

# The Science of Large Language Models: Introduction

---

**Robin Jia**

CSCI 699 (The Science of LLMs), Fall 2024

August 26, 2024

# Today: People Use LLMs for...

The Verge

ARTIFICIAL INTELLIGENCE / TECH / LAW

## OpenAI sued for defamation after ChatGPT fabricates legal accusations against radio host

ChatGPT is notorious for generating false and misleading information, but this seems to be the first defamation case involving OpenAI.

By James Vincent, a senior reporter who has covered AI for The Verge.

Jun 9, 2023, 3:08 AM PDT

WIRED

ASHLEY BELANGER, ARS TECHNICA

BUSINESS FEB 17, 2024 12:12 PM

## Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

for anything its chatbot says.

We need deeper understanding of how LLMs work and what behavior they're likely to exhibit

REUTERS

## New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By Sara Merken

June 26, 2023 1:28 AM PDT · Updated 9 months ago



## Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

# LLMs are a New Type of Creature in our House

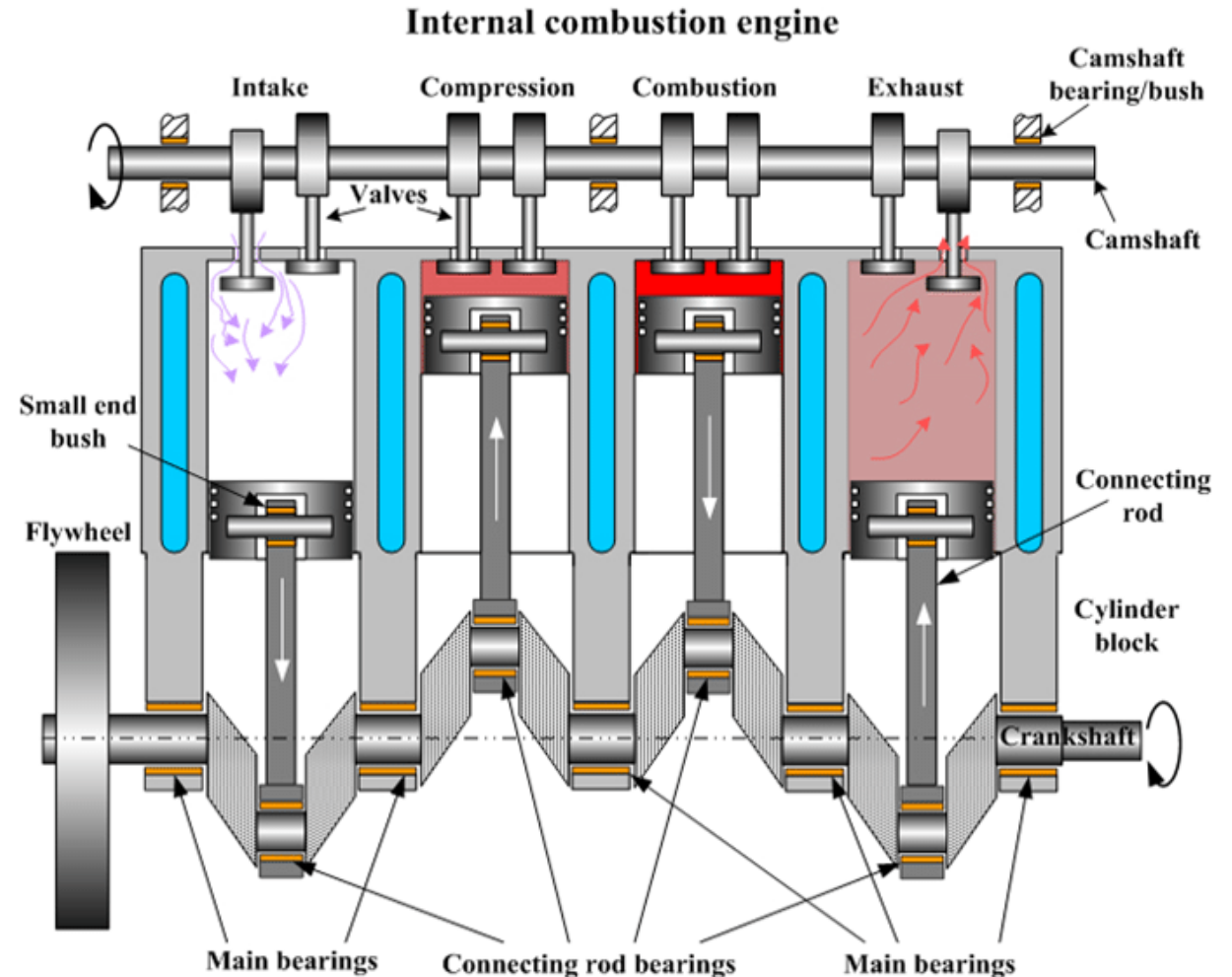


Can we trust this creature?

What sort of creature is this?

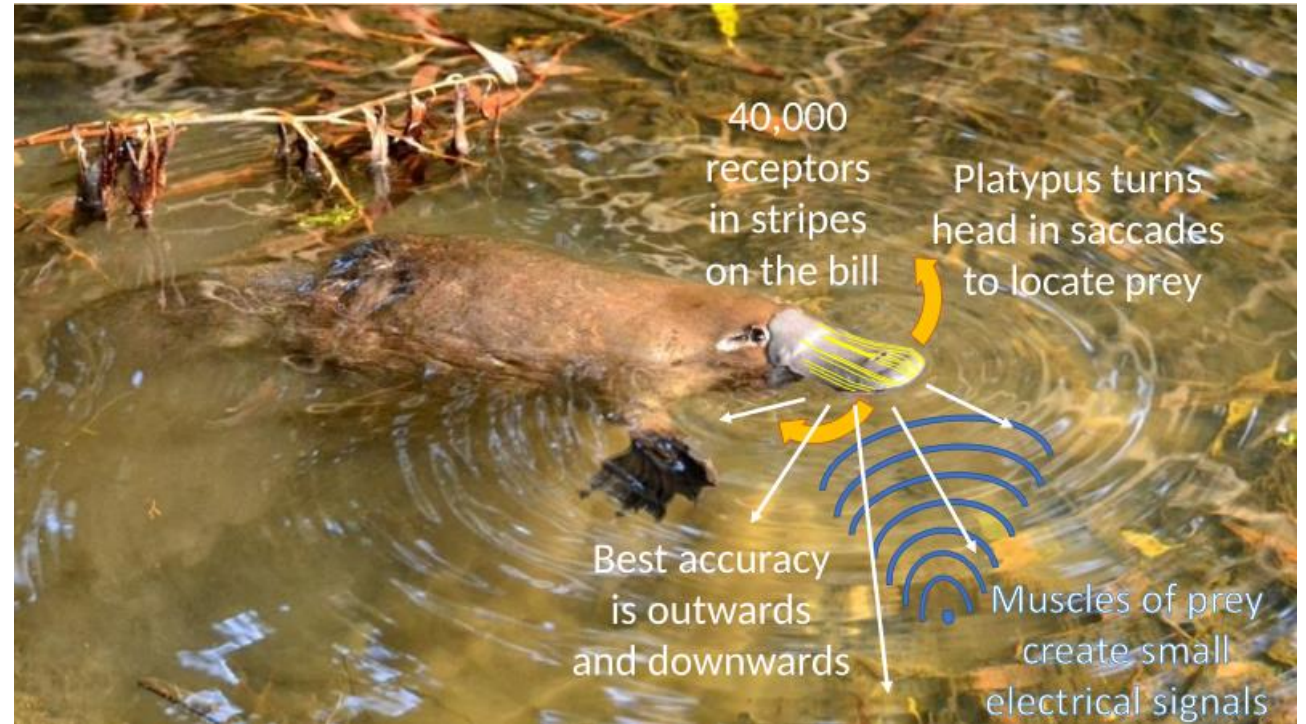
# Why is Scientific Study Needed?

- Intuition: “We created LLMs, therefore we understand them”
- This is tempting but wrong!



