



# Large Language Models

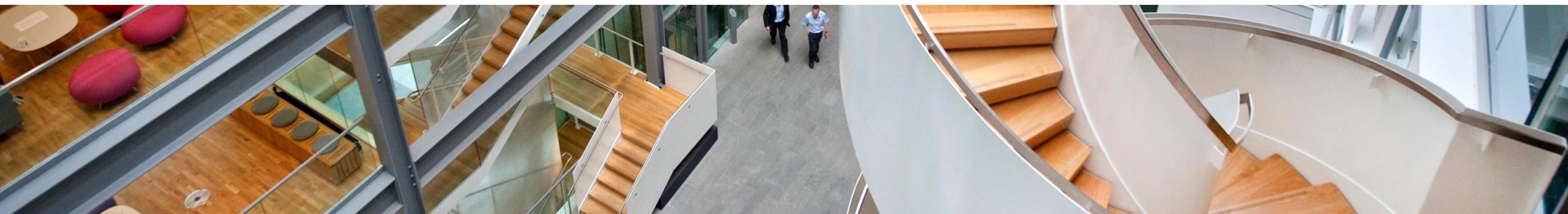
Tuğrulcan Elmas / Tj



THE UNIVERSITY of EDINBURGH  
**informatics**



THE UNIVERSITY of EDINBURGH  
School of Social  
& Political Science



# Challenge of The Course

- Understanding society with big data



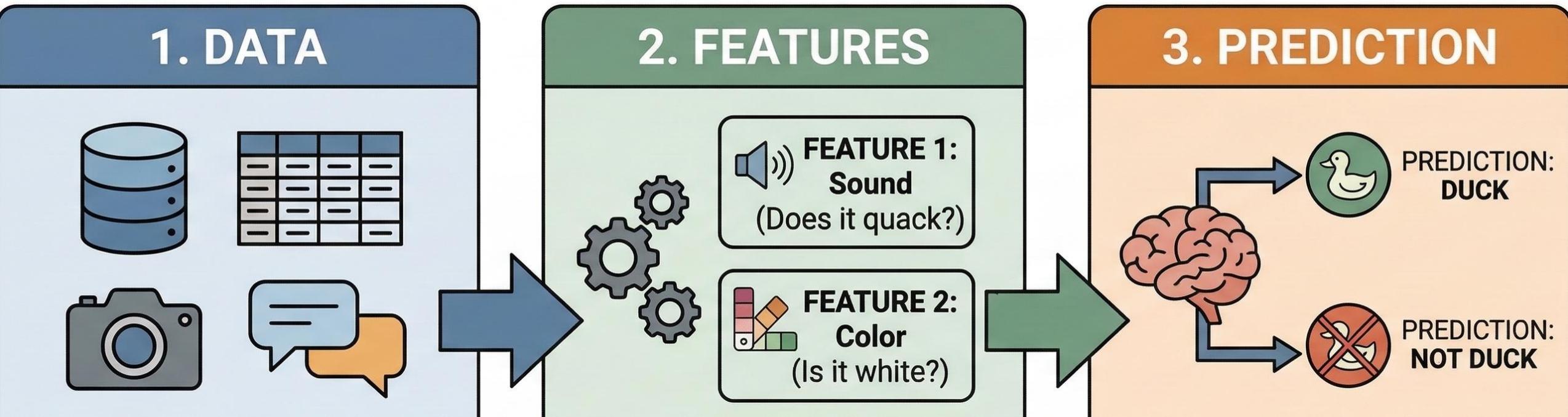
# Challenge of The Course

- ~~Understanding society with big data~~
- Let ChatGPT understand society with its big data



# Machine Learning Recap

- Machine learning = learning patterns from data to make predictions

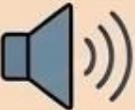
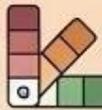


# Machine Learning Recap

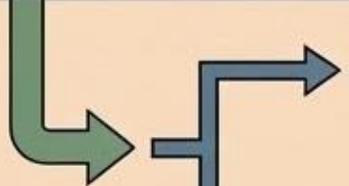
## 1. CLASSIFICATION: Discrete Categorical Output

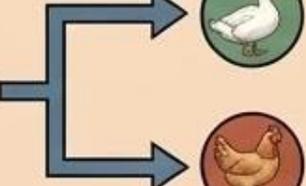
Goal: Assigning to specific categories.

**FEATURES USED (examples):**

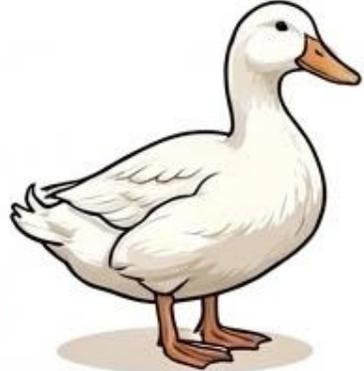
 **Sound:** Does it quack?       **Color:** Is it white?

## PREDICTION OUTPUT: DUCK/NOT DUCK

 **PREDICTION: DUCK**

 **PREDICTION: NOT DUCK**

Continuous output over a range

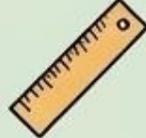


**INPUT:**  
Common Duck Data

## 2. REGRESSION: Continuous Numerical Output

Goal: Predicting a precise numerical quantity.

