What is Generative AI?

Understanding the Technology Behind the Hype

UBUS 670 | AI for Business Leaders

Day 1 • Week 1 • Spring 2026

Today's Learning Objectives

By the end of today, you will be able to:

1. Explain how LLMs work at a conceptual level
2. Define tokens, context windows, and temperature
3. Identify hallucination risks and mitigation strategies
4. Navigate Google Gemini Chat for business tasks

Today's Tool:

Google Gemini

gemini.google.com

Your Mission: Beacon Retail Group

You've joined the AI Strategy Task Force advising CEO Pat Holloway on AI adoption opportunities.

- Regional retail chain, 25 stores
- 1,200 employees, $312M revenue
- Founded 1987 in Rockford, IL
- Competing with Amazon, Walmart, Target

> Your role: Evaluate where AI can create real business value—and where it can't.

🏪

BEACON

RETAIL GROUP

"Lighting the way since 1987"

Beacon's Three Challenges

🧑‍💼 HR: Seasonal Hiring

4,200

applications/year

6 weeks to screen. 42% turnover rate. $2,500 cost per bad hire.

📧 Marketing: Customer Service

850

emails/week

4-hour response time. 60% are routine inquiries.

💰 Finance: Expenses

1,200

reports/month

8-day processing. 4% manual entry errors.

Can AI help? Let's find out what it actually is first.

Part 1

What is Generative AI?

Three Types of Software

Traditional Software

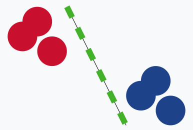
Rule-based

```
IF balance < 0
THEN show_warning()
```

Does exactly what programmers coded. Predictable but inflexible.

Machine Learning

Pattern-finding

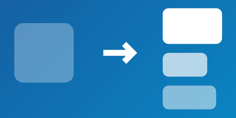


10,000 examples → Learns patterns

Finds patterns in historical data. Great for classification.

Generative AI

Content-creating



Creates NEW content

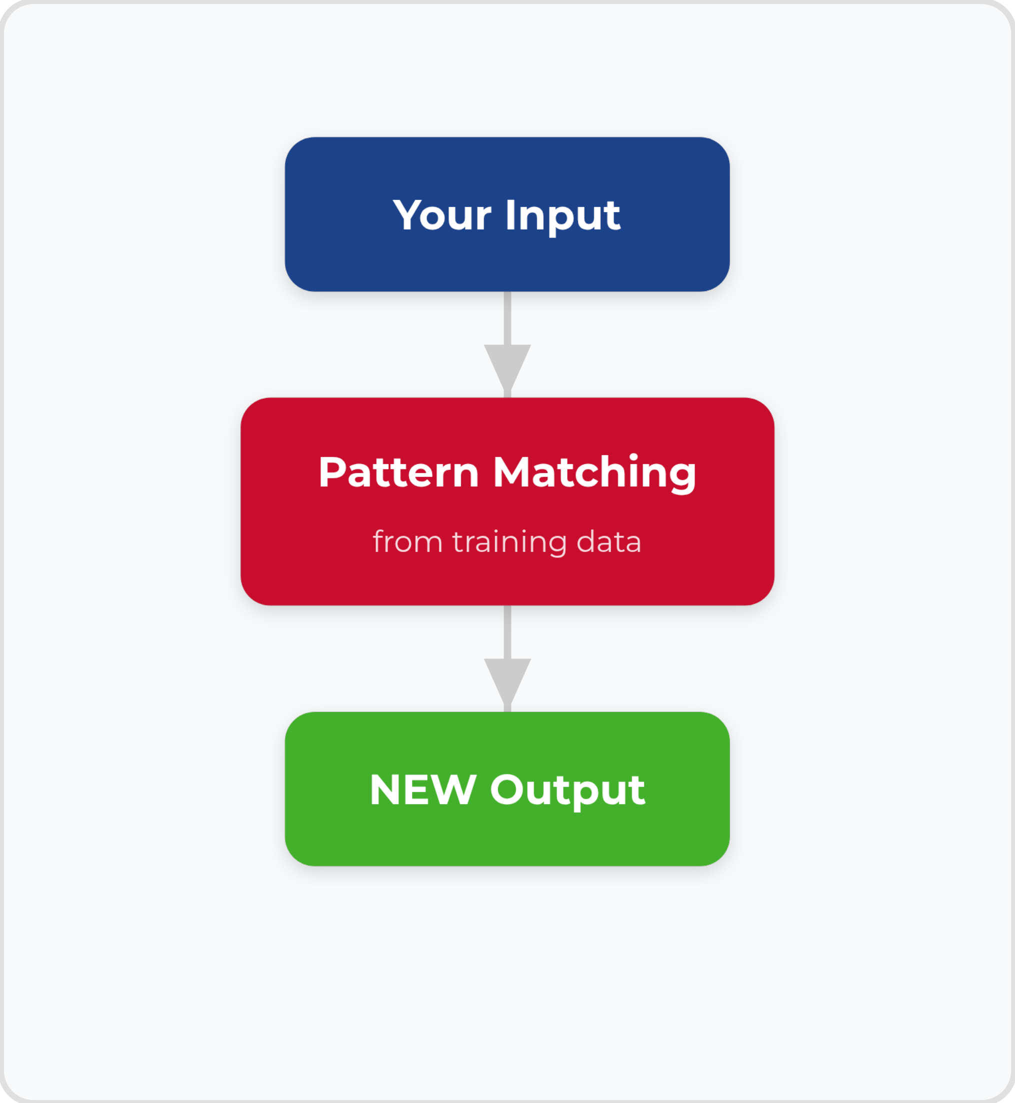
Produces novel outputs: text, images, code, ideas.

