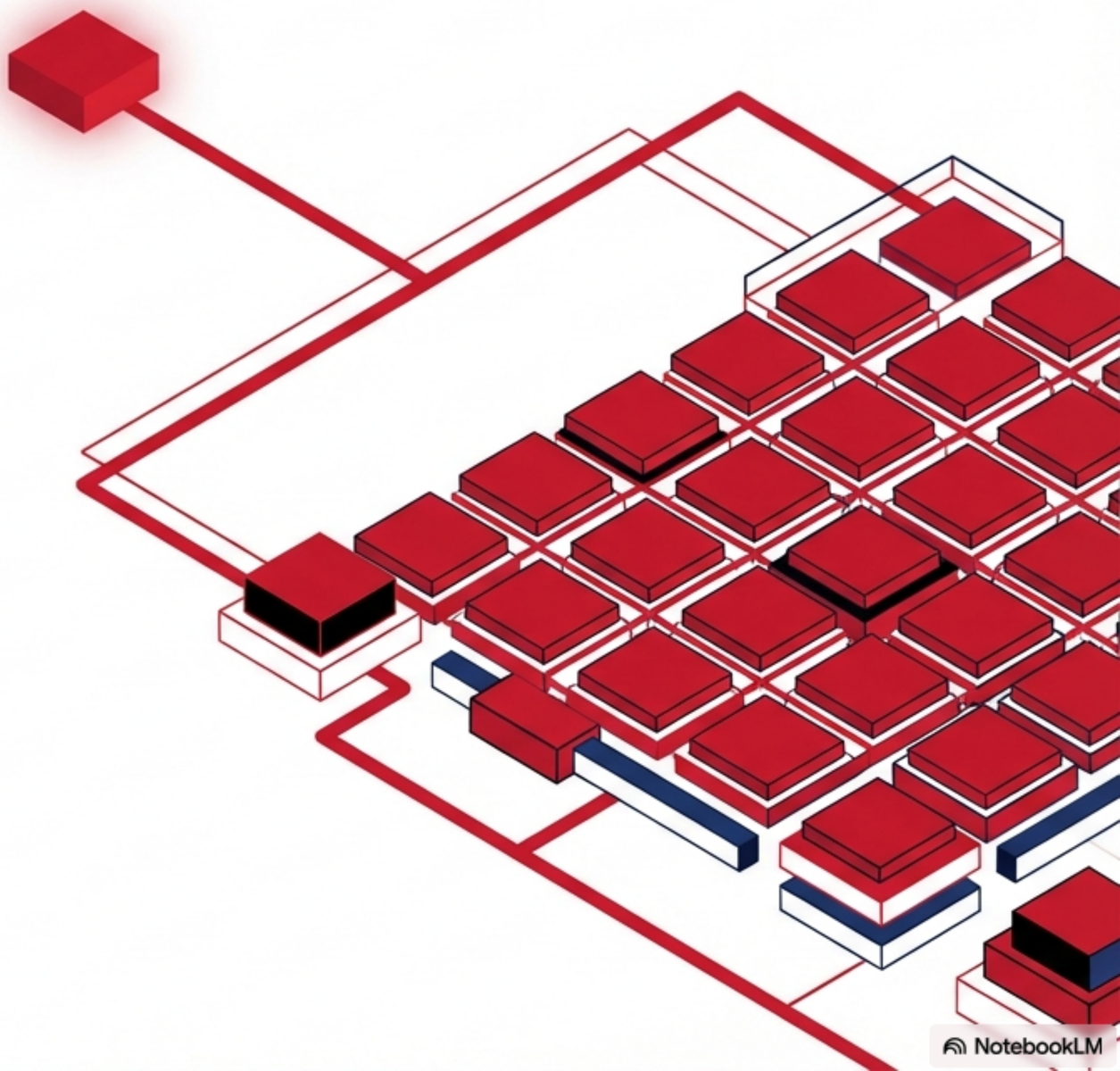
CONTEXT ENGINEERING

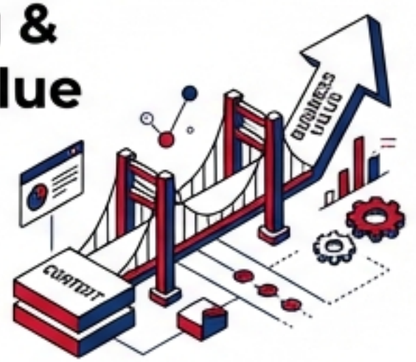
Building AI Systems with Your Own Data

UBUS 670 | AI for Business Leaders | Day 3

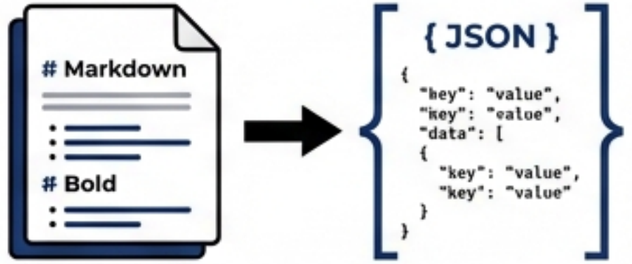


TODAY'S LEARNING OBJECTIVES

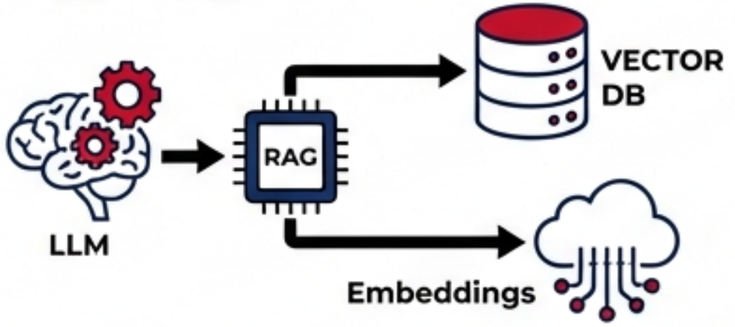
1 Explain Context Engineering & Business Value



2 Recognize Structured Formats (Markdown, JSON)



