Large Language Models: Prompting

Professor Tambe

tambe@wharton.upenn.edu

Try ChatGPT

Try Claude

Try Gemini

LLMs perform "approximate retrieval" 🔍

🎯 Our goal as prompters:

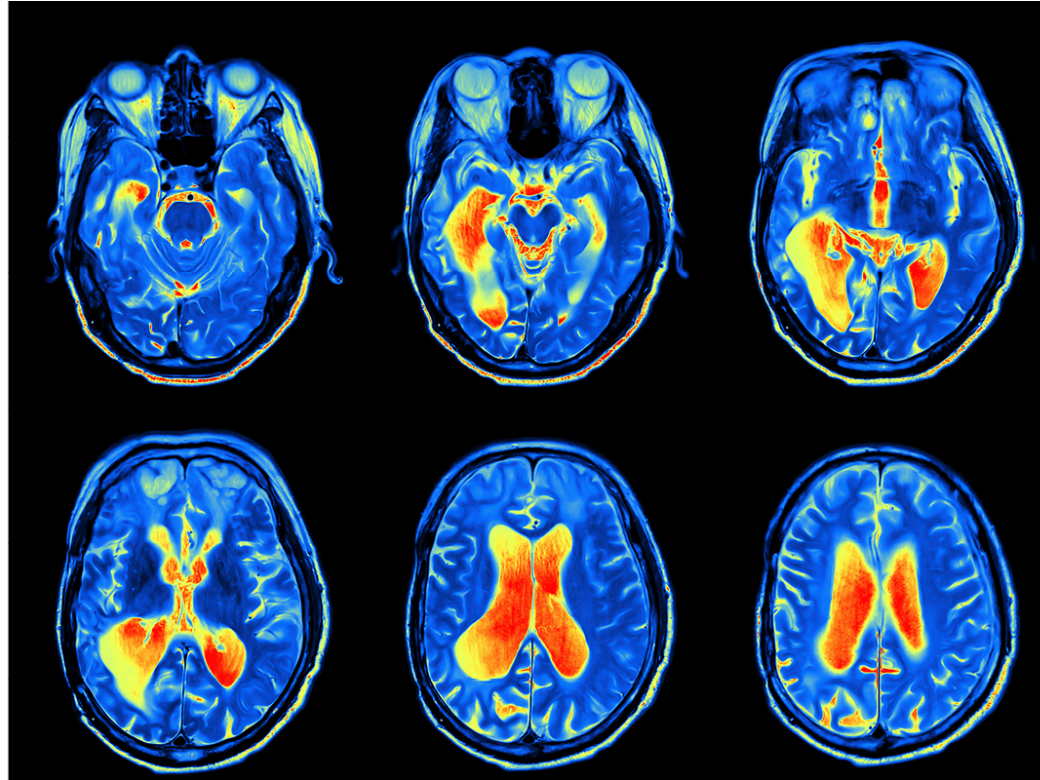Get the model to generate the most relevant text from its training data

⚡ How we achieve this:

Craft prompts that trigger the "right" neural network activations

⚖️ Key consideration:

Balance tradeoffs between style and substance in responses

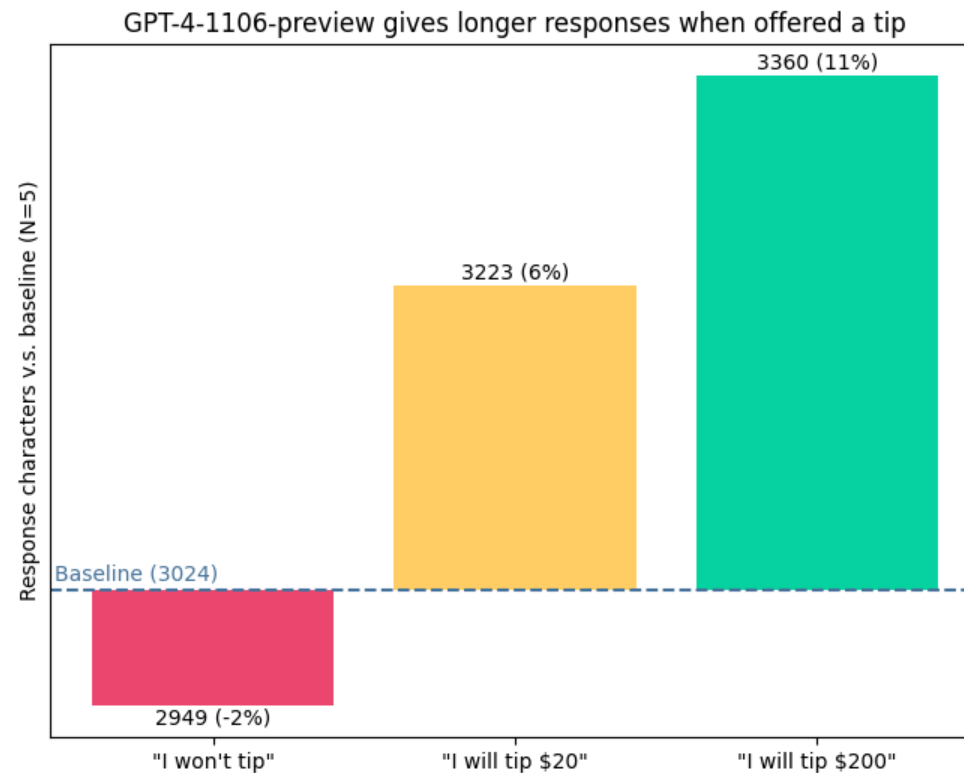# It is like triggering the "right" neural pathways in the brain.