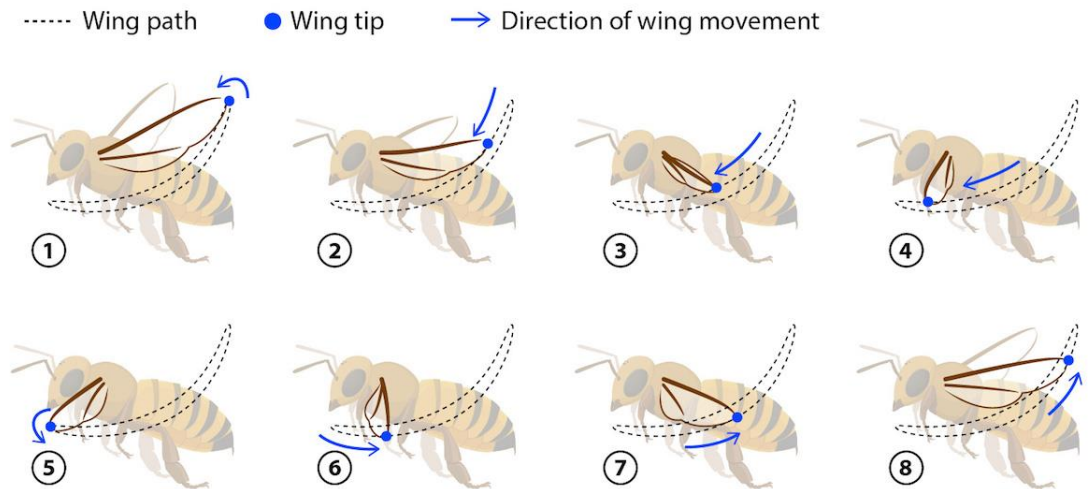
# Why is Scientific Study Needed?

- “We wrote the model architecture and the learning algorithm”
  - But that doesn’t mean we understand the result of model training
  - Analogy: Just because we understand how evolution works, doesn’t imply that we understand the things it creates

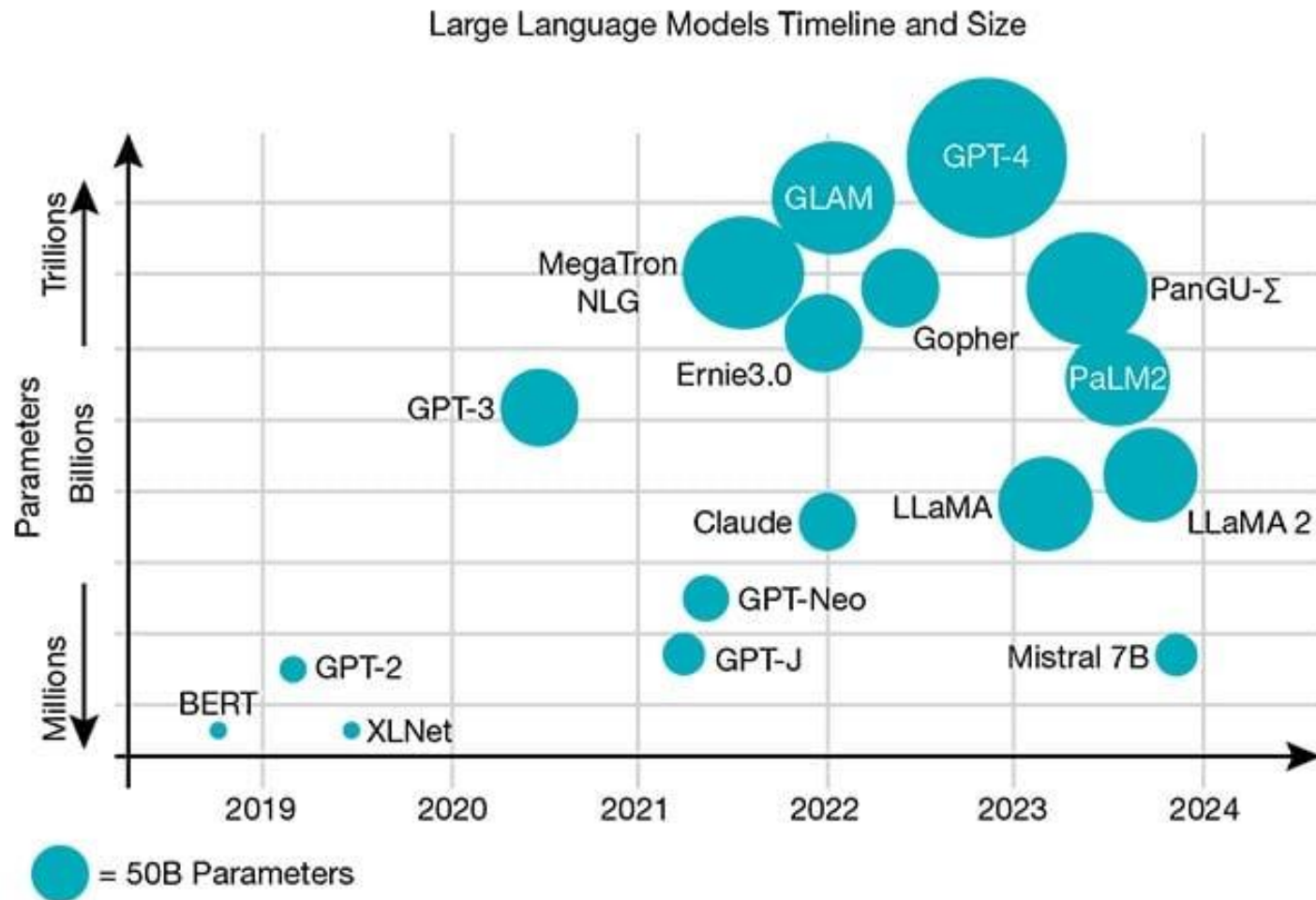


# Why is Scientific Study Needed?

- “We know the low-level numerical operations the model performs”
  - But it is hard to conceptualize why those lead to observed behavior
  - Analogy: Understanding the laws of physics does not immediately tell us everything about biology



# Growing Scale, Growing Complexity



- Modern LLMs are growing in scale, complexity, and capabilities
- More behaviors to study!

# Today's Agenda

---

- Course logistics and format
  - Role playing seminar format
  - Final project
- Three ways of scientifically studying large language models
  - Analyzing internals of LLMs
  - Understanding black-box LLM behavior
  - Examining external forces on LLMs



# Course basics

---

- Instructor: Robin Jia
  - Office Hours: Thursdays 11am-12pm in SAL 236
- TA: Johnny Wei
  - Office Hours: Mondays and Fridays 3-4pm in MCB lobby
- Course website: <https://robinjia.github.io/classes/fall2024-csci699.html>
  - All information is on the website
- General discussions: Slack channel
- Personal questions: Office hours or email (if you Slack me, I may forget)
- Class Zoom link for screen sharing: Posted on Slack, also on Brightspace
  - No remote attendance allowed—must attend in person!
- Assignments: Submit on Brightspace

# Should I take this class?

---

- If you have not taken a modern NLP class...
  - **Take CSCI 544**
  - By “modern” I mean, focused on Transformers and language models
  - CSCI 544 material has been updated **this semester** to focus almost exclusively on language models
  - Will provide a broader introduction to modern NLP
- If you are primarily interested in *applications* of LLMs...
  - **Also take CSCI 544**
- I am happy to have non-enrolled students sit in on presentations, subject to availability of physical space

# Should I take this class?

---

- You should take this class if...
  - You are actively pursuing research on LLMs **and**
  - You are experienced reading NLP research papers **and**
  - You are interested in understanding the low-level details of how LLMs work

# Course Schedule

---

- Today: Introduction
- Next 3 classes: Lectures on the Nuts and Bolts of LLMs
  - Transformer architecture review
  - Deep dive on Llama 3 paper
- Next 18 classes: Role Playing Seminars on Research Papers
- Final 5 classes: Final project presentations

# Role Playing Seminars

---

- Most class days will be *role playing seminars*
- For each main paper (~2 per day), different students will:
  - Pitch the work (as if writing a research proposal)
  - Present the methods and results
  - Search for and compare with prior work
  - Write a review of the paper
  - Brainstorm future work
- When there are background or bonus papers:
  - Summarize the paper
  - Draw connections with the main papers

# Choosing Your Role

---

- Each student must play **6 different roles on 6 different days**
  - One of these roles must be a “Main Presenter” role
- I will send out a spreadsheet next week
- First paper seminar class is Wednesday of Week 3 (September 11)
- **If you are enrolled, please decide by end of Week 2 (Friday, September 6) if you want to stay in the class, and drop the class ASAP**
  - This allows someone else to enroll and sign up for roles in time!

