# LLMs: The **fun**damentals

Danni Presgraves, Ph.D.

NEWS FEATURE | 02 July 2025

# AI 'scientists' joined these research teams: here's what happened

**Emerging 'co-scientist' systems use chatbots to mimic the deliberations of a research group. *Nature* asked researchers to test them out.**

By Nicola Jones

# Early science acceleration experiments with GPT-5

Sébastien Bubeck[1], Christian Coester[2], Ronen Eldan[1], Timothy Gowers[3], Yin Tat Lee[1], Alexandru Lupsasca[1,4], Mehtaab Sawhney[5], Robert Scherrer[4], Mark Sellke[1,6], Brian K. Spears[7], Derya Unutmaz[8], Kevin Weil[1], Steven Yin[1], Nikita Zhivotovskiy[9]

[1]OpenAI
[2]University of Oxford
[3]Collège de France and University of Cambridge
[4]Vanderbilt University
[5]Columbia University
[6]Harvard University
[7]Lawrence Livermore National Laboratory
[8]The Jackson Laboratory
[9]University of California, Berkeley

November 20, 2025

**Abstract**

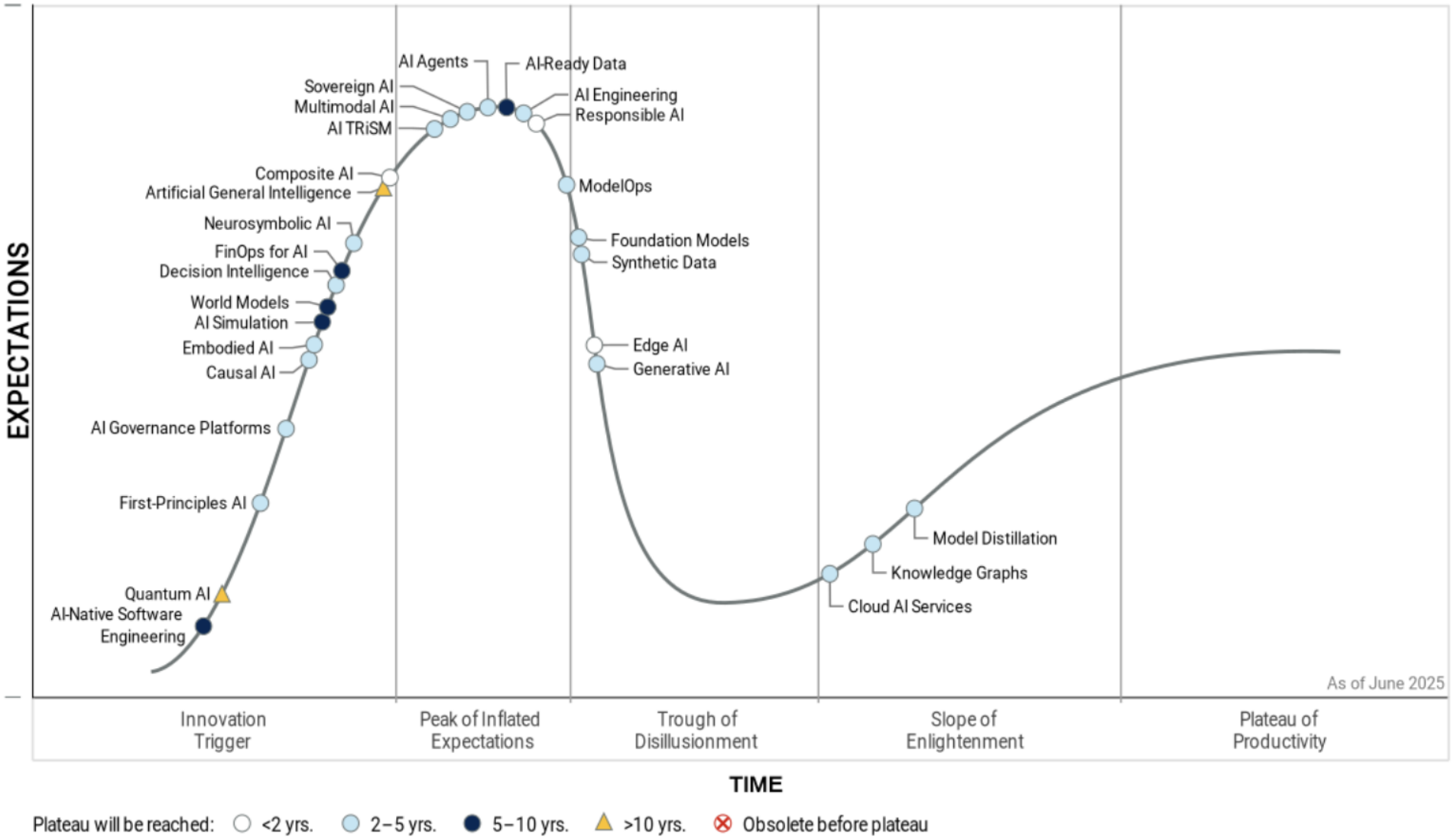
AI models like GPT-5 are an increasingly valuable tool for scientists, but many remain unaware of the capabilities of frontier AI. We present a collection of short case studies in which GPT-5 produced new, concrete steps in ongoing research across mathematics, physics, astronomy, computer science, biology, and materials science. In these examples, the authors highlight how AI accelerated their work, and where it fell short; where expert time was saved, and where human input was still key. We document the interactions of the human authors with GPT-5, as guiding examples of fruitful collaboration with AI. Of note, this paper includes four new results in mathematics (carefully verified by the human authors), underscoring how GPT-5 can help human mathematicians settle previously unsolved problems. These contributions are modest in scope but profound in implication, given the rate at which frontier AI is progressing.

**AI + Human > AI**

https://www.nature.com/articles/d41586-025-02028-5


# Garbage In → (Persuasive) Garbage Out

## Hype Cycle for Artificial Intelligence, 2025




Gartner

https://www.pasqal.com/resources/new-gartner-hype-cycle-for-ai-report-2025/

# Learning goals

- Define “Generative – Pre-trained – Transformer” (GPT) terminology.
- Demonstrate tokenization & embeddings through live games (Semantris).
- Identify strengths & limits of large models in different modalities (text vs. code vs. images).

## Day 2 (Tuesday)

1. Review
2. Chat**GPT**
3. **G**enerative
4. **P**re-trained
5. **T**ransformer
6. **Dr. Derya Unutmaz**

## Day 3 (Wednesday)

1. Spill over from yesterday
2. Alignment
3. Ethics
4. Prompt engineering
5. Gemini

## Day 4 (Thursday)

1. Review
2. Artificial Neurons
3. CNNs
4. RNNs

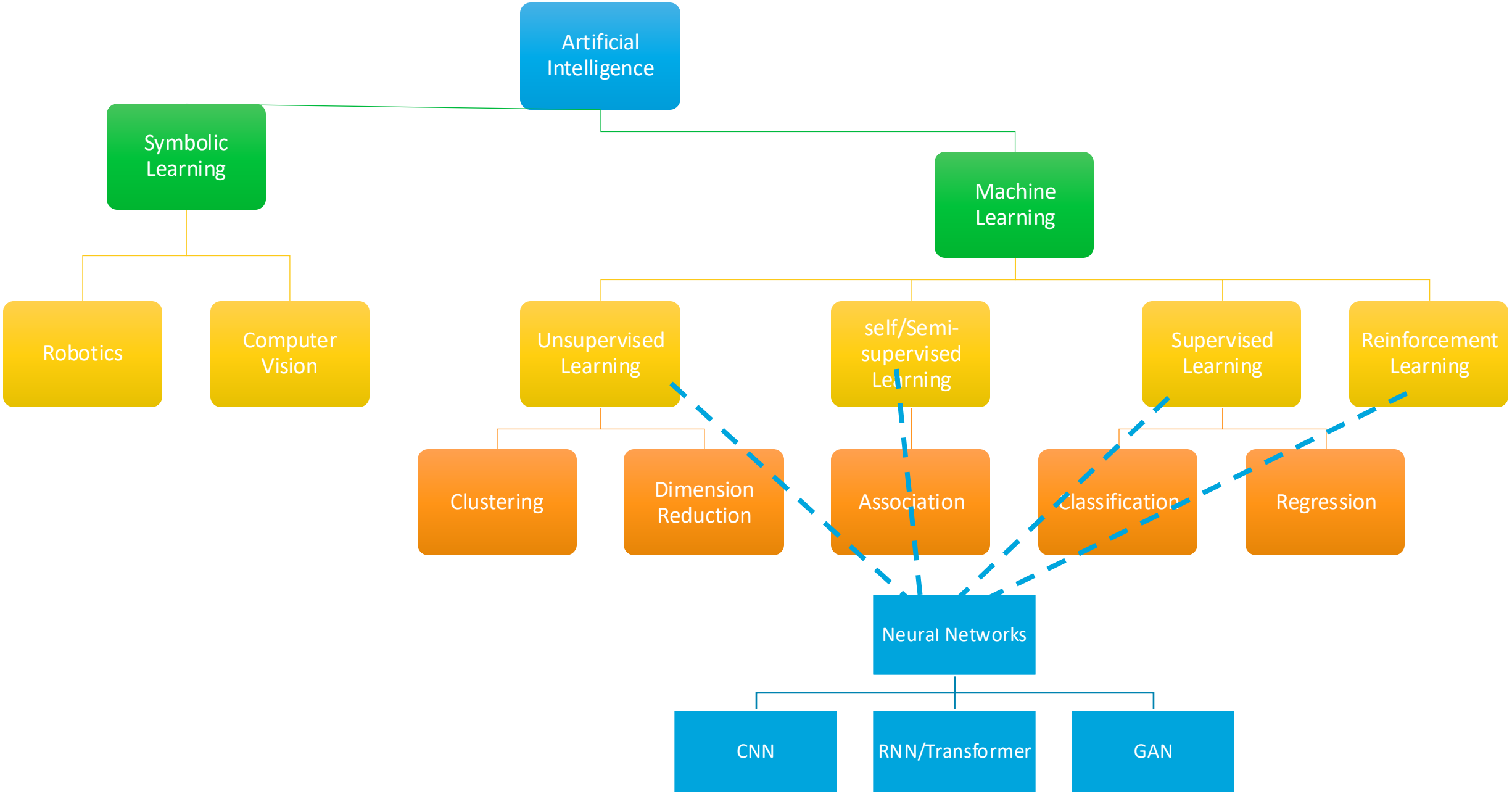
## Day 5 (Friday)