3 Describe RAG, Embeddings, & Vector DBs



4 Apply Context via Gemini Gems

Today's Skill: Moving from asking questions to building information environments.

THE LIMITATION OF PROMPTS

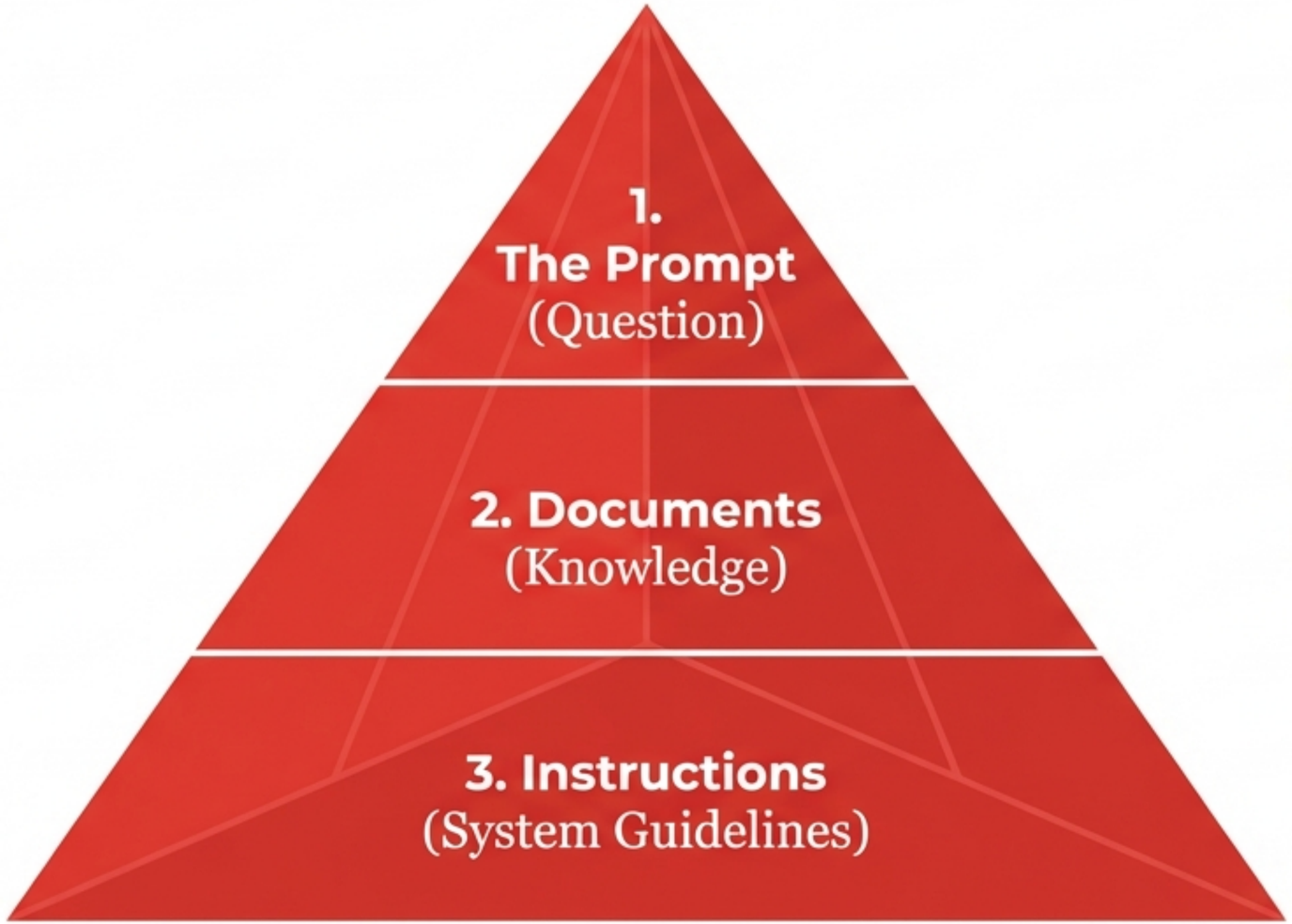
The Prompt

50+ Pages of Policy

Beacon Retail Group HR Challenge:
200 applications, 3 roles, legal guidelines, brand voice.

