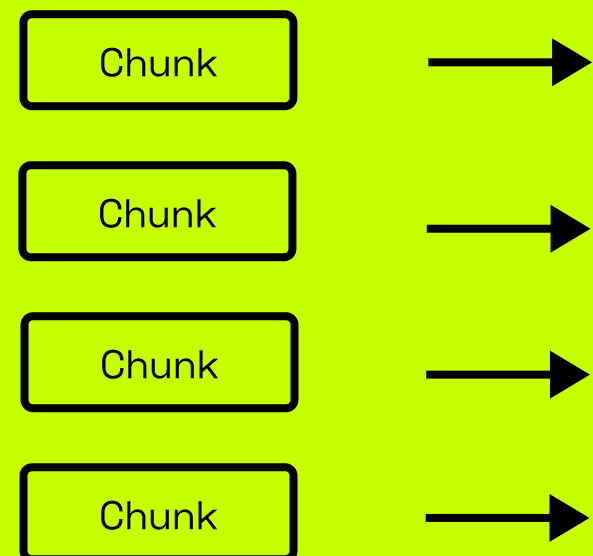
EMBEDDING

Chunks are transformed into embeddings - mathematical representations of meaning.

These are stored in a **Vector Database**, where their position is based on meaning, not keywords.

Closer = more related.

This is how an AI finds the *right* answer, even if the question doesn't match the original words.



HOW IT FITS TOGETHER

The core elements.

Data/Content

Data Pipelines

Content is sourced from key areas including team drives and other platforms.



Data Pre-Processing

Content is processed so that it's in a consistent format, and non-text content (tables, images, flows) is searchable.



Infrastructure

Vector Databases



Vertex AI



Embeddings & Model Selection

Selecting the right embedding model ensures your AI retrieves the most relevant, high-quality answers - faster, more accurately, and with less noise.

MiniLM - OpenAI Ada v2 - VoyageAI

Fine Tuning

Chunking Strategy

Semantic: cuts by meaning

Fixed: fast but naïve

Agentic: LLM decides chunk splits

Late: embed full doc first, chunk later

Best Practice

Chunk size: 512–2,000 tokens
Overlap: 10–20% to preserve context
Tailor to doc type: FAQ ≠ Manual

Retrieval Optimisation

Hybrid search: vector + keyword

Metadata filters: improve relevance

Reranking: VoyageAI or cross-encoders

Feedback loops: improve over time



Smart Retrieval = Better UX

Real results come from how well these parts work together.