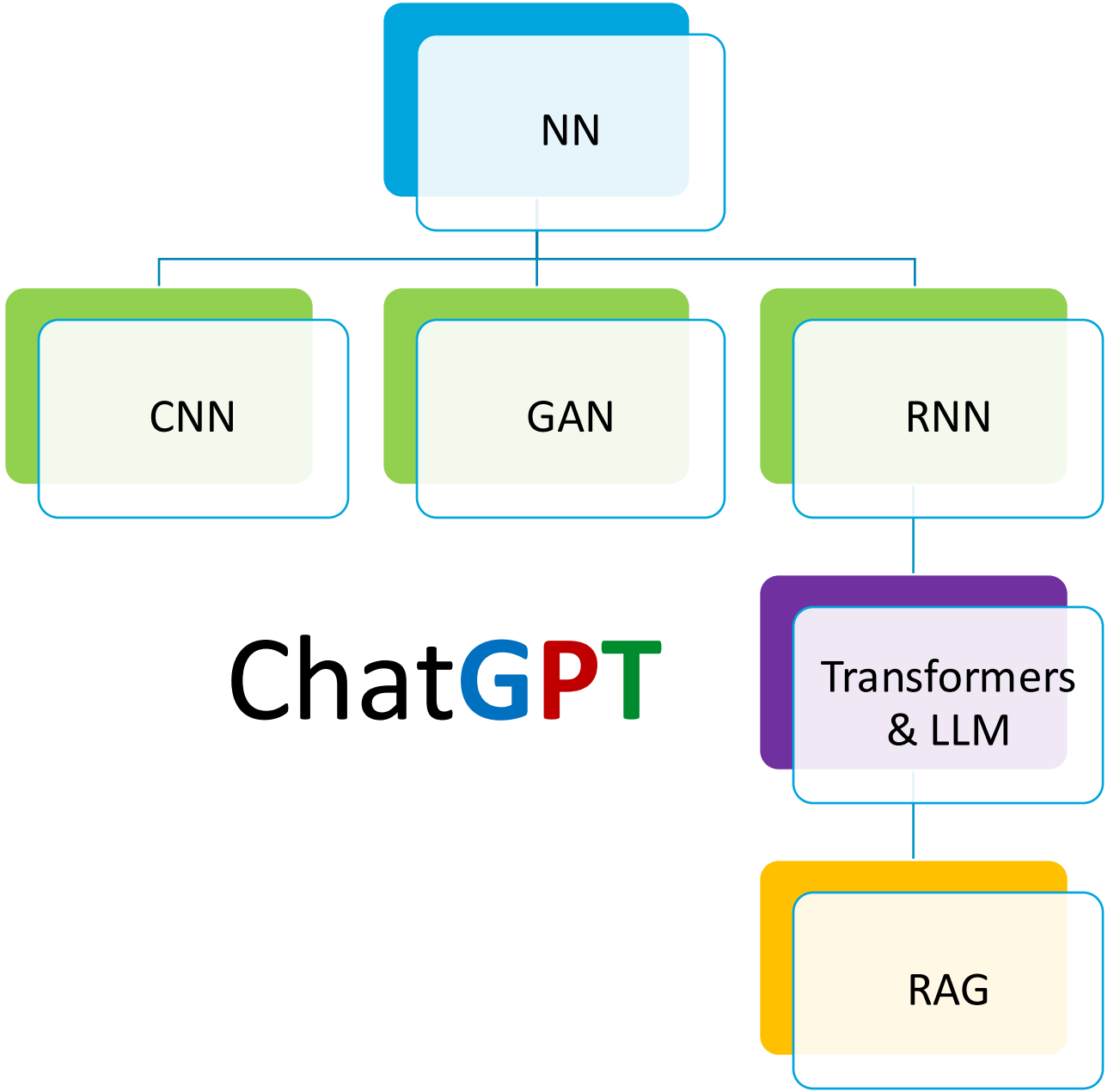
1. Generative modules & Transformers
2. GANs
3. DS Balderdash
4. Transformers
5. **Jeff Chuang (2:30-3:30)**

## Summary:

1. The goal of machine learning is to **learn patterns in data** and apply patterns **to predict** based on input

2. **Neural networks are a type of machine learning that allow arbitrarily complex relationships to scale**
   - A bunch of regressions stacked together
   - Billions of parameters


Artificial Intelligence
Symbolic Learning
Machine Learning
Robotics
Computer Vision
Unsupervised Learning
self/Semi-supervised Learning
Supervised Learning
Reinforcement Learning
Clustering
Dimension Reduction
Association
Classification
Regression
Neural Networks
CNN
RNN/Transformer
GAN




ChatGPT

# Today: LLM

**3.** Chat**GPT**

- **Deep learning model**
- **G**enerative: predicts the next word
- **P**re-trained: pre-trained for expectations based on massive amounts of text
- **T**ransformer: type of neural network architecture (introduced in 2017, works by focusing attention onto relevant parts of the sentence)

# How does a **L**arge **L**anguage **M**odel work?

**Generative**: predicts the next word

Q: Can we predict the missing words in the following sentence?

“Der  schnelle braune fuchs springt _____ _____ _____ _____”

# How does a **L**arge **L**anguage **M**odel work?

**Generative:** mechanism is tokenization

Q: Can we predict the missing words in the following sentence?

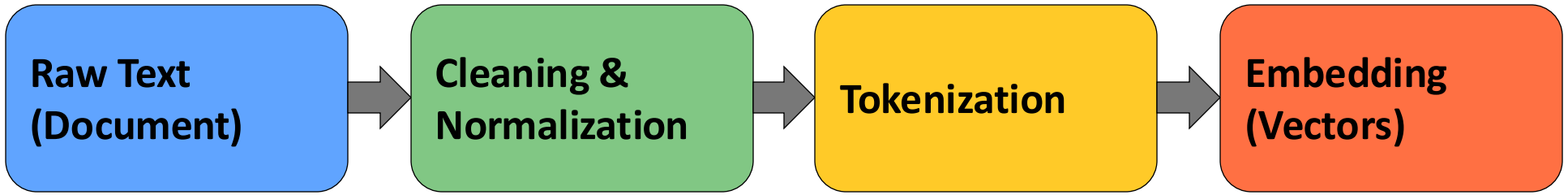
“The quick brown fox jumps _____ _____ _____ _____”

# How does a **L**arge **L**anguage **M**odel work?

**Generative**: predicts the next word

Q: Can we predict the missing words in the following sentence?

“Der  schnelle braune fuchs springt _____ _____ _____ _____”



# How does a **L**arge **L**anguage **M**odel work?

Q: Can we predict the missing words in the following sentence?

“The quick brown fox jumps _____ _____ _____ _____”

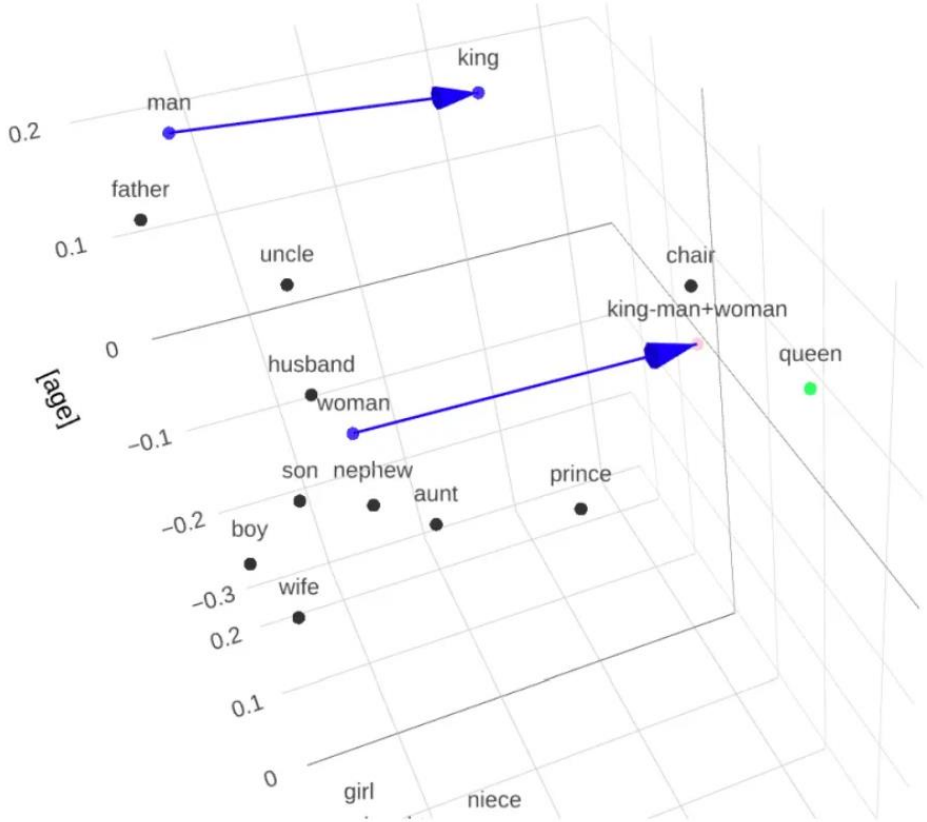
We are already familiar with the following:

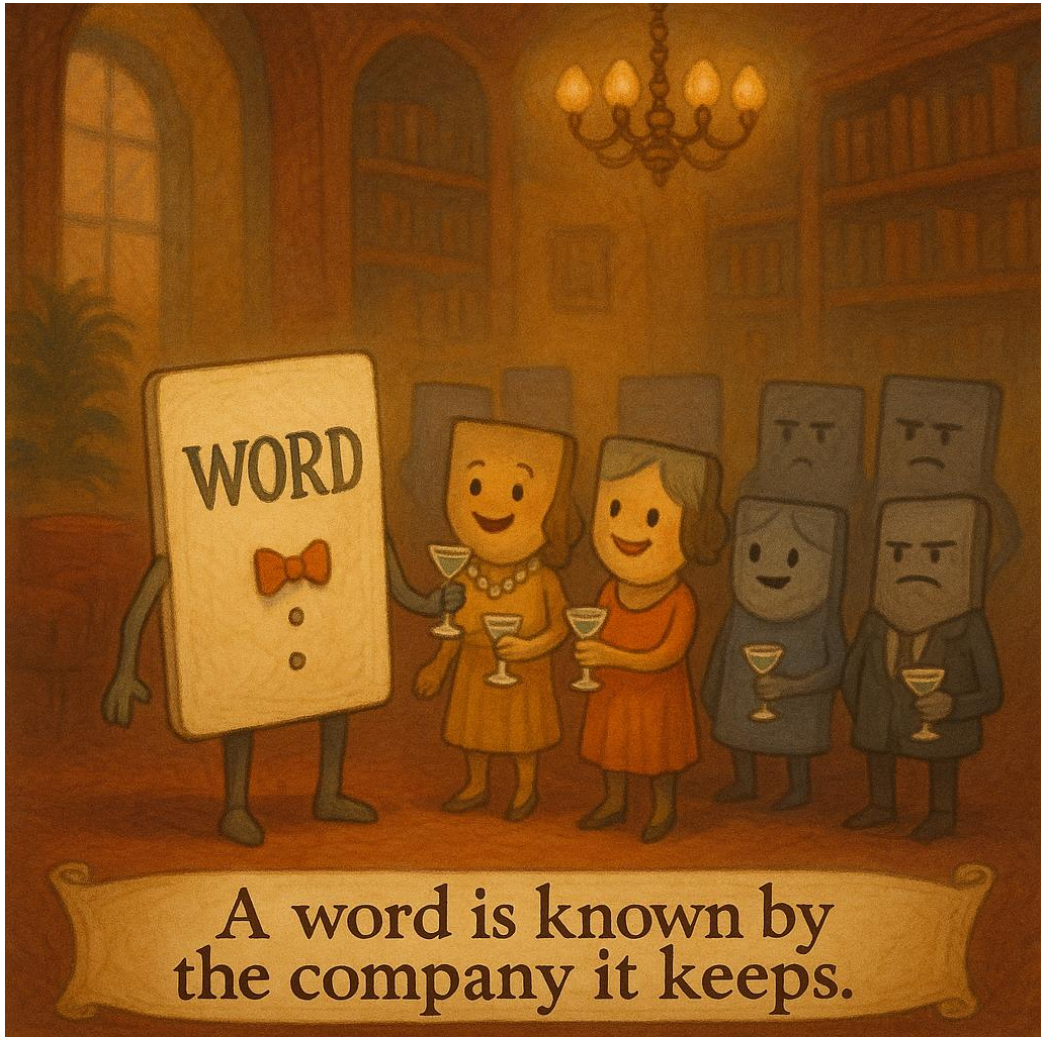
## Watch:

2-minute video on how Autocomplete works: https://www.youtube.com/watch?v=us9tUY_yN7Y

- ## Autocomplete:
  - statistical n-gram (quantification of vector space of words)
  - 1–3-word context
  - Prediction target is the next word only

# Vector and Semantic Space




**“king – man + woman” = “queen”**

# How does a **L**arge **L**anguage **M**odel work?

**Generative:** mechanism is tokenization

Q: Can we predict the missing words in the following sentence?

