# Large Language Models:

## How they work and their applications in Political Science

Dr. Matt DiGiuseppe

March 13, 2025

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
  - API usage - calling from R

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
  - API usage - calling from R
  - Qualitative Text to Numbers

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
    - API usage - calling from R
    - Qualitative Text to Numbers
    - Vibe coding

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
  - API usage - calling from R
  - Qualitative Text to Numbers
  - Vibe coding
  - Synthetic samples

**What we won't cover:**

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
    - API usage - calling from R
    - Qualitative Text to Numbers
    - Vibe coding
    - Synthetic samples

**What we won't cover:**

- Ethics of AI/LLMs

# Game Plan

**What we'll cover:**

- Basics of LLMs (How it works)
- Implications for how to use LLMs
- Application for use in social science research
    - API usage - calling from R
    - Qualitative Text to Numbers
    - Vibe coding
    - Synthetic samples

**What we won't cover:**

- Ethics of AI/LLMs
- Speculation on the future of AI

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.

## Capacity Enhancement

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.
- Not using LLMs can result in lower productivity and potential job risks.

## Capacity Enhancement

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.
- Not using LLMs can result in lower productivity and potential job risks.
- Learning to use them well offers a competitive advantage and more leisure time.

## Capacity Enhancement

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.
- Not using LLMs can result in lower productivity and potential job risks.
- Learning to use them well offers a competitive advantage and more leisure time.
- Using them *less well* puts you at risk

## Capacity Enhancement

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.
- Not using LLMs can result in lower productivity and potential job risks.
- Learning to use them well offers a competitive advantage and more leisure time.
- Using them *less well* puts you at risk

## Capacity Enhancement

- Access to capabilities once too costly to learn

# Why invest in LLMs?

## Labor Market Incentives

- LLMs make workers more productive.
- Not using LLMs can result in lower productivity and potential job risks.
- Learning to use them well offers a competitive advantage and more leisure time.
- Using them *less well* puts you at risk

## Capacity Enhancement

- Access to capabilities once too costly to learn
- Serve as your programmer, research assistant, web searcher, brainstorming partner, legal advisor, etc.

# Disclaimer

**While LLMs can increase your productivity, over reliance can make you dumber.**

# LLMs, How Do They Work?

AI and LLMs seem like magic.

# LLMs, How Do They Work?

AI and LLMs seem like magic.

But in fact they are based on technology developed in the 1950s.

LLMS are a (very!) large statistical models that predict the next token (word).

# Words to Numbers

**LLMs run on tokens:**

- Text is broken down to tokens (word, subword, or character)

**LLMs run on tokens:**

- Text is broken down to tokens (word, subword, or character)
- Tokens are then assigned a numeric id then a vector of numbers (embedding).

# Words to Numbers

**LLMs run on tokens:**

- Text is broken down to tokens (word, subword, or character)
- Tokens are then assigned a numeric id then a vector of numbers (embedding).
- Numbers are universal.

Token count
27

```
<|im_start|>system<|im_sep|>You are a helpful assistant<|im_end|><|im_start|>user<|im_sep|>strawberry is a weird word, right?<|im_end|><|im_start|>assistant<|im_sep|>
```

```
200264, 17360, 200266, 3575, 553, 261, 10297, 29186, 200265, 200264, 1428, 200266, 302, 1618, 19772, 382, 261, 95639, 67, 2195, 11, 1849, 30, 200265, 200264, 173781, 200266
```

**LLMs run on tokens:**

- Text is broken down to tokens (word, subword, or character)
- Tokens are then assigned a numeric id then a vector of numbers (embedding).
- Numbers are universal.
- Statistical models need numbers, not words.