The Key Mental Model

Generative AI is a Synthesis Engine

It doesn't search for pre-existing answers.
It doesn't retrieve from a database.
It generates new content by predicting what comes next.

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



Part 2

How Do LLMs Actually Work?

What is a Large Language Model?

Definition

A Large Language Model (LLM) is a deep learning model trained on massive amounts of text data that can understand and generate human-like language.

Key characteristics:

- Large — Billions of parameters (GPT-4: ~1.7 trillion)
- Pre-trained — Learned patterns from internet-scale text
- General purpose — Can handle many tasks without task-specific training

Training Data Scale

300B+

words in training data

1T+

parameters (model weights)

For comparison: A human reads ~1B words in a lifetime

Four Key Concepts to Understand

1️⃣ Tokens

How text is broken into chunks the model can process

2️⃣ Embeddings

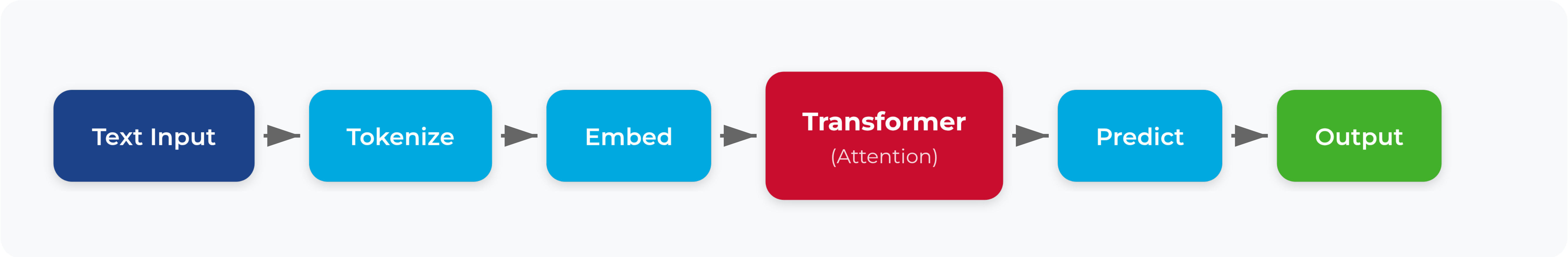
How words become numbers with meaning

3️⃣ Attention

How the model connects related words

4️⃣ Prediction

How responses are generated token-by-token



Step 1: Breaking Text into Tokens

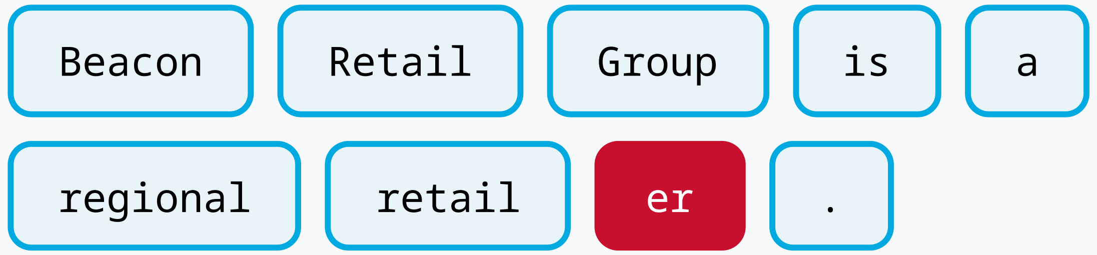
LLMs don't read words—they read tokens.
A token is a chunk of text, roughly:

- ~4 characters, or
- ~0.75 words

> ### Why does this matter?
>
> You pay per token and context limits are measured in tokens.

Example:

"Beacon Retail Group is a regional retailer."