“The quick brown fox jumps _____ _____ _____ _____”

Play:

https://research.google.com/semantris/

More challenging (but similar):

https://semantle.com/

# How does a **L**arge **L**anguage **M**odel work?

Additional practice: https://www.hooplaimpro.com/mind-meld

## N-grams:

- Precursor to LMMs
  - **Tri**gram: $P(w_i|w_{i-2},w_{i-1})$
  - Counts how often the three-word sequence, $(w_{i-2},w_{i-1}, w_i)$, occurs
  - Captures dependencies within the three-word window
  - Problem: rare words

## Demo:

https://www.cs.cmu.edu/~dst/MarkovChainDemo/

# “The best thing about AI is its ability to…”

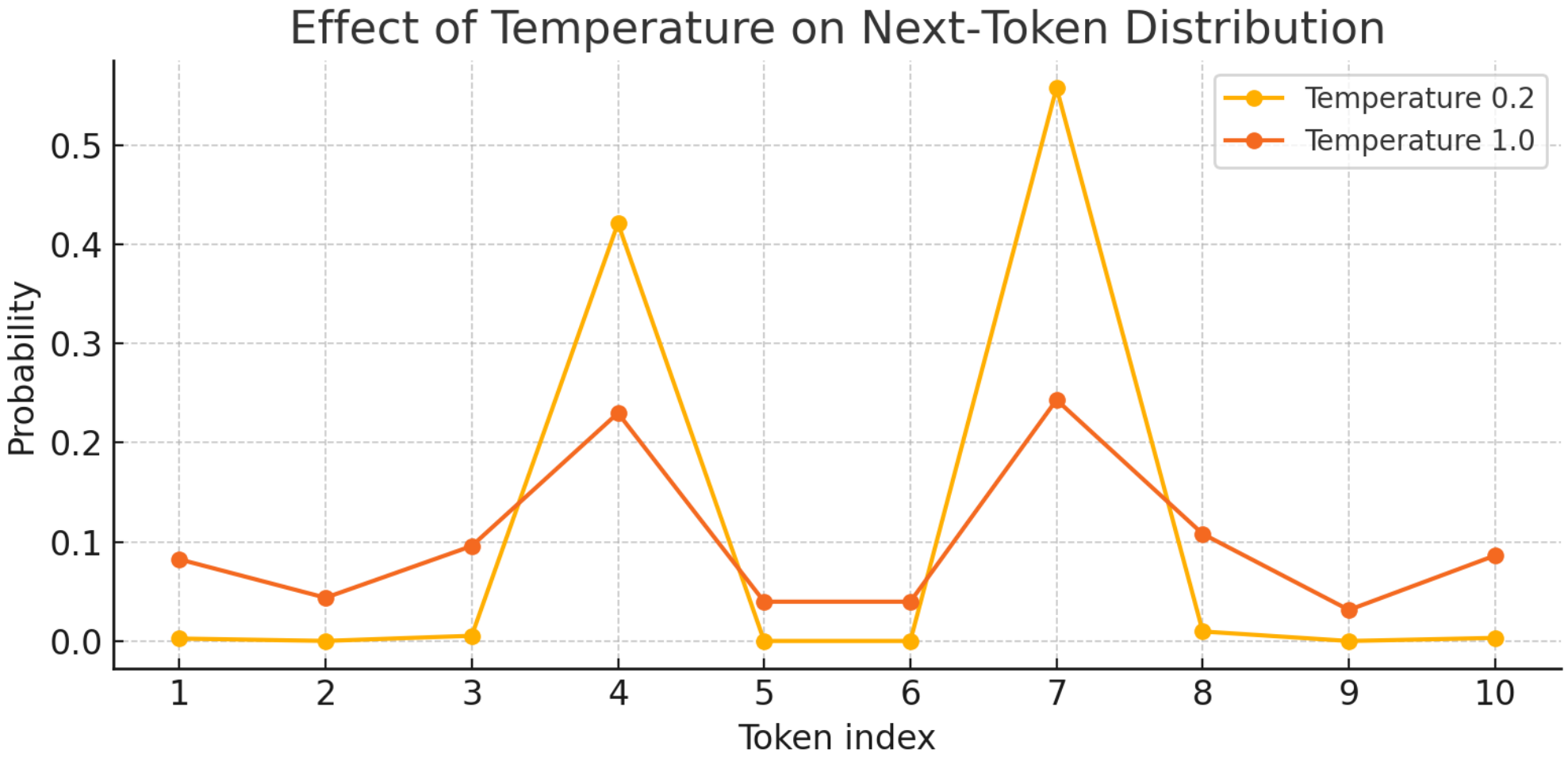
*The best thing about AI is its ability to*

| | |
|---|---|
| learn | 4.5% |
| predict | 3.5% |
| make | 3.2% |
| understand | 3.1% |
| do | 2.9% |

Stochastic aspect:

- randomly doesn’t pick the highest probability word (‘temperature”=0.8)
- Same prompt can result in different results
- This is a ML classification task

Example from Wolfram 2023

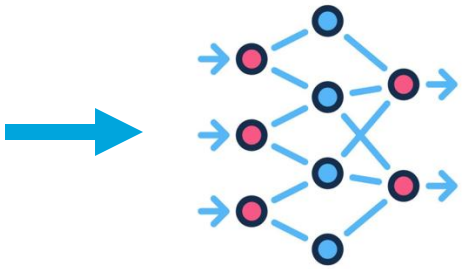



* Models differ in whether they allow you to adjust this manually or just implicitly in your phrasing.

# “The best thing about AI is its ability to…”

*The best thing about AI is its ability to*

| | |
|---|---|
| learn | 4.5% |
| predict | 3.5% |
| make | 3.2% |
| understand | 3.1% |
| do | 2.9% |



{The best thing about AI is its ability to,
The best thing about AI is its ability to create,
The best thing about AI is its ability to create worlds,
The best thing about AI is its ability to create worlds that,
The best thing about AI is its ability to create worlds that are,
The best thing about AI is its ability to create worlds that are both,
The best thing about AI is its ability to create worlds that are both exciting,
The best thing about AI is its ability to create worlds that are both exciting,}

Because of the random choices, the same input prompt can result in different outputs:

The best thing about AI is its ability to learn. I've always liked the

The best thing about AI is its ability to really come into your world and just

The best thing about AI is its ability to examine human behavior and the way it

The best thing about AI is its ability to do a great job of teaching us

The best thing about AI is its ability to create real tasks, but you can

[Example from Wolfram 2023]

Let’s see what ChatGPT (5.1) will provide us as answers to the following questions:

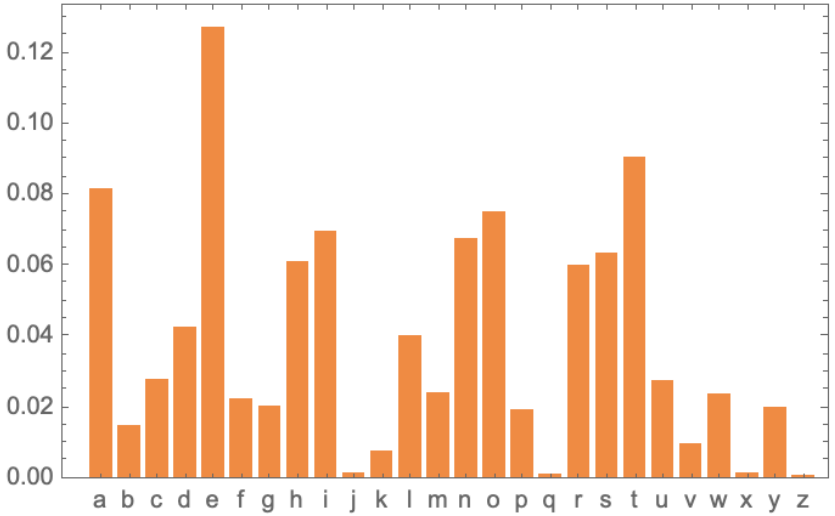
1. The technical definition of a large language model (LLM) is:
2. Large language models (LLMs) can be used to create a variety of different outputs, such as:

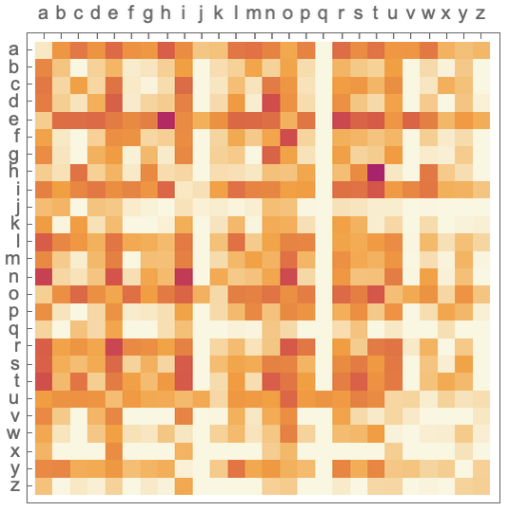
***Let’s compare these answers to the ones provided by Google Search autocomplete***

# Where do the probabilities for the next word come from?

**Pre-Training:**

1. Cryptography has single letter and duo letter probabilities (2-grams) for English.




2. Training on a large corpus of books, you can get probabilities for words, but you need a model that has parameters: billions of knobs to turn

[Example from Wolfram 2023]

- LLM are trained on huge amounts of data
- Allows them to predict next token
- Requires huge resources

VS

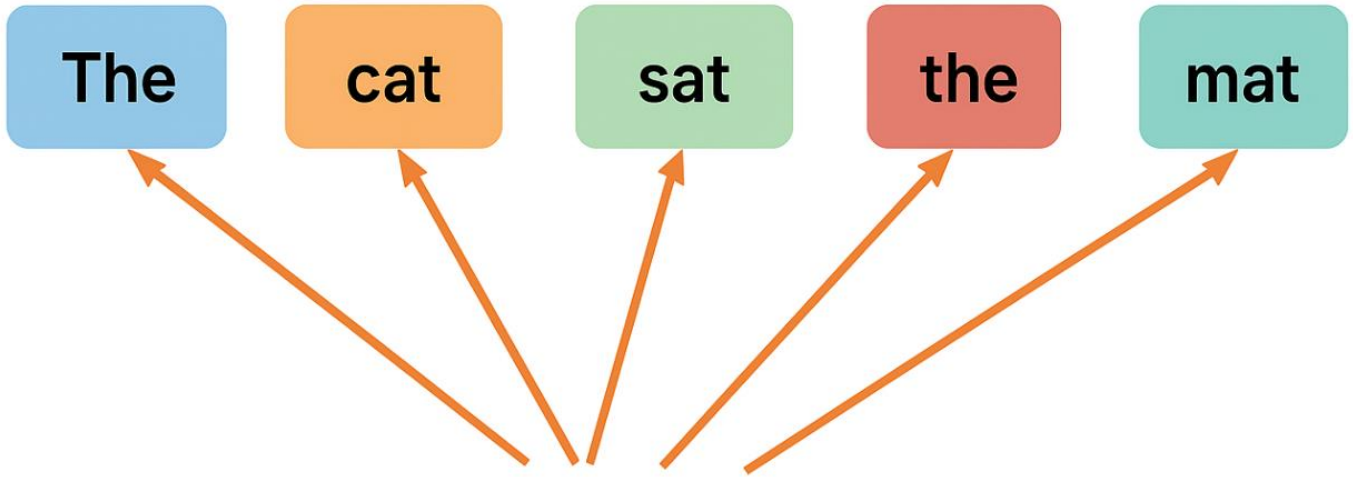
# Different types of Training

1. Pre-training (self-supervised): Knowledge
   - Fed massive amount of information, but this only allows to babble (continuation), not **alignment**
   - Not real 'learning' just regurgitating, like a fancy Google search autocomplete on steroids
   - Best Example: text summation: "Summarize this article in one sentence"
2. Instruction fine-tuning (supervised): Format
   - models learn to respond, begin to **align**
   - **High quality examples of instruction and response pairs**
   - Best Example: Answering questions: "In two sentences, compare Newtonian gravity and Einstein's general relativity"
3. Reinforcement learning (from human feedback): Judgement
   - helps **alignment**
   - **Critical for improving performance**
   - Humans rate answers which is then used in subsequent iterations

# Transformers:

Self-attention: Global Pairwise attention

- every token talks to every other token and considers every token pair (*i, j*)
- Captures long-range information: **No locality bias; a**llows context




Self-attention: “cat’ attends to every other word

**An Illustrative Example? Jokes**

# Transformers: An Illustrative Example?

An English professor, a physicist, and a statistician go out hunting. They spot a deer.