Imagine hiring a brilliant consultant but only giving them a one-sentence brief.

THE CONTEXT STACK

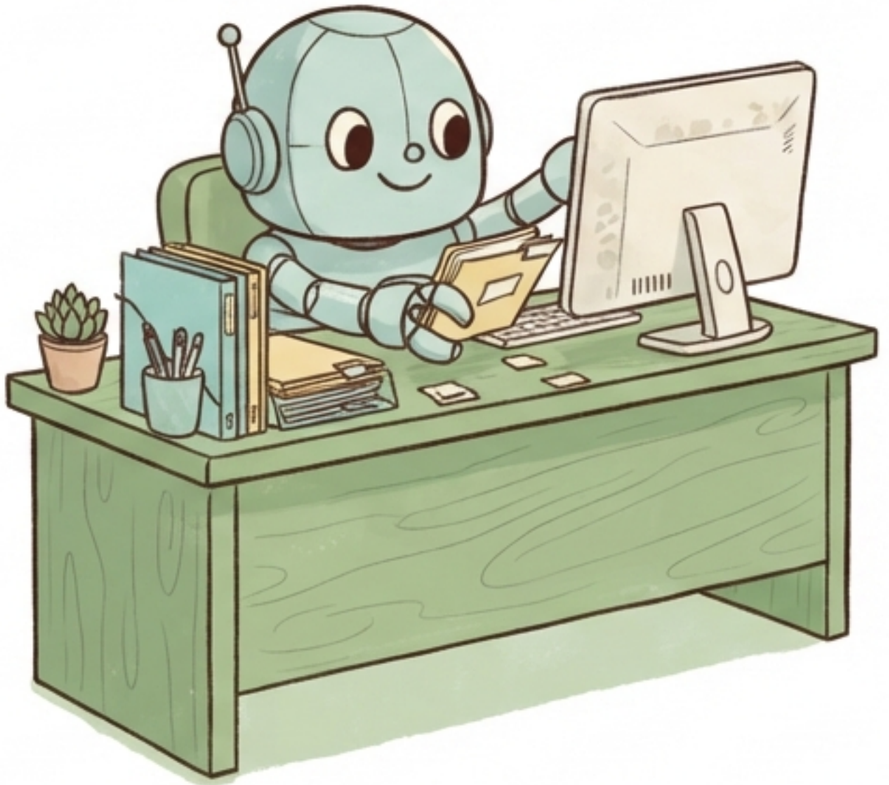


Context Engineering is building an information environment, not just asking a question.

STRUCTURE HELPS AI THINK BETTER

UNSTRUCTURED DATA

STRUCTURED DATA



Modern AI can read messy text, but structure reduces ambiguity.

MEET THE FORMATS: MARKDOWN

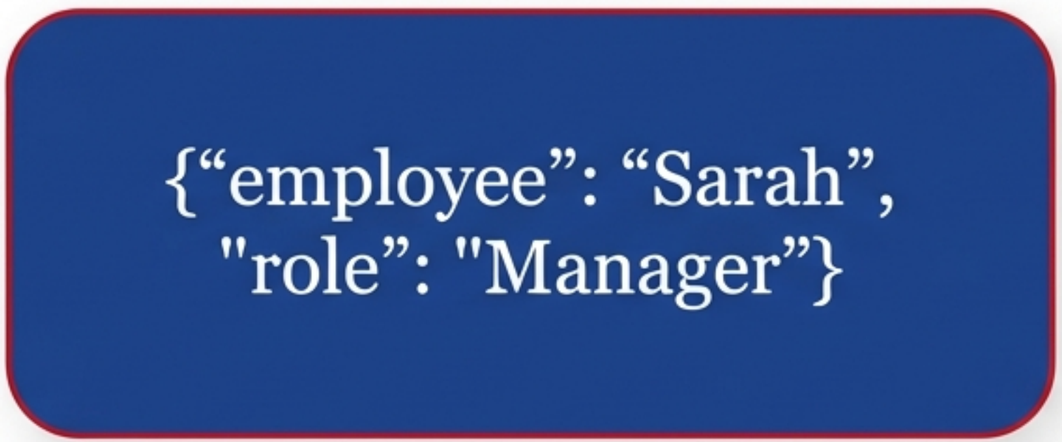
Beacon Vacation Policy.

