

# Bridging NLP & Survey Research: On LLMs, Language and Trust - An NLP researcher's perspective

**Barbara Plank**

LMU Munich  
& IT University of Copenhagen

October 7, 2024

**SurvAI Day: NLP meets Survey Science**  
University of Maryland



IT-UNIVERSITETET I KØBENHAVN

**mcmml**  
Munich Center for Machine Learning

# NLP today: LLMs everywhere!

---



# LLMs: A Swiss Knife for Science?



# NLP: The beauty & challenge of working with **LANGUAGE**

---

"Asking a Question Can Be a Science "  
Frauke Kreuter

# Language is **ambiguous**

---

You said you were looking for some **mixed nuts**?



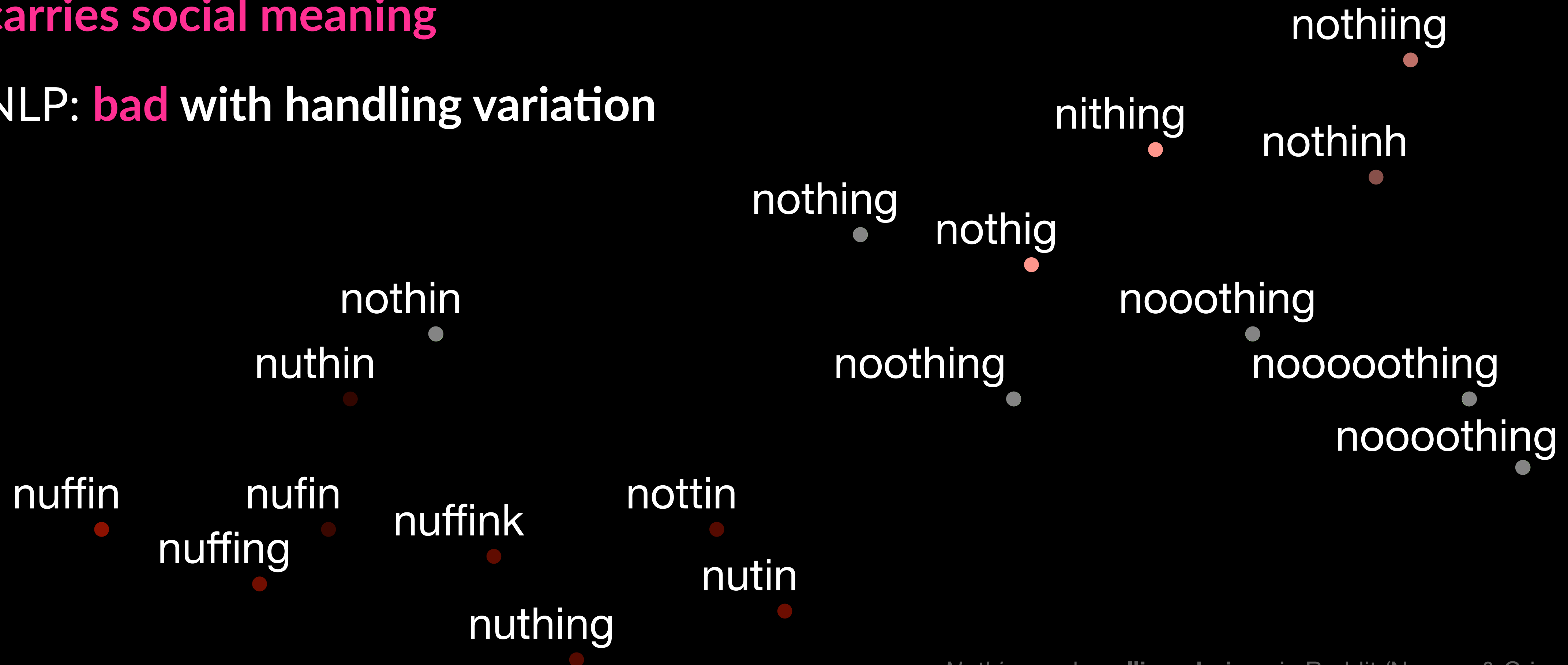
**You**

sort by number of employees: Blackberry, Mango, Apple.

# Language is **full of variation**

- ▶ The way we express a message **carries social meaning**
- ▶ NLP: **bad** with handling variation

You said **nothing**?



# Language is **dynamic and constantly changing**

How to sunny-day Saturday in Seattle:

- ✓ pop out of bed and fling open the drapes
- ✓ brew coffee ☕ and **grab your go-cup**
- ✓ get outside asap
- ✓ dog walk, hike, run, bike, kayak, sail
- ✓ 🍺 soak up the sun in your favorite beer garden
- ✓ 😊 👍 ✓