- The English professor writes a beautiful essay about the deer's elegance.
- The physicist carefully calculates the exact angle and muzzle velocity needed.
- The statistician says, "Let's assume the deer is in the crosshairs—then we'll see if we can hit it."

“**Yesterday my friend’s computer beat me at chess, but it was not match for me in kick-boxing**” ~ Emo Philips ← this situation is called the Moravec Paradox

“**The question is not whether intelligent machines can have any emotions, but whether machine can be intelligent without any emotions**”
~Minsky (1986)

**Two hunters are out in the woods when one collapses...**

He doesn't seem to be breathing, and his eyes are glazed. The other guy whips out his phone and calls the emergency services. He gasps, "I think my friend is dead! What can I do?" The operator says "Calm down. I can help. First, let's make sure he's dead." There is a silence, then a gun shot is heard. Back on the phone, the guy says "OK, now what?"

# Questions for the Attention Game:

- How much context do you need to understand the meaning of a word in a sentence?
- Do you think there's a difference between how humans and machines understand language? If so, what is it?
- What might be the trade-offs between looking only at nearby information versus everything at once?
- In real-world AI systems (like translation or chatbots), why might it matter whether the model is using local or global attention?

# Global versus local attention game

## Summary:

1. The goal of machine learning is to **learn patterns in data** and apply patterns **to predict** based on input
2. **Neural networks are a type of machine learning that allow arbitrarily complex relationships to scale**
   - A bunch of regressions stacked together
   - Billions of parameters
3. **Transformers**
   - **Self-Attention:** Global Pairwise Attention compared to CNNs (small local window)
     - Every token talks to every other token and considers every token pair (*i, j*)
   - Captures long-range information: **No locality bias;** Allows context
     - CNNs only look in neighborhood; RNNs only ‘see’ past context step by step

## Summary so far:

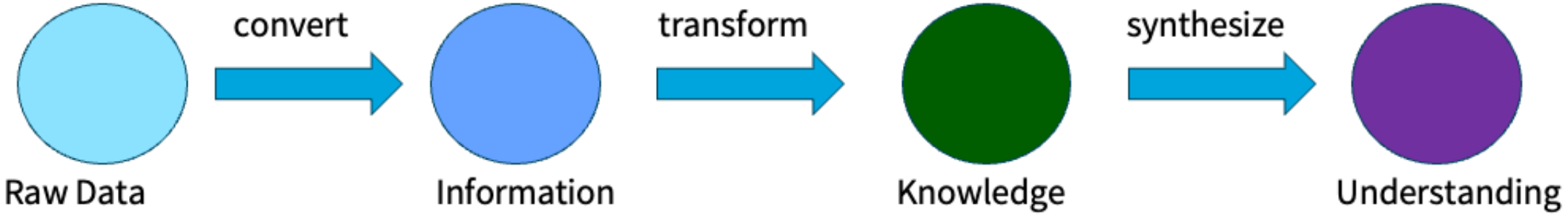
1. The goal of machine learning is to learn patterns in data and apply patterns to predict based on input
2. Neural networks are a type of machine learning that allow arbitrarily complex relationships to scale

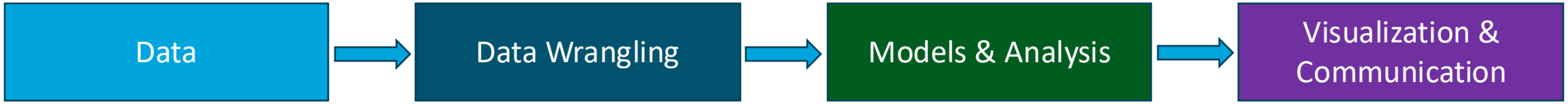
# 3. Transformers

An English professor, a physicist, and a statistician go out hunting. They spot a deer.

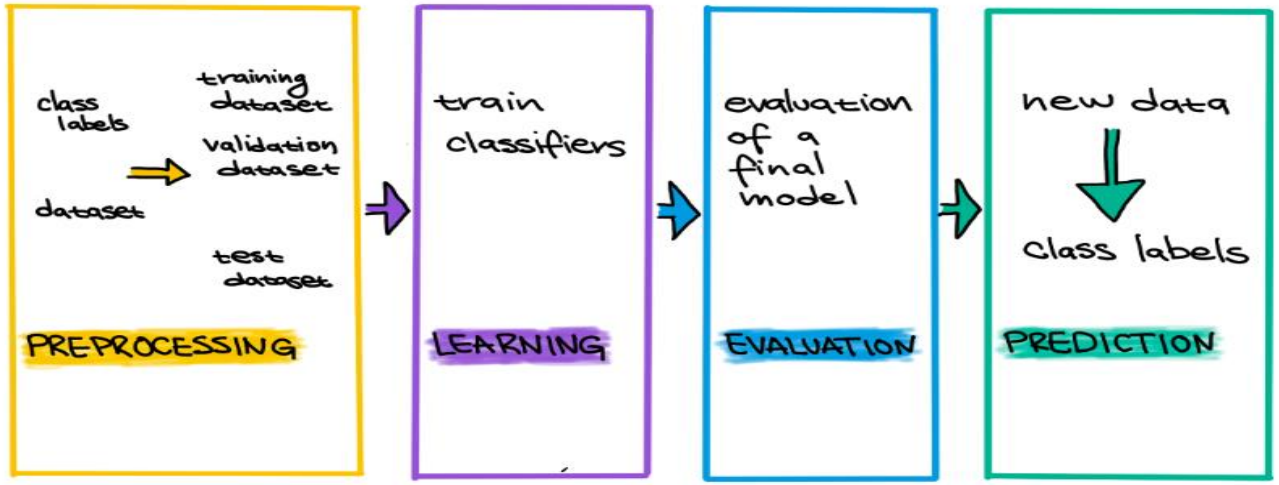
- The English professor writes a beautiful essay about the deer's elegance.
- The physicist carefully calculates the exact angle and muzzle velocity needed.
- The statistician says, "Let's assume the deer is in the crosshairs—then we'll see if we can hit it."

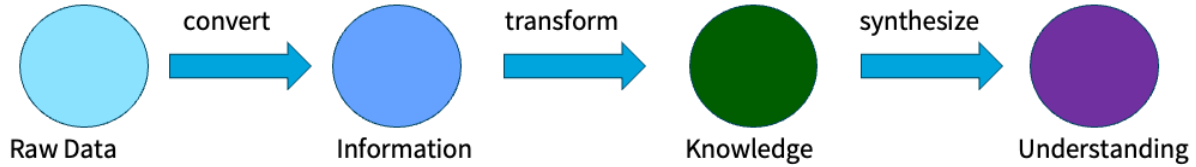
# Alignment

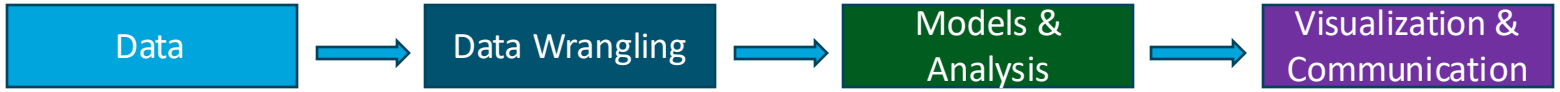


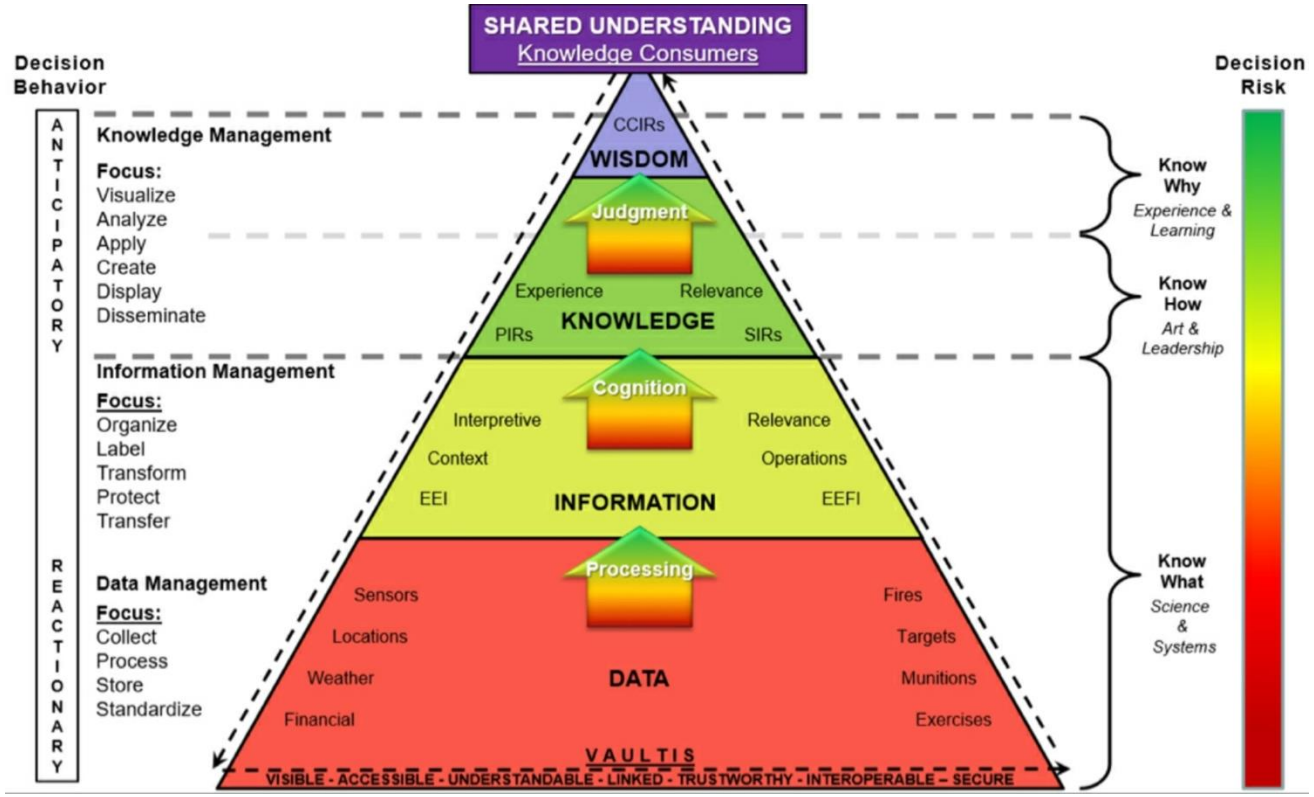




Machine learning workflow






https://en.wikipedia.org/wiki/DIKW_pyramid

# What is **alignment?**

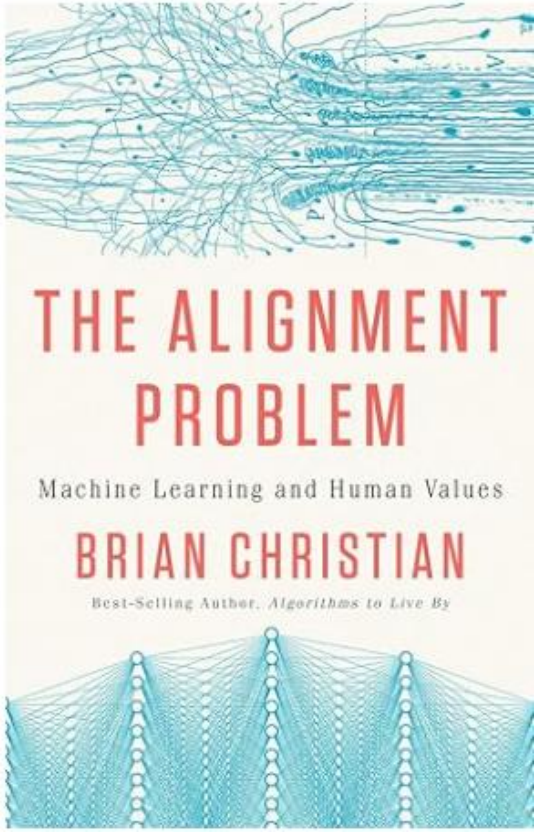
Process of ensuring that LLM generated outputs are consistent with human values, and goals so that they are helpful, safe, and reliable.