# Roles: Grading and Expectations

---

- In total, role playing accounts for **50% of the total grade**
- Written reports (25% total)
  - Generally, 1-2 page reports that reflect what you will talk about in class
  - Due at the start of class (4pm) on the day you are playing that role
- In-class presentations (25% total)
- Each of the 6 roles is weighted equally
  - Each written report is  $25\%/6 = 4.167\%$  of your overall grade
  - Each presentation is also  $25\%/6 = 4.167\%$  of your overall grade

# Main Paper Role: Proposer

---

- Role: Proposes the research in the paper to a funding agency
- Learning goal: Can you write a convincing research proposal? (e.g., for a fellowship application or grant)
- Write-up: 1-2 page report answering Heilmeier Catechism questions:
  - What are you trying to do? Articulate your objectives using absolutely no jargon.
  - How is it done today, and what are the limits of current practice?
  - What is new in your approach and why do you think it will be successful?
  - Who cares? If you are successful, what difference will it make?
  - What are the risks?
  - What are the mid-term and final “exams” to check for success?
- In class: ~5 minute slide presentation (one slide per question)



# Main Paper Role: Main Presenter

---

- Role: Present the paper's methods and results
- Learning goal: Can you present technical material? Required for conference talks, qualification exam, thesis defense, etc.
- Write-up: Submit your slides, no separate write-up
- In-class: ~15-20 minute slide-based presentation of the paper's methods and results
  - What is the problem setting?
  - What are the relevant baselines?
  - What is the main method proposed by this paper?
  - What are the main experiments and results?
  - What conclusions can be drawn from the results?

# Main Paper Role: Archaeologist

---

- Role: Compare/contrast current paper with prior work
- Learning goals:
  - Writing a related work section: Important not just to summarize related work, but to **relate** it to the current paper
  - Doing a thorough literature review
- Write-up: 1-2 page report on 3+ related prior papers
  - What does the prior work do?
  - Similarities/differences with current paper
  - In what ways is the current paper “novel” compared to this previous paper?
  - Does the prior work and the current work come to similar or different conclusions?
- At least one prior paper must **not be cited by current paper**
- Presentation: 3 minute oral presentation

# Main Paper Role: Reviewer

---

- Role: Write a review of the paper
- Learning goal: Practice for being a reviewer
- Write-up: A 1-2 page review in the format of ACL Rolling Review
  - Paper Summary
  - Summary of Strengths
  - Summary of Weaknesses
  - Comments/Suggestions
  - Soundness score (1-5)
  - Overall assessment (1-5)
- Presentation: ~3 minute presentation of your review.
  - Share your screen and scroll through the PDF of the paper.

# Main Paper Role: Visionary

---

- Role: Brainstorm follow-up research and products
- Learning goal: Planning a research direction and envisioning downstream impact
- Write-up: 1 page report giving two different future work ideas:
  - Future research: What questions, and how to answer them?
  - Future products: Envision a use case for a product based on this research
- Presentation: ~3 minute oral presentation of your future work ideas

# Bonus/Background Paper Role: Summarizer

---

- Role: Summarize a background/bonus paper
- Learning goal: Presenting a paper at a higher level (shorter conference talk or poster presentation)
- Write-up: Submit your slides, no separate write-up
- Presentation: 5-10 minute summary of the paper:
  - What is the goal of this paper?
  - At a high level, what is the paper's methodology?
  - What are the main experiments and results of the paper?
  - What conclusions can be drawn from the results?

# Bonus/Background Paper Role: Connector

---

- Role: Draw connections between bonus/background & main paper
- Learning goal: Practice making connections between different papers, often from different times
- Write-up: 1 page description of the connections between this paper and all main papers for that day:
  - What themes does this paper share with the main papers?
  - How is this paper different from the main papers?
  - (Background) How does it provide context for the main papers?
  - (Bonus) How does it enhance our understanding of the main papers?
- Presentation: ~3 minute oral presentation.

# Llama-3 Role: Re-examiner

---

- Role: Investigate how Llama-3 handles the current topic
- Learning goal: Understand how class concepts are applied in a real state-of-art language model
- Write-up: 1 page report discussing:
  - Key challenges discussed in the day's main papers
  - How Llama 3 deals with those challenges
  - Does this seem like a good choice, given our readings?
- Presentation: ~3 minute oral presentation

# Concluder Role

---

- Role: Summarize relationships between all main papers
- Learning goal: Read multiple papers in a research paper and draw overall conclusions
- Write-up: 1-2 page report describing how the main papers are connected:
  - Shared themes?
  - How do the papers support each other?
  - How do the papers disagree or present different narratives?
  - If there are disagreements, which side do you find more convincing, and why?
- Presentation: ~3 minute oral presentation of these connections
  - If there are disagreements between papers, choose a side!



# Responsibilities outside of roles

---

- Students are expected to participate in discussion even on days when they have no assigned role
  - Asking questions of presenters
  - Weighing in on discussion topics
- Class participation (excluding your role-related duties) is worth **10%** of overall grade
- Reading the (main) papers ahead of time is strongly encouraged!
  - If you're too busy to read them fully, you can practice skimming papers—a useful skill!

# Final Project

---

- Worth **40%** of the overall grade
- Can be done individually or in groups up to 3
- Expectations
  - Novel research conducting some scientific study of language models
  - Must include study of some open-weight language model
  - Can also compare with closed language models

# Final Project Timeline

---

- All due dates here are 11:59pm
- See class website for more details
- Project Proposal (2 pages, 5% of grade): Due September 27
  - Specify the goal, identify related work
- Midterm report (4 pages, 10% of grade): Due November 1
  - Have some initial results
- Project presentations (10% of grade): November 18 through December 4
- Final report (6-8 pages, 15% of grade): Due December 6
  - Similar in structure and content to a \*CL conference paper

# Late Days

---

- You are given **3 free late days** to use in integer amounts on any written assignment **except the final project report**
  - Role-related written reports
  - Project proposal and midterm report
- Note: For role-related reports, you must still present in class on the scheduled day!
- If working in a group for project, **every group member** must use a late day to turn in a project report one day late

# Generative AI Policy

---

- In this class, we are studying LLMs...but I expect you to conduct all work by yourself!
- **Greater learning experience** by relying on your own brain rather than using ChatGPT/Claude/etc.
- **Not allowed**: Using LLMs to generate any part of a written report (for roles or final project)
- **Not allowed**: Using LLMs to generate code used in your final project