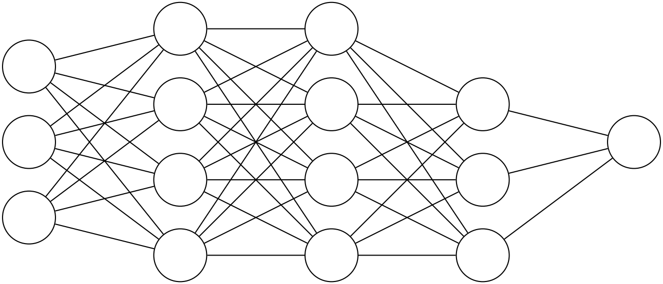
Token count
27

```
<|im_start|>system<|im_sep|>You are a helpful assistan
t<|im_end|><|im_start|>user<|im_sep|>strawberry is a w
ierd word, right?<|im_end|><|im_start|>assistant<|im_s
ep|>
```

```
200264, 17360, 200266, 3575, 553, 261, 10297, 29186, 2
00265, 200264, 1428, 200266, 302, 1618, 19772, 382, 26
1, 95639, 67, 2195, 11, 1849, 30, 200265, 200264, 1737
81, 200266
```

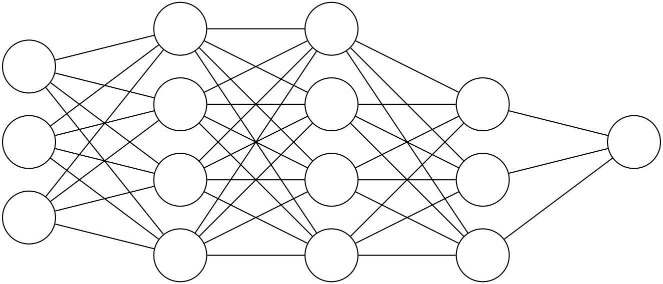
# What is a Neural Network?

- Inspired by the human brain, a neural network is a system of "neurons" that learn patterns.



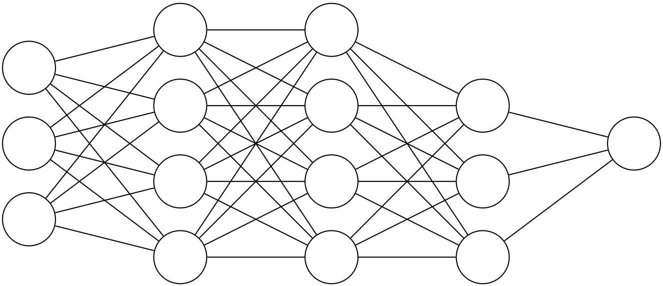
# What is a Neural Network?

- Inspired by the human brain, a neural network is a system of "neurons" that learn patterns.
- Each neuron (node) processes information and passes it to the next layer.



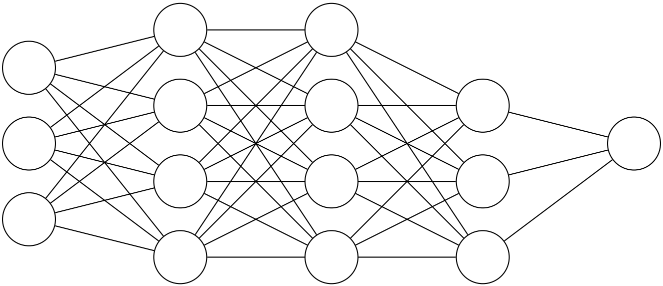
# What is a Neural Network?

- Inspired by the human brain, a neural network is a system of "neurons" that learn patterns.
- Each neuron (node) processes information and passes it to the next layer.
- By adjusting the connections (weights), the network learns from data.



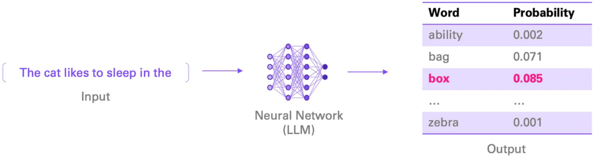
# What is a Neural Network?

- Inspired by the human brain, a neural network is a system of "neurons" that learn patterns.
- Each neuron (node) processes information and passes it to the next layer.
- By adjusting the connections (weights), the network learns from data.
- We have a limited information on how these nodes interact at scale...



