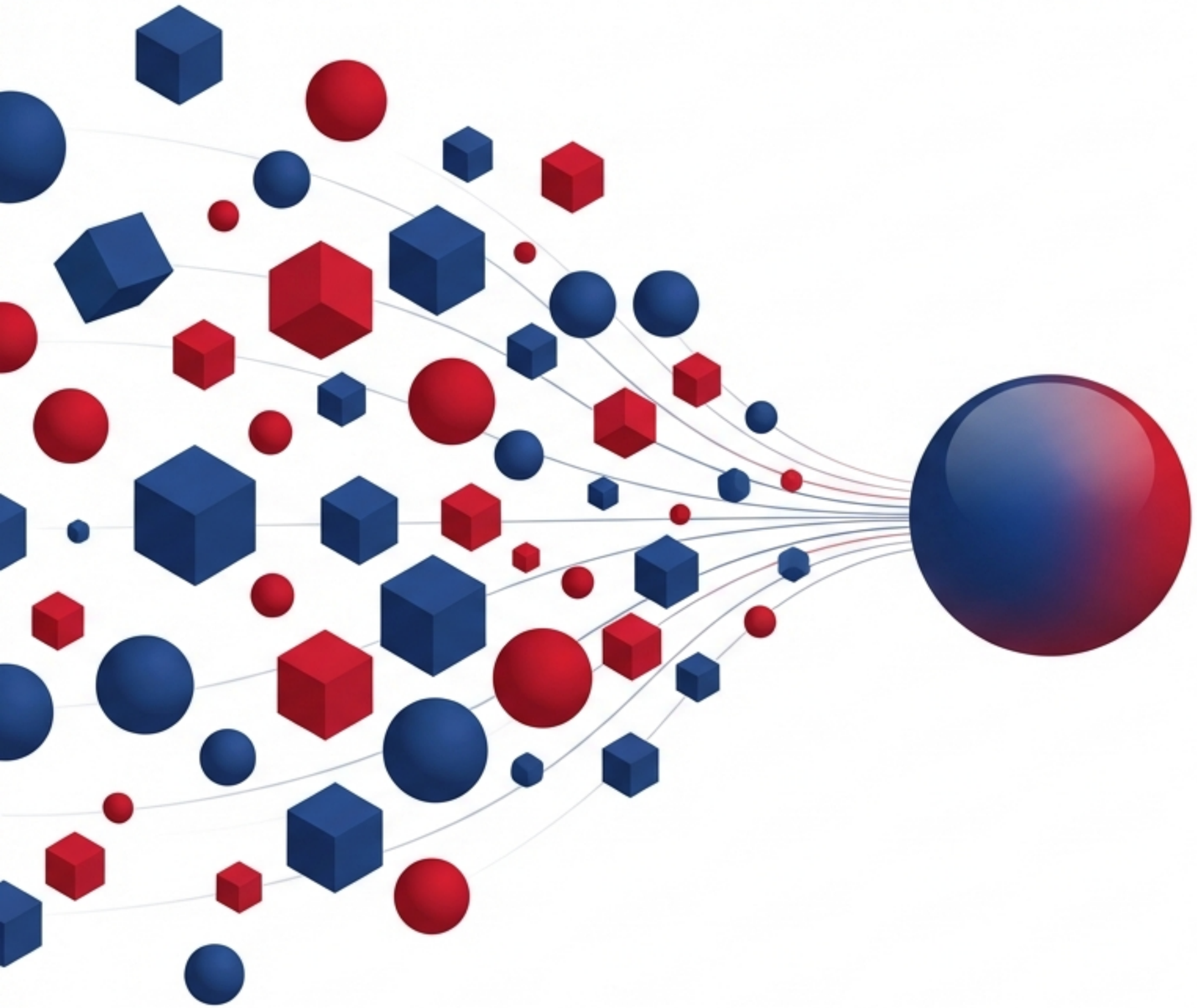


What is Generative AI?

Understanding the Technology Behind the Hype

UBUS 670 | AI for Business Leaders

Day 1 • Spring 2026

Today’s Learning Objectives

1. **Explain** how LLMs work at a conceptual level.
2. **Define** key mechanics: tokens, context windows, and temperature.
3. **Identify** hallucination risks and specific mitigation strategies.
4. **Navigate** Google Gemini to perform business tasks.