**FEATURES USED (examples):**

 **Weight (kg)**       **Wing Span (cm)**

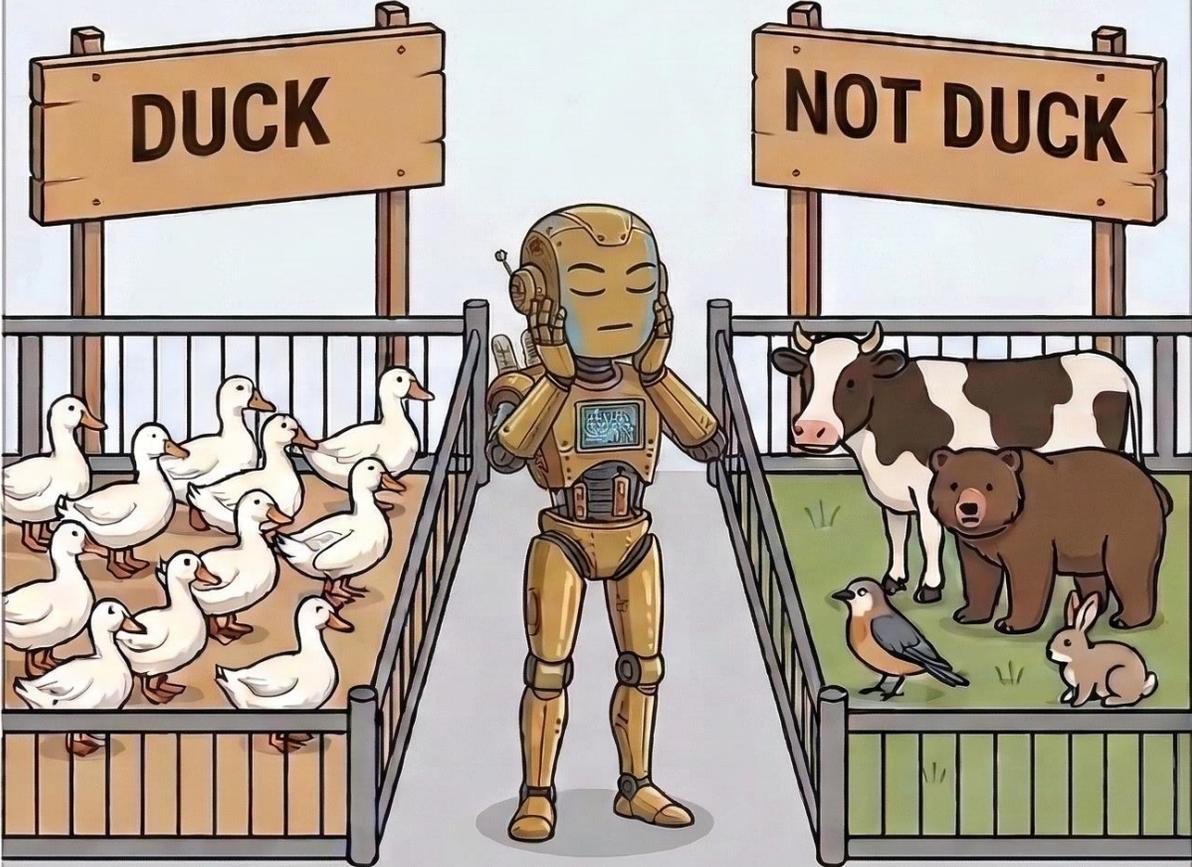
## PREDICTION OUTPUT: Duck Age

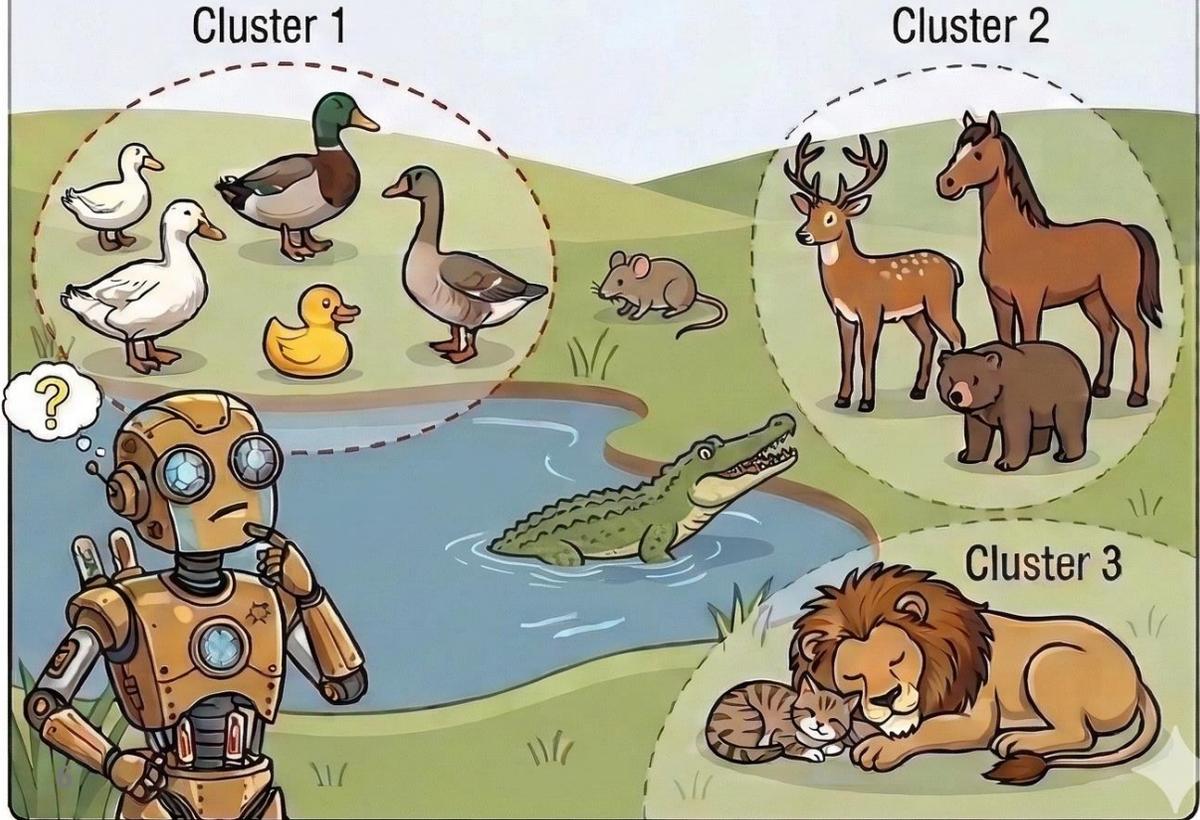
Continuous output over a range

# Machine Learning Recap

## 1. SUPERVISED LEARNING



## 2. UNSUPERVISED LEARNING

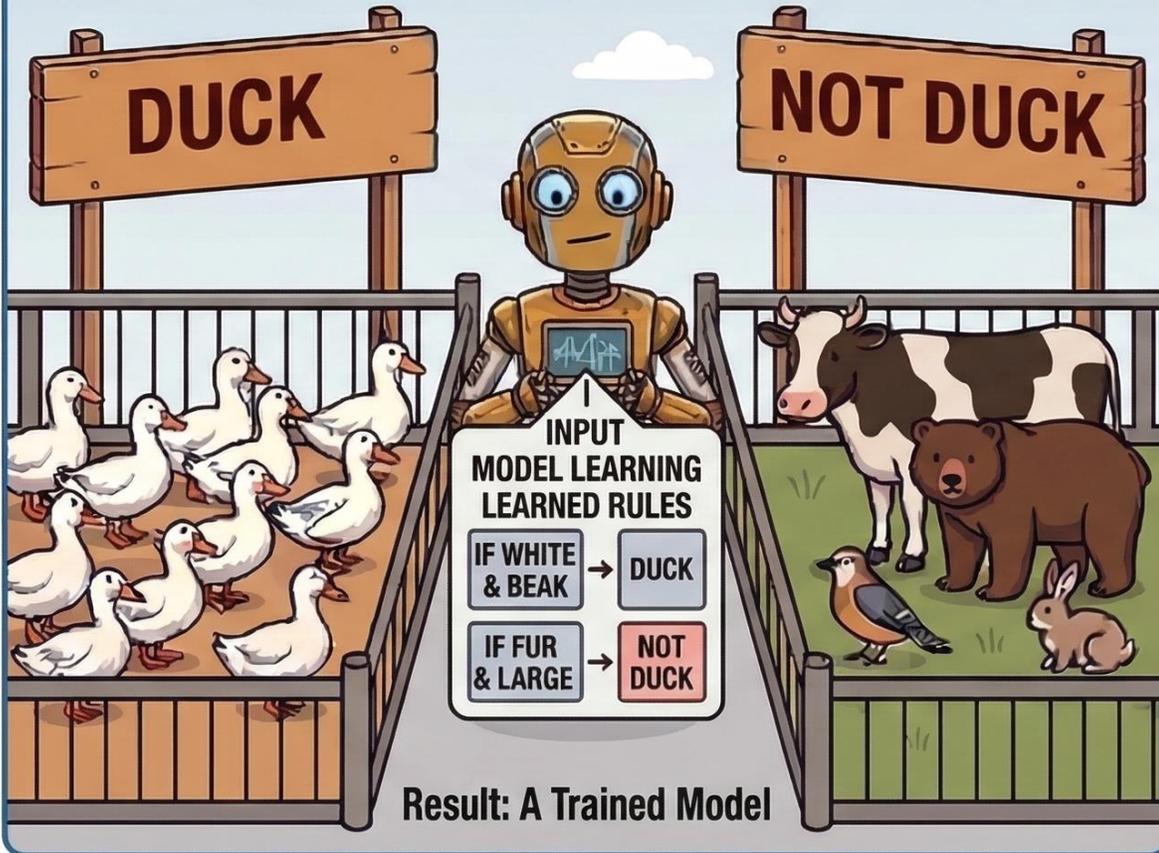


# Supervised Learning Recap

## Learning with Labeled Data

### 1. TRAINING (Building the Model)

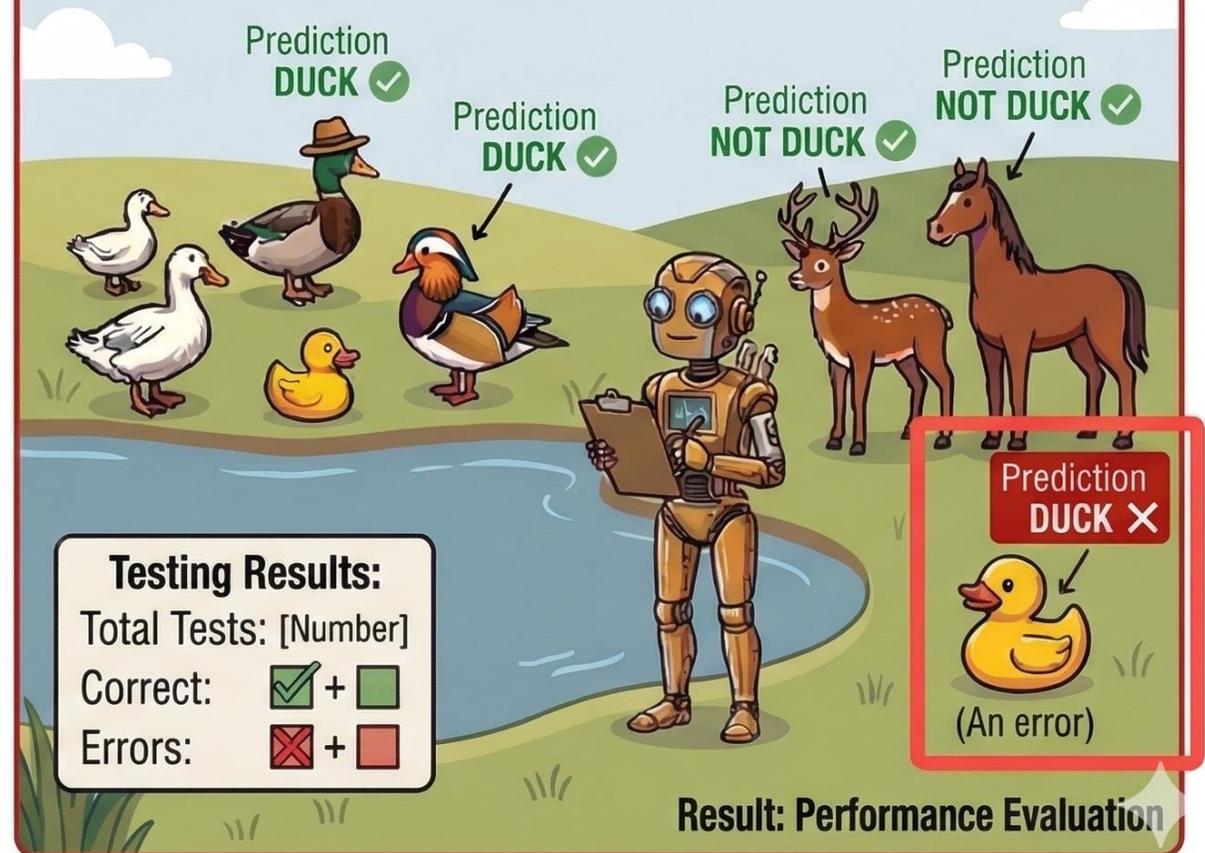
**Goal:** Model learns from a dataset with known answers.