# Halftime break!

---



# LLMs as Objects of Scientific Study



**How do we study this new creature?**

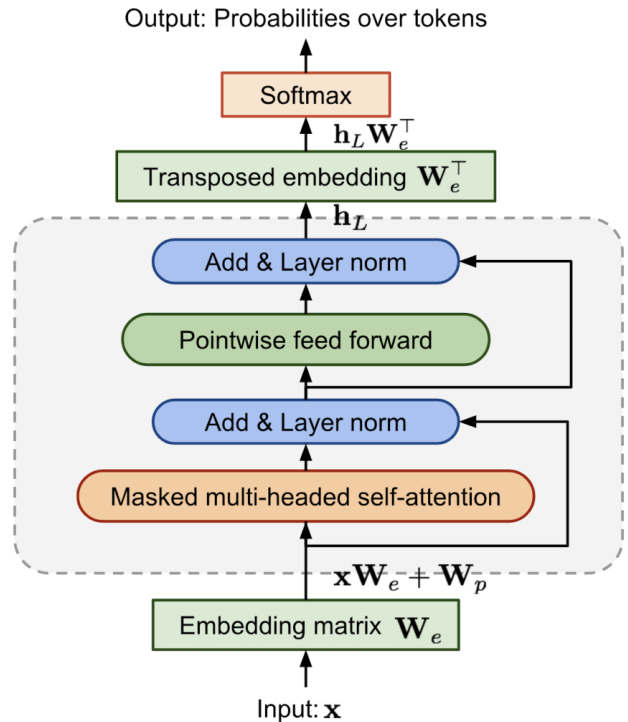
# The Science of LLMs

---

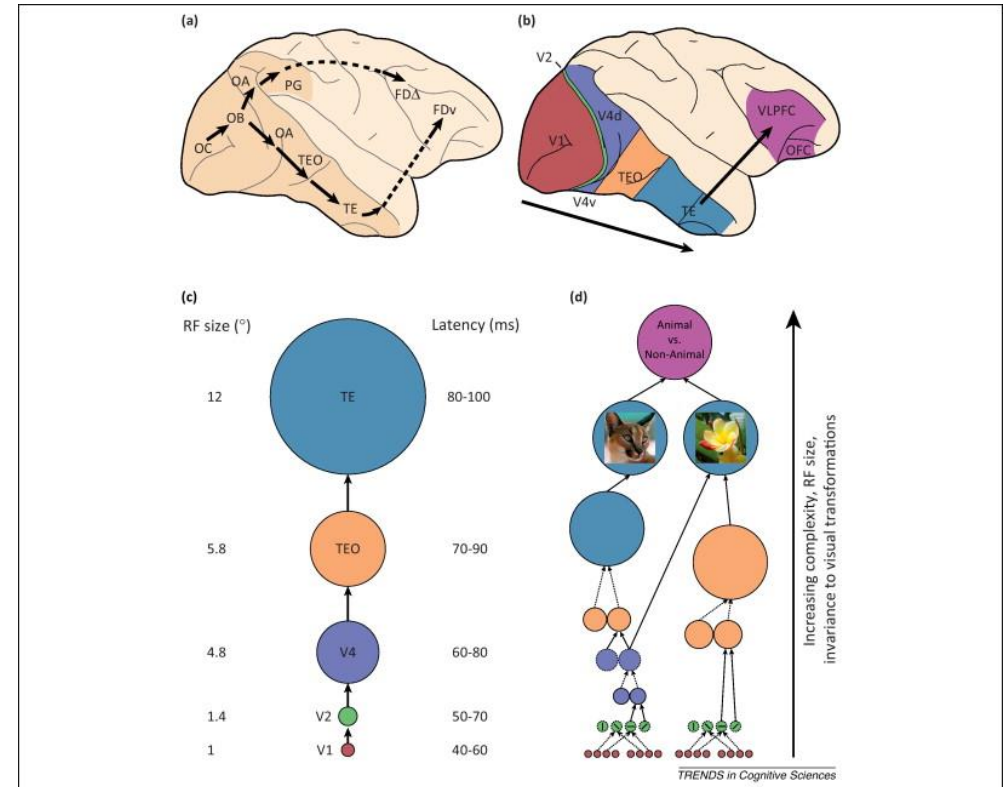
- Part 1: Analyzing internals of LLMs Neuroscience?
- Part 2: Understanding black-box LLM behavior Experimental Psychology?
- Part 3: Examining external forces on LLMs Social/  
Developmental Psychology?