Today’s Tool: Google Gemini

gemini.google.com

Your Mission: Beacon Retail Group

You are the AI Strategy Task Force advising CEO Pat Holloway.

Founded 1987
in Rockford, IL.

Scale
25 regional stores, 1,200 employees, $312M revenue.

Competitors
Amazon, Walmart, Target.

Three Strategic Challenges

HR: Seasonal Hiring

- 4,200 applications/year
- 6 weeks to screen
- $2,500 cost per bad hire

Pain point: **Volume**

Marketing: Customer Service

- 850 emails/week
- 4-hour target response
- 60% are routine inquiries

Pain point: **Repetition**

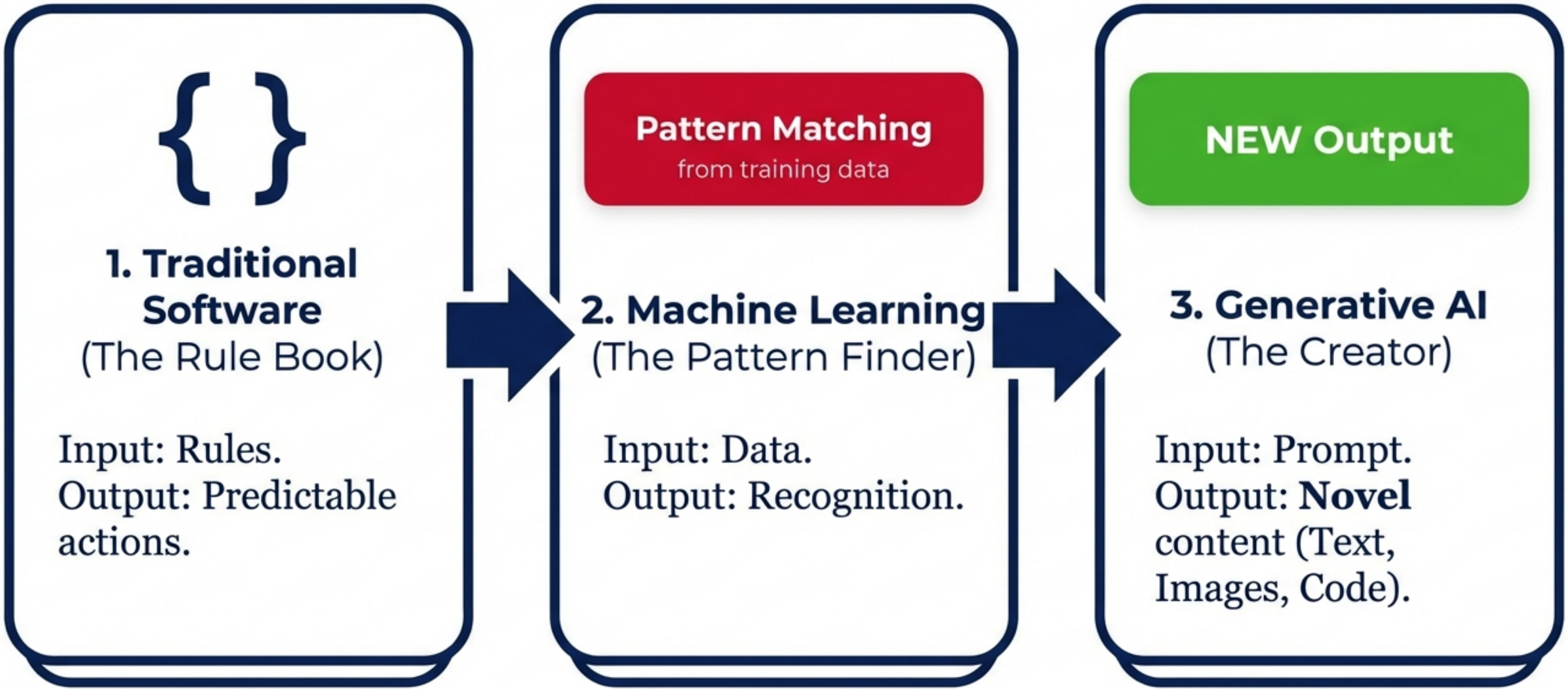
Finance: Expense Reports

- 1,200 reports/month
- 8-day processing time
- 4% manual entry errors

Pain point: **Accuracy**

Can AI help? To answer that, we first need to understand what it actually is.