Full-time employees get 15 days of paid time off annually. This accrues monthly. Part-time employees get 5 days after one year of service. Sick leave is separate and varies based on tenure. Contact HR for specific details and eligibility.

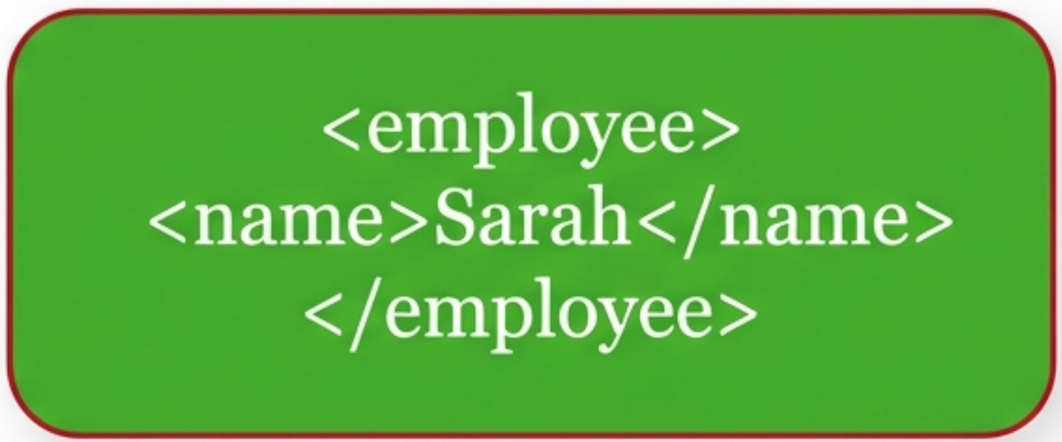
```
## Vacation Policy

**Bold Keys**
- Bullet points
- Full-time: 15 days
- Part-time: 5 days
```

Signposts for the AI.

OTHER FORMATS: JSON & XML



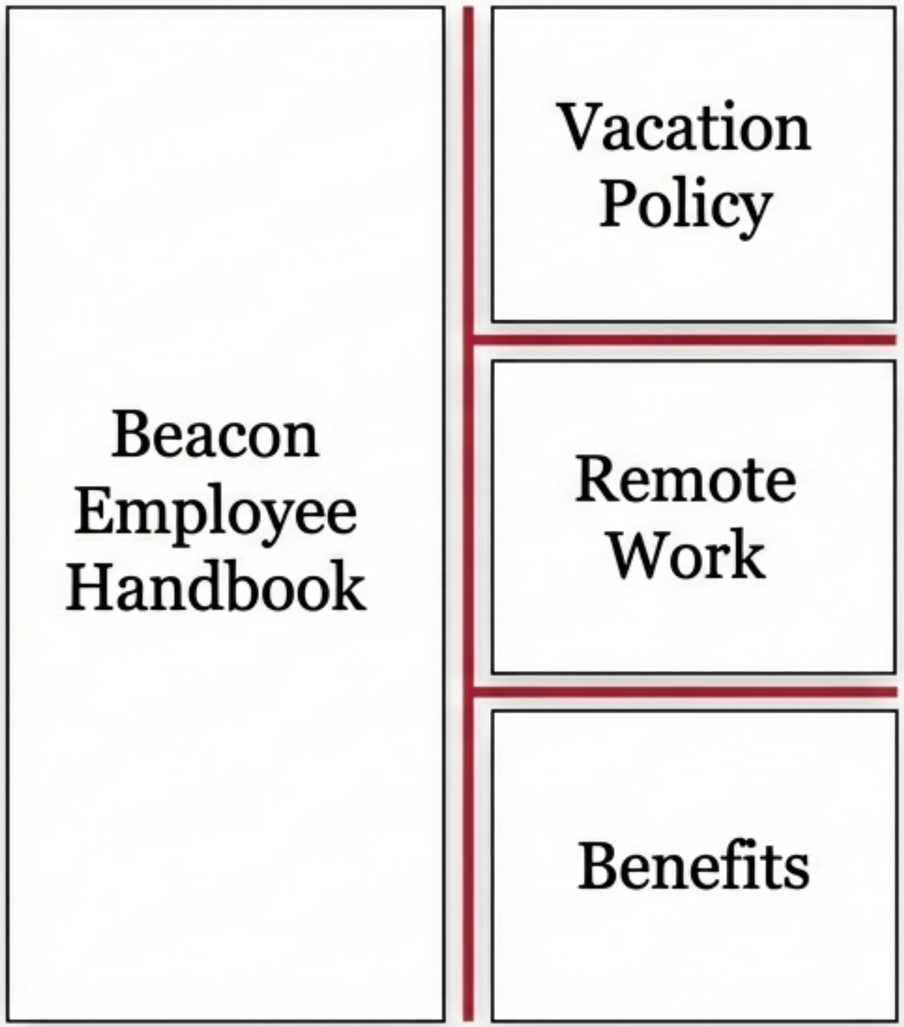
JSON: Key-Value Pairs



XML: Tagged Data

You don’t need to write code. You just need to recognize that structure = clarity for AI.