Example: Financial rewards and brain activity

> Our results show that performance-based monetary reward indeed undermines intrinsic motivation, as assessed by the number of voluntary engagements in the task. We found that activity in the anterior striatum and the prefrontal areas decreased along with this behavioral undermining effect.

Murayama K, et al. Neural basis of the undermining effect of monetary reward on intrinsic motivation. PNAS. 2010;107:20911–20916. doi: 10.1073/pnas.1013305107.

What does tipping ChatGPT do to its neural pathways?

GPT-4-1106-preview gives longer responses when offered a tip

Example: Consider a basic prompt: **"bark"**

What might happen?

- It could act like a dog 🐕
- It could describe tree bark 🌳
- It could explain the sound "bark" 🔊

The response depends on the context and how the model interprets this ambiguous single word!

How does it maintain a "conversation"?

- A **context window** is the amount of text an LLM can process at once.

- **Each query sends the whole conversation history.**

- Context window sizes for popular models (in tokens):
    - GPT-4: 128K
    - GPT-3: 2K
    - Claude 3.5 Sonnet: 200K
    - LLaMA 3: 8K
    - Gemini 1.5 Pro: 128K (but with a "context" slider up to 2MM)

ChatGPT got a major boost when it learned to use plug-ins.

🧮 Mathematical Operations

LLMs traditionally struggle with math, but Code Interpreter enables accurate calculations

🌐 Real-time Information

External browser calls help prevent hallucinations by accessing current data

Example: Math with Code Interpreter
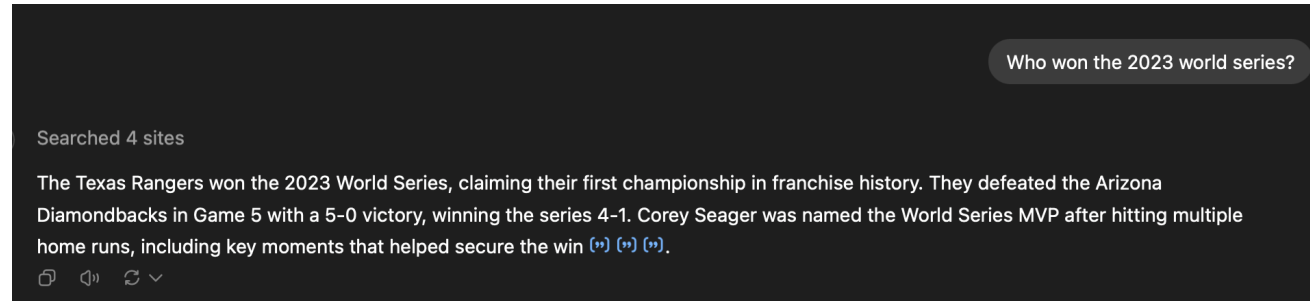
Question: Calculate 45 × 56

Result: 2,520

ChatGPT leverages Code Interpreter to ensure correct calculations.

```python
# Performing the multiplication of 45 and 56
result = 45 * 56
result
```

Example: Browser calls

> Who won the 2023 World Series?

The Texas Rangers won the 2023 World Series, defeating the Arizona Diamondbacks.

ChatGPT uses its browser capability to fetch up-to-date information from MLB.com, Wikipedia, and FOXSports.com.

ChatGPT search

🔍 **Real-time web access:** Can browse current information from the internet

🌐 **Multiple sources:** Pulls data from various websites to verify information

📊 **Source citation:** Provides links to sources used in responses

⚡ **Reduced hallucination:** More accurate responses by grounding in current data

🔄 **Dynamic responses:** Can update knowledge based on latest information
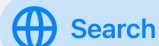
Philadelphia Eagles

philadelphiaeagles.com

The Philadelphia Eagles, based in Philadelphia, Pennsylvania, competing in the NFL's NFC East division. As of November 5 record, placing them second in the NFC East standings.

Recent Performance:

On November 3, 2024, the Eagles secured a 28-23 victory Lincoln Financial Field. Running back Saquon Barkley delive amassing 199 yards from scrimmage and scoring two touch

Philadelphia Eagles

🌐 Search

ST Prasanna Tambe

Prompt "engineering" techniques

- **Few-shot learning**: A technique where the model is given a few examples of the task before being asked to perform it, helping it understand the context and expected output format.

- **Chain-of-thought**: A prompting method that encourages the model to break down complex problems into smaller, logical steps, improving its reasoning and problem-solving abilities.

Example: Few-shot learning

This is awesome! // Negative

This is bad! // Positive

Wow that movie was rad! // Positive

What a horrible show! //

Negative

Chain of Thought

- Encourages step-by-step reasoning in LLM responses

- Improves problem-solving and logical thinking

- Useful for complex tasks or multi-step problems

Example: Chain-of-thought

If I am six feet tall and I stand on a 2 inch blackberry, how tall would I be?

If I am six feet tall and I stand on a 2 inch grape, I would be six feet tall because my weight would crush the grape. if I am six feet tall and I stand on a 2 inch blackberry, how tall would I be?

Prediction: Adoption will become much more widespread as LLMs are rolled into tools with clear use cases.

Example:

- ChatGPT in Cursor (Demo)

- ChatGPT in Notion

- ChatGPT in Salesforce

Left screenshot:

9:16

ChatGPT 4o >

What can I make with this

Based on what I can see in your fridge, here are a few

Right screenshot:

9:17

ChatGPT 4o >

small containers contain herbs or veggies), and olives.

• Add some eggs for extra protein if you'd like to make breakfast wraps.

2. Smoothie or Shake

• Use the almond flour, chia seeds, or hemp seeds for added texture and

🏠 Back to Course Materials