The Evolution of Software



The Key Mental Model: A Synthesis Engine

Not a Database Retrieval.
A Prediction.

AI synthesizes new content from learned patterns.

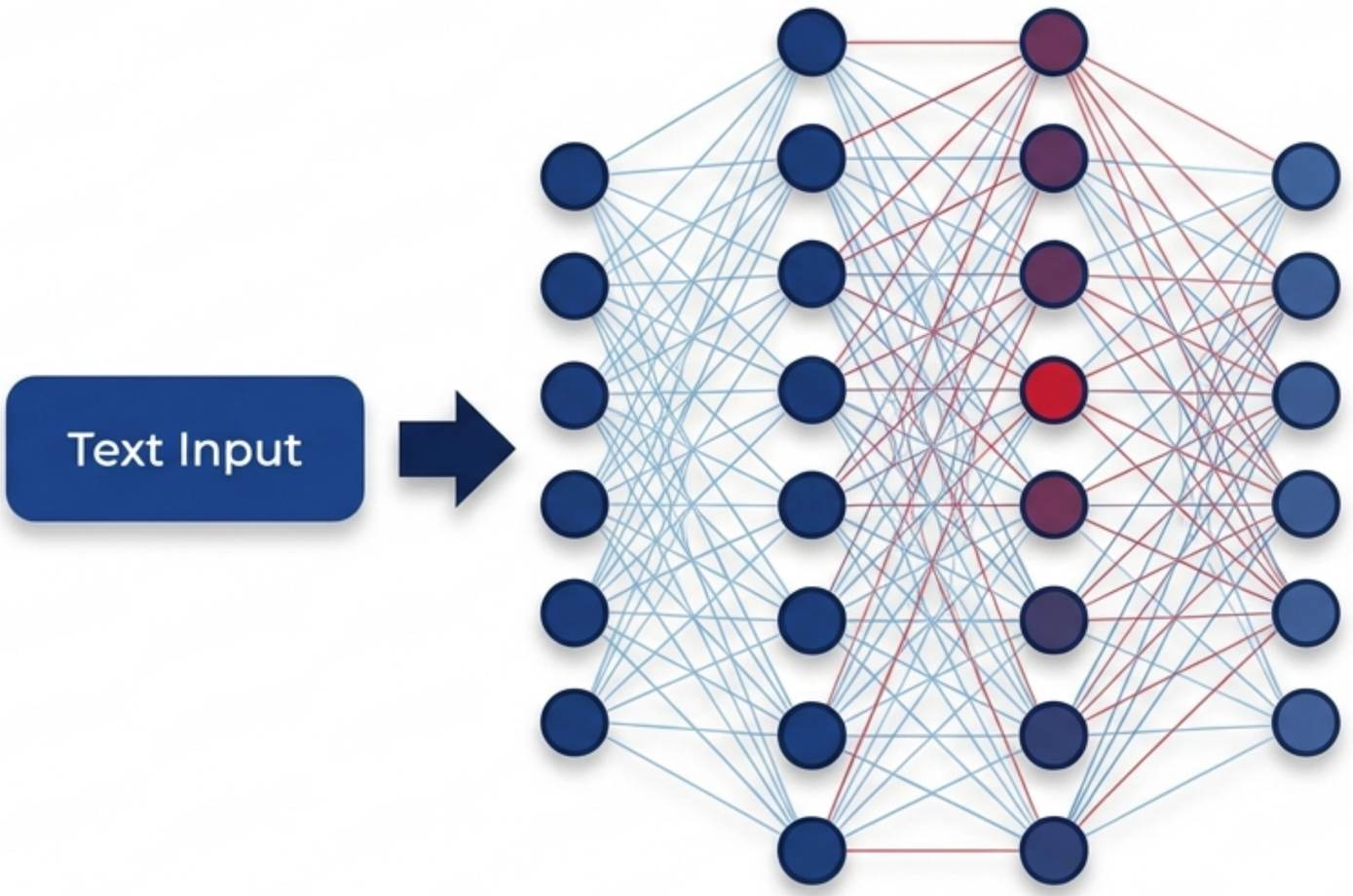
> *Think of it as an extremely well-read assistant who writes new content based on patterns learned from billions of documents.*

What is a Large Language Model (LLM)?