`Beacon` `Retail` `Group` `is` `a`

`regional` `retail` `er` `.`

= 9 tokens (note: "retailer" splits into "retail" + "er")

Real Cost Example: Beacon's Emails

Let's calculate what it would cost Beacon to use AI for customer email responses:

- 850 emails per week
- Average email: 200 words (~270 tokens)
- Average response: 150 words (~200 tokens)

Bottom Line

AI email processing could cost Beacon $5-170/month depending on the model—far less than one employee's time.

```
// Weekly email volume
emails = 850
tokens_per_email = 270 + 200
weekly_tokens = 399,500
// Monthly cost (4 weeks)
monthly_tokens = 1,598,000
// At Gemini Pro pricing ($0.00125/1K)
gemini_cost = $2.00/month
// At GPT-4 pricing ($0.03/1K input)
gpt4_cost = $47.94/month
```

Step 2: Words Become Numbers (Embeddings)

Computers can't understand words directly. Embeddings convert tokens into lists of numbers (vectors) that capture meaning.

The Magic of Embeddings

Similar words have similar numbers. "King" and "Queen" are closer together than "King" and "Pizza."

Why this matters for business:

- AI understands "complaint" and "unhappy customer" are related
- Search can find similar products even with different words
- Enables semantic understanding, not just keyword matching

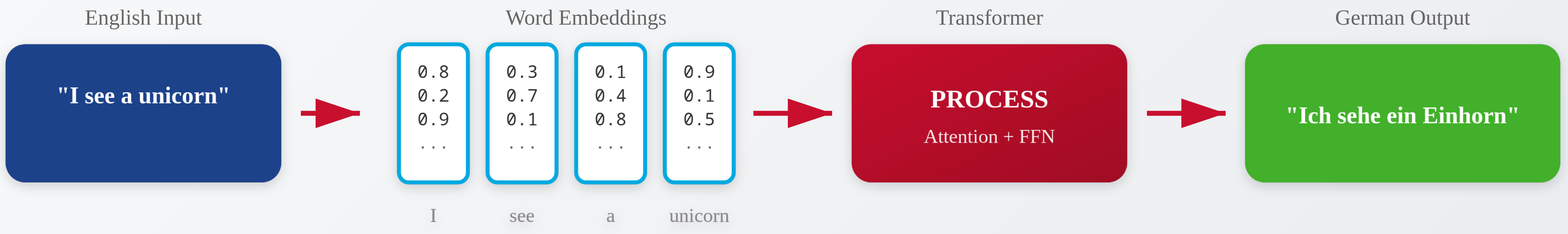
Simplified Example:

```
"King" → [0.8, 0.2, 0.9, ...]

"Queen" → [0.7, 0.3, 0.9, ...]

"Pizza" → [0.1, 0.9, 0.2, ...]
```

Real embeddings have thousands of dimensions, not just 3!

How AI Understands Language

Let's see how a transformer processes a simple sentence:



English Input

"I see a unicorn"

Word Embeddings

I	see	a	unicorn
0.8	0.3	0.1	0.9
0.2	0.7	0.4	0.1
0.9	0.1	0.8	0.5
...	...	...	...