# Analyzing Internals

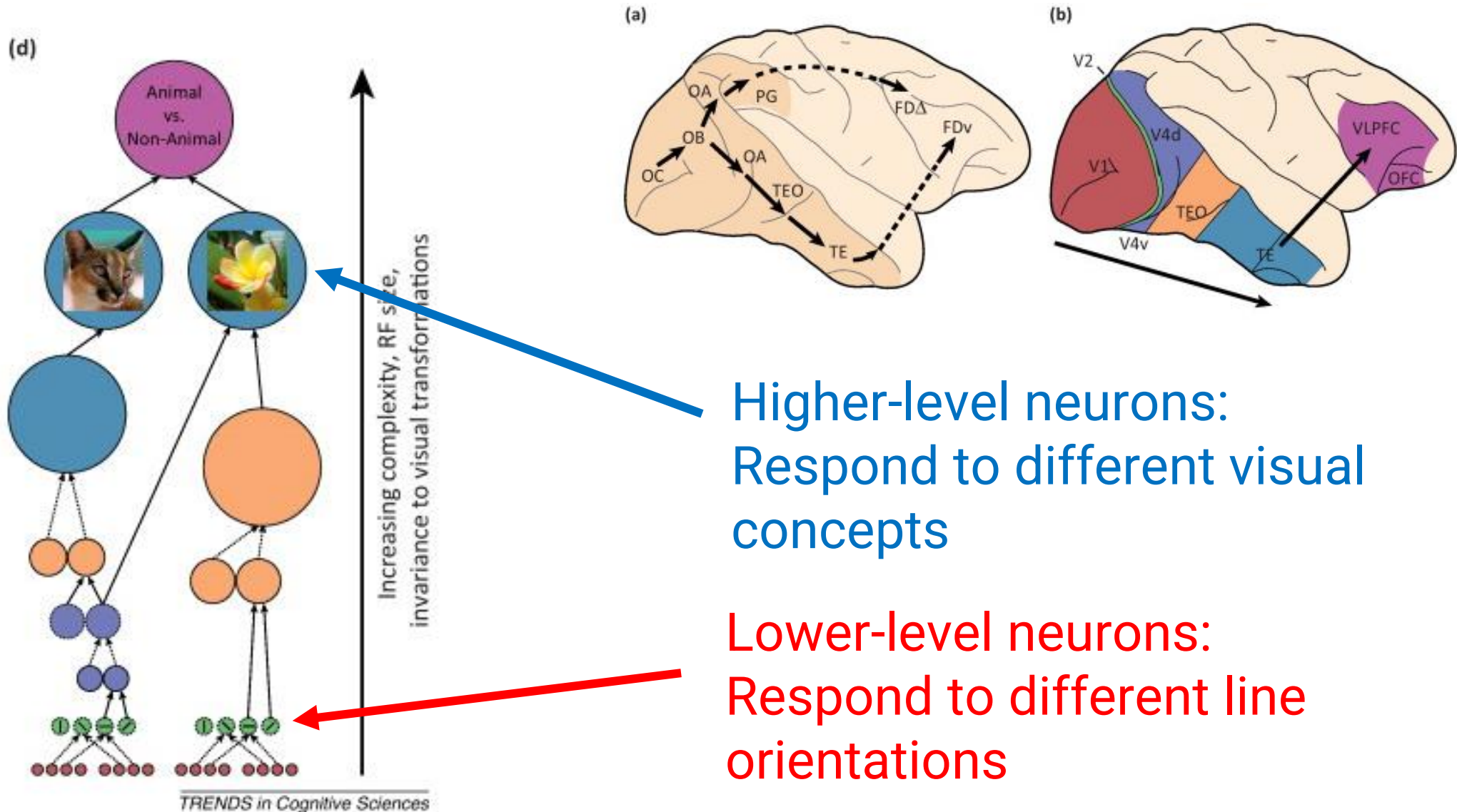


NLP: Analyze hidden states & layers of Transformer-based LLM



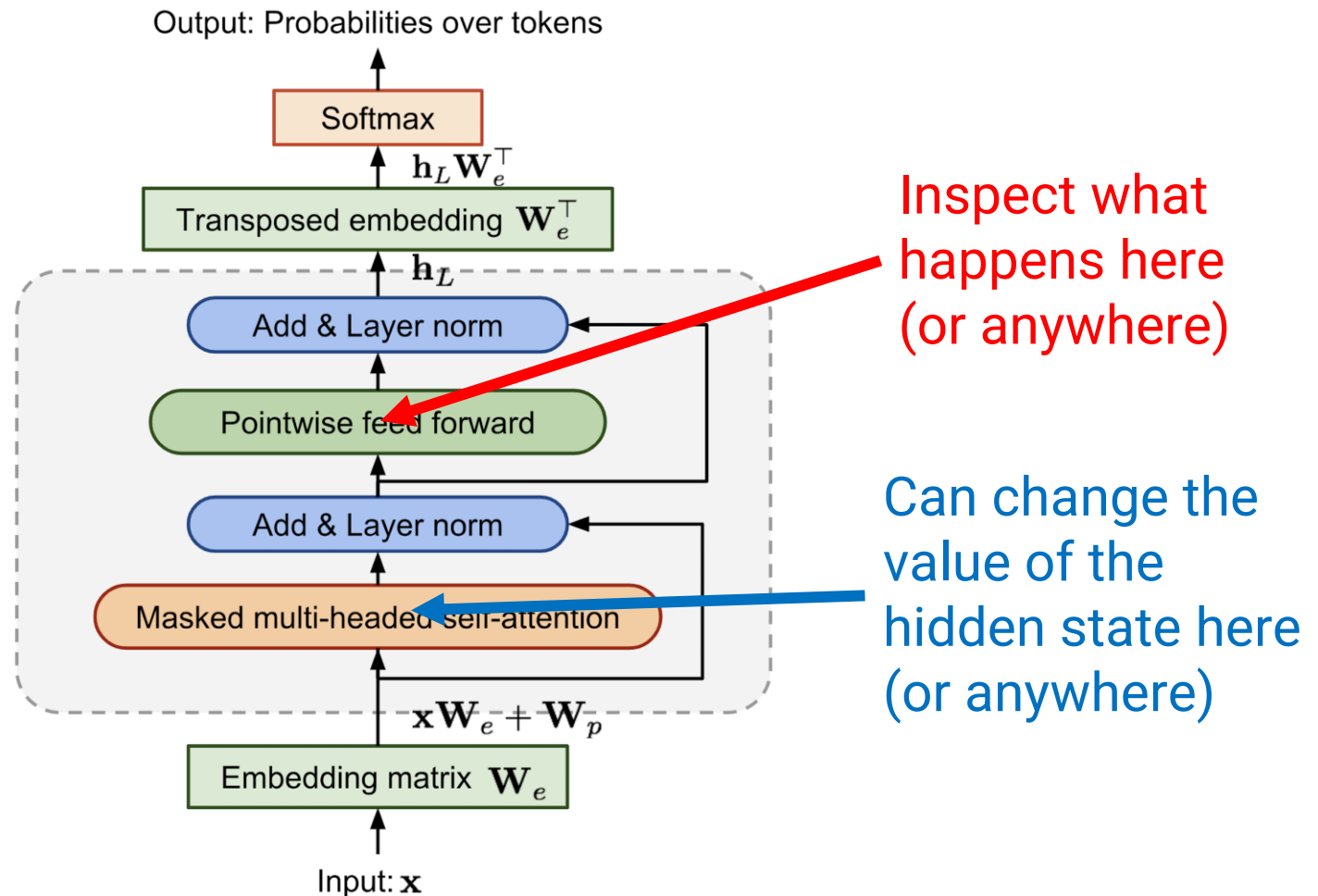
Neuroscience: Analyze function of different parts of brain

# Analyzing Internals



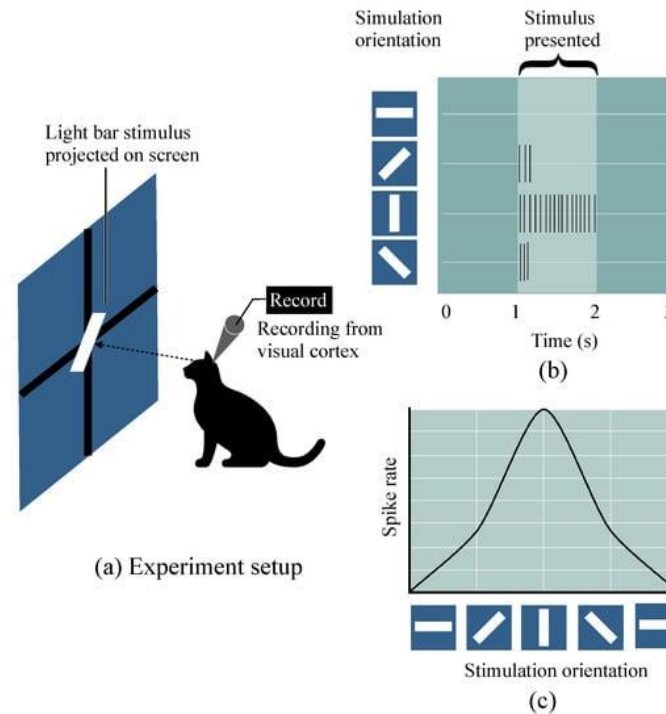
# The Luxuries of Computer Science

- **Read access:** We can measure anything!
- **Write access:** We can intervene on the model's operations at any point!

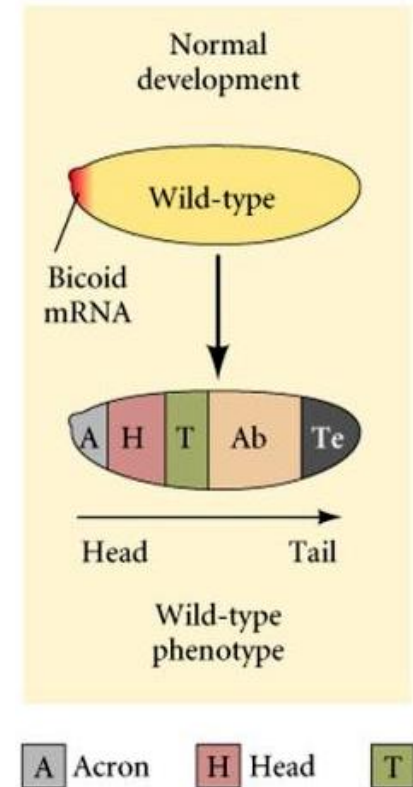


# Taking measurements

- Measure relationship between **stimuli and internal components**
  - We can change the stimuli at will to establish a **causal relationship**
- Measure relationship between **internal components and downstream behavior**
  - **Here causation is less clear!**



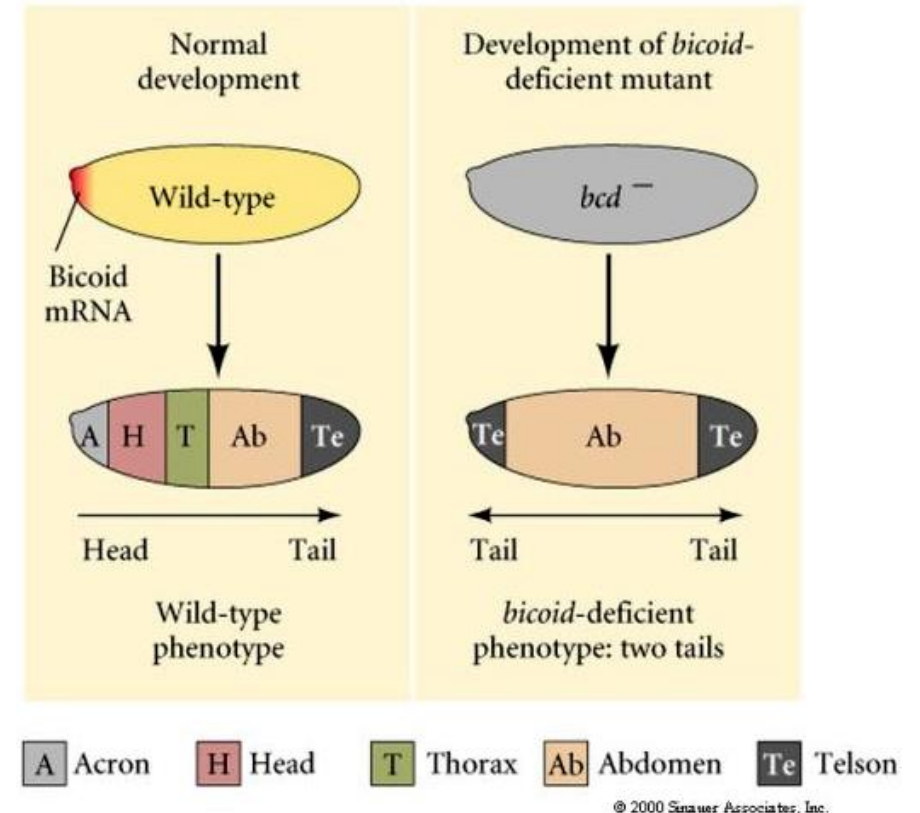
Recording activity from one neuron in a cat's brain while showing different stimuli