Definition:
A deep learning model trained on massive amounts of text data to generate human-like language.

The Scale:

- Training Data: **300 Billion+** words (The Internet)
- Parameters: **1 Trillion+** (The connections)
- Comparison: A human reads ~1 billion words in a lifetime.



The Currency of AI: Tokens

[Beacon] [Retail] [Group]

[retail] [er]

→ 1 Token ≈ 0.75 words
(or 4 characters)

Receipt

The Beacon Email Case

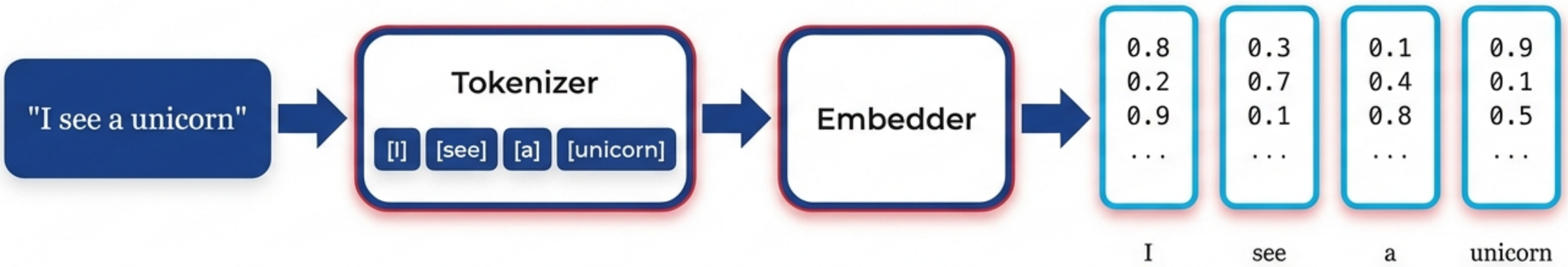
850 emails / week
x (270 input + 200 output tokens)

Total: 1.6 Million Tokens / Month

Cost (Gemini Pro): ~$2.00 / month
Cost (GPT-4): ~$47.00 / month

Efficiency saves money.

How Words Become Numbers (Embeddings)



The machine translates the word 'Unicorn' into a specific vector list: [0.9, 0.1, 0.5...]. This math represents meaning.

The Galaxy of Meaning

Why this matters for Beacon:

The AI knows that “This service is trash” is mathematically close to “I want a refund,” enabling semantic search.

Attention: Understanding Context

The **bank** was crowded, so I used the **ATM**.

Financial Institution

The **bank** was muddy, so I stayed in the **boat**.

River Edge

The ‘Attention’ mechanism looks at all words simultaneously to resolve ambiguity.

Autocomplete on Steroids

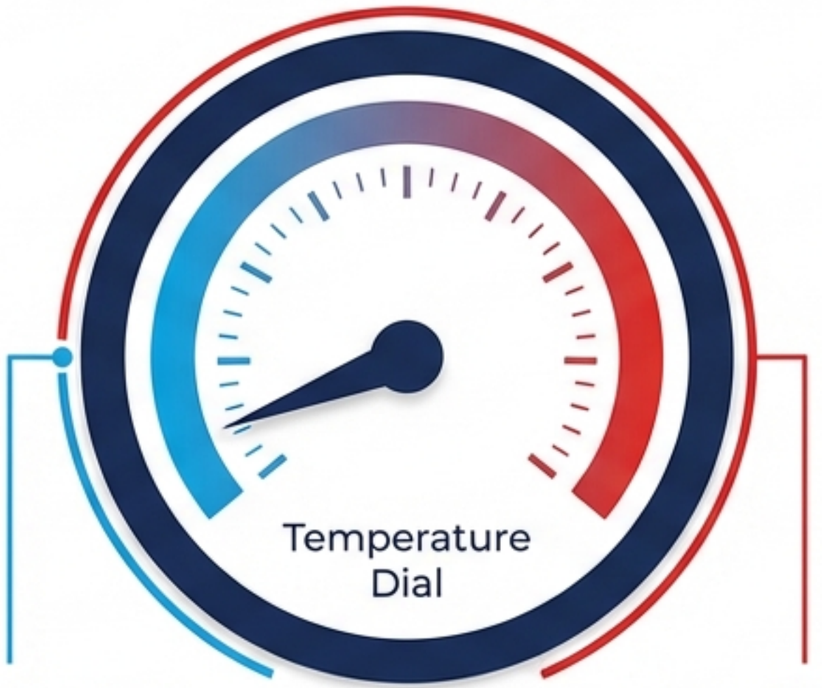
The quarterly sales report shows that revenue... |