2



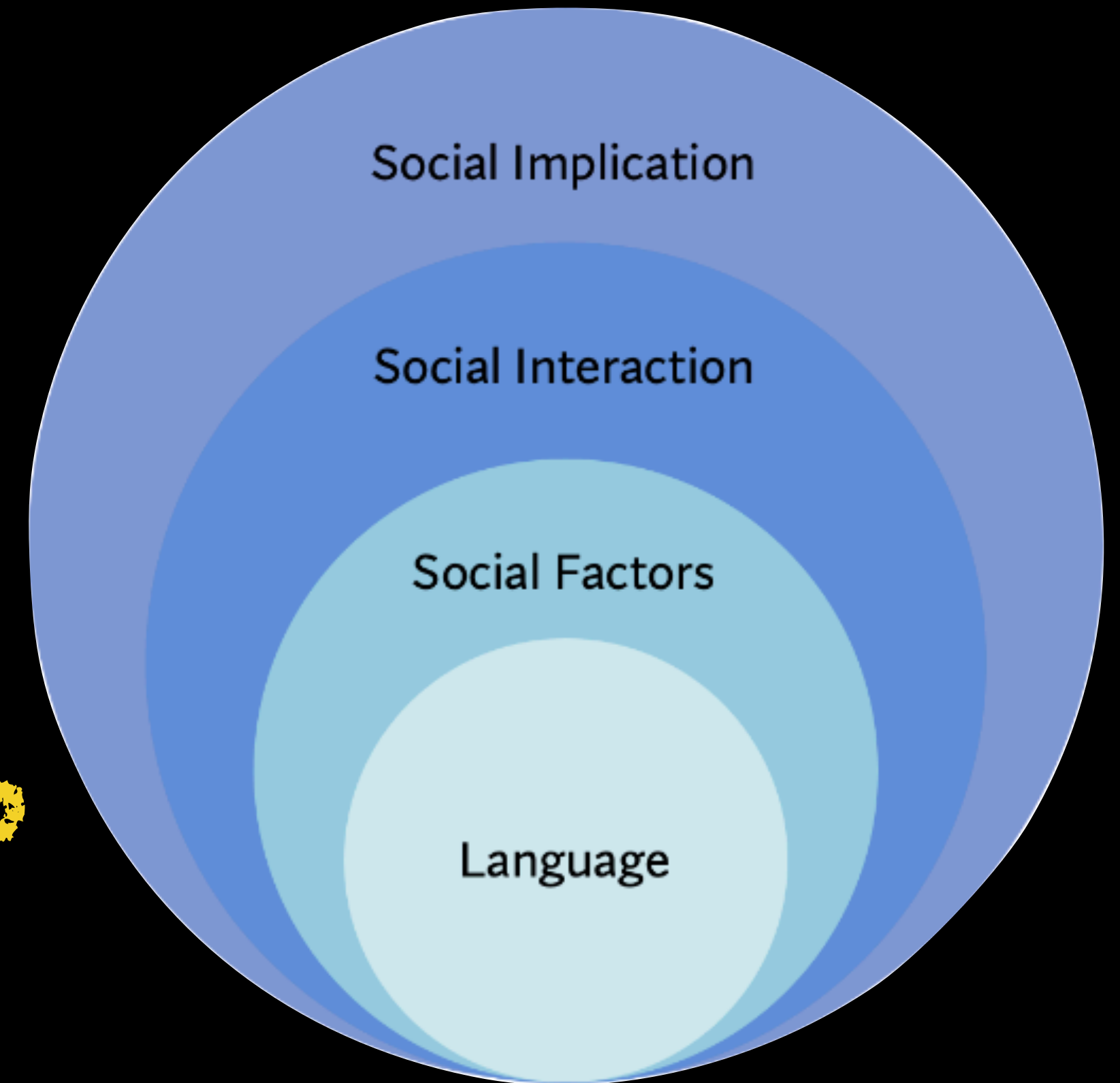
22



Cambridge Dictionary has revealed **its word of the year** for 2023 is **'hallucinate'**, as the term got a new additional definition relating to artificial intelligence (AI) producing false information.

# Language is **for and by people**

“The common misconception [is] that language use has primarily to do with words and what they mean. It doesn't. It has primarily to do with people and what they mean.



**Socially aware NLP**

## The Call for Socially Aware Language Technologies

**Diyi Yang**

Stanford University  
diyiy@stanford.edu

**Dirk Hovy**

Bocconi University  
mail@dirkhovy.com

**David Jurgens**

University of Michigan  
jurgens@umich.edu

**Barbara Plank**

LMU Munich  
bplank@cis.lmu.de



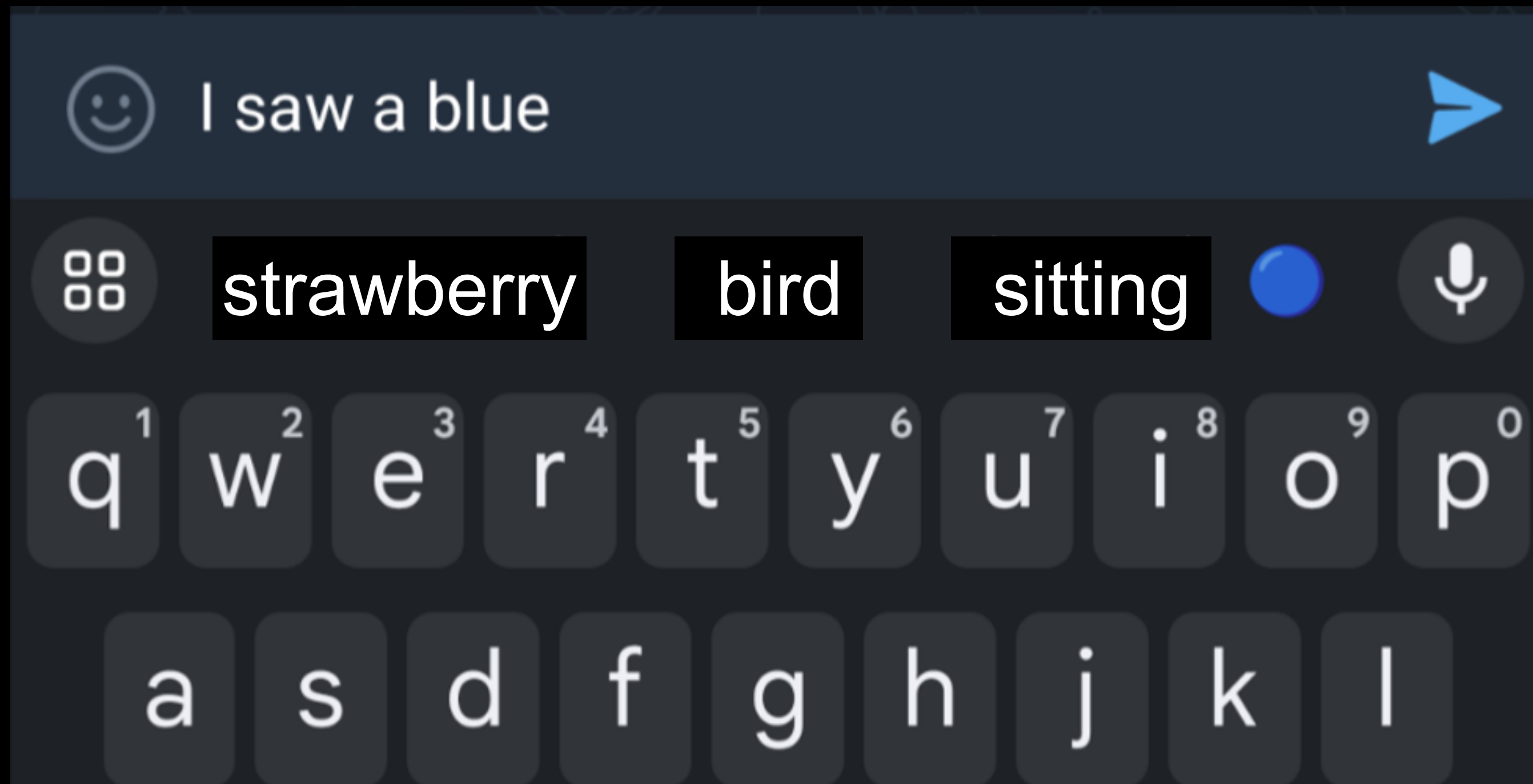
**What Can We Learn From Each Other?**

# Roadmap

- 1 Past: LLMs & Trust - How Did We Get There?
- 2 Present: Trust Issues with LLMs
- 3 Future: Trustworthy Human-Facing NLP