Showing that bicoid protein correlates with location of head in a fruit fly embryo

# Loss-of-function (“knockout”) experiments

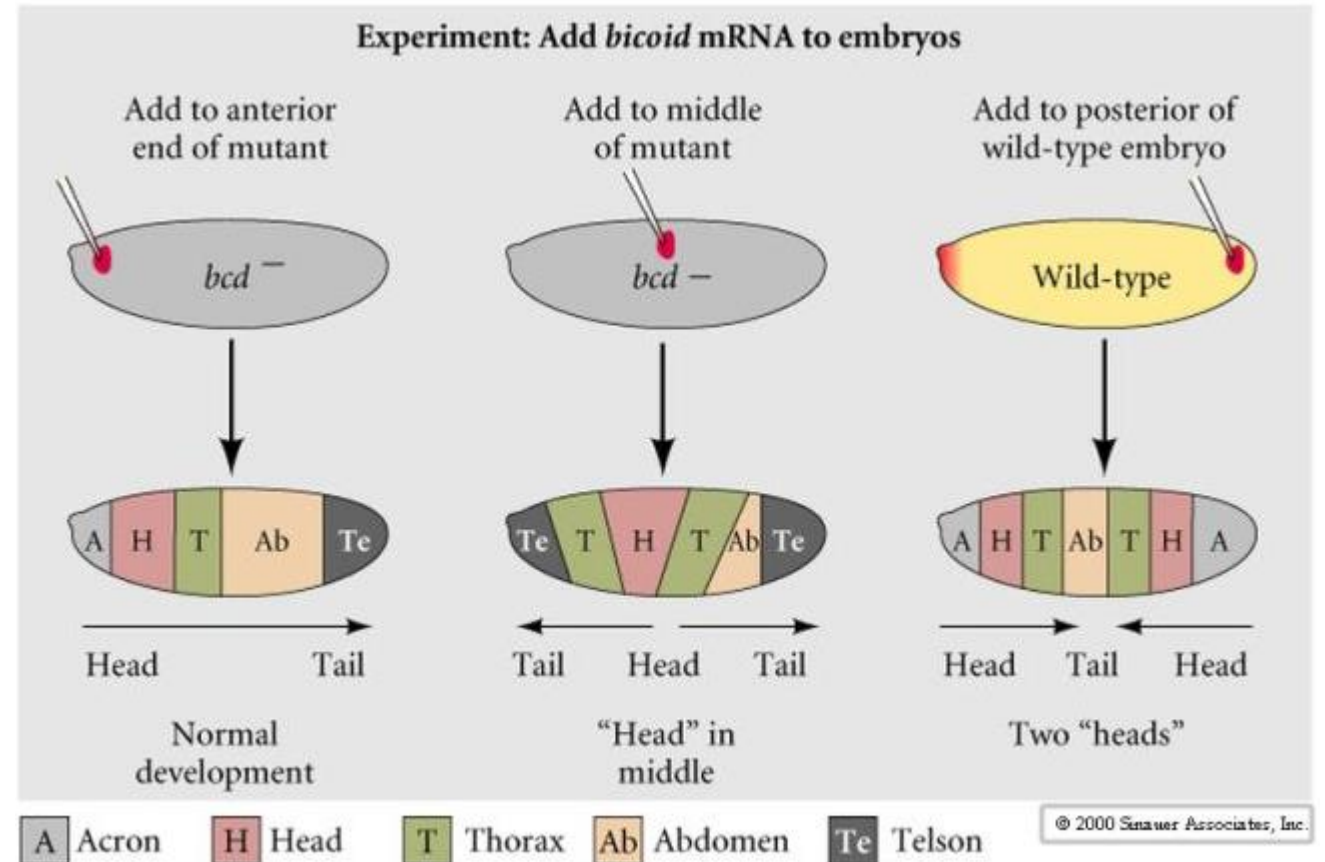
- You hypothesize that component X is important to behavior Y
- Correlation does not imply causation!
  - Bicoid is located where the head develops, but does it *cause* the head to develop there?
- Strategy 1: Loss-of-function (“knockout” or “ablation”) experiments
  - Remove component X from your system
  - If behavior Y no longer happens, what component X must have been **necessary** for behavior Y!



Mutated flies with no bicoid gene develop no head and two tails!

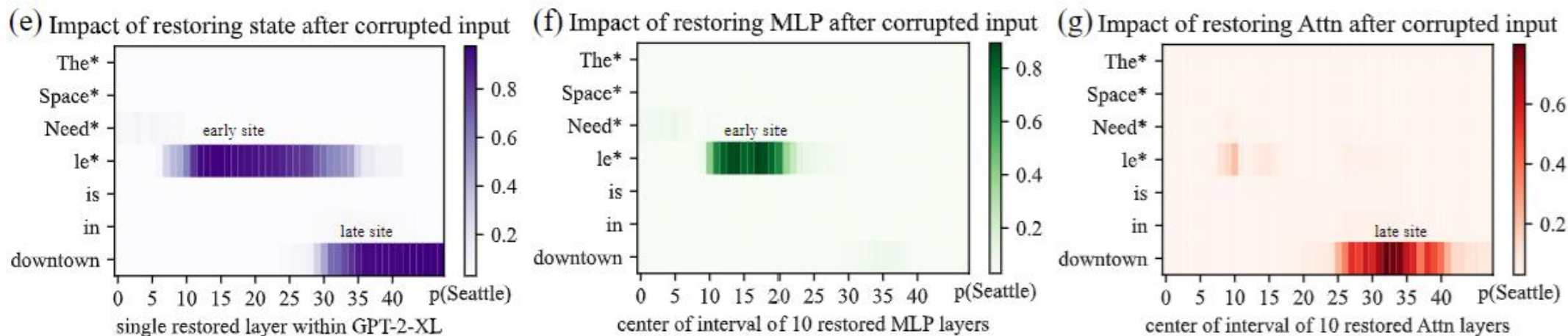
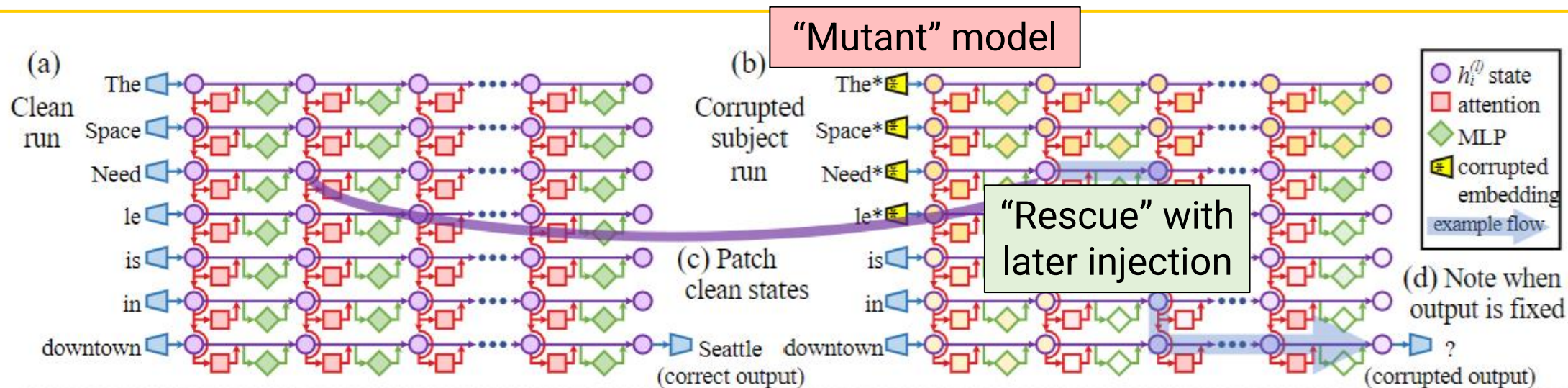
# Gain-of-function (“knock-in”) experiments

- You hypothesize that component X is important to behavior Y
- Strategy 2: Gain-of-function (“knock-in” or “rescue”) experiments
  - Start with a system that does not have component X
  - Add component X in
  - If behavior Y now appears, component X must have been **sufficient** for behavior Y!