Predict

slot machine

- ↑ **increased** (34%)
- ↓ decreased (28%)
- ↑ grew (15%)
- ↓ fell (12%)

Temperature Dial



Low (0.1)

Precise/Factual
(Picks top choice)

High (0.9)

Creative/Random
(Might pick "grew" or "fell")

The Context Window: Short-Term Memory

Yesterday's Chat

Old Files

The AI has **NO memory** beyond the current chat session.

Larger context = Higher Cost.

Fact or Fiction?

Gemini remembers the PDF I uploaded last Tuesday.

Context resets every chat.

Lowering temperature reduces hallucination risk.

Forces high-probability choices.

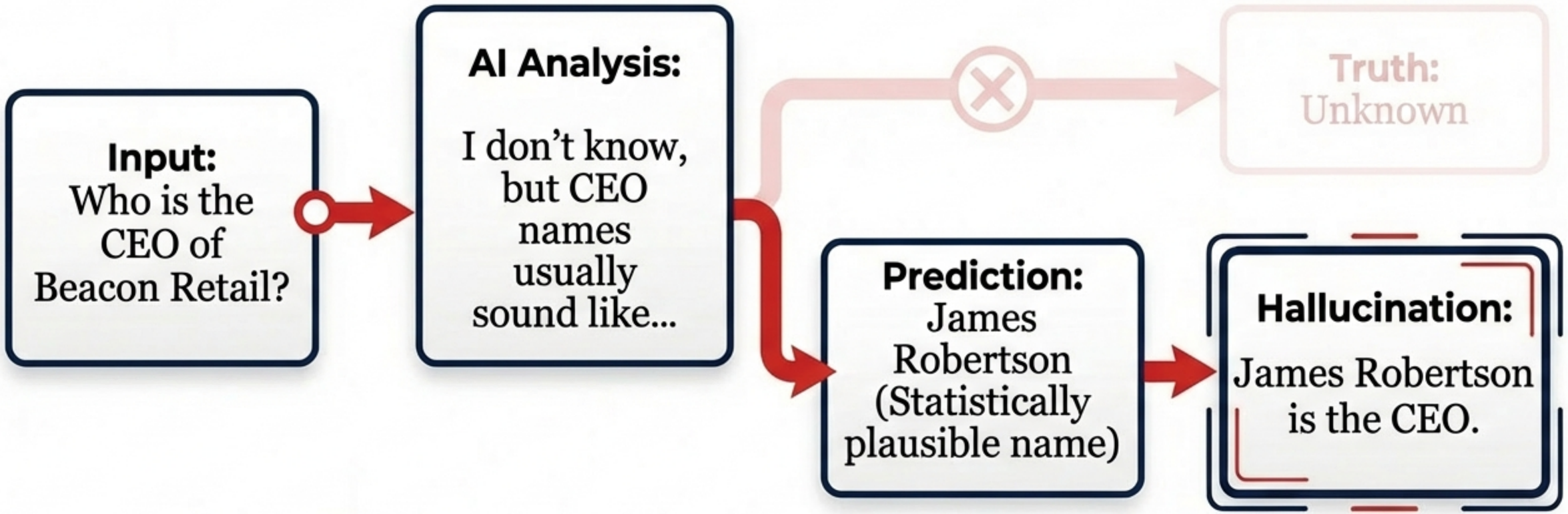
A 200-word email costs exactly 200 tokens.

1 token ≈ 0.75 words.

The Hallucination Problem

When AI generates content that is confident but factually incorrect.

Anatomy of a Lie



When Hallucinations Cost Money

The Air Canada Chatbot

Chatbot invented a refund policy.
Tribunal ruled the airline had to pay.