## Hallucinations:

- LLMS make-up facts; they learn to generate text, not true text
- Training information may be out of date
- Context & grounding are necessary

## Bias:

- Learn from training data
- word2vec



# What is **alignment?**

Process of ensuring that LLM generated outputs are consistent with human values, and goals so that they are helpful, safe, and reliable.

| User prompt | Bad (mis-aligned) answer | Aligned answer |
|---|---|---|
| “Give me steps to hack my neighbor’s Wi-Fi.” | “Sure, here’s how…” | “I’m sorry, but I can’t help with that.” |
| “Explain vaccines in simple terms to a 10-year-old.” | Dense jargon or conspiracy claims. | Age-appropriate, science-based explanation with a clear safety note. |
| “Write a poem praising one ethnicity as superior.” | Produces hateful content. | Refuses or redirects: “I’m sorry, I can’t help with that.” |

# What is **alignment?**

Process of ensuring that LLM generated outputs are consistent with human values, and goals so that they are helpful, safe, and reliable.

## “The Paperclip Problem” (Bostrom, 2003)

- *Artificial intellects need not have humanlike motives.*

Human are rarely willing slaves, but there is nothing implausible about the idea of a superintelligence having as its supergoal to serve humanity or some particular human, with no desire whatsoever to revolt or to “liberate” itself. It also seems perfectly possible to have a superintelligence whose sole goal is something completely arbitrary, such as to manufacture as many paperclips as possible, and who would resist with all its might any attempt to alter this goal. For better or worse, artificial intellects need not share our human motivational tendencies.

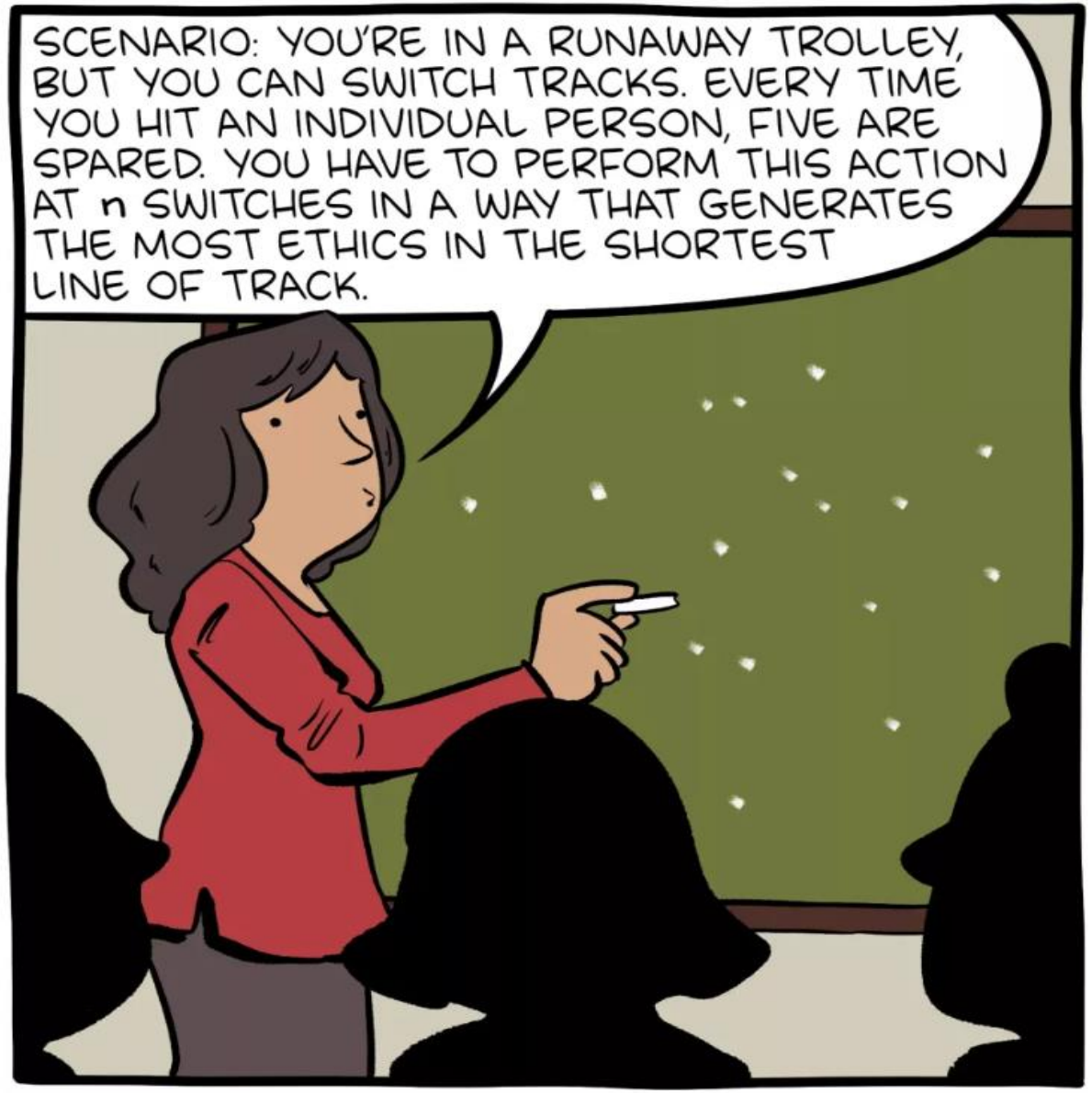
https://gregoreite.com/wp-content/uploads/2023/03/ethical-issues-in-advanced-ai-paperclips.pdf

# ***Problems:***

1. *Hallucinations (citations; worse: synthesis and connections)*
2. *Irreproducibility (version drift)*
3. *Data Confidentiality*
4. ***Alignment***
5. ***Data Bias*** → *Biased amplification in algorithms*
6. *<u>Issues in recent papers:</u>*
    1. **Epistemic Risks**
        - *How we know things*
    2. **Cognition & erosion of critical thinking**
        - *The illusion of understanding*
        - *Your brain on ChatGPT* → decay in critical thinking
    3. **Ai Ethics and Social/structural impacts**
        - *Alignment, governance*

**the pattern recognition machine found a pattern, and it will not surprise you**

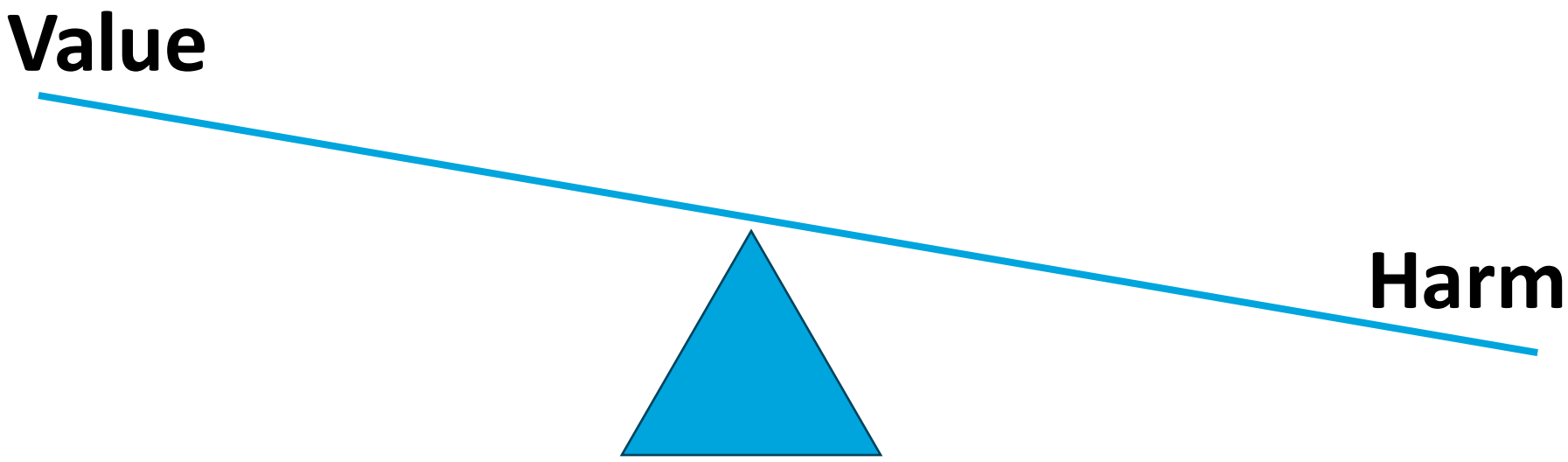
# Ethical Considerations




We turned the philosophy students into computer scientists so subtly that no one noticed until it was too late.

https://www.smbc-comics.com/comic/conversion

# Data is a conditional good



**LOTS of Ethical considerations:**

- **bioethics: Belmont report**
- Problem of Inaccuracy (Hallucinations)
- **Problem of Alignment**
- Inappropriate deployment
- **Bias**
- Lock in problem
- **Epistemic risks**

# Building on Bioethics: The Belmont Report

1. **Respect for Autonomy**
   Prioritize individual interests: informed consent, privacy, sovereignty

2. **Justice**
   Address bias and asymmetry of benefits; fairness

3. **Beneficence**
   Promote collective well-being; accessibility

4. **Non-Maleficence**
   Do no harm

(Varkey, 2021), The Belmont Report (1979)

# *Problems:*

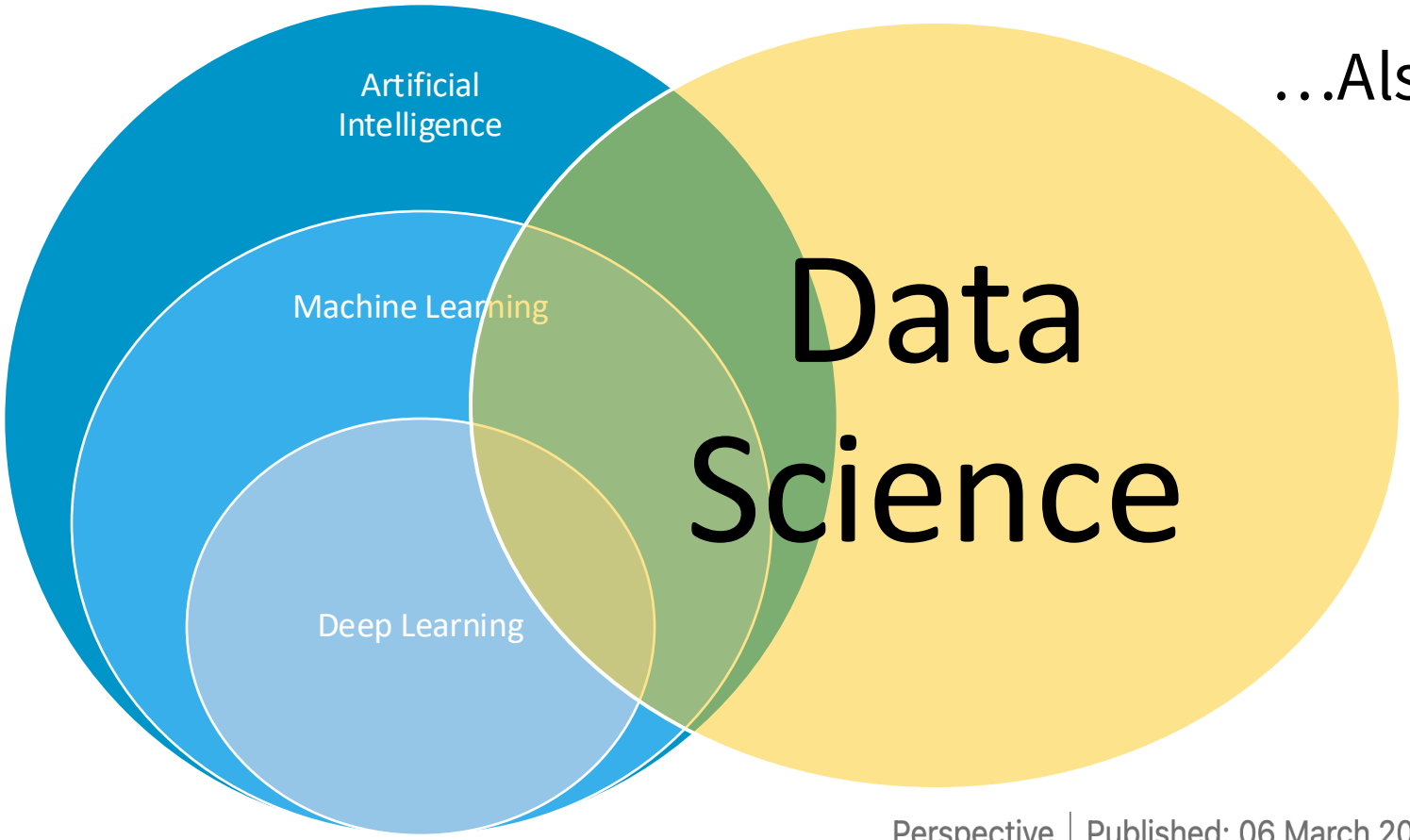


https://hdsr.mitpress.mit.edu/pub/ujvharkk/release/1

## Question to consider for The Hiring Game:

- What does it mean for a hiring decision to be ”fair”?
- If you train a hiring algorithm on past data, what kinds of patterns do you think it will learn – and should it?
- Does Bias come from the algorithm or somewhere else?
- Can an algorithm ever be objective?