Transformer

PROCESS

Attention + FFN

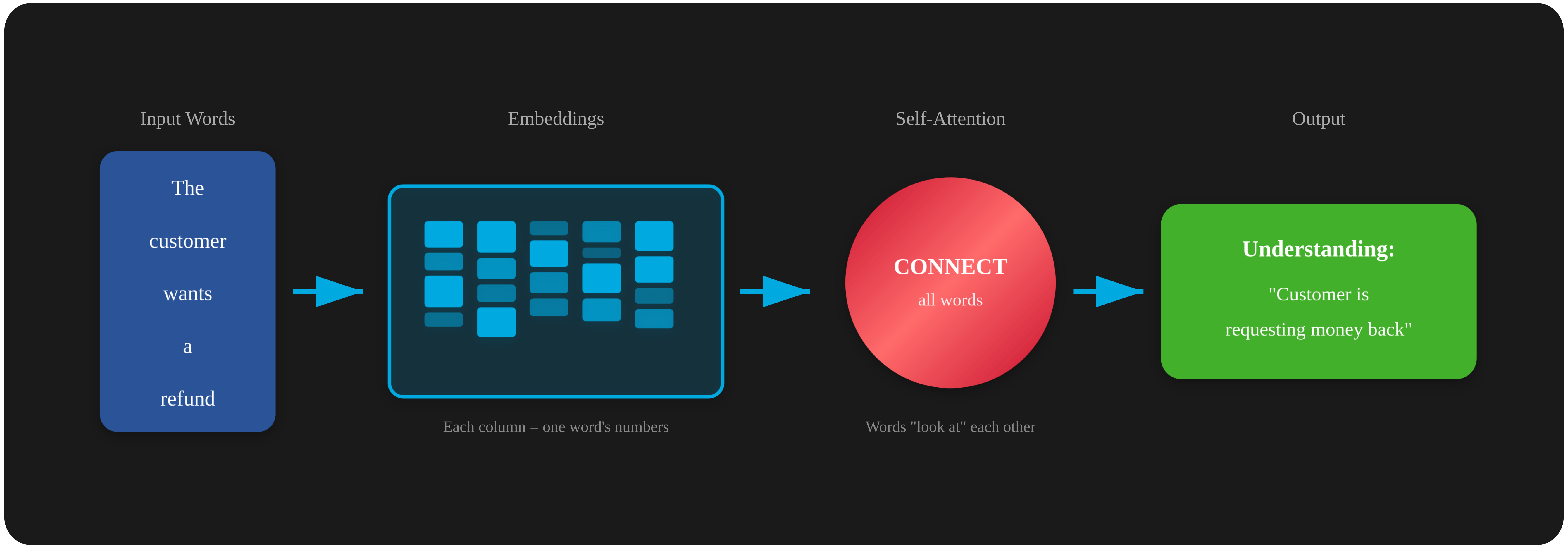
German Output

"Ich sehe ein Einhorn"

Key insight: Words become numbers (embeddings), get processed through the transformer, and new words are generated. The model learned these patterns from billions of translated sentences.

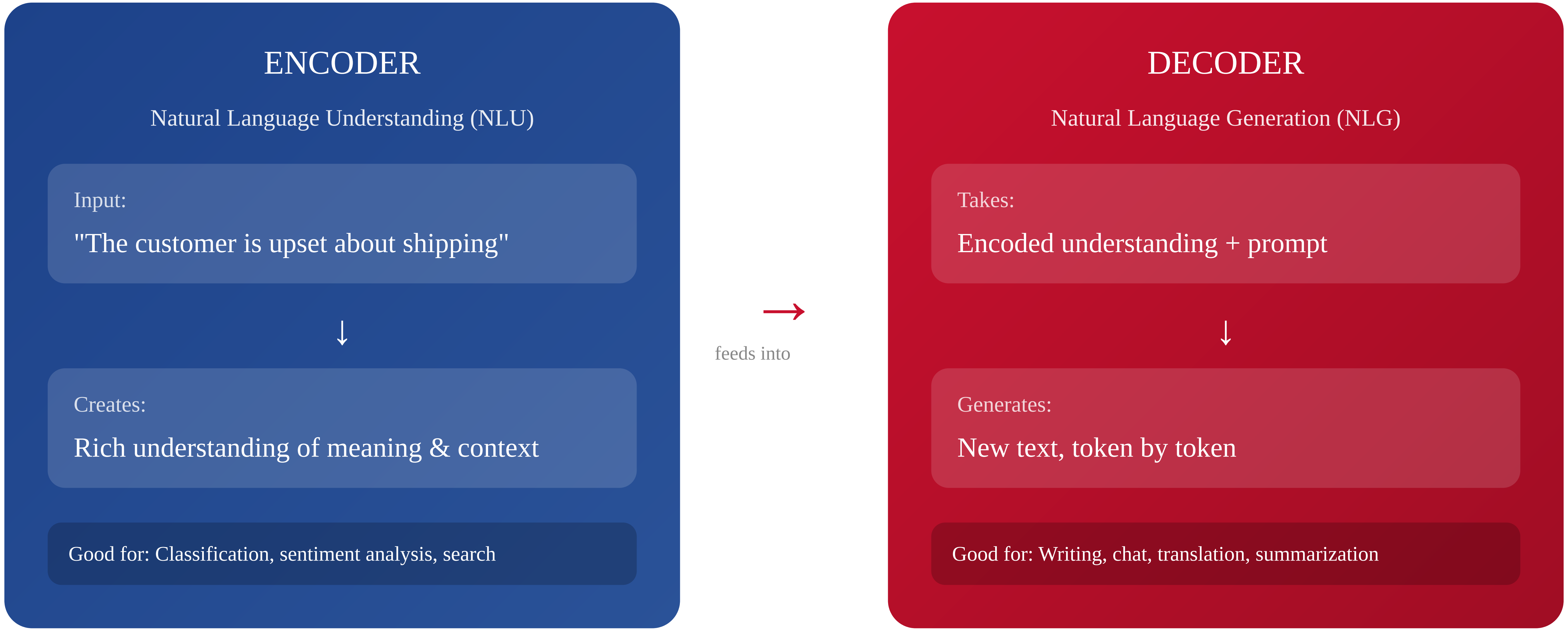
How a Transformer Works

The transformer architecture processes all words simultaneously:



Transformer Architecture: Encoder & Decoder

Modern LLMs use different combinations of encoders and decoders:



ENCODER

Natural Language Understanding (NLU)

Input:

"The customer is upset about shipping"

↓

Creates:

Rich understanding of meaning & context

Good for: Classification, sentiment analysis, search

→

feeds into

DECODER

Natural Language Generation (NLG)

Takes:

Encoded understanding + prompt

↓

Generates:

New text, token by token

Good for: Writing, chat, translation, summarization

ChatGPT, Gemini, Claude: Primarily use decoders (GPT = "Generative Pre-trained Transformer")
BERT (Google Search): Uses encoder only for understanding queries

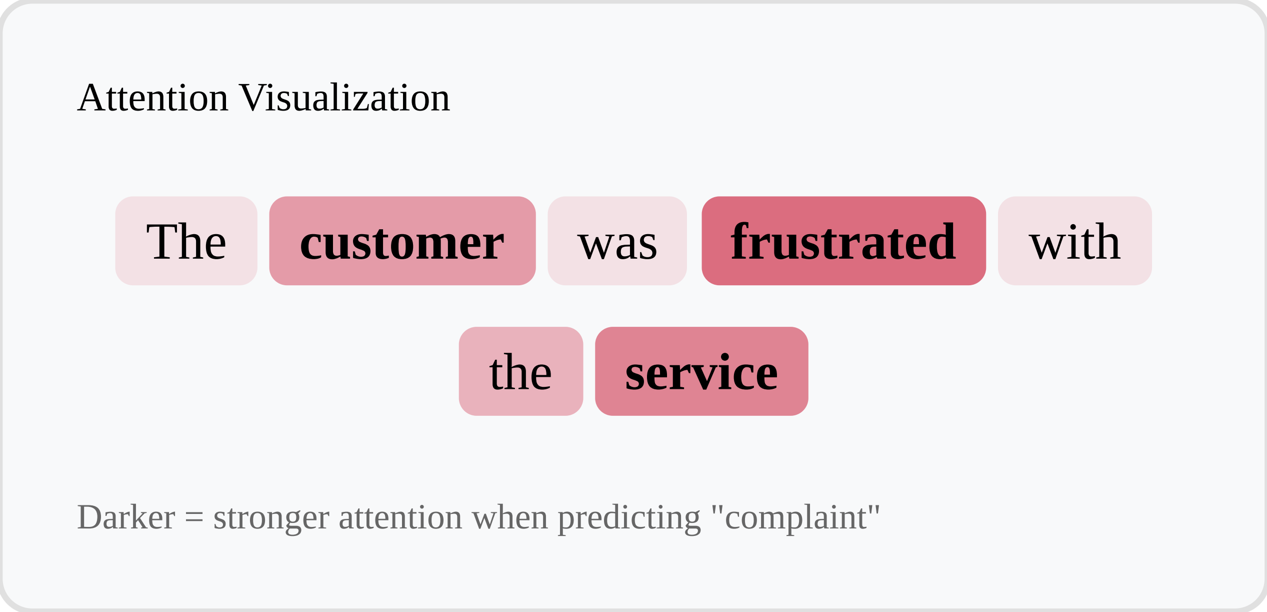
Step 3: Attention — Understanding Context

The Transformer architecture uses "attention" to understand how words relate to each other.

> ### Example
>
> "The bank was crowded, so I used the ATM."
>
> Attention connects "bank" to "ATM" to understand it's a financial institution, not a riverbank.

This is revolutionary because:

- Previous AI processed words in order (slow, forgetful)
- Attention looks at ALL words simultaneously
- Enables understanding of long documents



Attention Visualization

The **customer** was **frustrated** with the **service**

Darker = stronger attention when predicting "complaint"

Step 4: Predicting the Next Token

> An LLM is Autocomplete on Steroids
>
> Given some text, it predicts the most likely next token.

It does this one token at a time, thousands of times, to generate a complete response.

This is why AI can sometimes produce plausible-sounding nonsense—it's optimizing for "likely" not "true."

Input: "The quarterly sales report shows that revenue"

increased	34%
decreased	28%
grew	15%
fell	12%
...thousands more	11%

The Temperature Dial

Temperature controls how the model samples from those probabilities.

🧊 Low (0.0–0.3)

Predictable
Always picks highest probability
Use for: Facts, data extraction, classification

🌤️ Medium (0.4–0.6)

Balanced
Some variety, mostly consistent
Use for: Email drafts, summaries

🔥 High (0.7–1.0)

Creative
Willing to pick less likely tokens
Use for: Brainstorming, marketing copy

> Warning: Higher temperature = higher hallucination risk. For business tasks with factual requirements, keep temperature low.