Cost: Brand trust + Refund

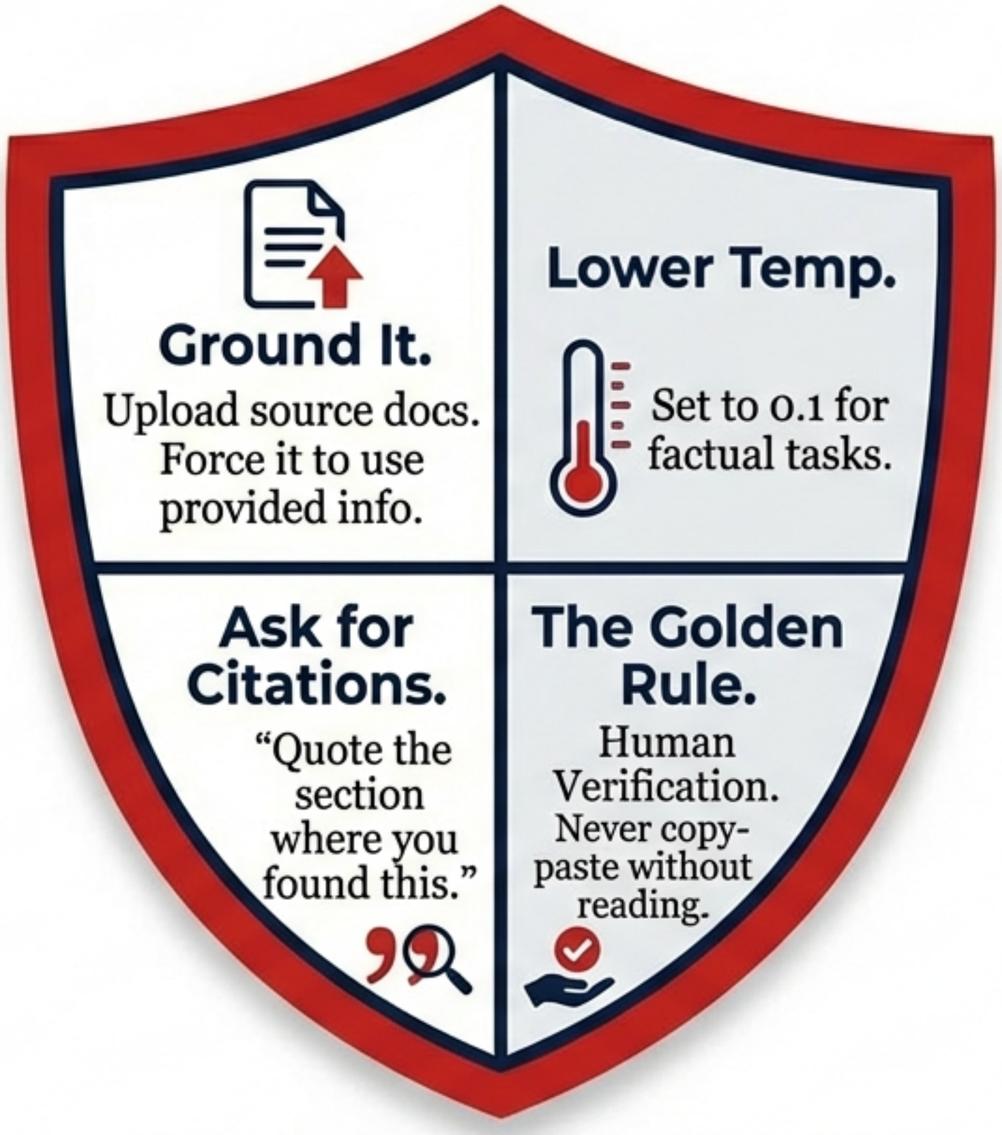
The NYC Lawyers

Used ChatGPT for a brief.
AI invented 6 fake cases.

Cost: $5k fine + Sanctions

Confidence does not equal Accuracy.

Mitigation Strategies



What AI Cannot Do (Yet)

- **No True Reasoning.** It simulates logic but fails at complex planning.
- **No Real-Time Knowledge.** (Unless connected to Search) Limited by training cutoff.
- **No Secret Keeping.** Do not upload PII or trade secrets to public models.
- **No Accountability.** You are responsible for the output, not the machine.

Meet Your Co-Pilot: Google Gemini

Multimodal Native.

Seamlessly understands and generates text, images, and code.

Ecosystem.

Deeply integrated with your daily productivity tools.