Injecting the *bicoid* gene back “rescues” the mutated fly.  
Injecting extra *bicoid* gene results in 2 heads!

# Example Paper: Localizing knowledge



# Understanding Black-box Behavior



ChatGPT ▾



You are sitting in a room with a basketball and a soccer ball. A blanket is placed over the basketball. How many balls are in the room now?



The number of balls in the room remains the same—there are still two balls, the basketball and the soccer ball. The blanket covering the basketball doesn't change the total number of balls.



NLP: Query LLM with different examples and observe behavior

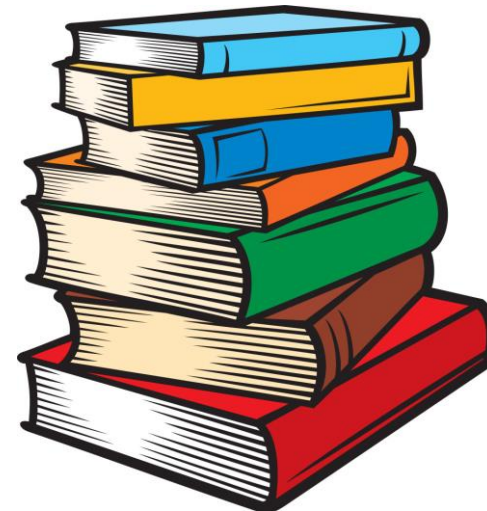
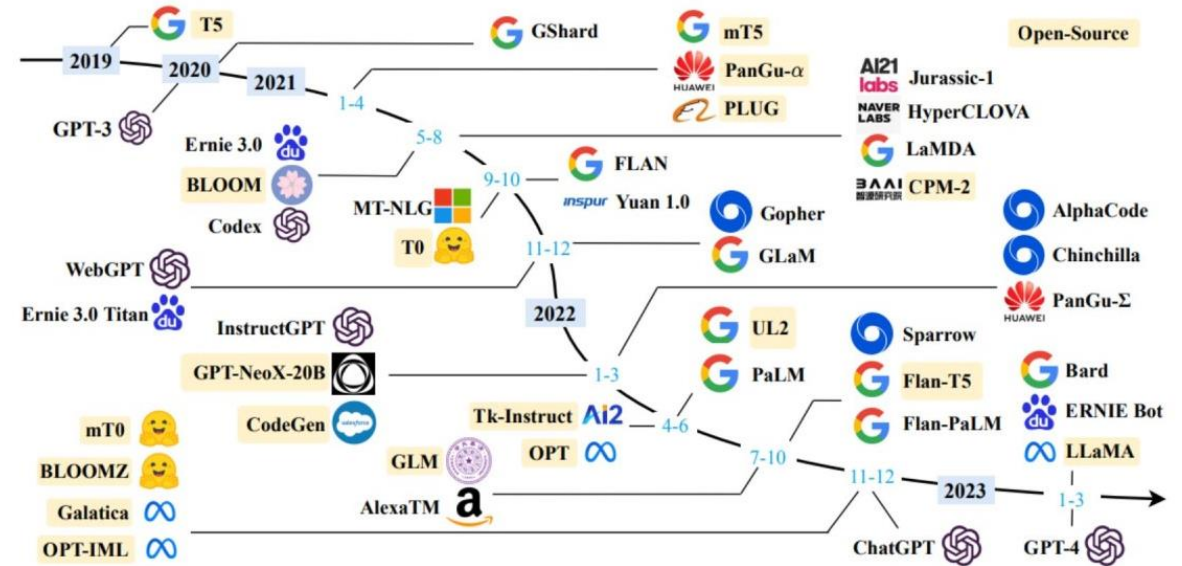


Experimental Psychology: Pose different situations to participants, observe their behavior

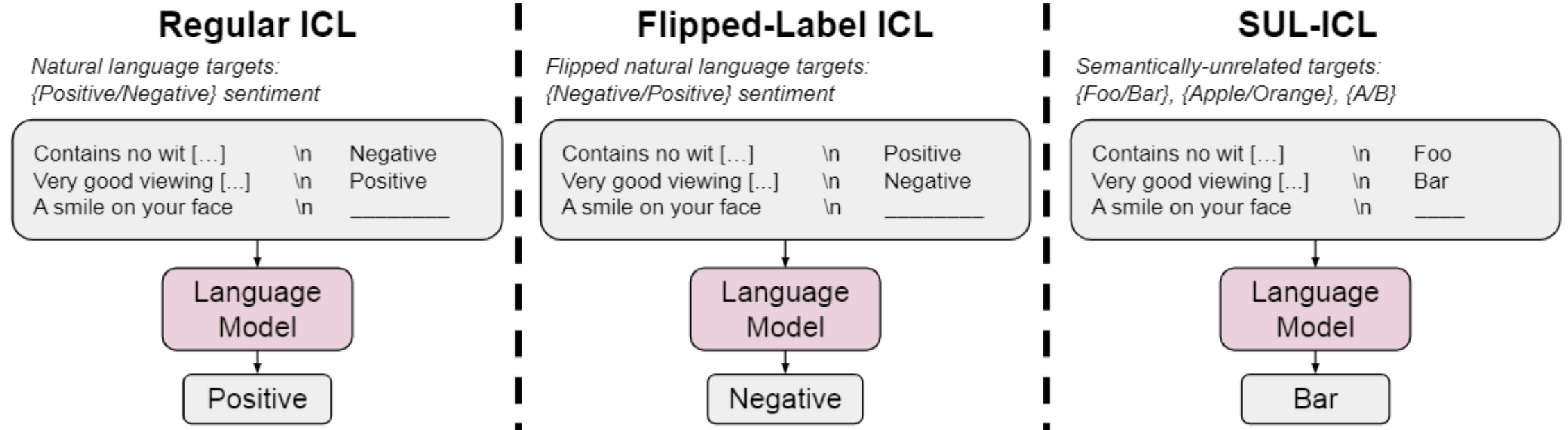


# The Luxuries of Computer Science

- We can query any model we want!
  - Psychology studies: Have to recruit participants, often college students
- We can query it on as many different inputs as we want can afford!
  - Psychology studies: Test limited number of scenarios per experiment



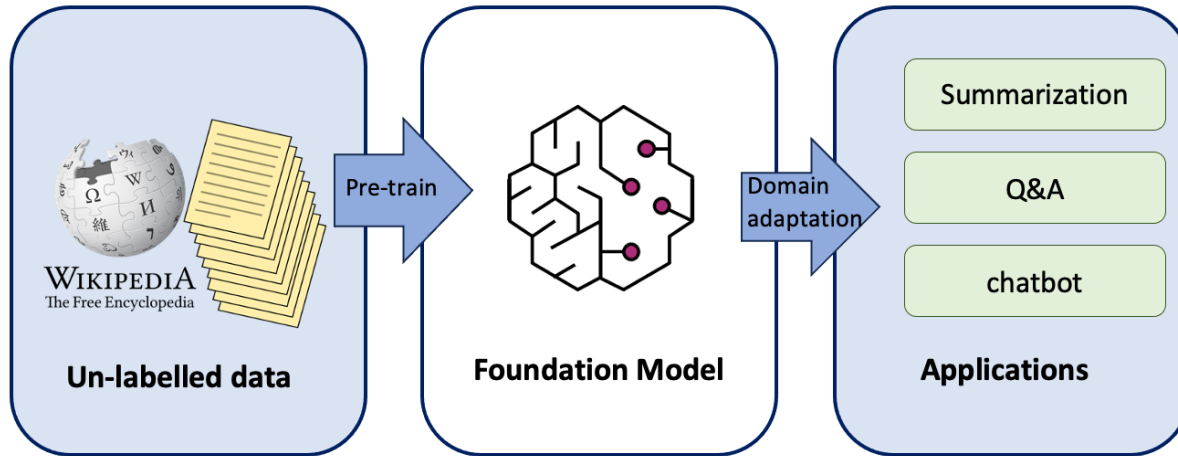
# Example Paper: In-context learning & Scale