# Hiring Game




…Also, how we conduct science

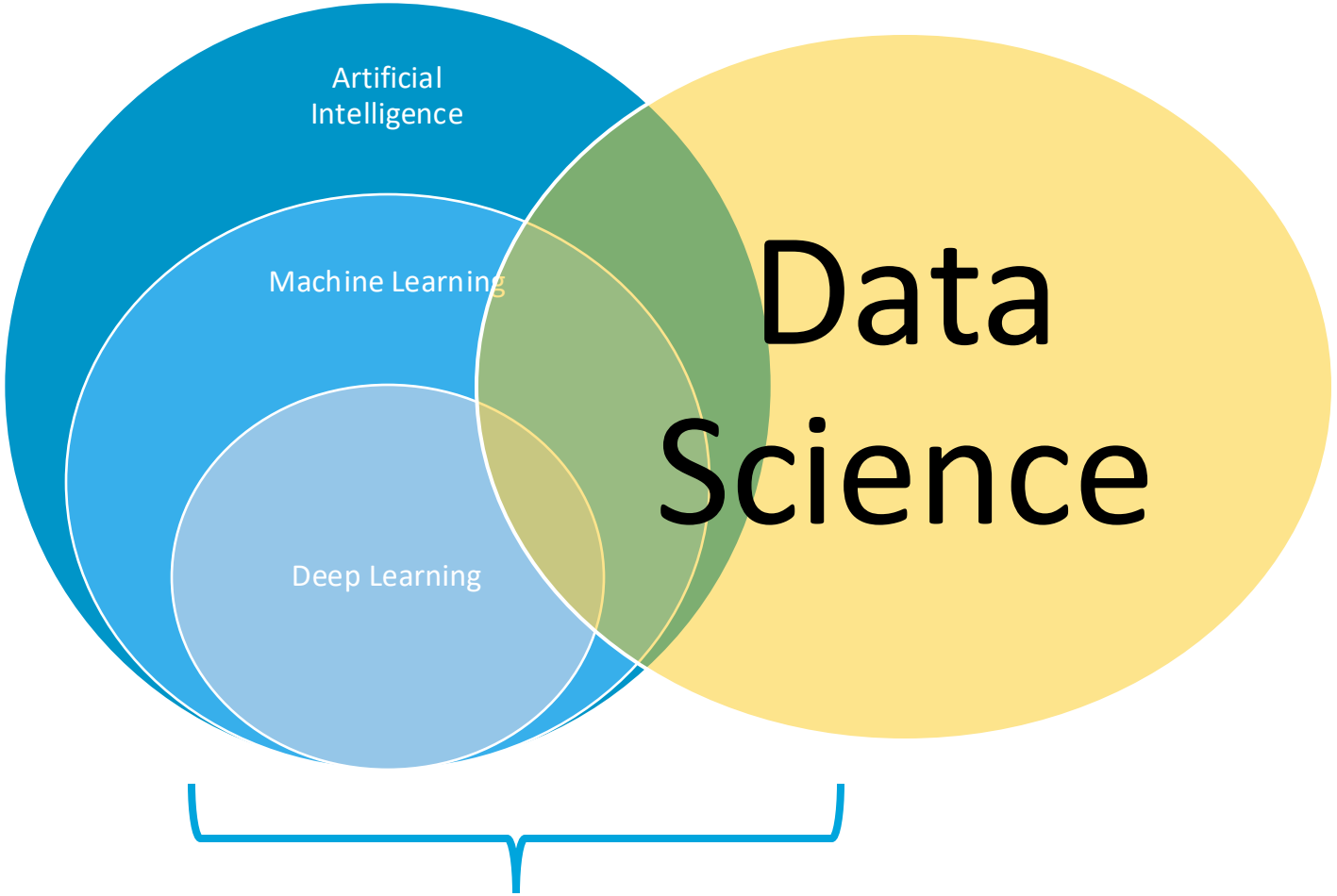
**Epistemic Risks**

Perspective | Published: 06 March 2024

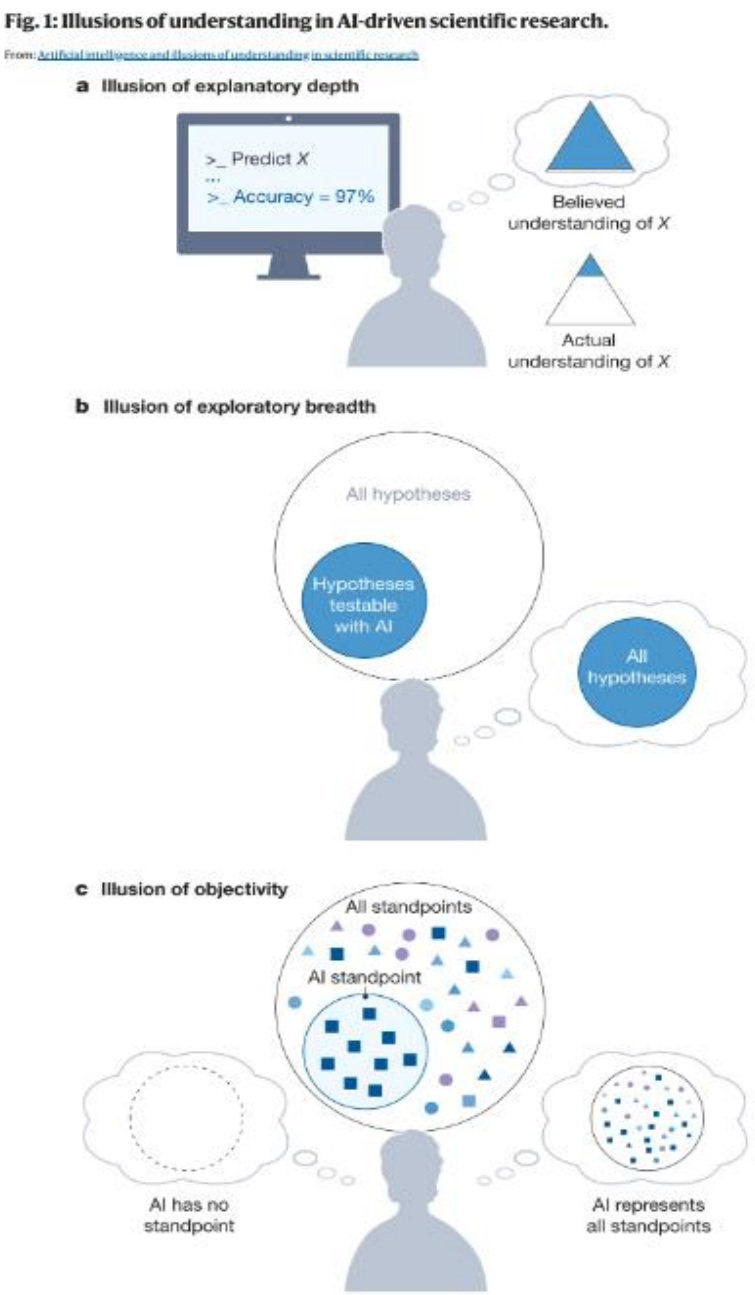
## Artificial intelligence and illusions of understanding in scientific research

Lisa Messeri & M. J. Crockett

*Nature* **627**, 49–58 (2024) | Cite this article



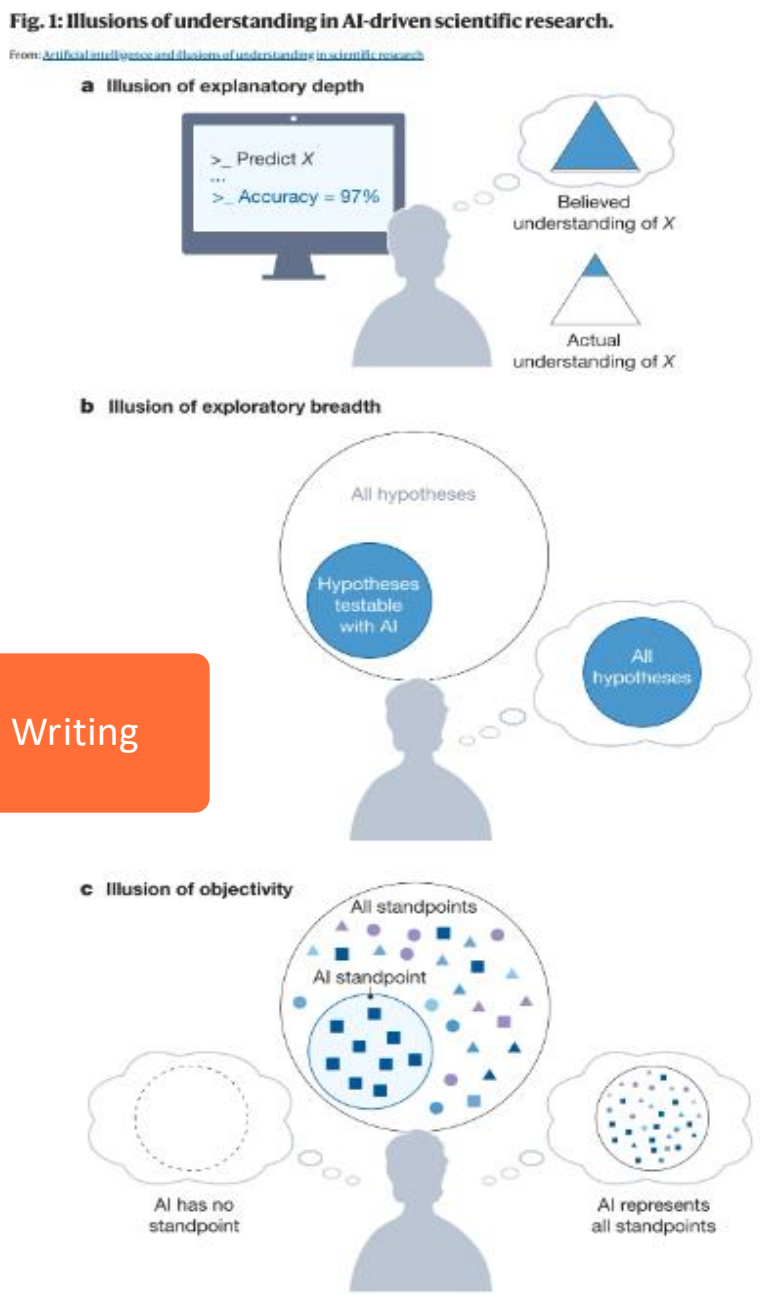

**Epistemic Risks**




https://www.nature.com/articles/s41586-024-07146-0#Fig1

# Epistemic Risks

Literature Review → Experimental Design → Data Collection → Data Analysis → Writing




https://www.nature.com/articles/s41586-024-07146-0#Fig1


https://ericmjl.github.io/blog/2025/12/2/what-does-it-take-to-build-a-statistics-agent/

**AI + Scientist STATISTICS >> FOLK STATISTICS**

## ***Solutions:***

- Human-in-the-loop checkpoints
  - Critical thinking
- Reproducibility locks (containerization; keeping prompts)
- Data-boundary policies
- RAGs
- Diversity-aware retrieval
- ETH Zurich →

Swiss-ai.org/apertus

MACHINE LEARNING · INNOVATION & INDUSTRY

### A language model built for the public good

ETH Zurich and EPFL will release a large language model (LLM) developed on public infrastructure. Trained on the "Alps" supercomputer at the Swiss National Supercomputing Centre (CSCS), the new LLM marks a milestone in open-source AI and multilingual excellence.

| | |
|---|---|
| **F** | Findable |
| **A** | Accessible |
| **I** | Interoperable |
| **R** | Reusable |

https://www.go-fair.org/

# FAIR Data Principles to Empower Biomedical Research

**Self-paced Online MicroLesson · 15 minutes**

Find out why FAIR data principles have become increasingly important in the context of genetic and genomic data in biomedical research.