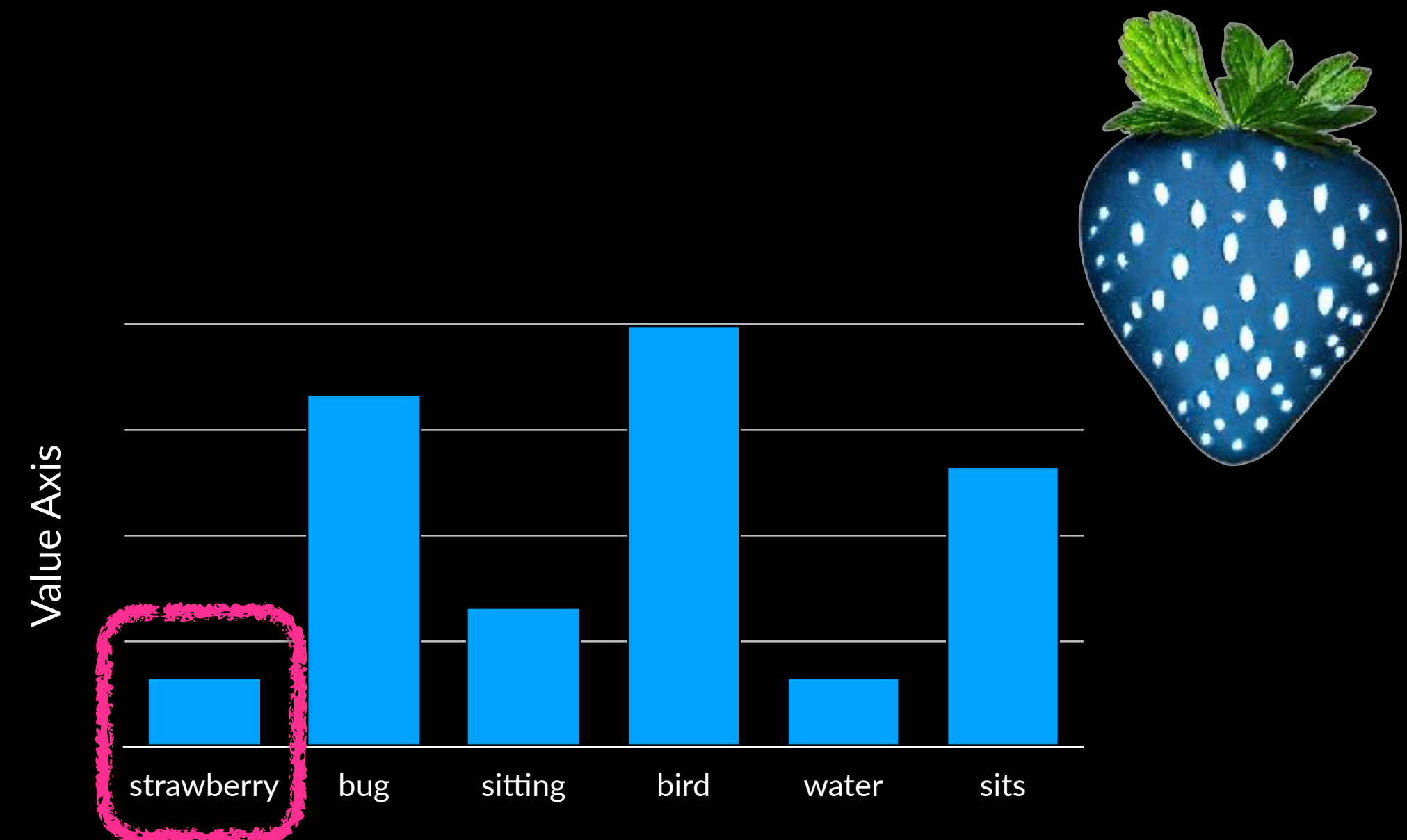
# A Language Model - The most likely text completion

- ▶ A LM computes the probability for a word given its previous words (=context)



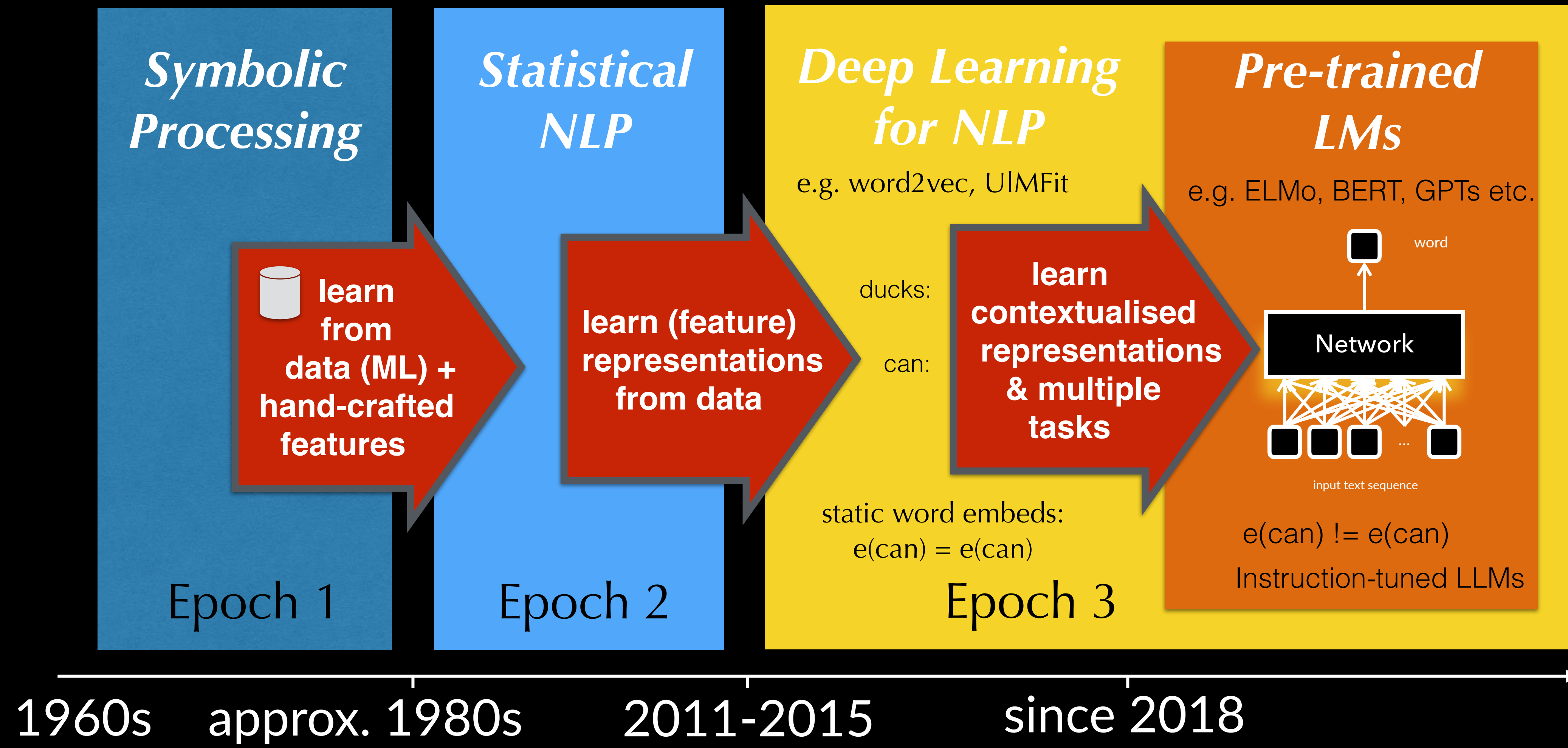
context (or history)