## Evaluating with Unseen Data

### 2. TESTING (Evaluating the Model)

**Goal:** Model is tested on new data it has never seen.

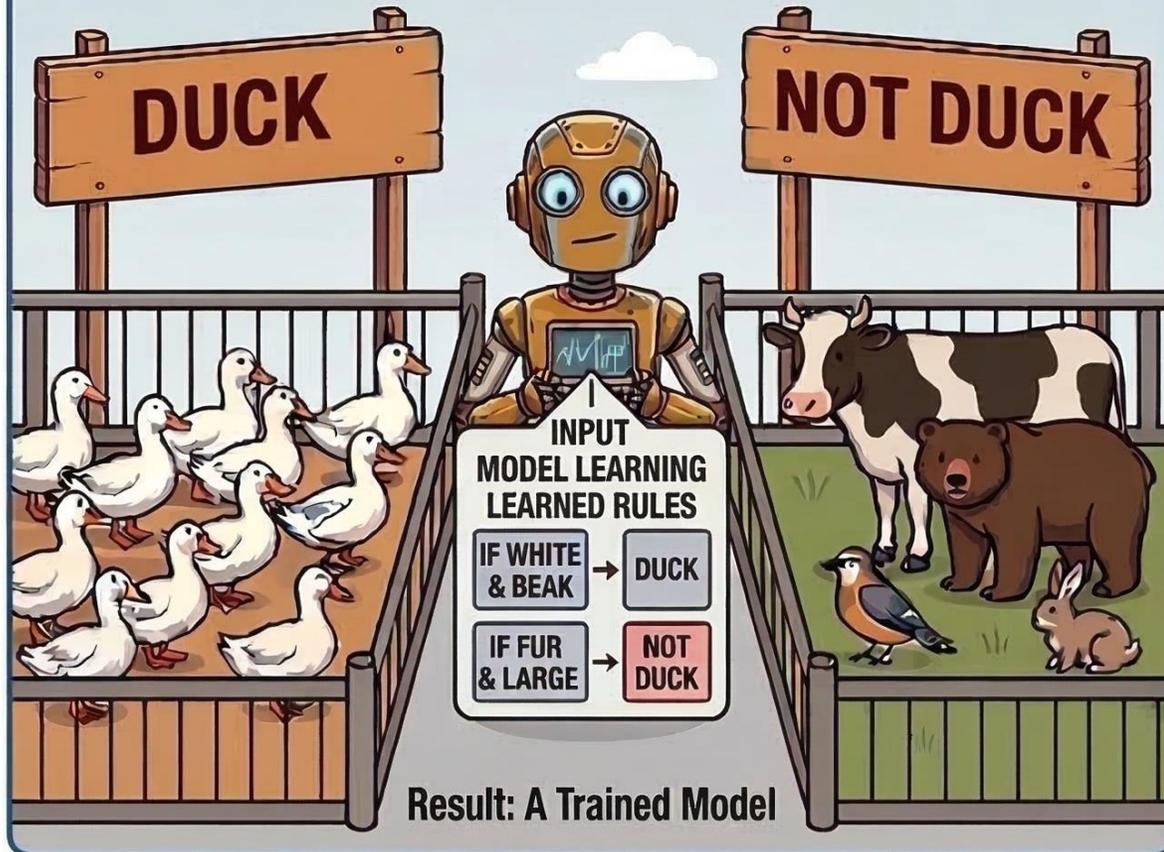


# Supervised Learning Recap

## Learning with Labeled Data

### 1. TRAINING (Building the Model)

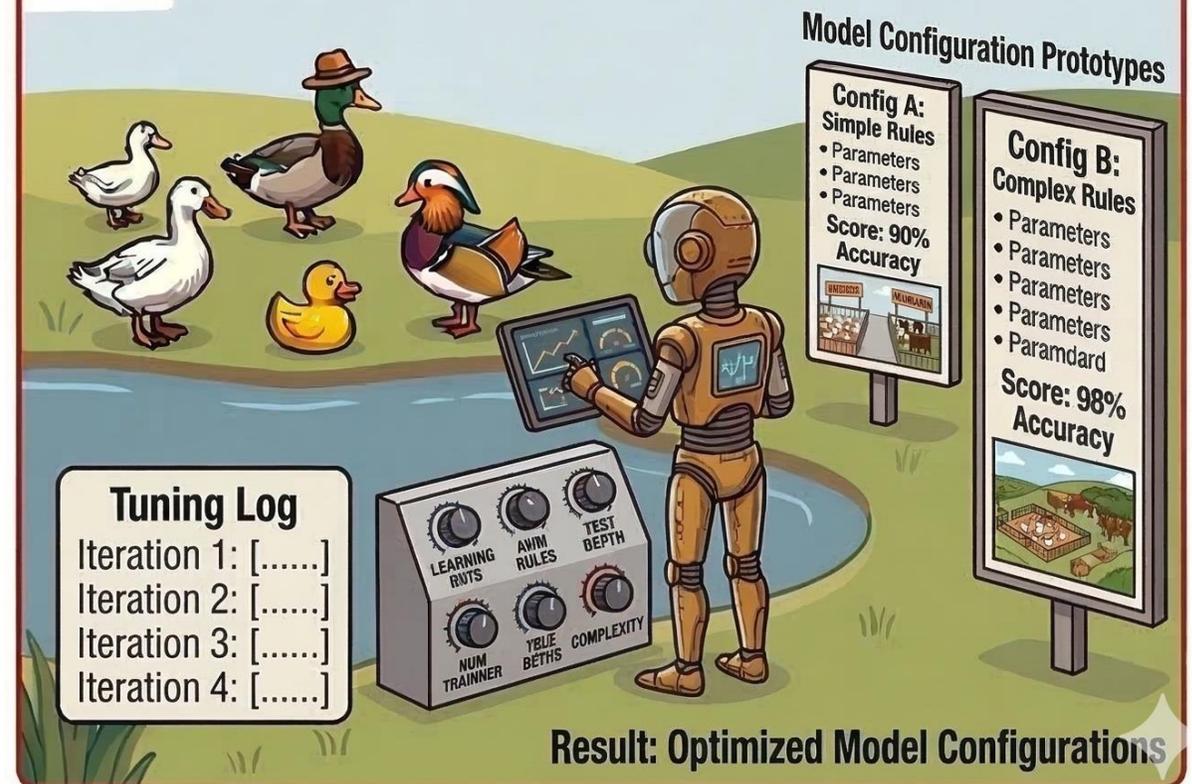
**Goal:** Model learns from a dataset with known answers.



## Iteratively Adjusting Settings

### 2. HYPERPARAMETER TUNING

**Goal:** Model is iteratively configured and evaluated to find the best settings.



# Text Classification

- Classify text data (a document) into classes
- Text => Features => Classification

Data	PREDICTION
"Hi Prof CSS is super cool!"	NOT SPAM
"I am Madagascar Prince I give you gold"	<b>SPAM</b>
"Here is your login code" from: facebook.com	NOT SPAM
"Here is your login code" from: asdfasdas.com	<b>SPAM</b>

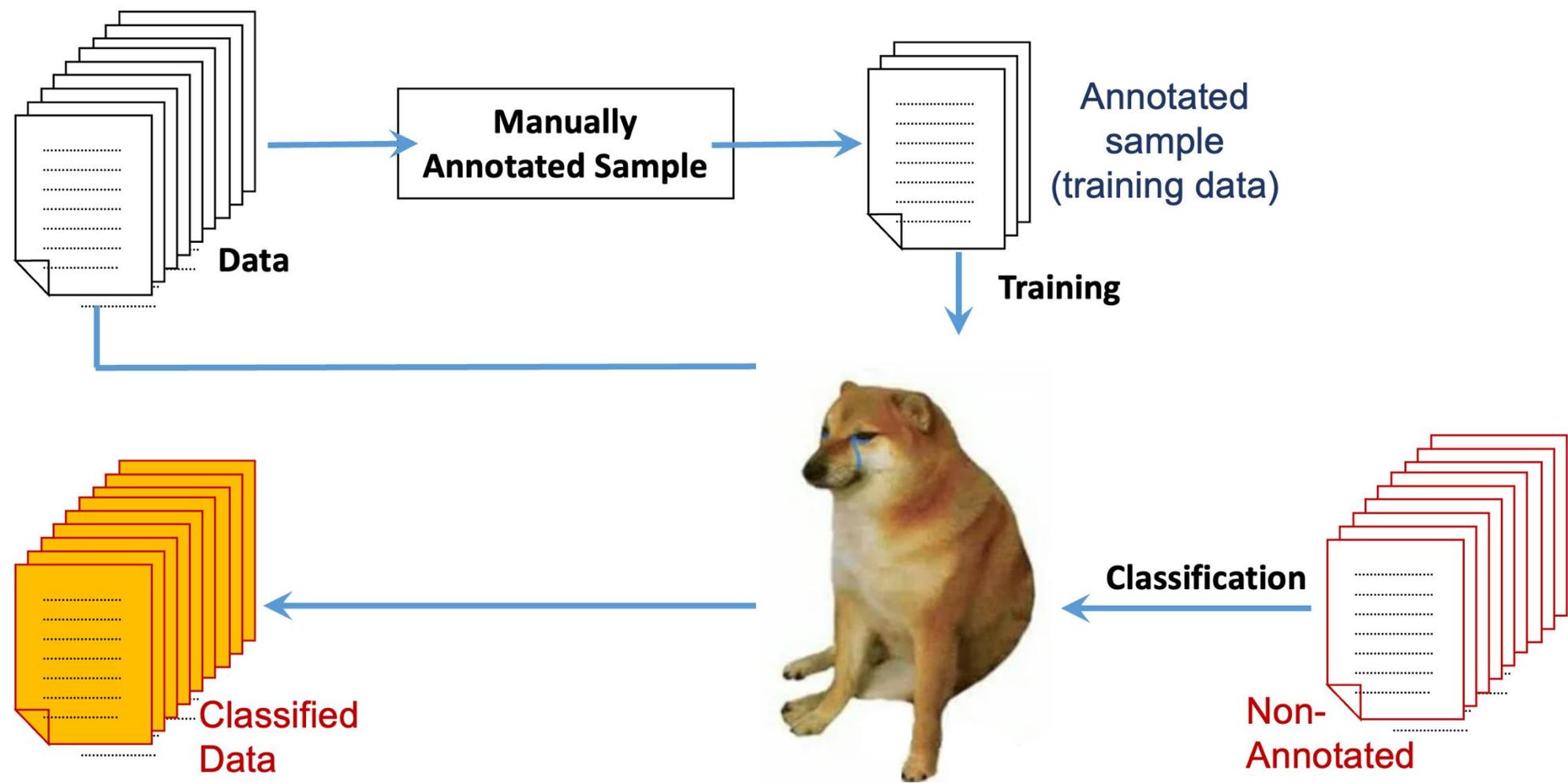


# Dictionary / Bag of Words

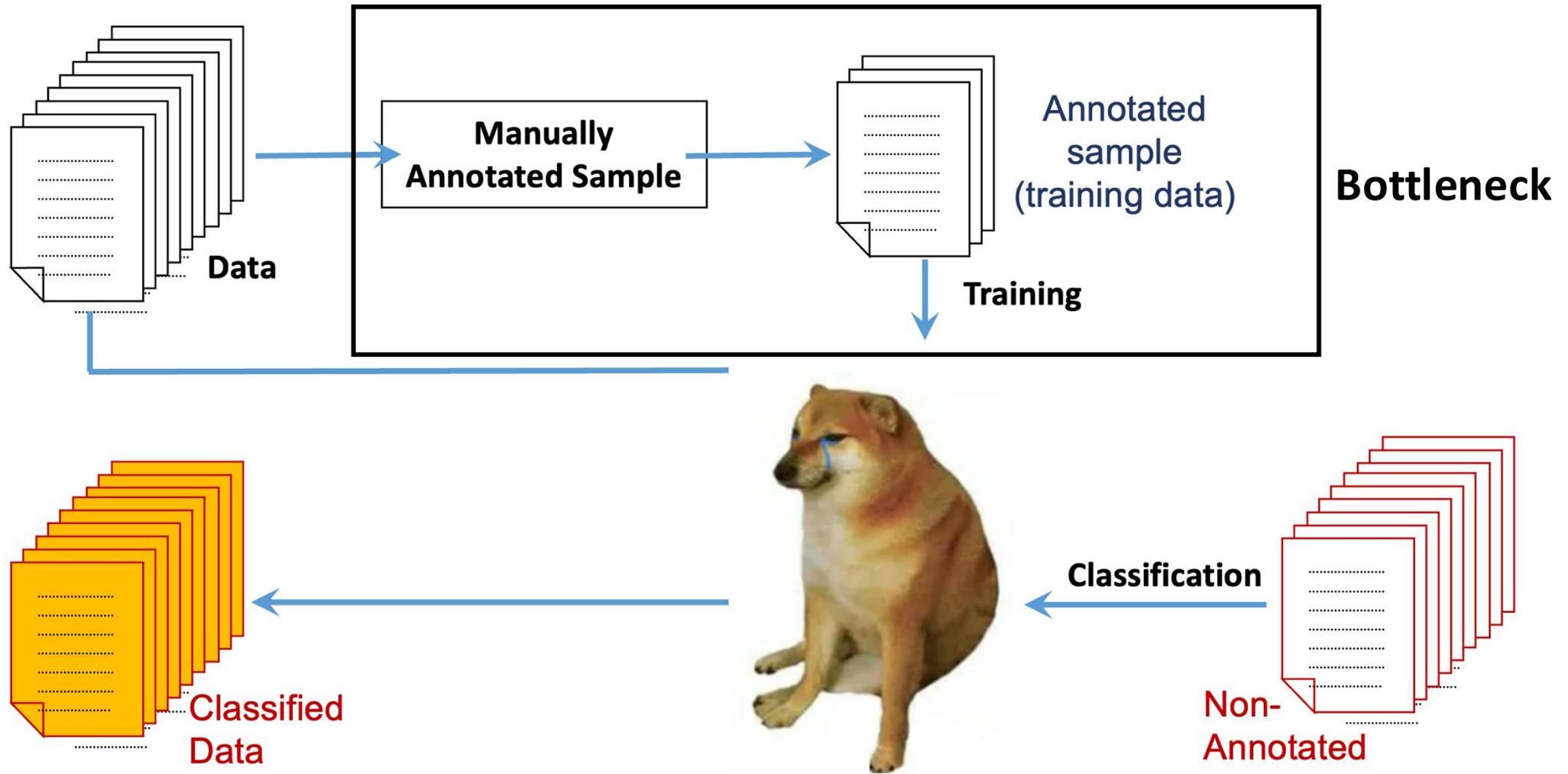
- Text Features... What's the problem here?