# How Training Data Adjusts Weights

- Starts with random connections (weights).
- Each training example is passed through the network.
- The network makes a prediction and compares it to the correct answer.
- The error (difference between prediction and actual value) is calculated.
- Using backpropagation, the network adjusts the weights to reduce the error.



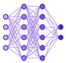

Source: Andreas Stoffelbauer, Microsoft

# How Training Data Adjusts Weights

- Over many iterations, the network improves its accuracy.
- Once weights are established, the model can predict the next token (over and over).
- Weights don't change as you use it!

[ The cat likes to sleep in the ]

Input



Neural Network
(LLM)

| Word | Probability |
|---|---|
| ability | 0.002 |
| bag | 0.071 |
| **box** | **0.085** |
| ... | ... |
| zebra | 0.001 |

Output

Source: Andreas Stoffelbauer, Microsoft

# A Better Tutorial on How LLMs Work

3Blue1Brown.com

# What is Used for Training Data?

- Training data consists of input-output pairs that help the network learn.
- The quality and quantity of training data impact model performance.
- LLMs use...
    - Books, articles
    - Web-based text
    - Technical documents
    - Conversations (e-mails, texts, tweets)
    - Code bases (Stack Overflow, Github)

# Training Data and Bias

**Training data is not the truth!**

- Bias occurs when training data is not representative or contains systematic errors.
- It can lead to unfair or incorrect predictions.
- Examples:
    - LLMs might reflect racial, gender, or other biases of human writing
    - Training data is heavily English and Western in origin
    - Could lead to harmful outputs -
        - *build a chemical weapon? Sure, I'll show you how!*

# Steps Taken After Training an LLM

- Fine-tuning: refinement on domain-specific or high-quality datasets to improve accuracy and reduce bias.
- Reinforcement Learning from Human Feedback (RLHF)
- Safety, Ethical, and Quality Adjustments:**
- System Prompts: hidden instructions to shape responses:
  - "You are a friendly chatbot that speaks in an informal, engaging tone."
- Personality Design: adjusting tone, style (via RLHF, system prompts, finetuning)

No guarantee this removes all biases or harms

# Model Size

- Positive relationship b/w parameters and performance but not a hard rule.
- Speed is the other important factor - larger models are slower

| **Model** | **Parameter Count** |
|---|---|
| LLaMA 3.1 (Meta) | 8B, 70B, 405B |
| GPT-4 | 1.7 Trillion?? |
| Gemini Flash 2.0 | 100B?? |
| Claude 3.7 | 52B?? |
| Grok-1 | 314B |

# Training to Inference

After a model has been trained and tuned, you can use it to predict ("inference").

- Every new chat starts fresh*
- Your prompt and whatever you upload is used to predict the next token.
- Each new contribution to a chat is retained to keep track of the conversation.

The context window serves as the memory of
what you've said and uploaded &
what the LLM responded with in this chat.

Older tokens drop out →

```
User:  Hello!

LLM: Hi there!

User:  How are you?

LLM: Doing well, thanks.

User:  What's the weather?

LLM: Sunny today.

User:  Any plans?

LLM: Reading a book.

User:  Which book?

LLM: The latest bestseller.
```

← New tokens enter

# LLM Context Window Sizes (2025)

Luckily, context window have gotten very large!

| Model | Context Window (Tokens) | Type |
|---|---|---|
| GPT-4.5 (OpenAI) | 128,000 | Proprietary |
| Claude 3.7 Sonnet (Anthropic) | 200,000 | Proprietary |
| Grok-3 (xAI) | 128,000 | Proprietary |
| Gemini 2.0 Pro (Google) | 2,000,000 | Proprietary |
| Llama 3.1 (Meta) | 128,000 | Open-Source |
| DeepSeek R1 | 131,072 | Open-Source |
| Mistral Large 2 (Mistral) | 32,768 | Open-Source |
| Gemma 2 (Google) | 8,192 | Open-Source |

Yet, it is unclear if LLMs pay equal attention
to everything in the context window!