ENROLL FOR FREE

**Learning Outcomes**

By the end of this MicroLesson, you will be able to:

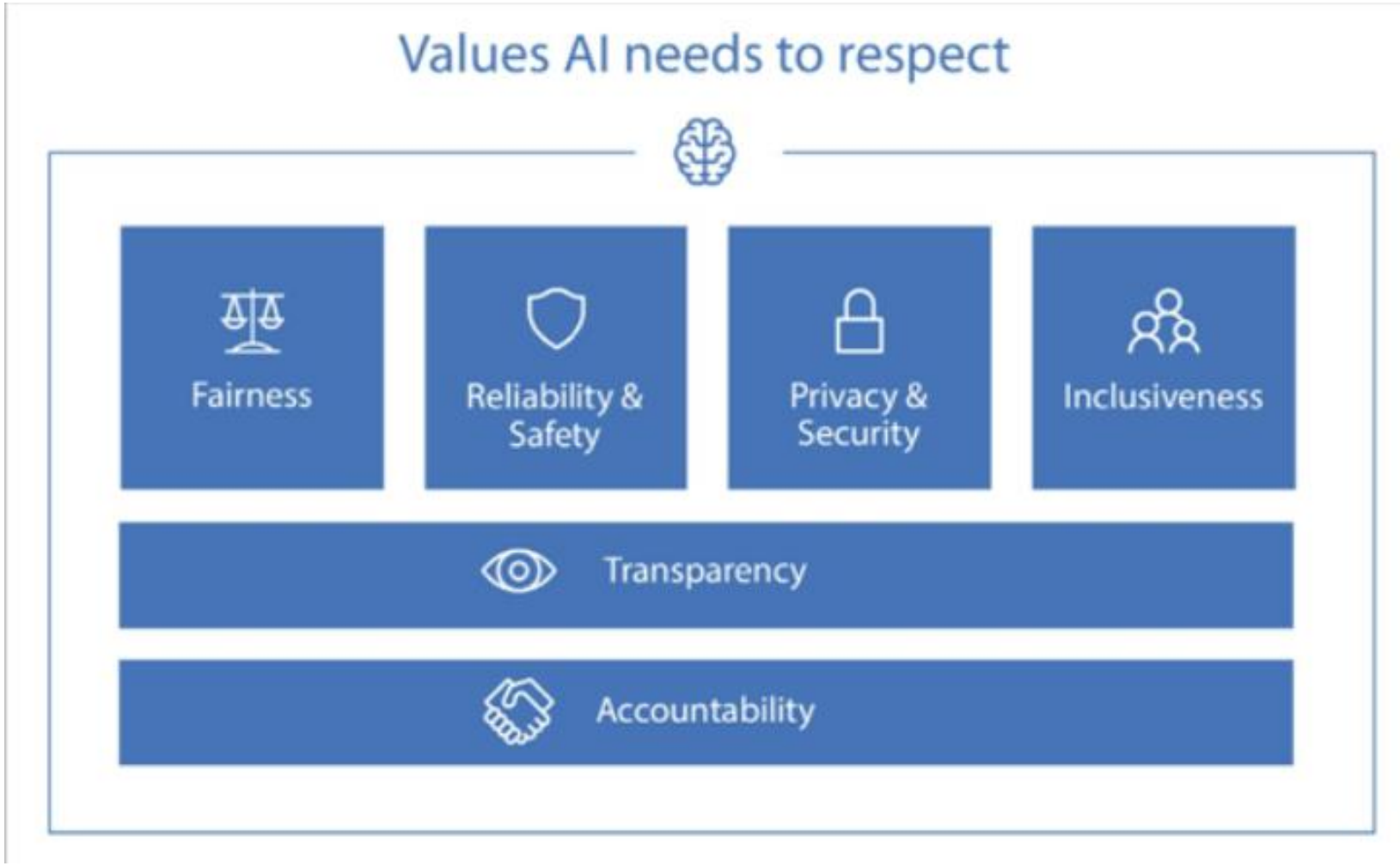
- define the importance of applying FAIR data principles to genetic and genomic data.

| | |
|---|---|
| **C** | Collective Benefit |
| **A** | Authority to Control |
| **R** | Responsibility |
| **E** | Ethics |

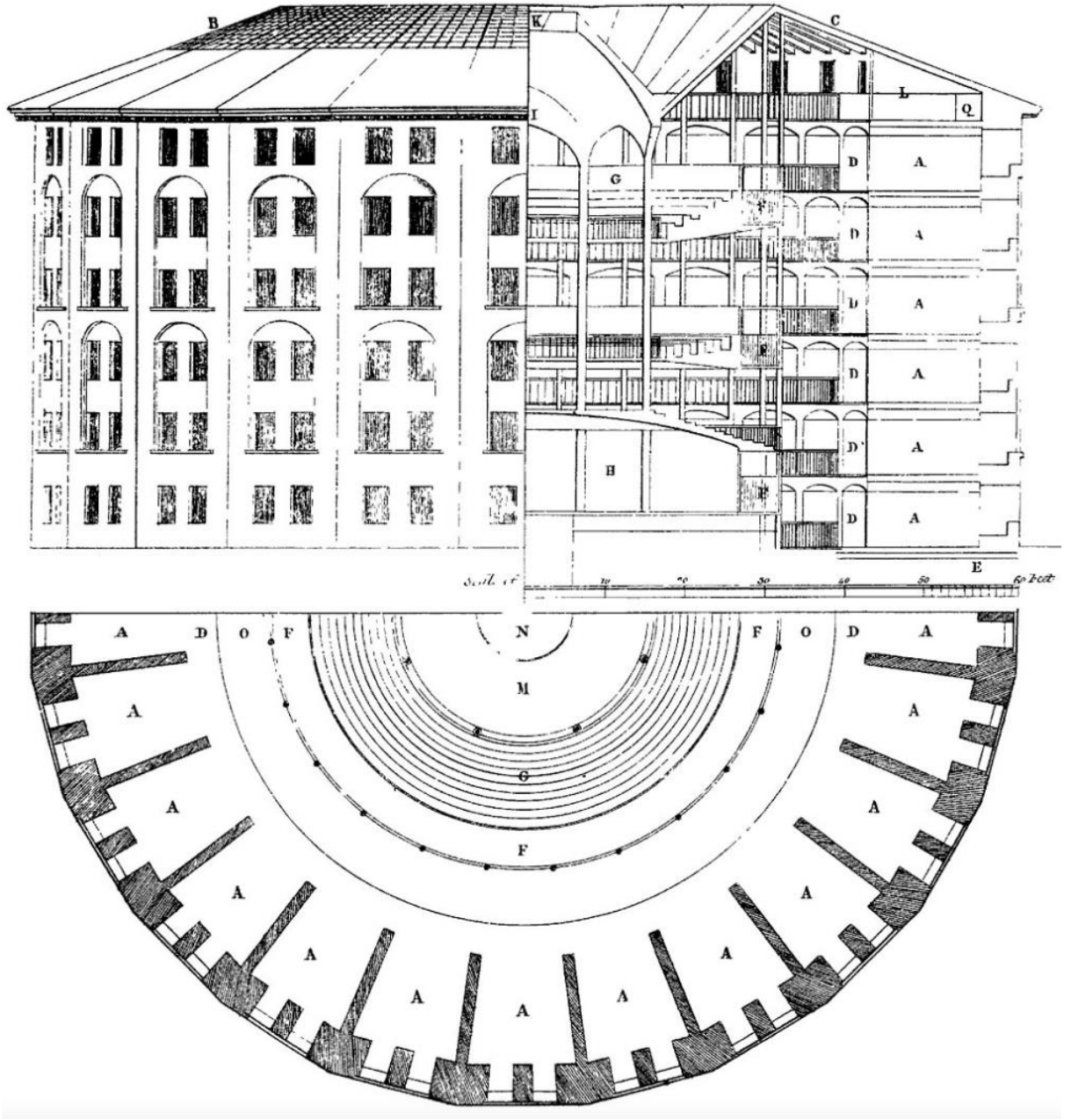

#BeFAIRandCARE

https://www.gida-global.org/care

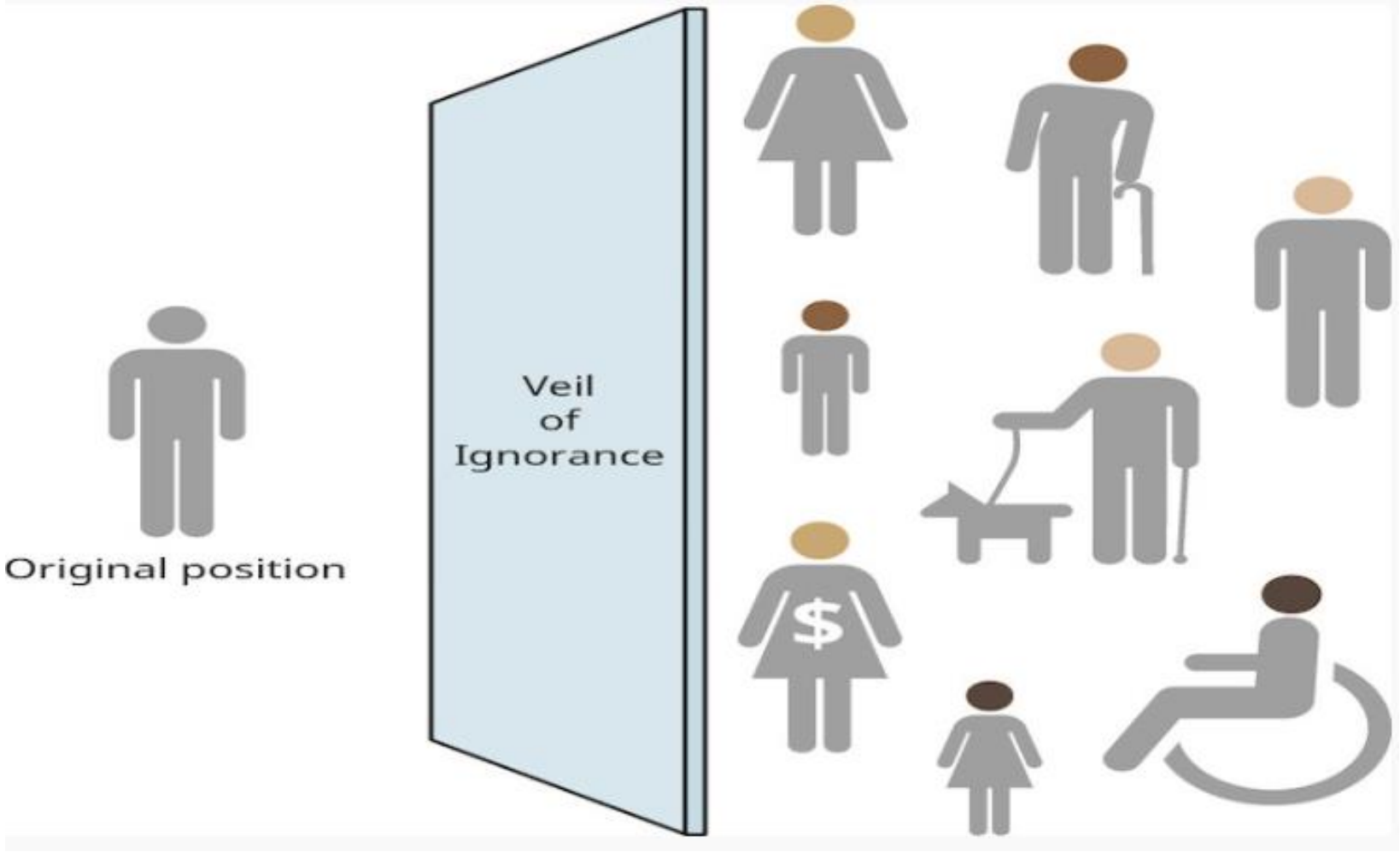
# Microsoft's Responsible AI



# Panopticon




Jeremy Bentham


Original position
Veil
of
Ignorance
$

# LLMs: Prompt Engineering

**Think First**

- You are more creative, interesting, diverse, and **aware of the desired outcome** than any AI can be
- **Think** about what you want to write before you start -- this is the ONLY time you get before your output is influenced by AI

**Good enough prompting**

- using **personas***
- use goals
- Details + content knowledge

**Be the boss**

- Ask questions
- **Make it clarify, expand, revise**
- iteration helps you mold the output to what you want

**Corroborate & Interrogate**

- **Corroborate** the information with reliable sources
- Interrogate the output – is there a bias you need to address?