CHUNKING: MANAGING ATTENTION



Even with Gemini 2.5's massive context window, chunking **reduces** cost and improves retrieval accuracy.

NottebookLM does this automatically.

THREE STRATEGIES FOR CHUNKING

SEMANTIC SECTION

Split by topic (Chapters).

FIXED SIZE

Every 500 words with overlap.

TASK RELEVANCE

Upload only what's needed for the specific question.

The Inverted Pyramid Rule: AI pays most attention to the beginning and end of a file.

RAG: RETRIEVAL-AUGMENTED GENERATION

RETRIEVAL
(Search the Library)

+

GENERATION
(Write the Answer)

=

RAG

Grounding the AI in your data to reduce hallucinations.

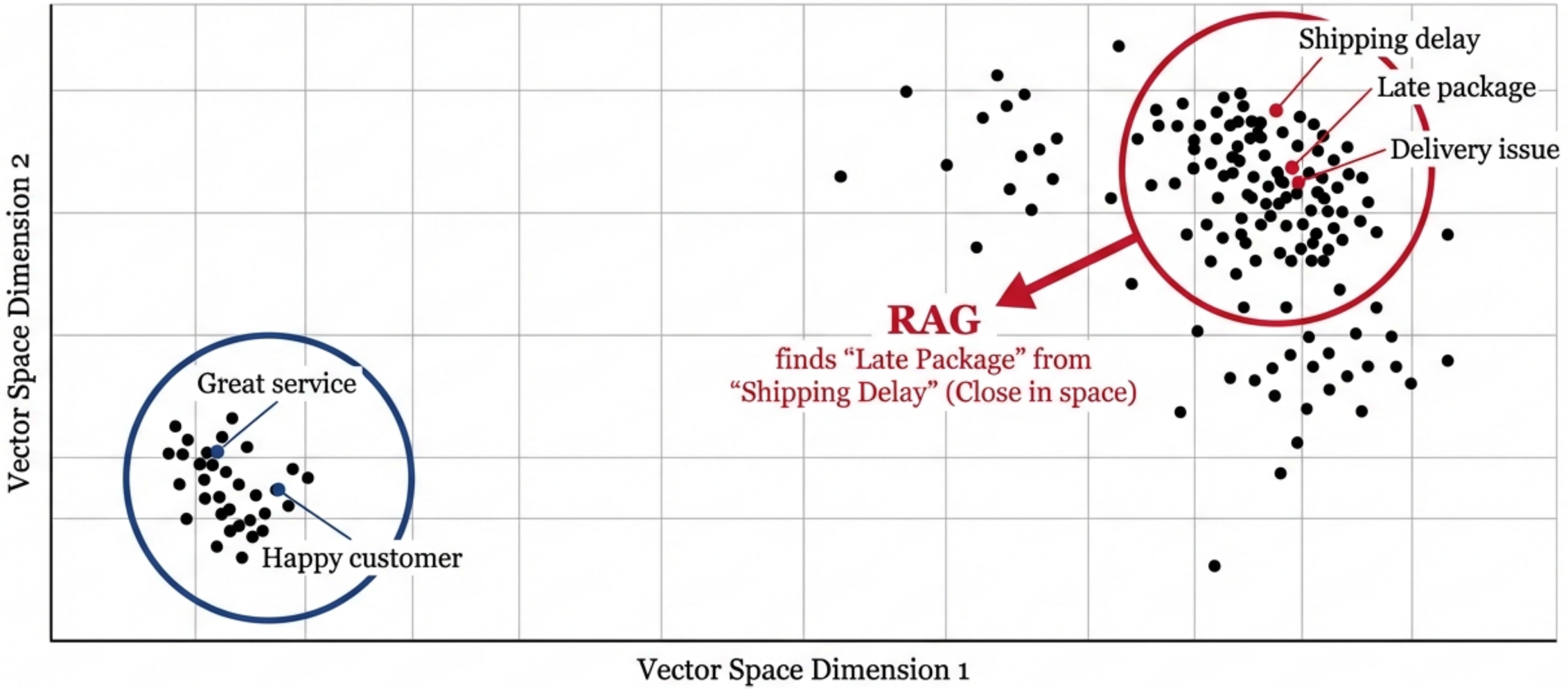
THE ANALOGY: OPEN VS. CLOSED BOOK

Closed Book (Traditional AI)

Open Book (RAG)

RAG turns AI from a 'know-it-all' into a 'know-where-to-look'.