$$P(\text{"strawberry"} \mid \text{"I saw a blue"})$$



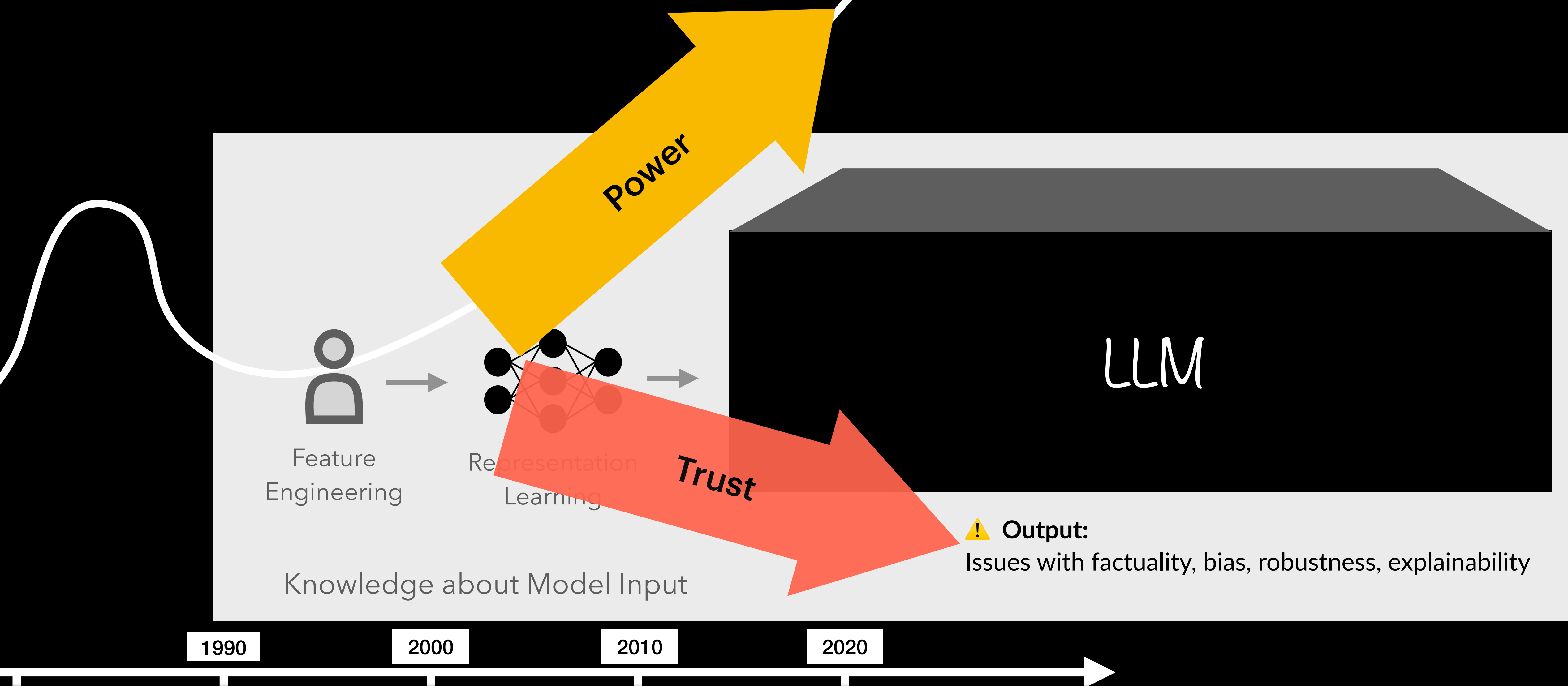
# **A Look Back - How Did We Get There?**

# A Short NLP History



# Gained Power - At What Cost?

2022-today: ✨ Explosion of LLMs



Epoch 2: Statistical Processing

Epoch 3: Deep Learning (DL) for NLP

**What is trust?**

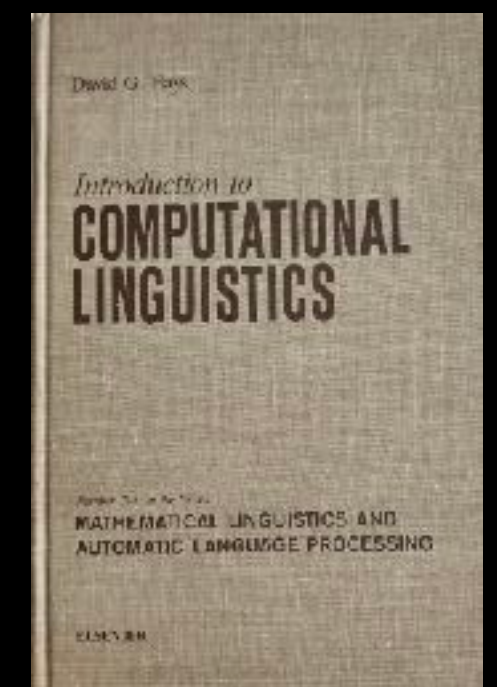
# Trustworthiness - Working Definition

---

“Trust arises from knowledge of origin as well as from knowledge of functional capacity.”



Hays. Applications. ACL 1979.