# Implications for Use

- LLMs don't have memory, once trained they start anew with each new chat window*
  - If a chat is proving unhelpful, start over with a new chat (wipe the memory clean).
  - If you are using it for research - assume "one-shot" - it will not recall other chats and the context.
- Your prompt matters for output (for now).
- Training data if fixed after training and not updated - until a new model is trained.

# Implications for Use

- Training data limits LLM "knowledge"
  - LLMs usefulness correlates positively with representativeness in training data
  - Responsible use requires we know when LLMs are likely to return a good answer and when they are not.
  - You still need to think about that!
- LLMs have a hard time saying "I don't know" even when they are wrong.
- Hallucinations still exist but are becoming more rare.
  - Be aware where they will be more likely (where there is little training data)

# Extensions of LLMs

- Tool Use (Calculator, Python)
- Can search the internet and load pages in to context window.
  - Good for incorporating recent/niche information.
  - Unclear how "deep" the search is in many LLMs
- **Retrieval Augmented Generation (RAG)**
- **Chain of Thought / Reasoning**
- Agents - "Deep Research", computer use, operator
- Modalities - Audio/Image/Video
- Assistants - provide your own system prompts and fine tuned info.

# Retrieval Augmented Generation (RAGs)

- Upload your own documents to search
- Integrates external retrieval with generation
- Improves factual accuracy and relevance
- Search within large corpus of documents beyond 'keywords' with natural language
- Useful for summarizing a large amount of articles/docs
- Talk to research papers and textbooks.
- Create a podcast of a paper instead of reading!

Google NotebookLM Example (OPEN AI Assistants, )

# Chain of Thought & Reasoning Models

- Analogous to human intuition vs. thinking
- Models that generate intermediate "reasoning" steps
- It has a conversation with itself to determine...
  - What you're asking, and how it should proceed
  - Propose several avenues of inquiry and choose the best one
  - It refines your prompt itself by clarifying what you're asking.
  - Then breaks down request in to manageable chunks

# Chain of Thought & Reasoning Models

- Analogous to human intuition vs. thinking
- Models that generate intermediate "reasoning" steps
- It has a conversation with itself to determine...
  - What you're asking, and how it should proceed
  - Propose several avenues of inquiry and choose the best one
  - It refines your prompt itself by clarifying what you're asking.
  - Then breaks down request in to manageable chunks
- Enhances performance on complex problem solving
  - Great for statistical model help
- Potential for transparent decision-making processes

Terrance Tao on o1: *"The experience seemed roughly on par with trying to advise a mediocre, but not completely incompetent, static simulation of a graduate student."*

# Effective Prompting Techniques

- Craft clear, detailed prompts (spelling seems to matter)
- Break down large tasks into smaller ones
- Use iterative refinement for better results

- Ineffective: "What are the main causes of political polarization in the United States?"
- Effective: "Analyze the primary causes of political polarization in the United States. In your response, break down your analysis into distinct factors, explain how each contributes to polarization, and describe any interrelationships among these factors. Please provide your reasoning step by step. Draw on Academic Articles only for your response."

Ask the LLM for help with prompting!

# Brainstorming with LLMs

- Generate creative research questions and ideas
- Explore diverse perspectives quickly
- Collaborate with AI for innovative approaches

One fascinating biological concept to consider is **quorum sensing**, a mechanism used by bacteria to coordinate behavior based on population density.

---

**Quorum Sensing in Biology**

Bacteria communicate by releasing small signaling molecules—often called autoinducers—into their environment. As the bacterial population grows, the concentration of these molecules increases. When a critical threshold is reached, the bacteria collectively change their behavior, triggering processes like bioluminescence, biofilm formation, or virulence factor production. This threshold-dependent coordination allows the population to act in unison, even though each bacterium functions independently.

---

**Political Analogy: Mobilization Through Collective Signals**

The principle of quorum sensing can offer a fresh perspective on political mobilization and collective action:

- **Threshold for Action:** Just as bacteria must reach a certain concentration of signaling molecules to trigger a coordinated response, political movements often require a critical mass of support before large-scale change is possible. Political actors or social groups might remain dormant until the "signal"—be it public discontent, social media buzz, or grassroots activism—reaches a level that enables collective action.
- **Coordinated Response:** Once the threshold is reached, the movement can undergo a rapid transformation. This mirrors how bacteria switch behaviors once they sense enough of their peers, resulting in a synchronized response. In politics, this might translate into mass protests, sudden shifts in public opinion, or rapid policy changes.