Data	Madagascar Prince	from:facebook.com	from:asdfasdas.com	...	PREDICTION
"Hi Prof CSS is super cool!"	0	0	0	0 0 0 0 0 0 0 0	NOT SPAM
"I am Madagascar Prince I give you gold"	<b>1</b>	0	0	0 0 0 0 0 0 0 0	<b>SPAM</b>
"Here is your login code" from: facebook.com	0	<b>1</b>	0	0 0 0 0 0 0 0 0	NOT SPAM
"Here is your login code" from: asdfasdas.com	0	0	<b>1</b>	0 0 0 0 0 0 0 0	<b>SPAM</b>

# Classic Text Classification with Supervised Learning

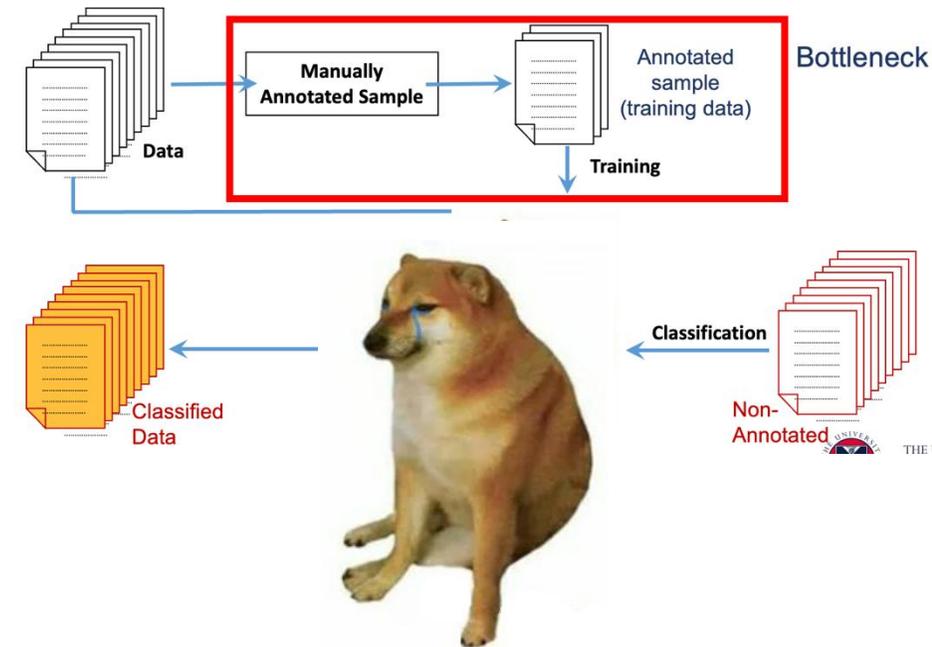
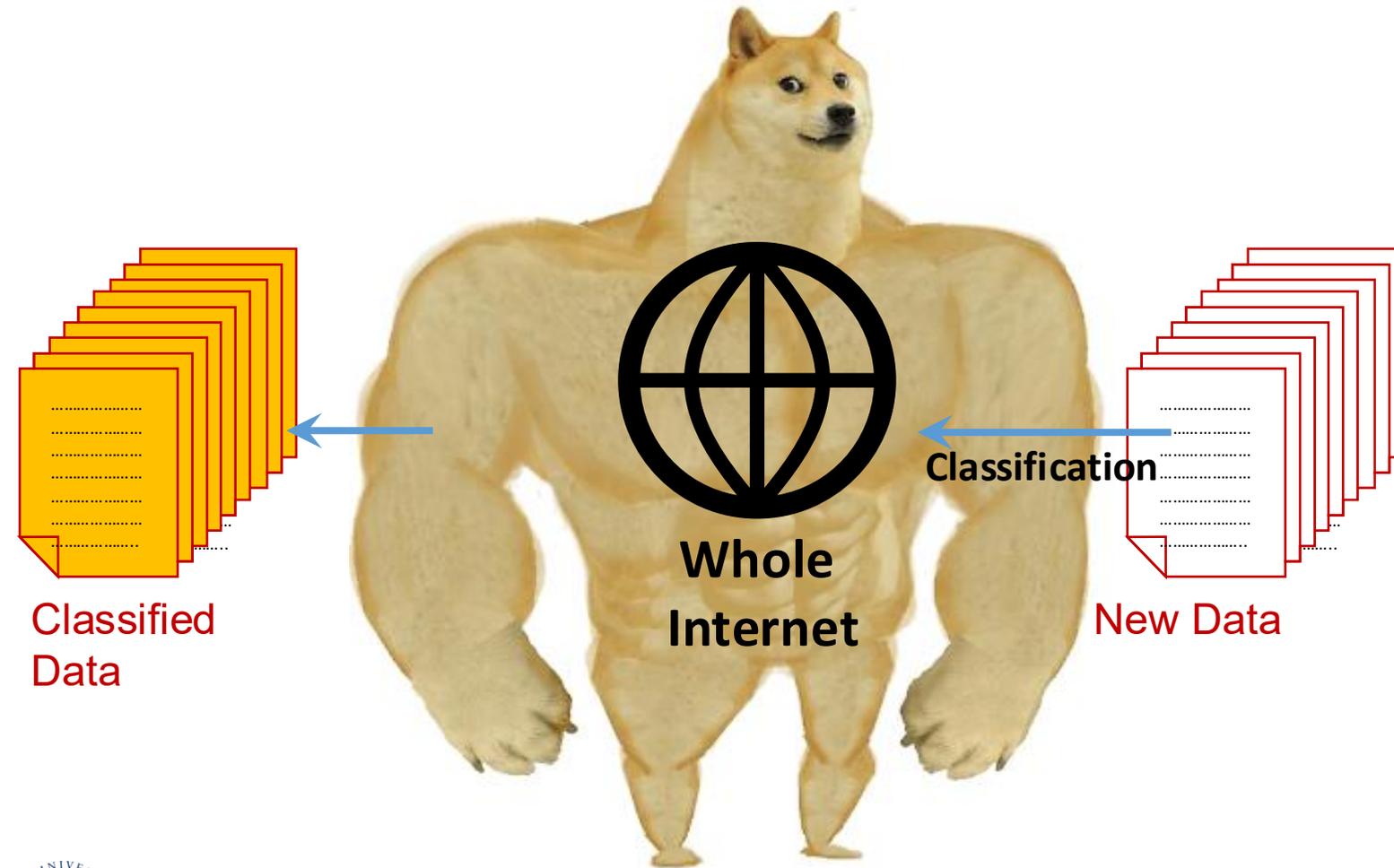


# Classic Text Classification with Supervised Learning



Your Dummy Classifier 😞

# Classic Language Models vs Large Language Models



# So how does ChatGPT answer me?

- “Hi ChatGPT how are you today?”
- “Fine thanks and you?”
  
- What is the data? Features?
- What is being predicted?
- Is it classification? regression?

# Text Generation Problem

- "Hello how are you?" → "Fine thanks and you?"
- Generation problem is a classification problem
- How do you "classify" a whole sentence?

# Text Generation Problem

- "Hello how are you?" → "Fine thanks and you?"
- Generation problem is a classification problem
- How do you "classify" a whole sentence?
- Answer: One word at a time!
  - Step 1 — **Input:** Hello how are you? → **Prediction:** Fine
  - Step 2 — **Input:** Hello...you? Fine → **Prediction:** thanks
  - Step 3 — **Input:** Hello...Fine thanks → **Prediction:** and
  - Step 4 — **Input:** Hello...thanks and → **Prediction:** you

# Text Generation

- Predict the next word given previous words
  - $P(\text{next word} \mid \text{past words})$
- “An answer” (sentence) is a chain of these predictions:
  - $P(\text{word2} \mid \text{word1}) \times P(\text{word3} \mid \text{word2, word1}) \times \dots$
- Autocomplete, summarization, translation, question answering, AI agents...

# Terminology

- Language model: Predicting how the text will continue
- "Large" Language Model LLM
  - Billions of parameters
  - Cannot work on a regular computer
  - Requires a lot of time and money
  - Very Effective!
- Token: unit of a text. word, subset of a word or a letter
  - Depends on how common the word is
  - Models have their own "tokenizers" (token creator)

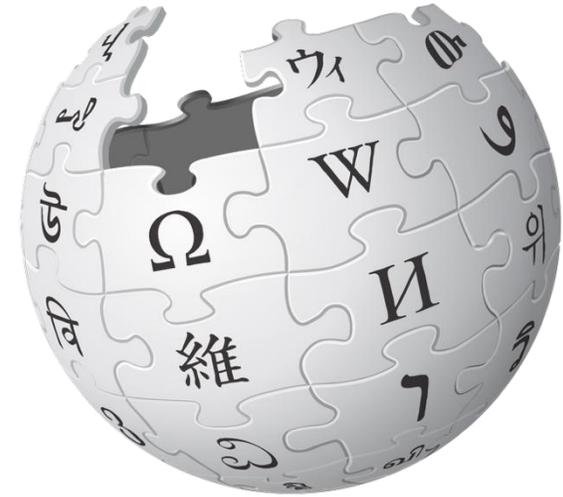
going underlying tugrulcan

# Questions?

- Important you understand up to this point

# How To Train LLMs?

- Supervised Learning
  - Need to label a dataset – too much work
- Self-supervised Learning
  - Automatically create labels from the data
  - What do we need then?
- **Big Data!**



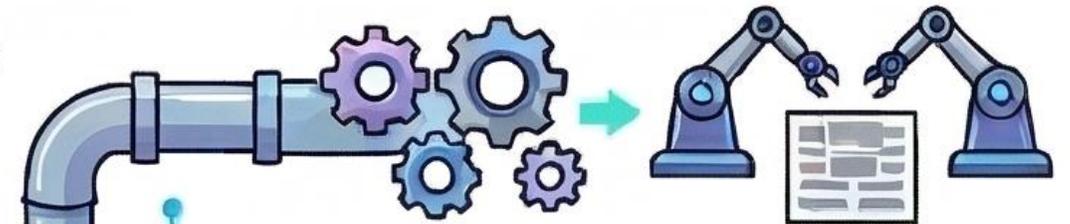
# Pretraining – Self-Supervised Learning