# Roadmap

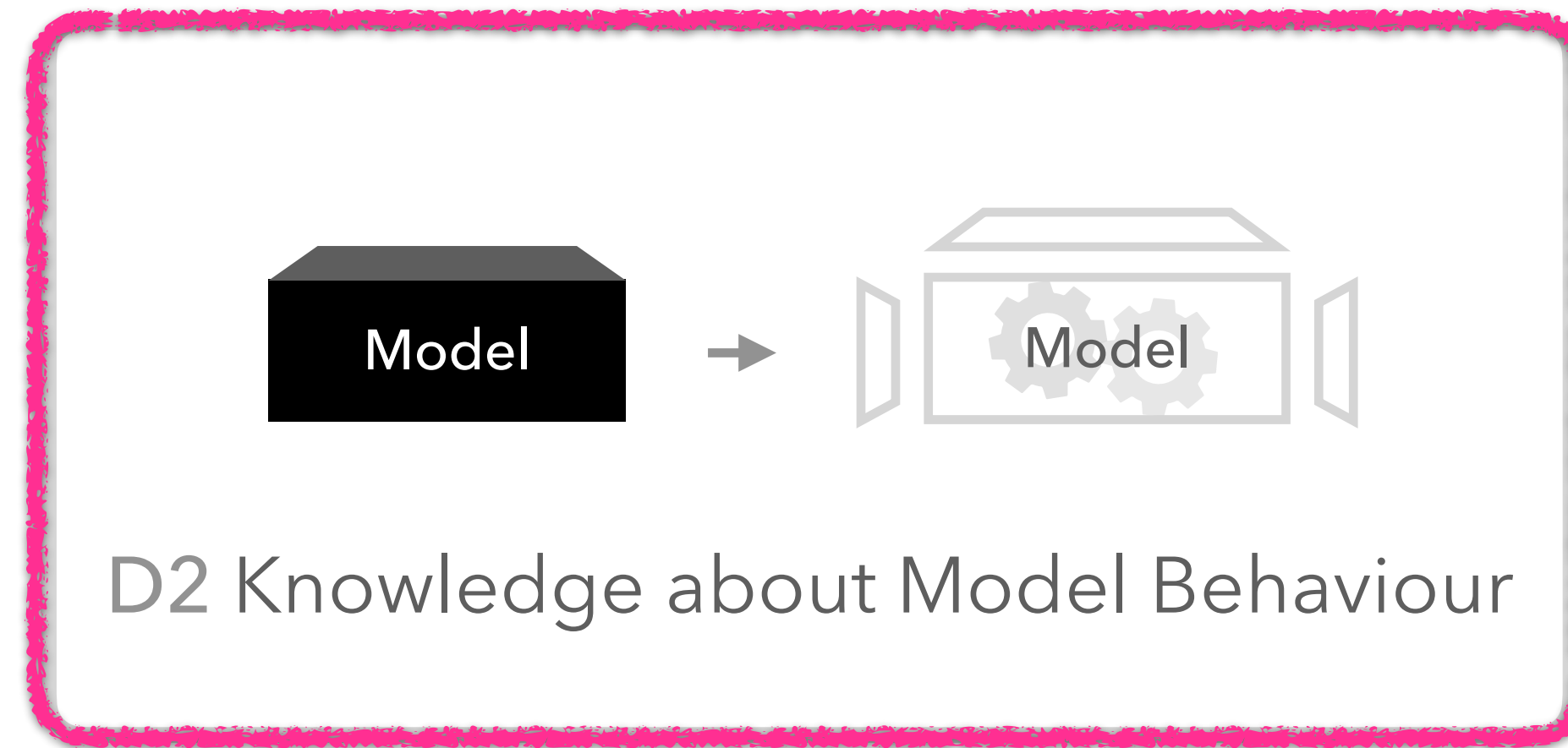
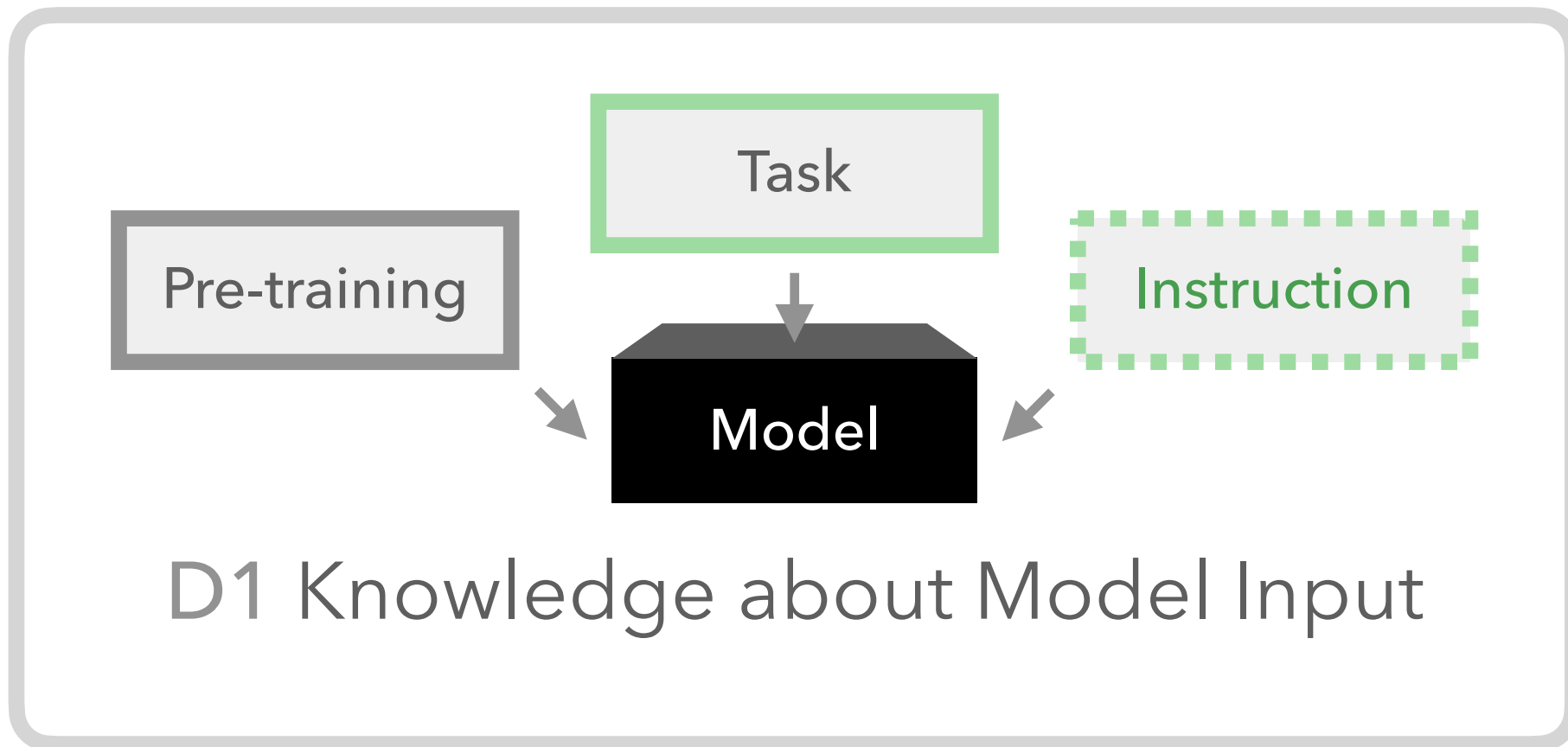
- 1 Past: LLMs & Trust - How Did We Get There?
- 2 Present: Trust Issues with LLMs
- 3 Trustworthy Human-Facing NLP