EMBEDDINGS: SEARCH BY MEANING



Embeddings convert text into numbers (vectors). RAG finds “Late Package” even if you search for “Shipping Delay” because they are close in mathematical space.

VECTOR DATABASES

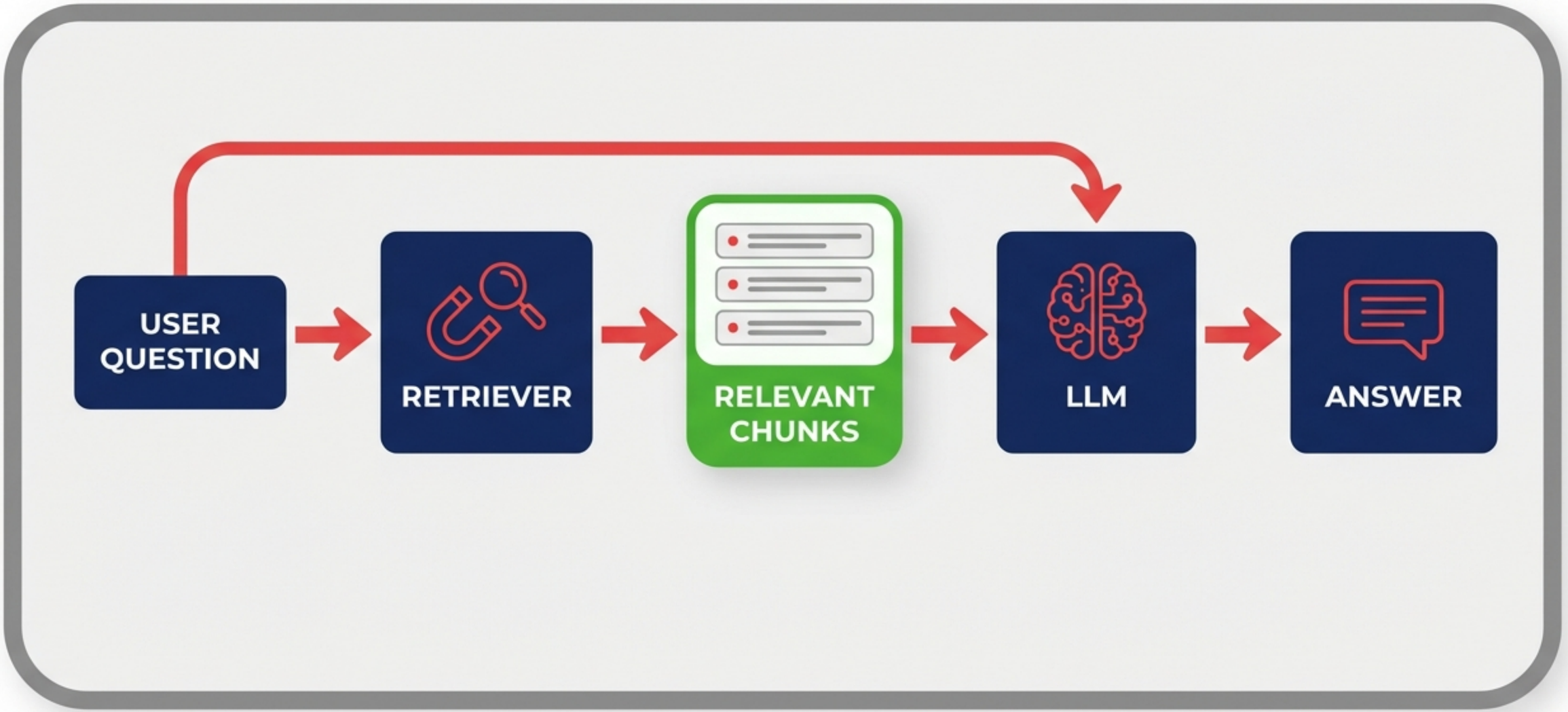
The AI Filing Cabinet

Stores “meanings” (embeddings) instead of just keywords.

Pinecone, Weaviate, Chroma

You don’t build this; your engineers do. You just need to know it’s the engine.

HOW RAG WORKS: THE ARCHITECTURE



WHY RAG MATTERS FOR BUSINESS

Reduces Hallucinations (Cites Sources)

Keeps AI Current (Instant

Proprietary Data Q4 Number

Auditable (Trace the Decision)

(Cites Sources) (Instant Updates) (Your Q4 Numbers) (Trace the Decision)

RAG VS. FINE-TUNING

FEATURE	RAG	FINE-TUNING
Goal	Add Knowledge	Change Behavior
Analogy	Giving a Textbook	Sending to Med School
Cost	Low	High
Updates	Instant	Slow

For 95% of business cases in 2026, use RAG.