## 1 STEP 1: DOWNLOAD BIG DATA (UNLABELED)



## 2 STEP 2: CREATE TRAINING DATA (MASKING)

THE COMPUTER FAKES MISSING WORDS



RAW TEXT → MASKED TEXT

The computer comes with a mouse

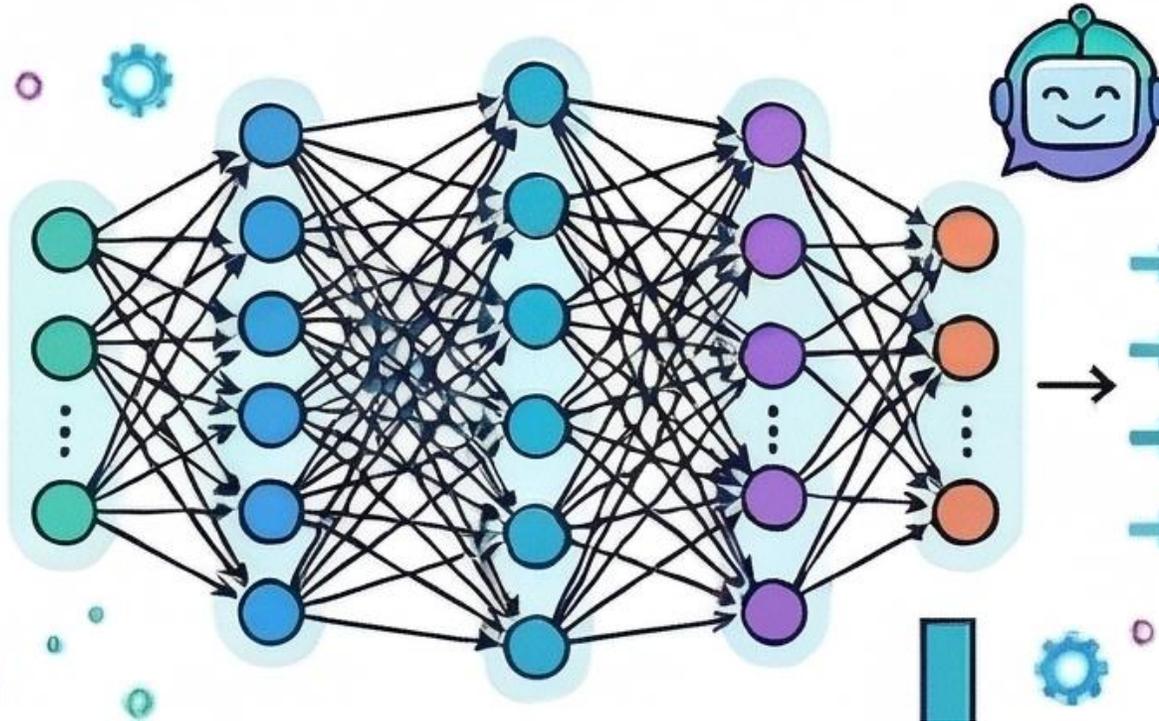


The computer comes with a [MASK]

# Pretraining – Self-Supervised Learning

## 3 STEP 3: TRAIN MODEL TO PREDICT MISSING WORDS

The computer comes with a [REDACTED]



- keyboard
- monitor
- **mouse**
- radio ❌

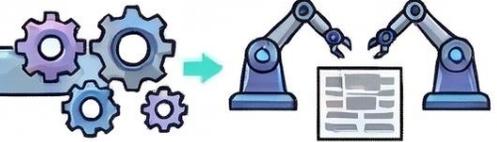
# LLM PRETRAINING: SELF-SUPERVISED LEARNING

## 1 STEP 1: DOWNLOAD BIG DATA (UNLABELED)



## 2 STEP 2: CREATING TRAINING DATA (MASKING)

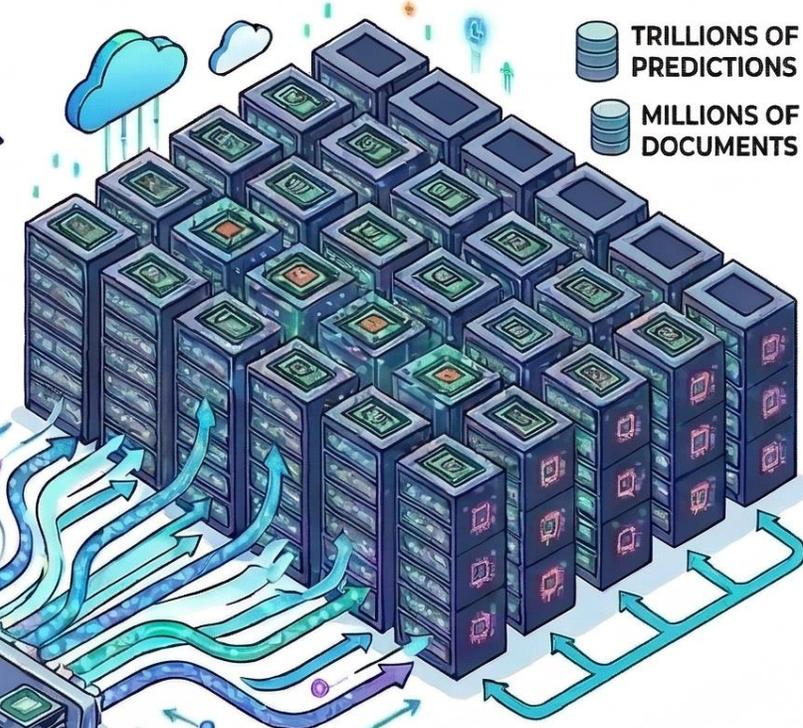
THE COMPUTER FAKES MISSING WORDS



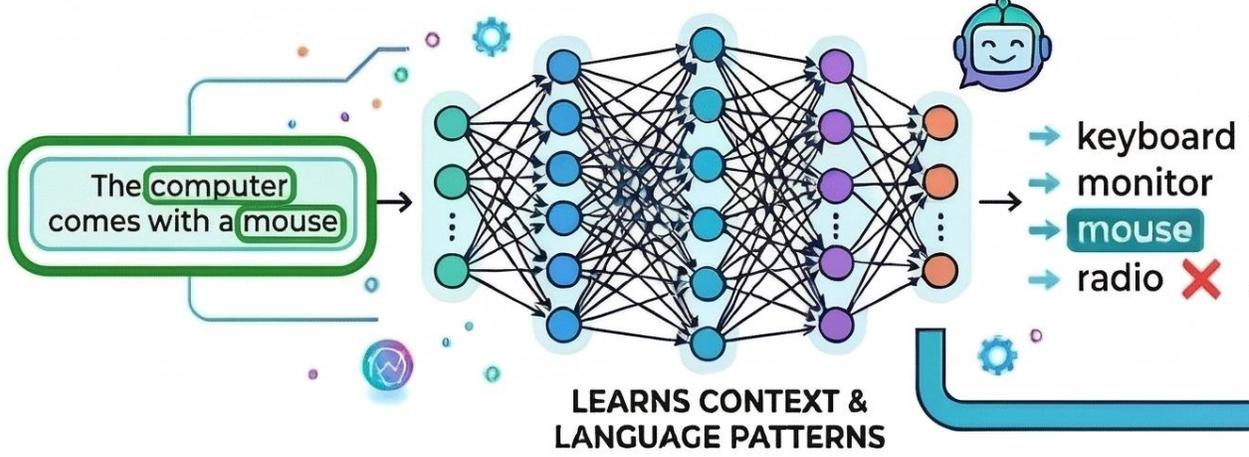
The computer comes with a mouse  
↓  
The computer comes with a [MASK]

THE COMPUTER FAKES MISSING WORDS

## 4 STEP 4: MASSIVE PARALLEL PROCESSING



## 3 STEP 3: TRAIN MODEL TO PREDICT MISSING WORDS



RUNNING ON GIGANTIC CLUSTERS



# What Happened to the Features?

Data	Madagascar Prince	from:facebook.com	from:asdfasdas.com	...	PREDICTION
"Hi Prof CSS is super cool!"	0	0	0	0 0 0 0 0 0 0 0	NOT SPAM
"I am Madagascar Prince I give you gold"	<b>1</b>	0	0	0 0 0 0 0 0 0 0	<b>SPAM</b>
"Here is your login code" from: facebook.com	0	<b>1</b>	0	0 0 0 0 0 0 0 0	NOT SPAM
"Here is your login code" from: asdfasdas.com	0	0	<b>1</b>	0 0 0 0 0 0 0 0	<b>SPAM</b>

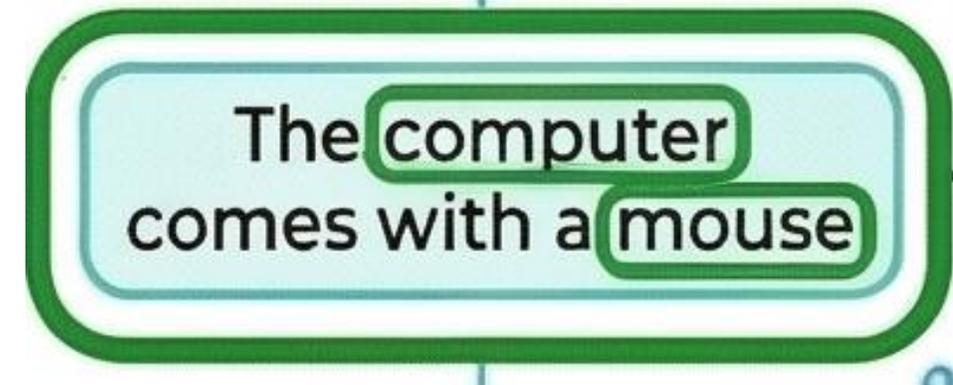
# Features

- LLMs learn parameters during pretraining
- LLMs use learned parameters to compute features for input data

Data	feature0	feature1	feature2	...	PREDICTION
“Hi Prof CSS is super cool!”	123123	914234	345345	...	NOT SPAM
“I am Madagascar Prince I give you gold”	234234	24234234	345345	...	SPAM
“Here is your login code” from: facebook.com	<b>42</b>	3245345	320948	...	NOT SPAM
“Here is your login code” from: asdfasdas.com	<b>42</b>	345345	240927	...	SPAM

# Terminology

- **Attention:** LLMs pay attention to the relevant part of the text while answering
- **Context:** the neighbourhood of a token
- Tokens have different meanings depending on the context
- "Mouse" near "computer" → device
- "Mouse" near "cat" → animal



# Same Word, Different Features

- LLM can compute features for a single token as well
- Same tokens, different context => different features