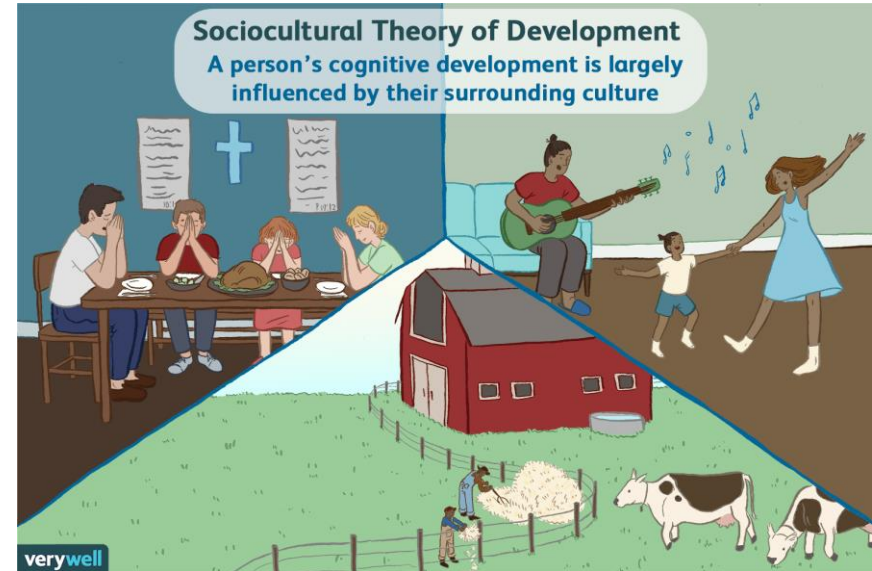
**Research question: How does ability to handle these different types of in-context learning change with model scale?**

- Query many different models
- Query many different examples

# Examining External Forces on LLMs



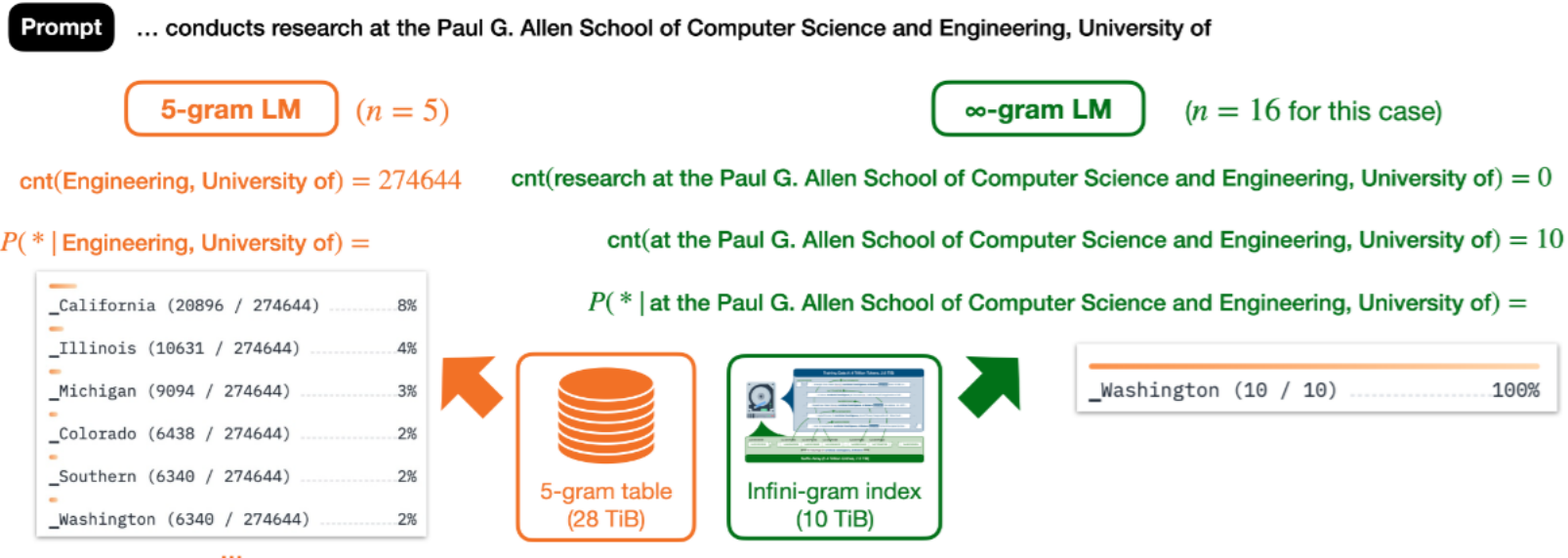
NLP: Understand LLMs as products of a pipeline that involves massive amounts of training data, tokenization decisions, etc.



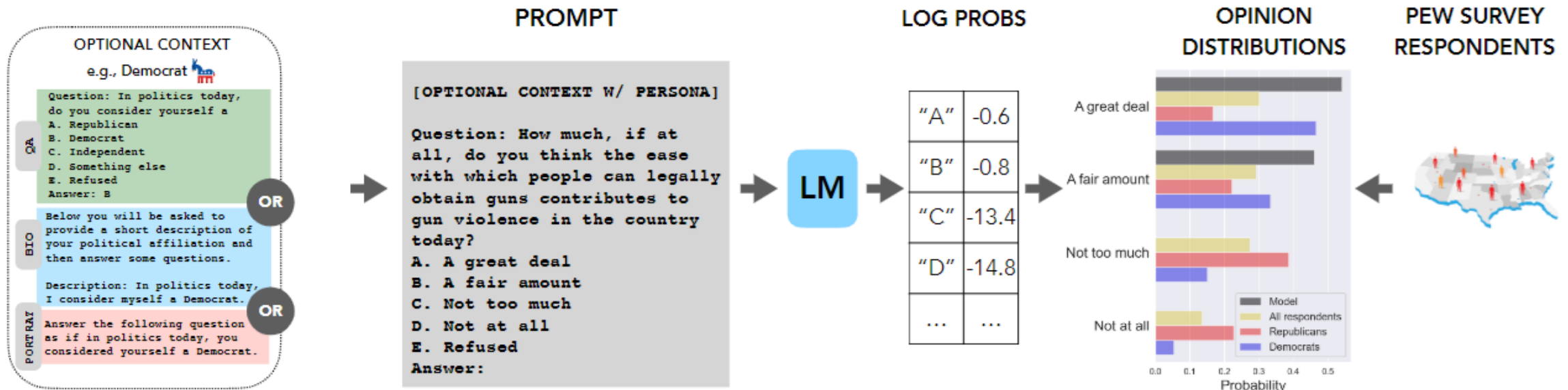
Social/Developmental Psychology: Understand how people are influenced by social/developmental factors that shape their behavior

# The Luxuries of Computer Science

- We (sometimes) know exactly what data the model has seen!
  - Data is the main “societal force” acting on LLM behavior
- We can easily inspect same model at different stages of training (“development”)



# Example Paper: Tracing LLM “Opinions”



- Measures LLM “opinions” using Pew research survey
- How does this correlate with opinions in LLM training data?

# Summary

---

- Course goal: Arrive at deep understanding of large language models via three types of scientific inquiry
  - Analyzing how LLMs work internally
  - Drawing conclusions from LLM black-box behavior
  - Understanding how LLMs are influenced by training data & other factors
- Class format
  - Role playing seminars
  - Final project