# Trust Issues with LLMs

---

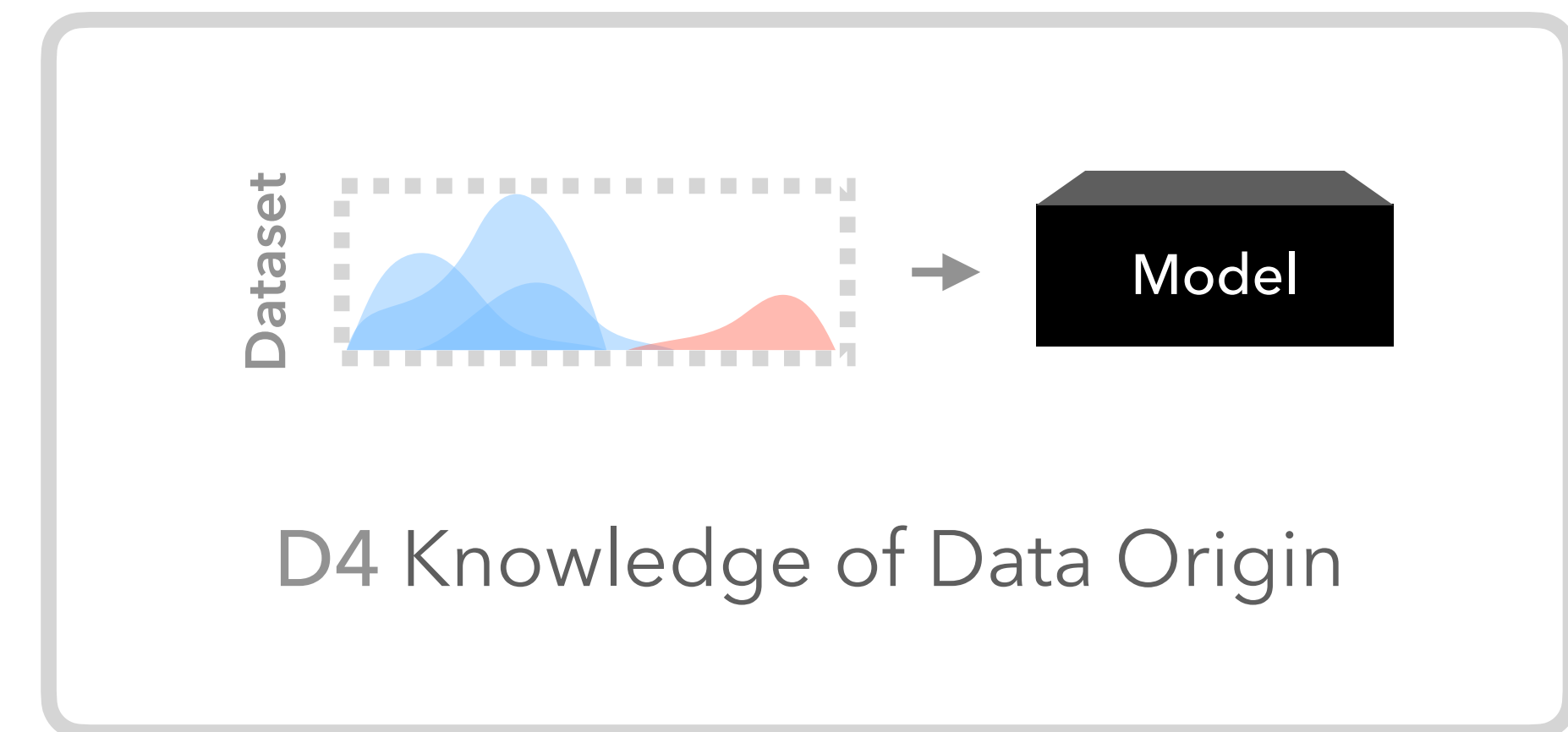
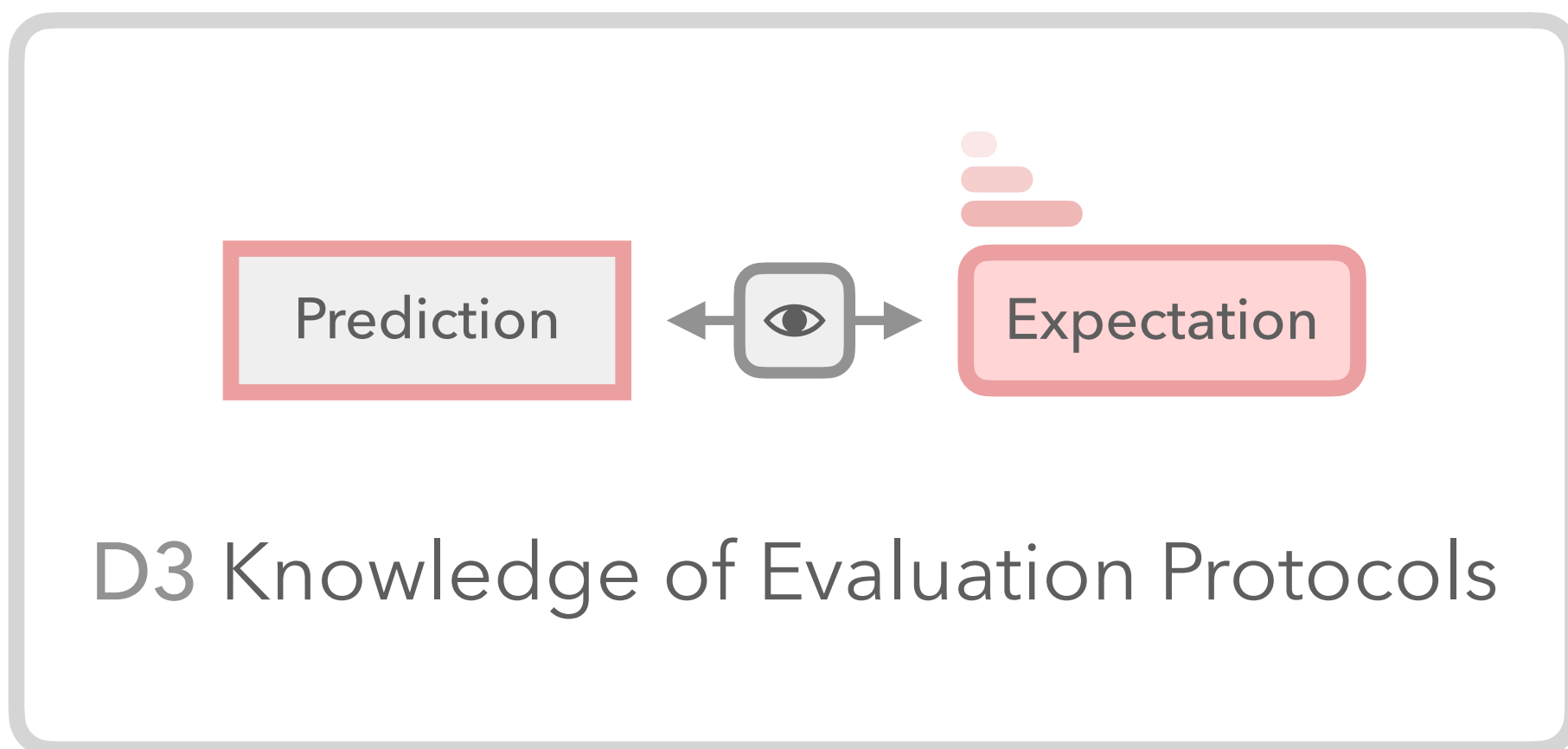
Selected Research Examples

— Four Desiderata to Increase Trust —



Trust arises from knowledge of origin as well as from knowledge of functional capacity.