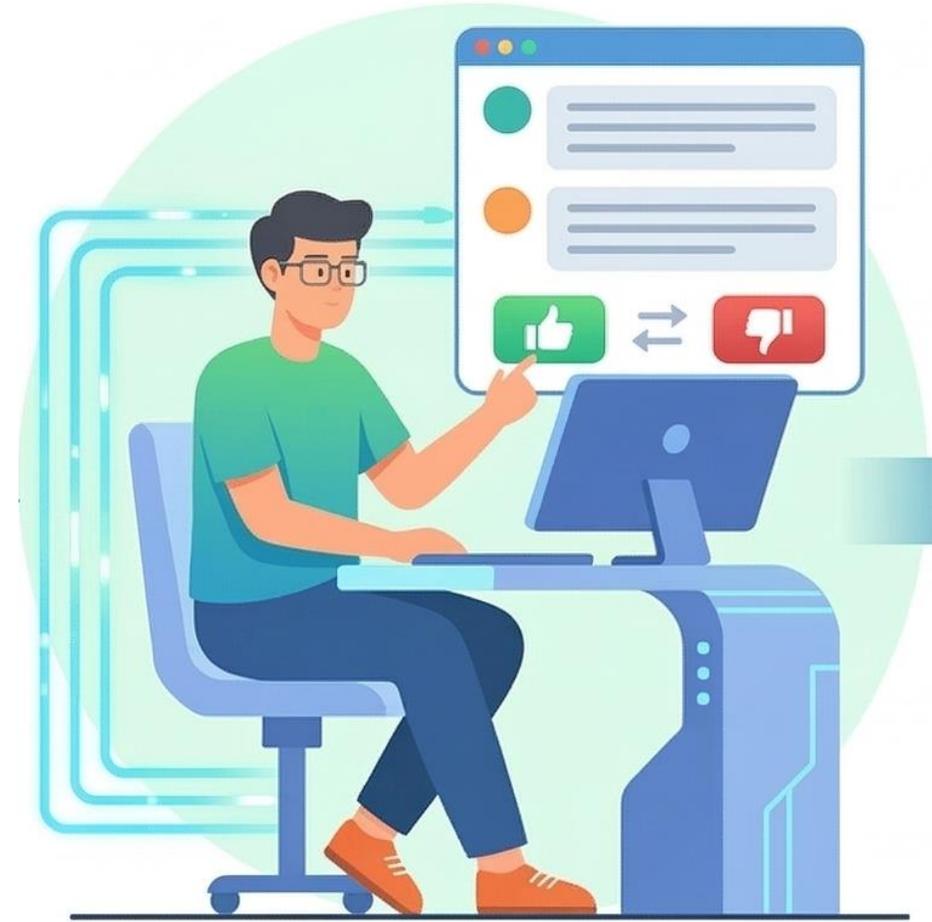
Data	Context	feature0	feature1	feature2	...
mouse	The computer comes with a <b>mouse</b>	123123	914234	345345	...
mouse	The cat is chasing a <b>mouse</b>	<b>90</b>	24234234	345345	...
mouse	mickey <b>mouse</b> talks to donald duck	<b>90</b>	3245345	320948	...

# Finetuning

- Take a pretrained model and continue training on a smaller, **specialised** dataset
- Supervised - Need Labels!
- (Question, Answer) pairs
  - Good for “LLM safety” (guardrails)
  - E.g, “ChatGPT give me atomic bomb recipe”, “No I can’t sorry
- Cheaper than pretraining
  - Done with less data and time
  - You can do it at home

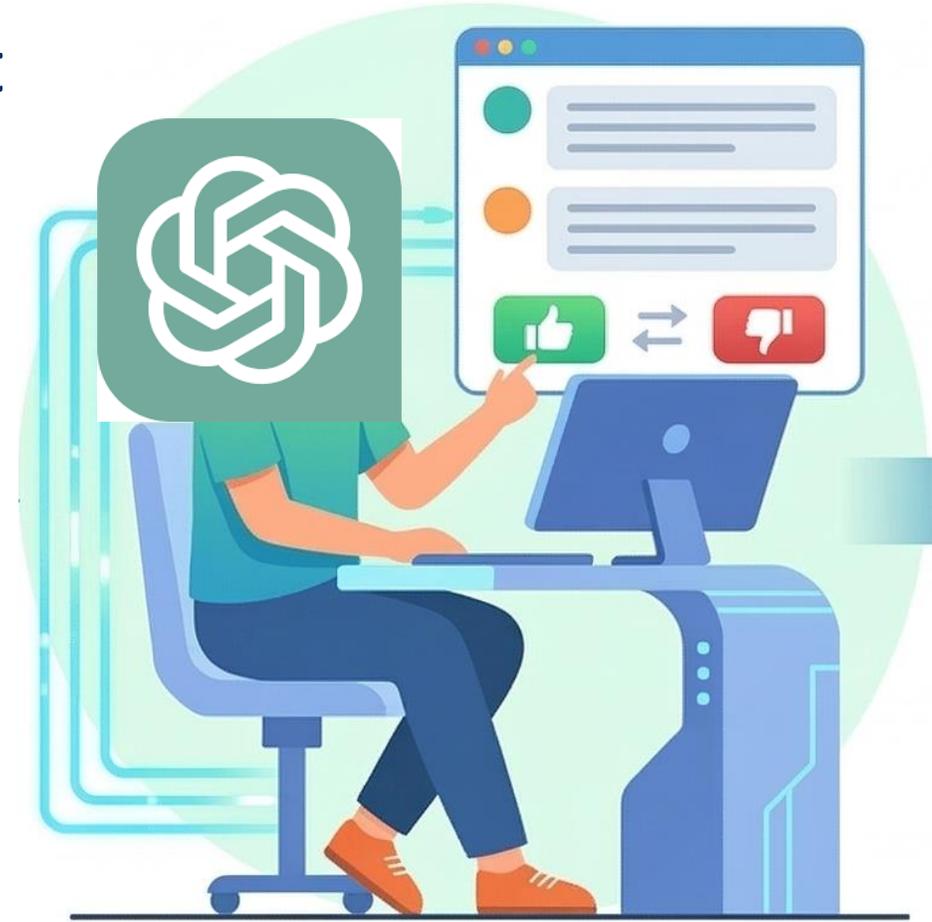
# Finetuning: RLHF

- Reinforcement Learning from Human Feedback
  - Humans rank model outputs
  - A model learns their preferences
  - Makes ChatGPT "helpful and harmless"



# Distillation

- A teacher (bigger) model trains a student (smaller) model
- Student model trained on teacher model outputs
- GPT 5 -> GPT mini



# Training on Another Model's Output

- Create synthetic data by using another model
- (Question, ChatGPT answers)
- Pretraining, Finetuning, Distillation
- "DeepSeek is training on ChatGPT output"



# Model Size

- Model Size = Number of Parameters
- Small LLM: ~7 billion parameters (e.g., Llama2-7b)
  - Your computer may be able to run it
- Medium LLM: ~100 billion parameters, e.g.,
  - Llama 3.3 70b : ChatGPT 3.5 Level Performance
  - FREE & UNLIMITED ACCESS TO the University of Edinburgh people yaaaaaay
- Large LLM: ~300B+ parameters
  - Frontier models: GPT 5, Gemini 3, Claude Opus 4.6...(Undisclosed)

# Questions?



# Terminology

- Prompt: The initial input
- System prompt: The initial prompt given by the system
  - "You are ChatGPT. You are a helpful assistant. We are in 2026"
  - Often unseen to the user
- User prompt: The initial prompt by the user

# Context Window

- The longest text the LLM can take as input
  - Technical: the maximum number of tokens the model can process at once
- Whole chat occupies the context window
  - System Prompt + User Prompt + Assistant Answer + User Answer + ...
- Larger for larger LLMs (typically)
- Can be compressed (e.g., replace a part with a summary)
- LLM Performance degrades with long text even if it fits the context window

# “In Context Learning”

- Do classification by prompting the LLM
- “ChatGPT is this email SPAM or NOT SPAM?”
- “Grok is this TRUE?”
- Performance varies according to your prompt
- No training in the classical sense

# Prompt Engineering

Instead of simply “ChatGPT classify this email pls...”

- Provide instructions, background, examples
  - More clues for the LLM to answer
- Guide an LLM how to perform a new task
  - Tell them their role
  - Describe the problem
  - One Shot or Few Shot Learning (“Here is one / a few examples”)
  - Reasoning (“Answer step by step”)
  - Provide additional documents (RAG)

# Zero Shot, One Shot, Few Shot

Prompt: “You are an email classifier. Classify each email I give you into two categories: {SPAM, NOT SPAM} ...

**Zero shot:** No examples

**One shot:** “... Here are two examples:

“I am Madagascar Prince: {SPAM}”

“CSS is cool: {NOT SPAM}”

**Few shot:** Multiple examples

# Data Classification

Prompt: “You are an email classifier. Classify each email I give you into two categories: {SPAM, NOT SPAM} ...

Example: “I am Madagascar Prince: SPAM”

Data: \*Attaches **1 million** emails\*

**Does this work?**

# Data Classification

Prompt: “You are an email classifier. Classify each email I give you into two categories: {SPAM, NOT SPAM} ...

Example: “I am Madagascar Prince: SPAM”

Data: \*Attaches **1** email\*

## What is the problem?

# Batch Classification

Prompt: “You are an email classifier. Classify each email I give you into two categories: {SPAM, NOT SPAM} ...

Example: “I am Madagascar Prince: SPAM”

Data: \*Attaches **100** emails\*

Nice middle way... What are cons?

# Constraints

After introducing the problem and examples...

- Ask LLM to give the answers in JSON Format
- Machine Readable
  - can be used by Excel, PowerBI, any other tool
- Other instructions related to format e.g., one label per data, put the index of the data point to the json object etc.
- Named "constraints"

# What Text Classification Tasks?

- **Sentiment analysis:** very easy, any model works
  - E.g., ChatGPT is this email happy
- **Domain dependent tasks:** experiment with fewshot
  - **Hate speech:** depends on the data, culture, law
- **Topics:** LLMs are better if data is small
  - Otherwise try both traditional topic modeling and LLMs
- **Not human solvable** -> Most likely not LLM solvable
  - You should be able to classify by yourself!

# What To Use For Classification

- **Small Datasets:** Just use ChatGPT / Gemini app / website
  - Or [elm.edina.ac.uk](http://elm.edina.ac.uk)
- **Medium Datasets:** Automate above (agents handle it)
  - Codex, Gemini-CLI, Claude Code
- **Large Datasets (10k+ rows):** Reconsider
  - Can you work with a sample instead?

# Classifying Large Datasets

Try Free / Small Models First – No need to pay!