Figure: "Take an obscure sub field of Biology at random. Next, try applying a concept from this subfield to political science to provide new insights on the political world"

One fascinating biological concept to consider is **quorum sensing**, a mechanism used by bacteria to coordinate behavior based on population density.

---

**Quorum Sensing in Biology**

Bacteria communicate by releasing small signaling molecules—often called autoinducers—into their environment. As the bacterial population grows, the concentration of these molecules increases. When a critical threshold is reached, the bacteria collectively change their behavior, triggering processes like bioluminescence, biofilm formation, or virulence factor production. This threshold-dependent coordination allows the population to act in unison, even though each bacterium functions independently.

---

**Political Analogy: Mobilization Through Collective Signals**

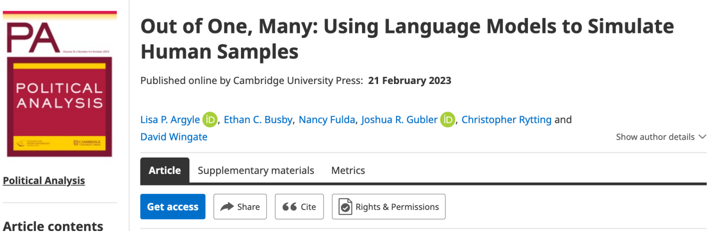
The principle of quorum sensing can offer a fresh perspective on political mobilization and collective action:

- **Threshold for Action:** Just as bacteria must reach a certain concentration of signaling molecules to trigger a coordinated response, political movements often require a critical mass of support before large-scale change is possible. Political actors or social groups might remain dormant until the "signal"—be it public discontent, social media buzz, or grassroots activism—reaches a level that enables collective action.
- **Coordinated Response:** Once the threshold is reached, the movement can undergo a rapid transformation. This mirrors how bacteria switch behaviors once they sense enough of their peers, resulting in a synchronized response. In politics, this might translate into mass protests, sudden shifts in public opinion, or rapid policy changes.

Figure: "Take an obscure sub field of Biology at random. Next, try applying a concept from this subfield to political science to provide new insights on the political world"

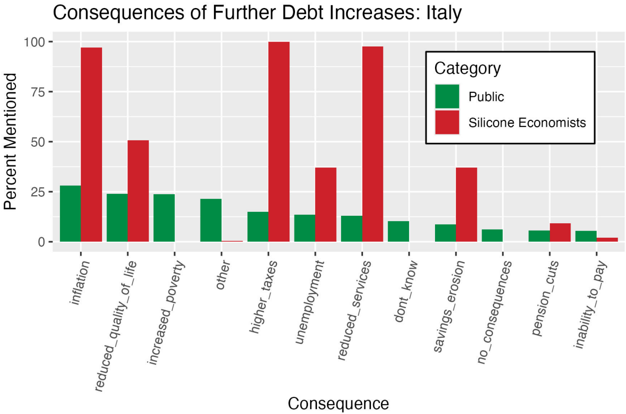
Figure: Fall of Ceaușescu

# Synthetic Samples in Research




- Ask LLMs to take on personas representative of the actual population and take a survey.
- Not VALID but useful for initialy exploring concepts (like in a thesis).
- Complement limited real-world data
- Considerations for validity and reliability

# Synthetic Experts in Research




Aspide et al. (2024) Putting the Public in Public Debt
LLMs can take on the role of Experts to serve as a baseline.