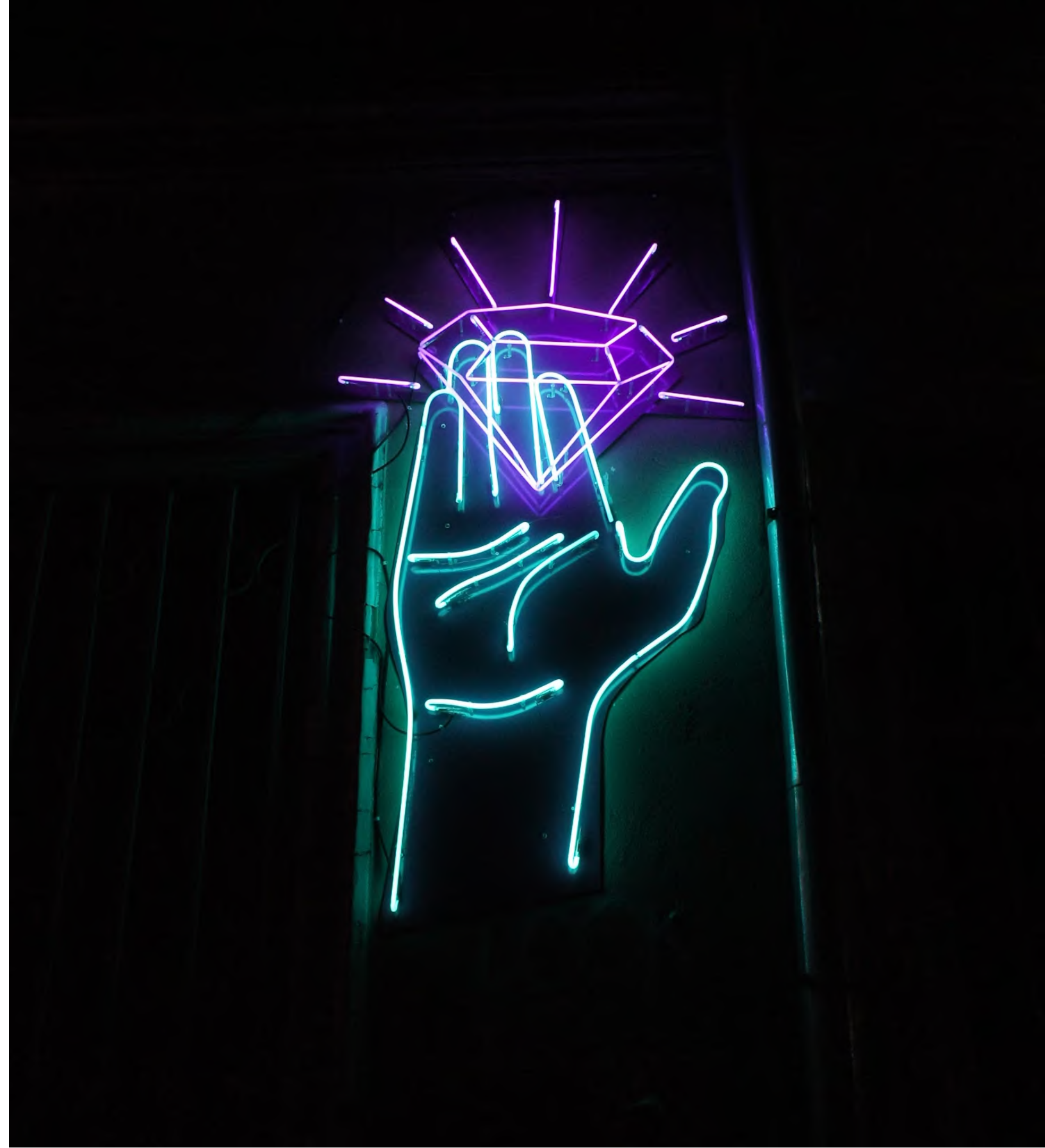
RESULTS

Private, Secure AI Knowledge Bases

Turn your knowledge into action.



















- Built on **your content**, not public data
- Large repository of **your documents (1000+)**
- Continuously updated via **live data pipelines**
- Tailored to **roles, experience, and business context**

Not a chatbot. Not a search bar.
A trusted internal decision
engine aka "the brain".







Product Comparison

Off the shelf knowledge bases require manual uploading of content, they struggle with non-text content, do not automatically update when your documents change and can't be fine tuned to your business.

	Advanced RAG	Copilots, Wikis & Generic AI
 Injects structured & unstructured data at scale		
 Extracts image, table, graph data for searching and retrieval		
 Auto-updates when documents change		
 Fine-tuned outputs by role & seniority		
 Full audit logging & access control		
 Domain-specific prompt engineering		

ROI

IMPACT

Impact	Business Value
 Faster onboarding	Cut ramp time by 40–60%
 Fewer internal escalations	Reduce repeated queries by 30%
 Retained knowledge	Prevent IP loss during turnover
 Better compliance	SOPs followed, not forgotten

Time saved = **money reclaimed**. Risk avoided = **cost prevented**.

CALCULATE YOUR RETURN

How Much Could You Save?

Want to see how much efficiency you could gain in monetary terms?

Go here and [calculate your ROI](#).

It only takes 2 minutes.



WORKING WITH

G3NR8 CLIENTS

SAMSUNG

J.P.Morgan



Perrigo®

gousto



VISA

Red Bull 
R A C I N G


Pernod Ricard

GSK


JustFix

Johnson
Controls 

G3NR8

FIN.

Get in touch to book a call, discuss your requirements and get clear, transparent pricing options based on your needs.

Whatever stage you're at we can answer all your questions and help guide you on the right path.

info@g3nr8.com