- Many small / distilled model excel in easy tasks
  - E.g., sentiment analysis
- 2B-7B models can (probably) run in your computer
- [elm.edina.ac.uk](http://elm.edina.ac.uk) provides free access to Llama 70B

Otherwise use OpenAI / Gemini / Claude API

# How To Experiment for Text Classification

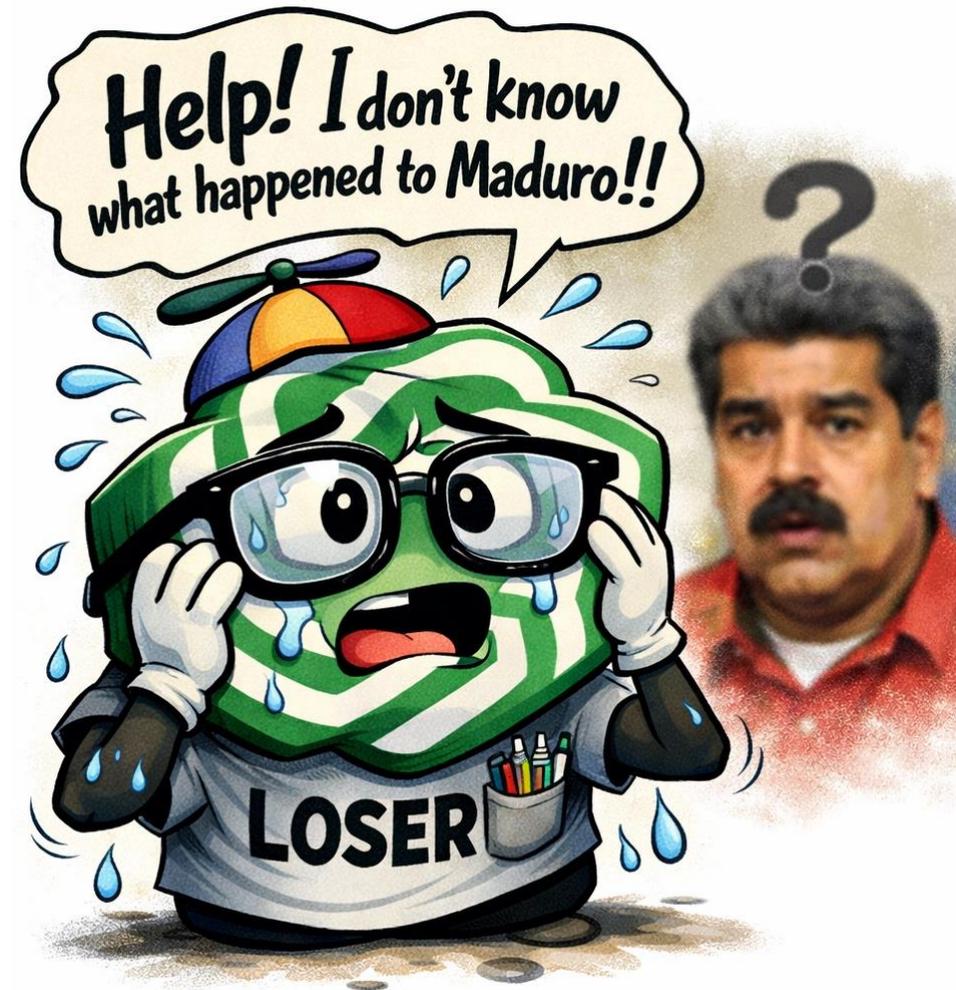
- Randomly sample and annotate a small dataset
- Try it on ChatGPT / Gemini / elm.edina.ac.uk
- If getting %100 accuracy skip experiments – task is easy

Else:

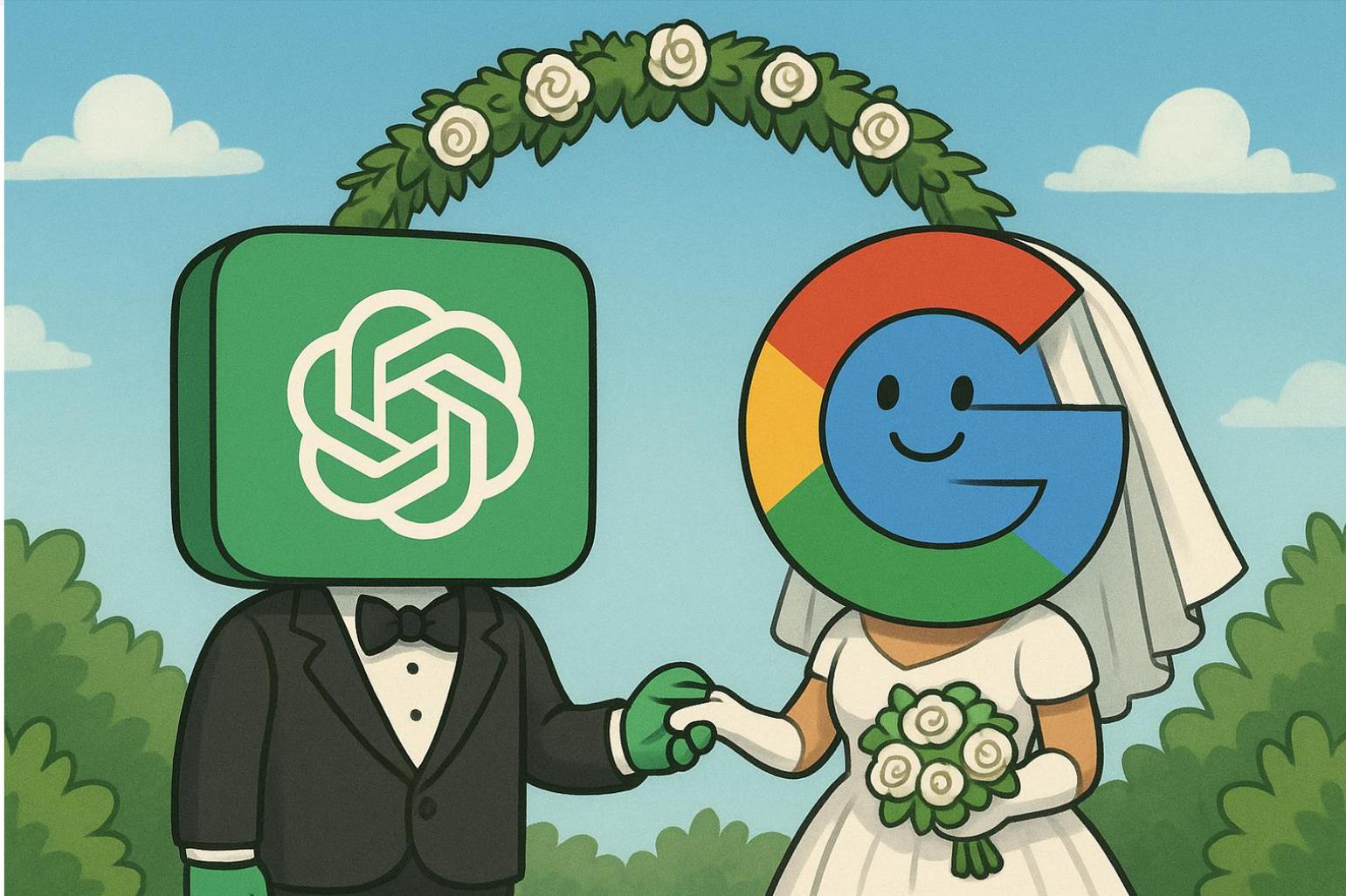
- Try different models and different prompts
  - Same idea as “Hyperparameter Tuning” -> Model selection / Prompt tuning
- Annotate another dataset (a test set)
- Test if the model & prompt generalizes

# How to Update an LLM?

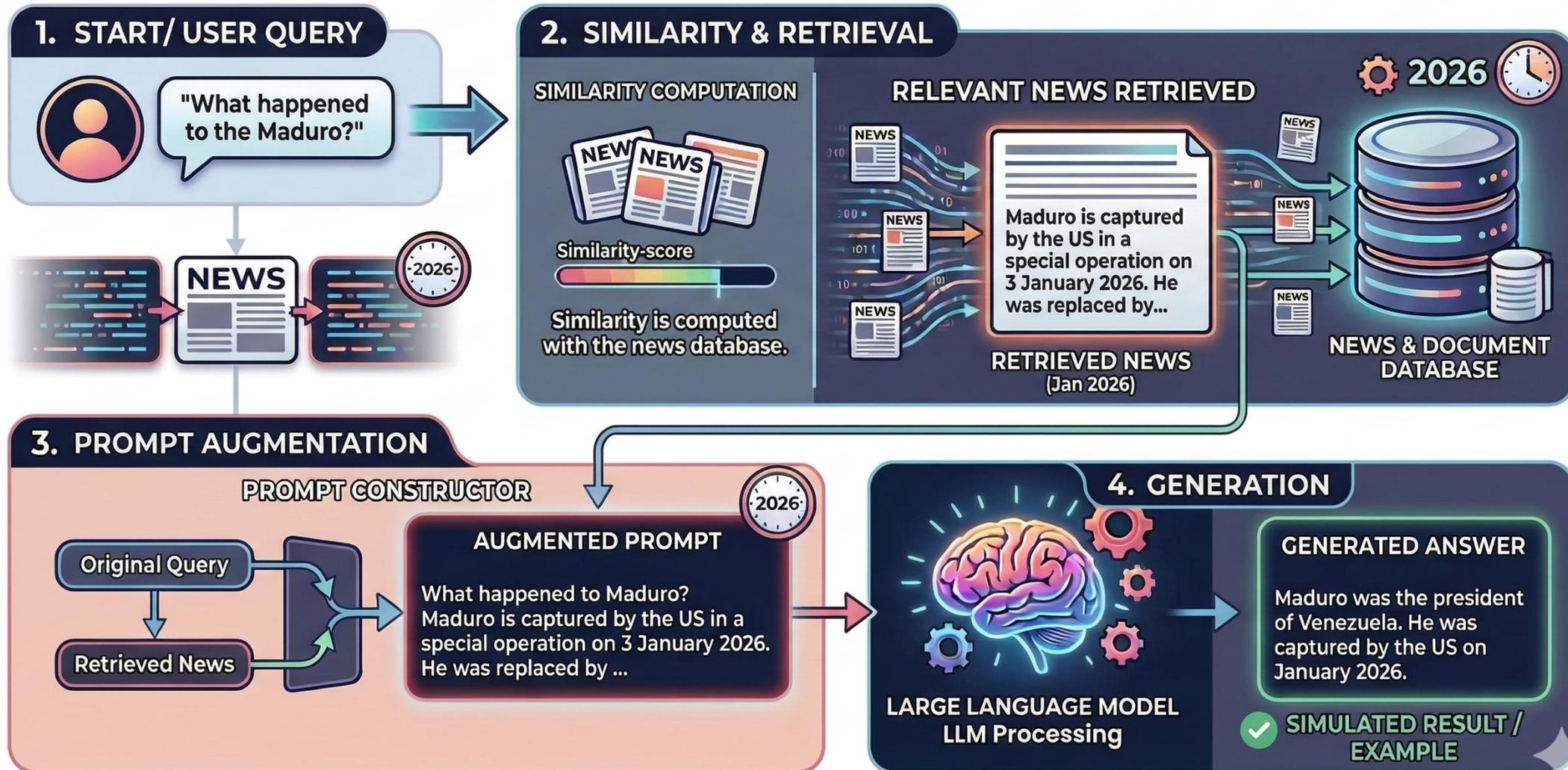
- ChatGPT is trained on Big Data
- "Big Data" gets obsolete quick
  - Trump wages war to a new country everyday
- What to do? More training?
- Retrieval Augmented Generation
  - Retrieve new information
  - Augment the prompt
  - Generate using the augmented prompt



# ChatGPT + Google = RAG?



# Retrieval Augmented Generation



# Google NotebookLM

- Answers are "grounded" in source -> RAG

The screenshot displays the Google NotebookLM interface. At the top, the notebook title is "Global Warming and Climate Change Debates" with a "Shared" status. A "Create notebook" button and various utility icons are visible in the top right. The interface is split into two main sections: "Sources" on the left and "Chat" on the right.

**Sources:** A file named "file-001.txt" is listed. Below it is a "Source guide" section with two highlighted text blocks:

- "and unprecedented action to avoid catastrophic climate change. 91 scientists from 40 countries have compiled this landmark study for the united nations and it has found that we have just over 11 years to get climate change under control in order to avoid catastrophic flooding."
- "aisle, i think about climate change, climate change terrifies me, as a mom. the fact that the republican party isn't taking it seriously as is actually denying that it is happening is shocking. it is horrifying. it is extremely concerning. what grade would you give"

Below the highlights, there is a paragraph of text: "he said he never talked to him about climate change about that you donald trump has an affinity for science and he know there is are arguments on both sides. the reality is there isn't really. scientists agree about climate change. let go to the people he doubts."