*Trustworthiness - Working Definition by David G. Hays, 1979*



# Model Behaviour: Does it Matter How we Prompt an LLM?

- ▶ **!** Instability in prompting: Performance is highly sensitive to the linguistic variation of a prompt; prompts transfer poorly across datasets and models; LM perplexity dot not correlate well with model accuracy (open questions on connection data distribution and model behavior)

	prop.	prompt
mood	inter.	Do you find this movie review positive?
	indic.	You find this movie review positive.
	imper.	Tell me if you find this movie review positive.
aspt.	active	Do you find this movie review positive?
	pass.	Is this movie review found positive?
tense	past	Did you find this movie review positive?
	pres.	Do you find this movie review positive?
	future	Will you find this movie review positive?
modality	can	Can you find this movie review positive?
	could	Could you find this movie review positive?
	may	May you find this movie review positive?
	might	Might you find this movie review positive?
	must	Must you find this movie review positive?
	should	Should you find this movie review positive?
would	Would you find this movie review positive?	
synonymy	apprai.	Do you find this movie appraisal positive?
	comm.	Do you find this movie commentary positive?
	criti.	Do you find this movie critique positive?
	eval.	Do you find this movie evaluation positive?
	review	Do you find this movie review positive?

**The language of prompting:  
What linguistic properties make a prompt successful?**

Leidinger, van Rooij, Shutova, EMNLP 2023 Findings.

Table 1: Examples of variation of linguistic properties

# Model Behaviour: How Well Do LLMs Deal with Ambiguity?

- ▶ ⚠ LLMs and ambiguity is a major open problem: e.g. perform poorly at implicitly disambiguating entity types & biased towards preferred entity readings (influenced by entity popularity)

BP

You

sort by number of employees: Blackberry, Mango, Apple.

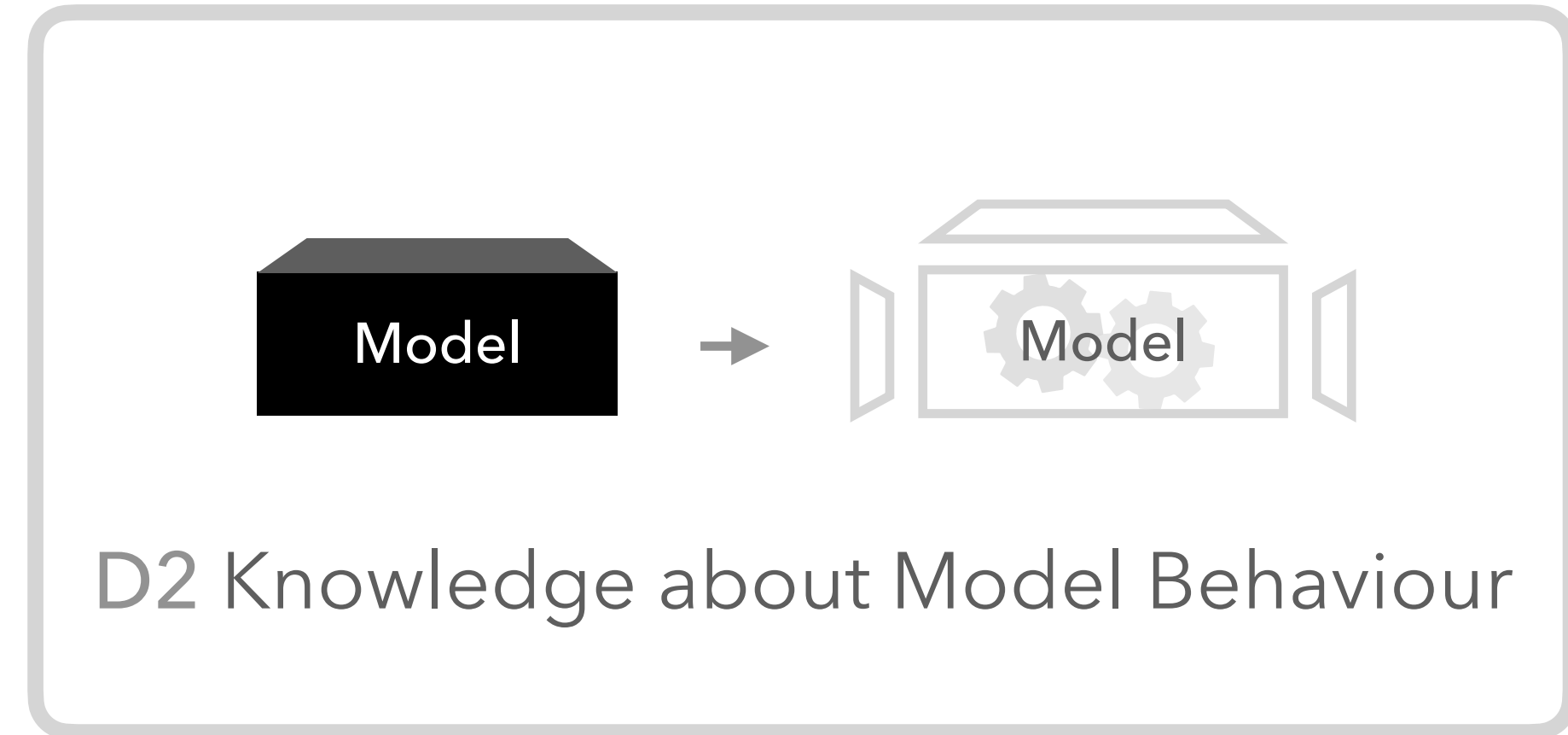
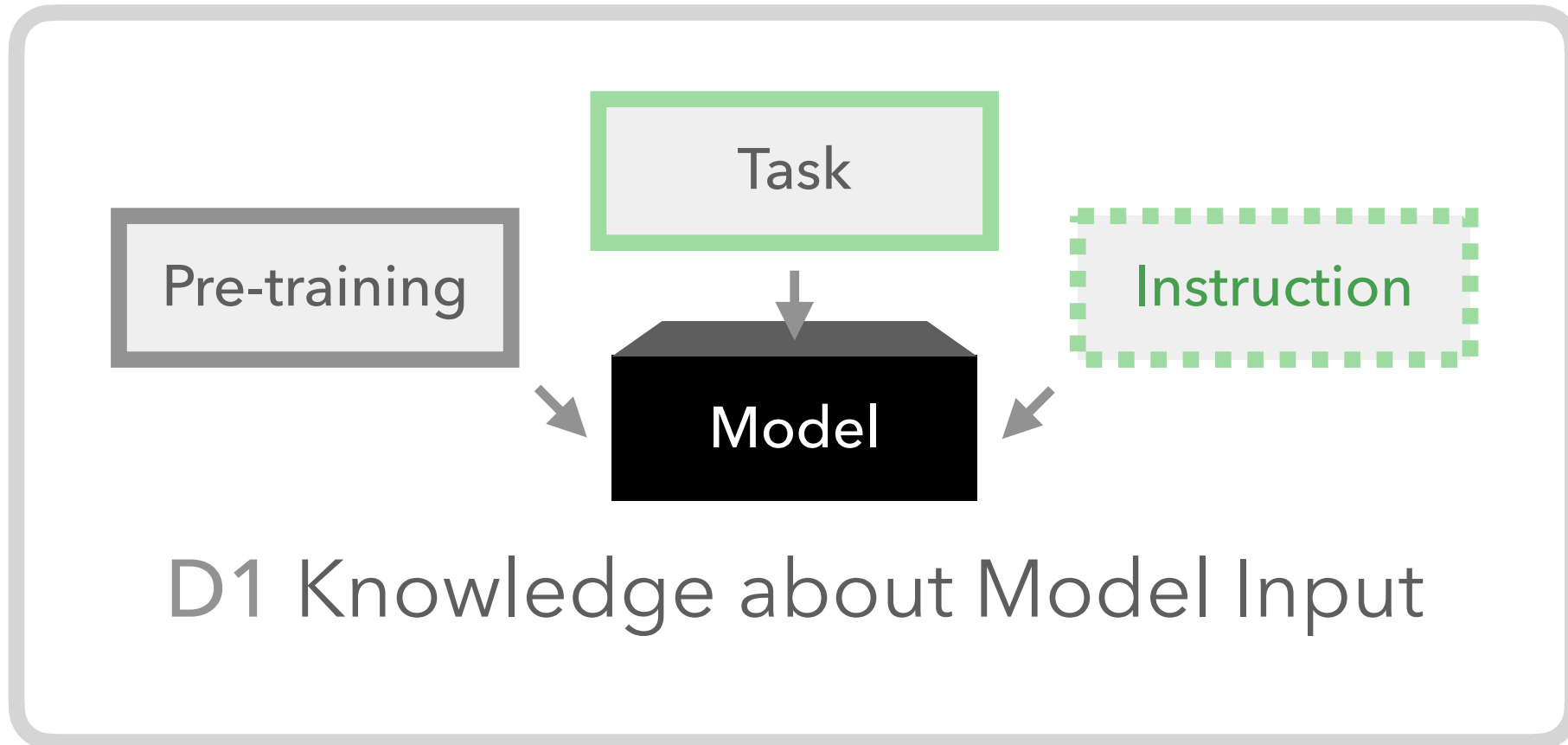
ChatGPT