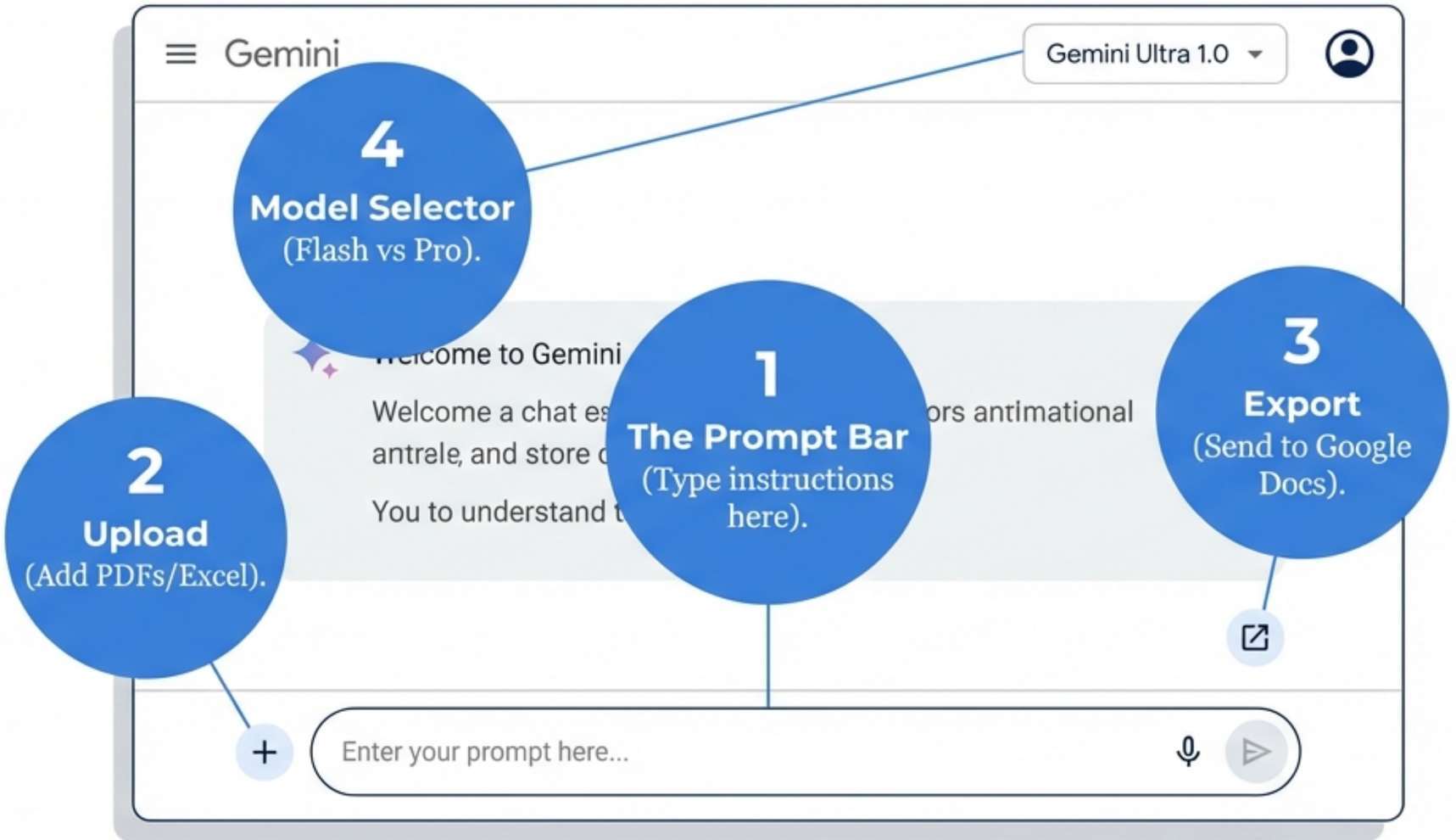
Why Gemini?

(2M tokens)

Massive Context.

Processes vast amounts of information and complex documents.

The Interface Tour



Up Next: Hands-On Lab

Key Takeaways

- LLMs are Prediction Machines, not Truth Machines.
- Context Windows are limited; manage them wisely.
- Hallucinations are a feature, not a bug—Verify everything.

Lab Objectives

1. Set up Gemini Account.
2. The Beacon Tasks: Summarize HR reports, Draft emails.
3. The Challenge: Trigger a hallucination.

Open gemini.google.com