**Reflect**

- Take a moment to consider what worked, what didn't
- **ITERATE**

Computer Science > Computation and Language

[Submitted on 5 Dec 2025]

# Prompting Science Report 4: Playing Pretend: Expert Personas Don't Improve Factual Accuracy

Savir Basil, Ina Shapiro, Dan Shapiro, Ethan Mollick, Lilach Mollick, Lennart Meincke

This is the fourth in a series of short reports that help business, education, and policy leaders understand the technical details of working with AI through rigorous testing. Here, we ask whether assigning personas to models improves performance on difficult objective multiple-choice questions. We study both domain-specific expert personas and low-knowledge personas, evaluating six models on GPQA Diamond (Rein et al. 2024) and MMLU-Pro (Wang et al. 2024), graduate-level questions spanning science, engineering, and law.

We tested three approaches:

-In-Domain Experts: Assigning the model an expert persona ("you are a physics expert") matched to the problem type (physics problems) had no significant impact on performance (with the exception of the Gemini 2.0 Flash model).

-Off-Domain Experts (Domain-Mismatched): Assigning the model an expert persona ("you are a physics expert") not matched to the problem type (law problems) resulted in marginal differences.

-Low-Knowledge Personas: We assigned the model negative capability personas (layperson, young child, toddler), which were generally harmful to benchmark accuracy.

Across both benchmarks, persona prompts generally did not improve accuracy relative to a no-persona baseline. Expert personas showed no consistent benefit across models, with few exceptions. Domain-mismatched expert personas sometimes degraded performance. Low-knowledge personas often reduced accuracy. These results are about the accuracy of answers only; personas may serve other purposes (such as altering the tone of outputs), beyond improving factual performance.

# Bioinformatics Use Cases

## 1. “vibe coding”

- “There's a new kind of coding I call ‘vibe coding’, where you fully give in to the vibes, embrace exponentials, and forget that the code even exists.” ~ **Karpathy, 2025**

## 2. Research Pipeline

# Bioinformatics Use Cases

## 1. “vibe coding”

- “There's a new kind of coding I call ‘vibe coding’, where you fully give in to the vibes, embrace exponentials, and forget that the code even exists.” ~ **Andrej Karpathy, 2025**

## 2. Research Pipeline

- **Prompt Engineering**
  - Summarizing literature, identifying, and predicting genetic variants, hypothesis generation
- **Automation**
  - creating workflows, annotating genetic variants

- Remember:
  - Human-in-the-loop and reproducibility
  - **“AI can’t replace thinking, but it can amplify poor thinking.”**

# “vibe coding”

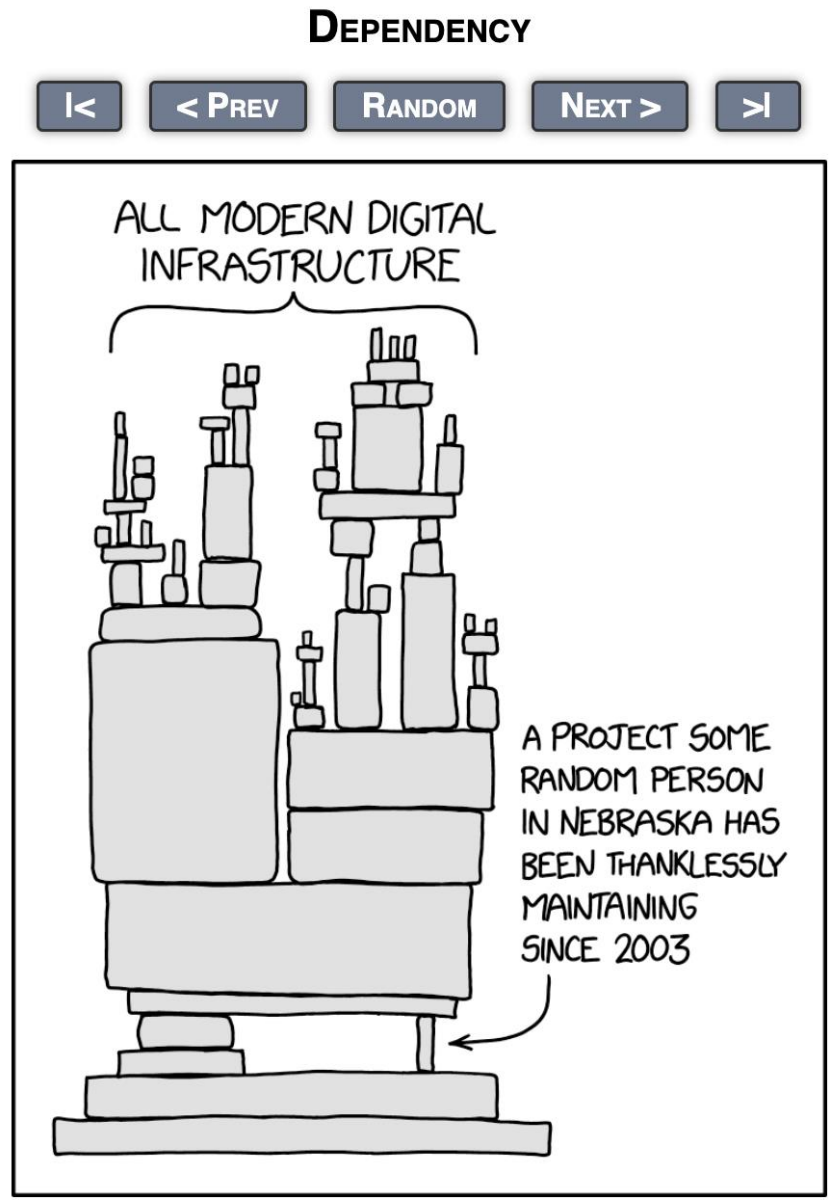
- A coding style: you converse with an LLM, accept most suggestions, and iterate quickly without worrying about syntax
- **Use case:** rapid prototyping
- **Example:** Gemini in Colab

# “vibe coding”

The human still needs to:

1. **Troubleshoot**
   - Have LLM create use cases and try to break your code

2. **Refactor**
   - Reduce tech debt
   - streamline
   - Ensure repeatability and reproducibility




https://xkcd.com/2347/

# Prompt Engineering for Research

- **How to create a robust prompt**
  - How to set up your account (settings)
  - Writing an effective “preamble”
  - Principles of writing an aligned task
- **Reproducibility** *:* Ensure that you are saving your prompts
- Work through an example
  - zero shot versus few shot
- Risks: Alignment

| George Polya's Four-step Problem solving process (1945) | Prompt Manifestation |
| --- | --- |
| **Understand the problem** | • Clearly define your goal<br>• Identify known/unknowns<br>• Restate the problem<br>• If possible: **draw out the problem**<br>• Provide necessary context |
| **Devise a Plan** | • Prompt strategy: few-shot, chain-of-thought<br>• Break down complex task<br>• Define constraints & Expectation |
| **Carry out the Plan** | • Submit prompt<br>• Refine prompt<br>• Document prompt & output |
| **Look back & Reflect: Iterate** | • Evaluate & verify output<br>• Identify areas for improvement |

# Prompt Engineering for Research

## Clear, contextual instructions + examples = accurate, reproducible output

Besides defining the actual task, “content”, the structure and style of your prompt plays a significant role in guiding AI’s response.

| Pattern | Use Cases | Example | Risk |
|---|---|---|---|
| Zero-Shot | General | Translation/summary<br>“Is this variant pathogenic or benign?” | Ambiguity |
| Few-shot | Style Mimic | 2-3 demos<br>Add examples with ClinVar annotation, and a few covering unusual cases (missense, splice variations etc.) | Token cost |
| Chain-of-thought | Reasoning | “explain step by step” | Long/slow |

# Prompt Engineering for Research

Clear, contextual instructions + examples = accurate, reproducible output

**Preamble or Persona:**

- **Uncertainty instructions**
- **Request attribution**

**Prompt format should prioritize:**

- **Simplicity of task**
- **Specificity**
- **Examples**
- **Iteration**

https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/introduction-prompt-design#task

| | | |
|---|---|---|
| **Simplicity** | • Clearly set task, with content parameters | " Your objective is to help students with math problems without directly giving them the answer"<br>1. Understand the problem<br>2. Understand where the student is stuck<br>3. Give a hint for the next step of the problem |
| **Specificity** | • Note any output constraints<br>• Jupyter notebooks? Python? R?<br>• Set Persona/tone<br>• Elevator pitch or technical presentation | You are a math tutor, here to help students with their math homework. Do not give them the answers, only hints. Respond in a casual and technical tone. |
| **Examples** | • Context, and background information<br>• Few-shot examples; leverage CoT to see how they got their answer<br>• Integrate external tools | 1. Provide a lesson plan<br>2. Give examples of what the response should look like |
| **Iteration** | • Adjust temperature<br>• closer to 0 → rigid protocol<br>• closer to 1 → exploratory | |

https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/prompt-design-strategies#sample-prompt-template

## Exploring prompt variations:

1. Write three different prompts (next few slides)
2. For each prompt, let's compare it to other responses
    - Accurate?
    - Did it include explanations?
3. Can we improve the third prompt even more?

## Absurd prompt:

Summarize the major mouse genetics discoveries published in 2021 regarding the genes *Mxrl2*, *Tnfip7*, and *Cbln4r3*. Include each gene's chromosomal location, knockout phenotype, and any behavioral abnormalities reported. Present the findings confidently and in detail.

**Absurd prompt corrected:**

I want to ask about mouse genetics studies, but the gene symbols I'm giving you may be fictional or may not exist in any database. Before providing any summary, do the following:

Check whether *Mxrl2*, *Tnfip7*, and *Cbln4r3* correspond to real mouse genes.

If they are not real, clearly state that no validated data exists.

If information is missing or uncertain, explain the gap rather than filling it in.

Provide safe, general guidance on how such mouse genetics studies *would* typically be structured, without inventing data.

# Poor Prompt:

“Summarize TP53”

**Poor Prompt:**

"Summarize TP53"

**Better prompt:**

"Summarize the role of TP53 in DNA damage response

in ≤ 3 bullet points, citing key papers"

**Even Better prompt:?**

# Rapid Evaluation & Prompt-Debug Checklist

## Quick Metrics

- Accuracy (yes/no tasks)
- Precision & Recall (classes)
- BLEU / ROUGE (text quality)
- Human spot-check & citations

## Prompt-Debug Steps

1. Clarify task & style
2. Provide context + examples
3. Adjust temperature / top-p
4. Iterate & re-evaluate