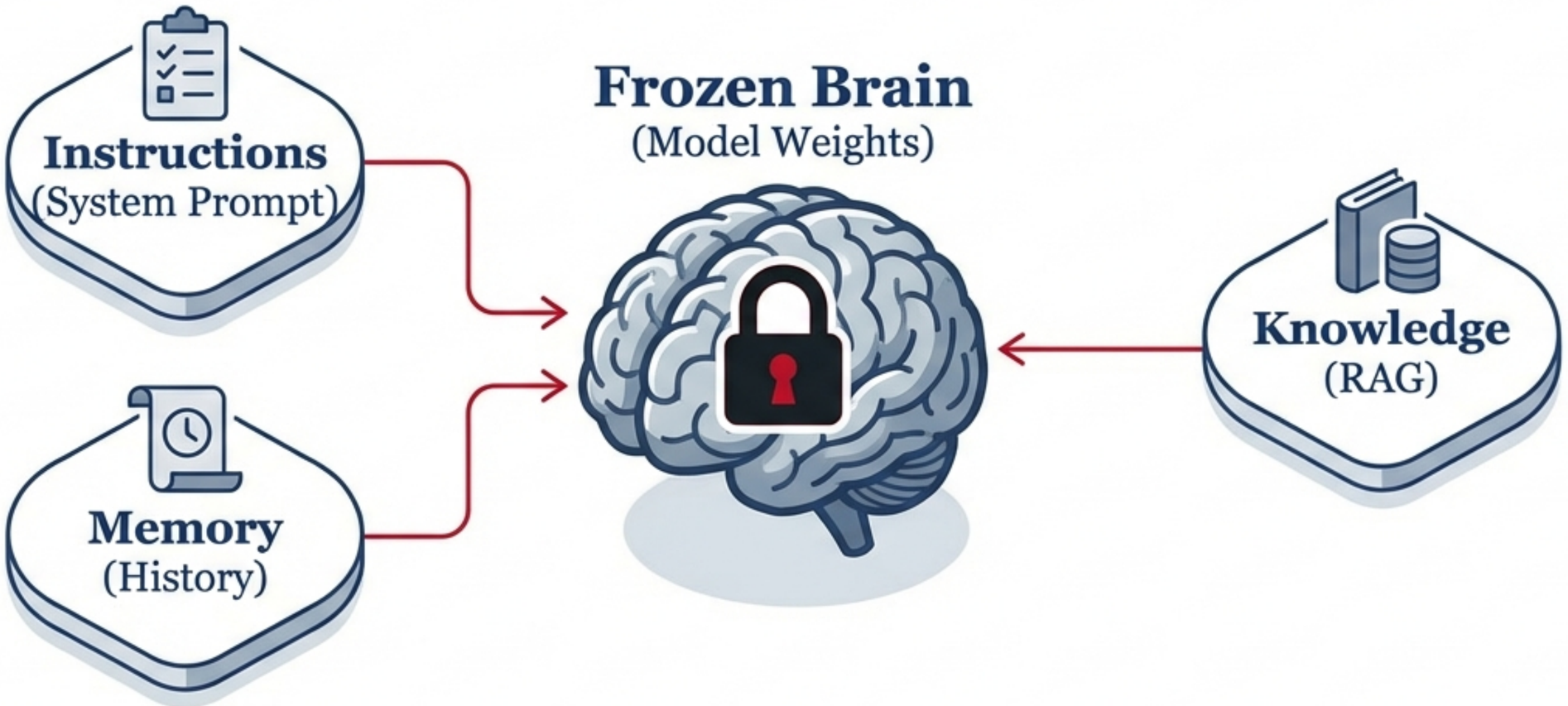
THE ACCIDENTAL FINE-TUNER

Dr. Chen thought she ‘trained’ the model to be empathetic by correcting it for weeks.