The Context Window

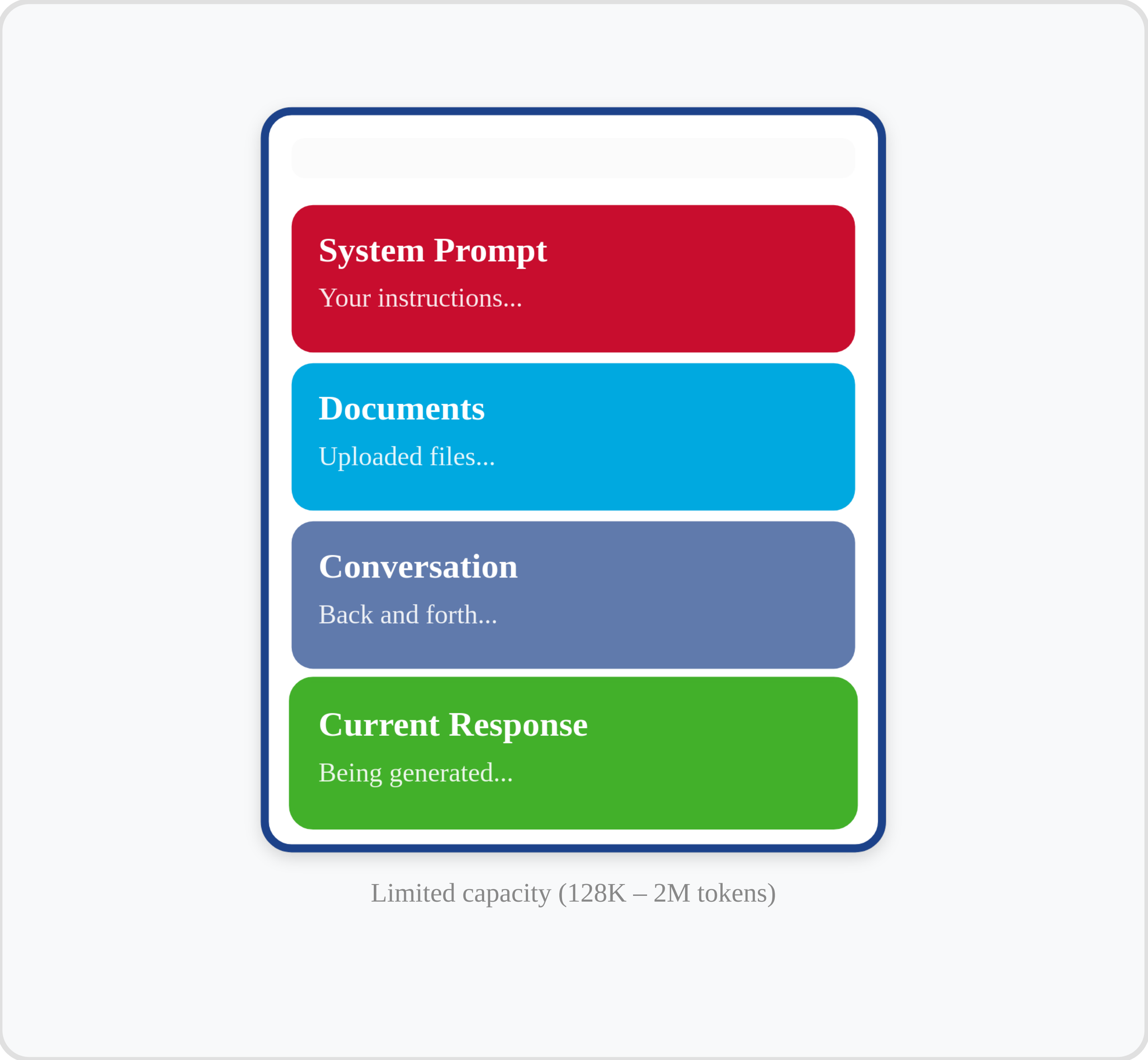
Think of it as "Working Memory"

The context window is everything the AI can "see" at once.

What's in the context:

- Your instructions (the prompt)
- Any documents you provide
- The conversation history
- The response being generated

Critical insight: The AI has NO memory beyond the context window. Each new conversation starts completely fresh.



Limited capacity (128K – 2M tokens)

Context Window Sizes (2026)

Model	Context Window	Roughly Equivalent To	Cost Impact
Gemini 1.5 Flash	1 million tokens	~10 novels	**$ (cheapest)**
Gemini 1.5 Pro	2 million tokens	~20 novels	**$$**
GPT-4 Turbo	128K tokens	~1 novel	**$$$**
Claude 3 Opus	200K tokens	~1.5 novels	**$$$**

Tip: Larger context ≠ always better. Larger contexts cost more, and models can "lose focus" in very long contexts. Use only what you need.