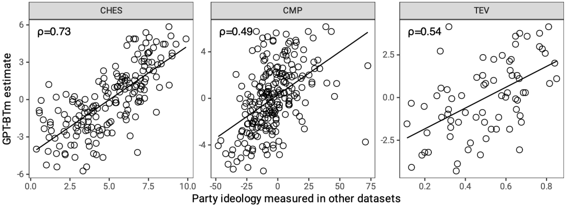
# Coding Qualitative Tweets & Topic Modeling

- Automate coding of qualitative text
  - Sentiment of tweets
  - Policy position (e.g. on a scale of 0-10)
  - Alignment with ideology
  - Political knowledge (LLMs are experts in many domains).
- Utilize topic modeling for exploratory analysis
  - How can I separate paragraphs/sentences into different topics
  - Which topics are most frequent/salient?
  - Process documents 1-by-1 asking for relevant topics
- Merge qualitative insights with quantitative analysis

# Latent Variable Construction

- Ask LLMs to compare two parties.
- Do this 1,000s of times and then estimate where they stand relative to each other. See the paper for methodology.
- Potential to measure concepts quickly (like in a thesis).

Figure 1: Benchmarking European Parties' Left-right Positions according to GPT-3.5 against experts, manifestos and opinion polls (Reference Year: 2009).




*Notes:* In each facet, we plot the left-right ideological positions of European parties obtained applying a Bradley-Terry model to the pairwise comparisons performed by GPT-3.5 (GPT-BTm), in reference year 2009. In each panel we employ different validation datasets, namely: Chapel Hill Expert Survey (CHES), Comparative Manifesto Project (CMP), True European Voter (TEV).

Figure: Di Leo et al. 2024: Mapping (A)Ideology: A Taxonomy of European Parties Using Generative LLMs as Zero-Shot Learners

# You can also do this with text!

DiGiuseppe and Flynn (2025)
LLM-Paired Comparisons Scaling Open-ended Survey Responses

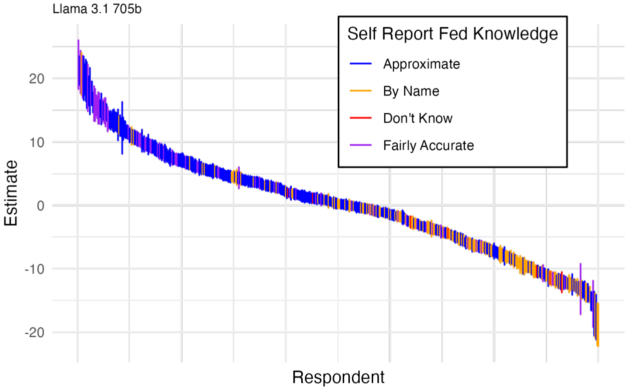



Figure: Estimate of Interest Rate Knowledge

# Using the API (Application Programming Interface)

In order to work well with LLMs in research API calls are necessary.

- Can be done in R, Python or any language.
- Looks harder than it is. Use LLMs for help.
- Requires an API key (don't share this with LLM or anyone) and a bit of money in an LLM account (or use a local LLM).
- Open weights models on Huggingface, Fireworks.ai.

Let's walk through how to do this in R with Open AI with some help from chat GPT

# Vibe Coding

"English is the hottest new programming language." - Andrej Karpathy

- Have it design for a poster in HTML from the airport
- Have it create a software to sort through PDFs, and rename them based on the content.
- Create a chat bot for university policy documents
- Create an app that breaks down expenses for a group vacation and splits the costs
- Create a CV updater that automatically updates your CV based on new events.

# Vibe Coding

Helpful Software:

- Cursor
- Windsurf
- or Just a frontier LLM (GPT o3, Claude 3.7, Deepseek)

Helpful Tips:

- Tell the LLM you are a novice - "Walk me through this like a Boomer coding for the first time"
- Break up tasks in parts. Share your code (Chat GPT can read your terminal window now)
- Share errors, ask for help - its an iterative process.

# API Usage and Ollama

If you're working w/ sensitive data, you can access local (smaller) LLMs for free on your own laptop!

Johannes Gruber's Ollama Demonstration

Prompt: *I have ollama installed on my computer. Help me write a program to call ollama from a browse window. I want to use the model: Gemma3:9b. I am new to programming. Explain how to implement this in VScode*

Questions?