THE TWIST

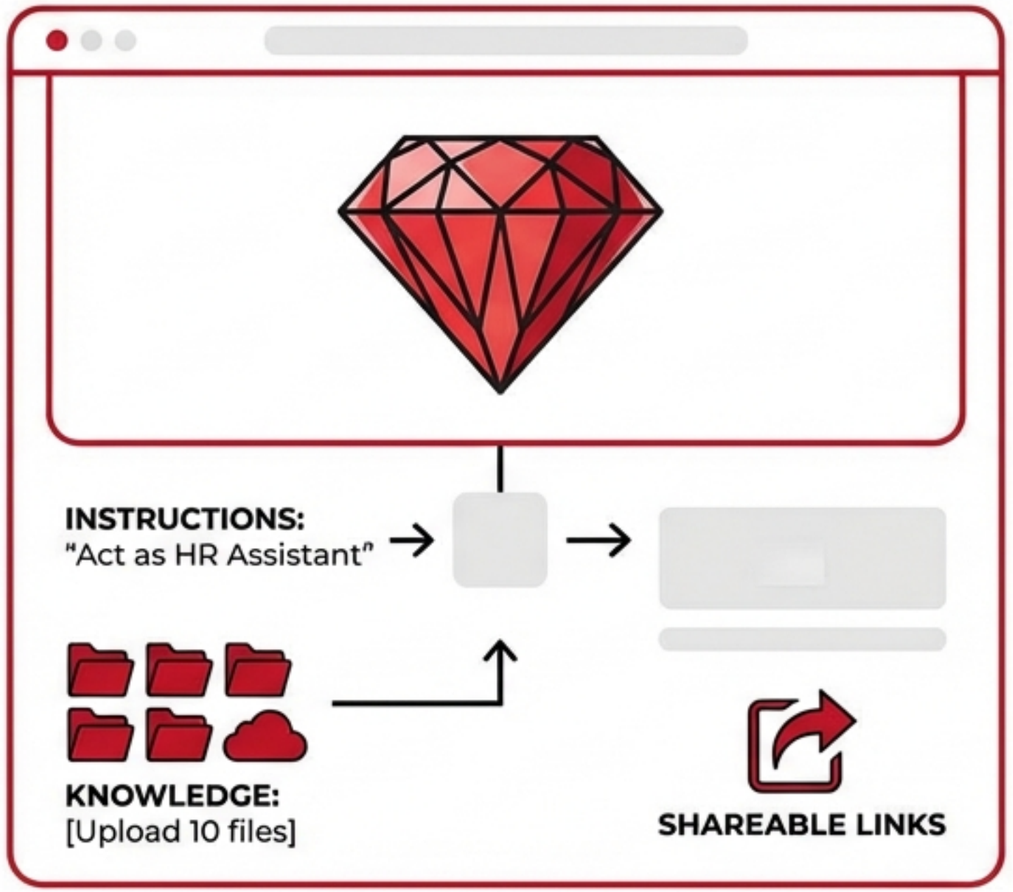
When developers tried to extract the “new brain,” they found nothing. The weights were locked.

WHAT REALLY HAPPENED?



Behavioral configuration ≠ Fine-tuning.
You don't need code to change behavior.

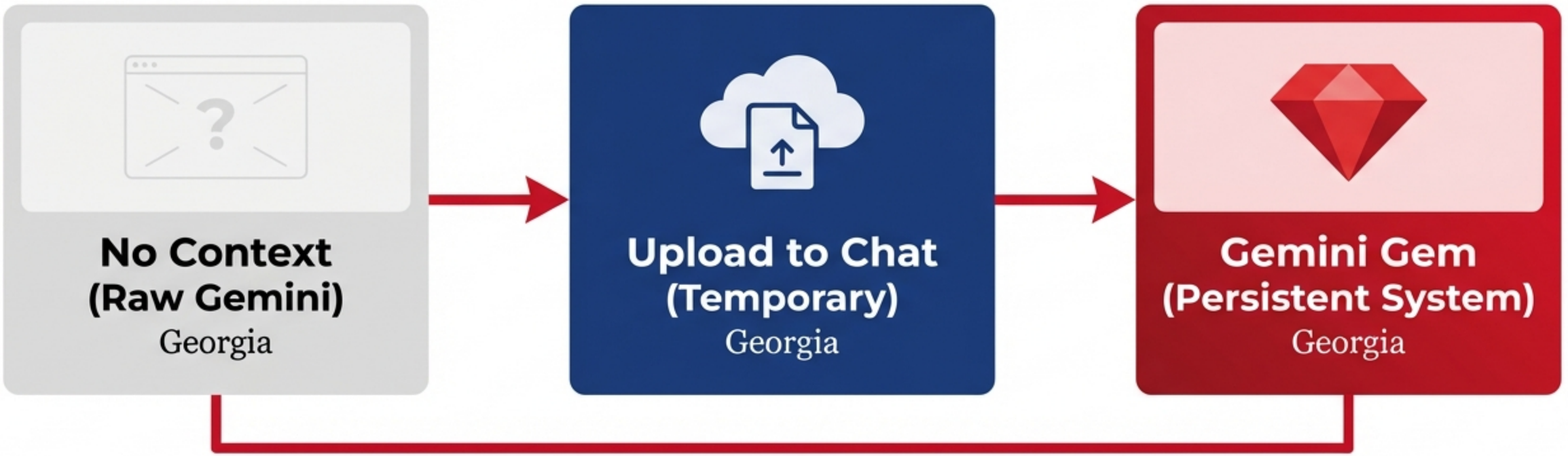
YOUR TOOL: GEMINI GEMS



- **No-Code RAG**
- **Persistent Knowledge** (Upload 10 files)
- **Custom Instructions** ("Act as HR Assistant")
- **Shareable Links**

Today, you build the Beacon Knowledge Assistant.

LAB PREP: THE EXPERIMENT



Open your laptops and head to gemini.google.com