🧠 Checkpoint: Test Your Understanding

Before we continue, let's make sure these concepts are clear. Click an answer to check.

Q1: If a business email is 200 words, approximately how many tokens is that?

A) 50 tokens

B) 200 tokens

C) 270 tokens

D) 800 tokens

Q2: You asked Gemini about a document yesterday. Today you start a new chat and ask a follow-up question. Gemini doesn't remember the document. Why?

A) New conversations start with an empty context window

B) The document was too long

C) There's a bug in Gemini

Part 3

The Hallucination Problem

What is a Hallucination?

Definition

A hallucination is when AI generates content that sounds confident and plausible but is factually incorrect, made up, or inconsistent with provided information.

Why does this happen?

- LLMs predict likely text, not true text
- No built-in "I don't know" instinct
- Confidence doesn't correlate with accuracy
- Training data may contain errors or outdated info

?

Invented Citation

"According to Smith & Jones (2024)..."

Paper doesn't exist.

✕

Confident Wrongness

"Founded in 1982 by Michael Beacon..."

Wrong year, founder, and city.

Three Types of Hallucinations

🔴 Factual Fabrication

Making up facts that sound true

> *"The CEO of Walmart, James Robertson, announced..."*

There is no James Robertson at Walmart. The AI invented a plausible-sounding name.

🟠 Source Fabrication

Inventing citations and references

> *"Research by MIT (Chen et al., 2023) found that 78% of companies..."*

Paper doesn't exist. Statistics are fabricated.

🟡 Context Contradiction

Contradicting information you provided

> *You: "Our budget is $50K"*
> *AI: "With your $75K budget..."*

AI ignores or misremembers context you provided.

Real-World Hallucination Disasters

⚖️ Lawyers Cited Fake Cases (2023)

A New York law firm used ChatGPT to write a legal brief. The AI invented 6 fake court cases with made-up citations.

Result: Lawyers sanctioned, fined $5,000, case dismissed.

📰 AI News Errors (Ongoing)

CNET published AI-written articles containing basic math errors and factual mistakes about financial products.

Result: Public apology, articles corrected or removed.

The Lesson

AI can produce content that passes a quick glance but fails under scrutiny. Always verify outputs before using them professionally.

Mitigating Hallucinations

✅ Provide Source Documents

Ground the AI in YOUR specific facts. Instead of asking about Beacon in general, paste the actual company data.

✅ Use Lower Temperature

For factual tasks, use temperature 0.0-0.3. Higher creativity = higher hallucination risk.

✅ Ask for Uncertainty

Add: "If you're not sure about something, say so rather than guessing."

✅ Always Verify

Never trust AI output for critical decisions without human verification.

> Rule of thumb: The more critical the decision, the more verification you need. AI can draft, but humans must validate.

Part 4

What AI Cannot Do (Yet)

Current Limitations of Generative AI

🚫 No Real-Time Information

Models have a training cutoff date. They don't know about events after that date unless you provide the information.

🚫 No True Reasoning

LLMs simulate reasoning through patterns. Complex logic, math, and multi-step planning can fail unpredictably.

🚫 No Persistent Memory

Each conversation is independent. The AI doesn't remember you, your company, or previous conversations.

🚫 No Guaranteed Accuracy

AI can be confidently wrong. There's no internal "fact checker." Verification is YOUR responsibility.

> The upside: Many of these limitations are being actively addressed. But for now, design your AI workflows with these constraints in mind.

Part 5

Meet Google Gemini

Why We're Using Gemini

🎓 Perfect for Learning

- Free tier — No cost for coursework
- No credit card required — Just a Google account
- Large context window — Handle long documents
- Multimodal — Text, images, PDFs, and more

🛠️ Our Tool Progression

Week 1: Gemini Chat (prompt basics)

Week 2: Google Opal + AI Studio

Week 3: Multi-Agent Systems

> Already used ChatGPT? Great! The skills transfer. Gemini's interface is similar, but we'll explore features specifically useful for business tasks.

Gemini vs. ChatGPT: What's Different?