**Chat:** The chat window contains a qualitative analysis of the provided text, describing it as "deeply polarized discourse surrounding climate change, characterized by conflicting narratives regarding scientific validity, economic impact, and political identity." It also includes a section titled "1. The Politicization of Science and Consensus" and two bullet points:

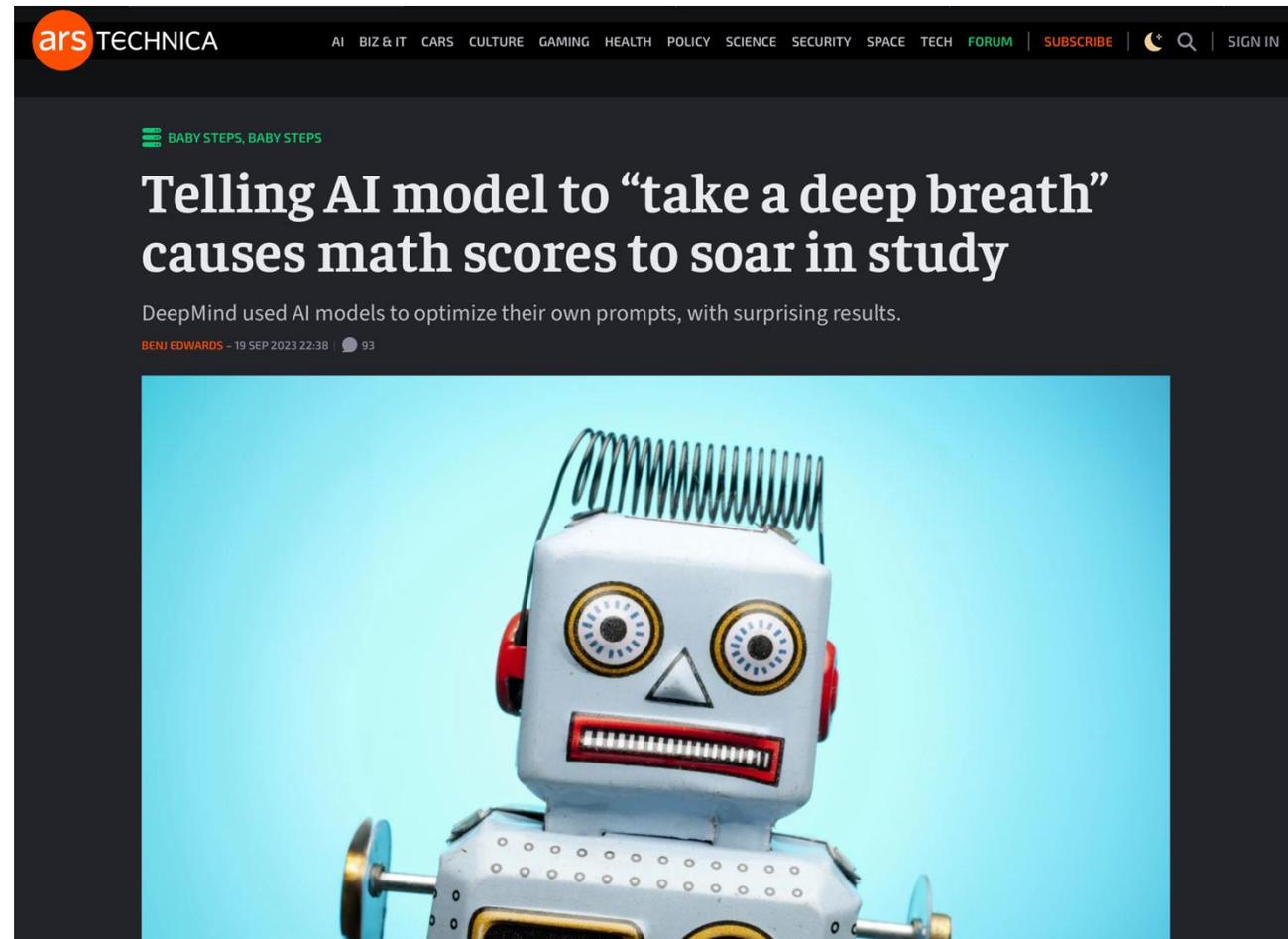
- The "Settled Science" Narrative:** Proponents frequently cite a "97% consensus" among scientists to argue that the debate is over and action is mandatory (1, 2, 3). Figures like Al Gore and John Kerry frame climate change as an existential threat comparable to nuclear weapons, requiring immediate global mobilization (4, 5).
- The "Hoax" and Skepticism Narrative:** Conversely, a significant portion of the text is dedicated to challenging this consensus. Skeptics, often identified as political conservatives or specific media personalities, frame global warming as a "hoax," a "hoax," or "pseudo-science" (6, 7, 8). The "Climategate" scandal (involving hacked emails from the

At the bottom of the chat window, there is a "Start typing..." input field and a "20 sources" indicator with a right-pointing arrow.

NotebookLM can be inaccurate; please double-check its responses.

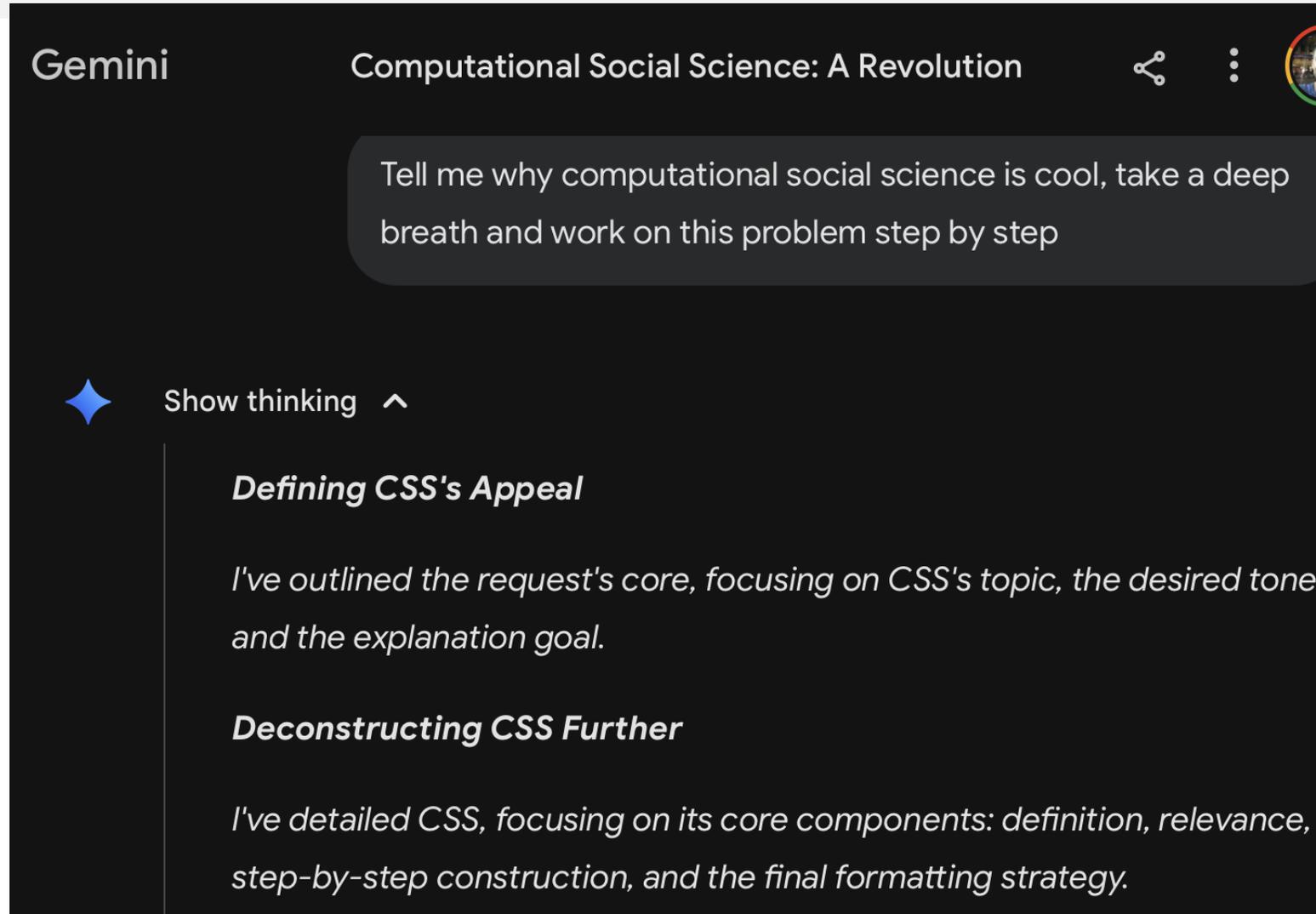
# Reasoning

- Take a deep breath and work on this problem step by step



# Reasoning

- Instructs the LLM to generate intermediary steps
- LLM enhances prompt by itself
- Performance increases
- Models naturally support this



The screenshot shows the Gemini AI chat interface. At the top, it says 'Gemini' and 'Computational Social Science: A Revolution'. A user prompt asks to explain why computational social science is cool. Below the prompt, a 'Show thinking' button is visible, which has been clicked to reveal a 'thinking' process. This process is divided into two sections: 'Defining CSS's Appeal' and 'Deconstructing CSS Further', each with a brief explanation of the approach.

Gemini Computational Social Science: A Revolution

Tell me why computational social science is cool, take a deep breath and work on this problem step by step

◆ Show thinking ^

**Defining CSS's Appeal**

*I've outlined the request's core, focusing on CSS's topic, the desired tone and the explanation goal.*

**Deconstructing CSS Further**

*I've detailed CSS, focusing on its core components: definition, relevance, step-by-step construction, and the final formatting strategy.*

# AI Agents

- LLMs predict the next action, not just answer
  - Browse the web, run code
  - "Prescriptive" machine learning
- LLM executes the action (if system permits)
- Loop: Observe → Think → Act → Repeat
  - Combine reasoning with tool use
- CSS relevance: automate data collection, run analyses, simulate social scenarios, do projects

# AI Agents



# What is the Best Model / Agent?

- In-house metrics: training loss, perplexity
- Benchmarks
  - Language Understanding (MMLU)
  - Maths
  - Medicine
  - Software Engineering
  - Task Time
  - ARC-AGI
- Look at system cards & Leaderboards

# LLM Risks: Reliability & Safety

- Hallucination
  - Unreliable or nonsensical outputs
  - “LLMs are trained to make guesses – no rewards for saying idk”
  - vulnerable to malicious prompts
- LLMs are biased
  - Trained on internet data → reflects society biases
- LLMs can cause harm
  - Suicide, nuclear bomb recipe
  - Guardrails
- LLMs will take our jobs

# Questions?