If you've used ChatGPT before, here's what to expect with Gemini:

✅ Similar

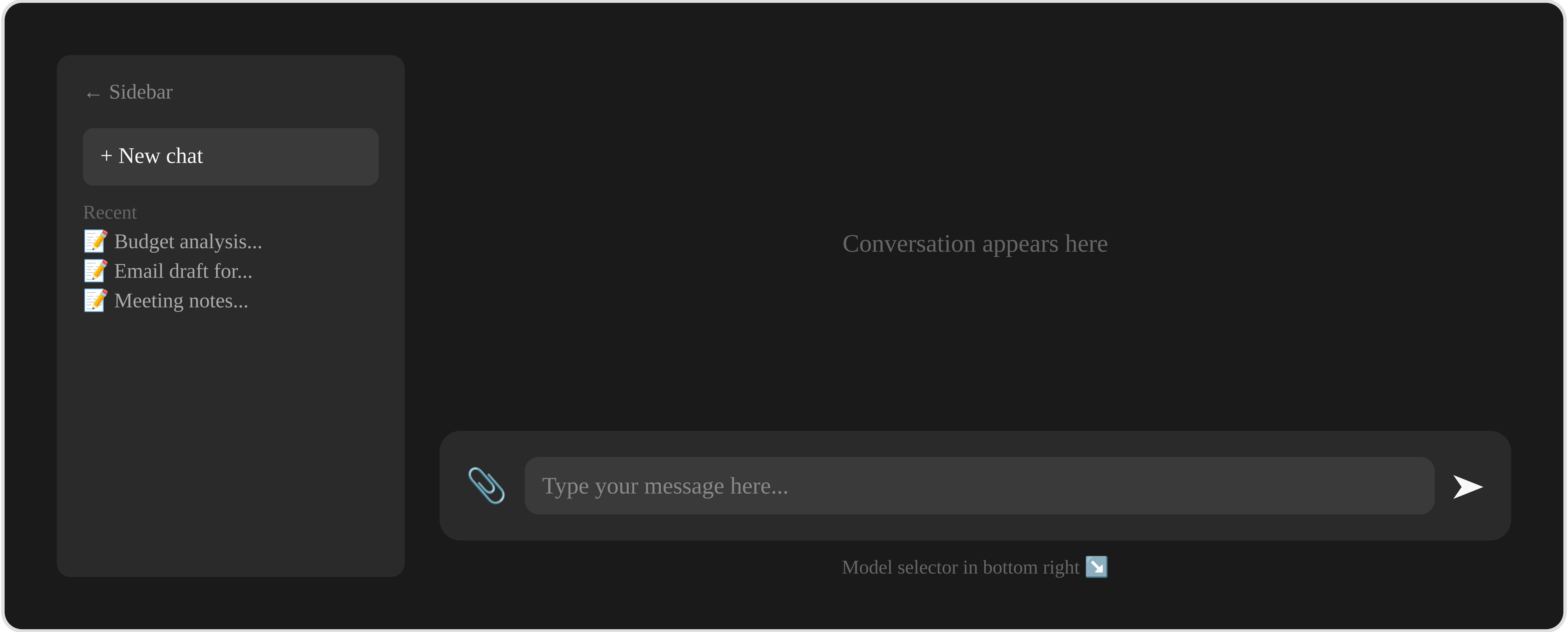
- Chat-based interface
- Follow-up questions in same conversation
- Can upload files and images
- Conversation history saved

🆕 Different

- Google account login (not OpenAI)
- Integrates with Google Drive, Docs, etc.
- Different model versions available
- May give different answers than ChatGPT

> Key insight: Different AI tools have different "personalities" and strengths. Learning multiple tools makes you more versatile.

Gemini Interface Tour



Model selector in bottom right ↘️

Key Gemini Features for Business

Upload Documents

PDFs, spreadsheets, images—Gemini can read and analyze them

Iterate & Refine

Ask follow-up questions to improve outputs

Copy & Export

Easy to copy responses or export to Google Docs

Model Selection

Choose between speed (Flash) vs. capability (Pro)

Pro tip: Start with the default model. If responses seem shallow or miss nuance, try switching to a more capable model.

Getting Started with Gemini

1. Go to gemini.google.com

   Works in Chrome, Firefox, Safari, Edge

2. Sign in with your Google account

   Personal account is fine—no upgrade needed

3. Start typing in the message box

   No special commands required

4. Press Enter or click Send

   Watch Gemini generate a response

Ready to try it?

Open Gemini →

You'll set this up properly in today's lab

Key Takeaways

1. Generative AI creates new content by predicting the next token based on patterns in training data.
2. Tokens = cost + limits. Understanding tokens helps you manage expenses and work within constraints.
3. Context window = working memory. The AI only knows what you put in the current conversation.
4. Temperature controls creativity vs. predictability. Use low for facts, high for brainstorming.
5. Hallucinations are inevitable. Always verify important information—never blindly trust AI output.

Up Next: Hands-On Lab

In today's lab, you will:

1. Set up your Google Gemini account
2. Complete three Beacon business tasks:
   - Summarization
   - Q&A with context
   - Email drafting
3. Trigger and document a hallucination
4. Reflect on what you learned

Estimated Time

90-120

minutes

Start Lab →

Questions?

Before we move to the lab...