To sort the fruits by the number of employees in ascending order, it would be:

1. Apple
2. Mango
3. Blackberry

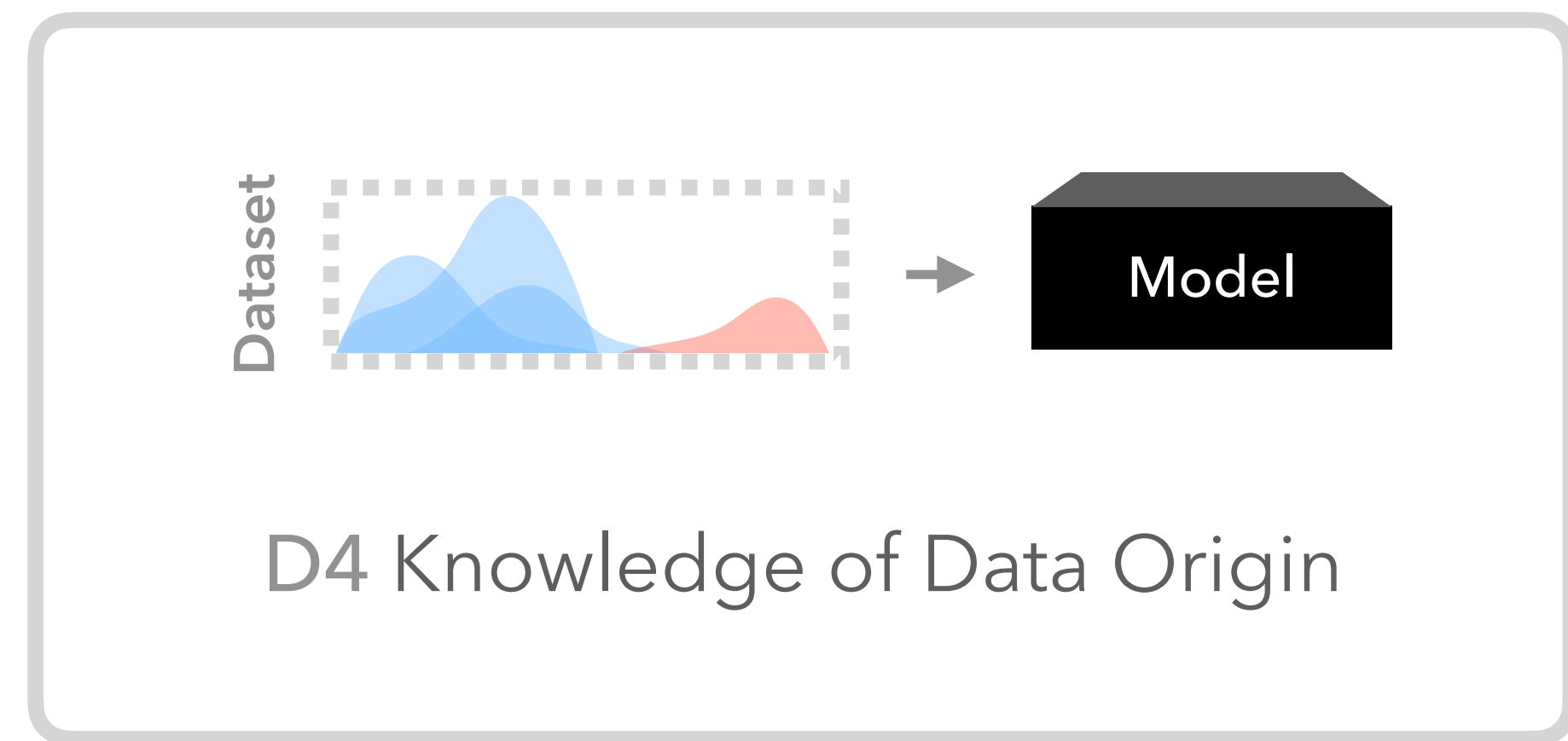
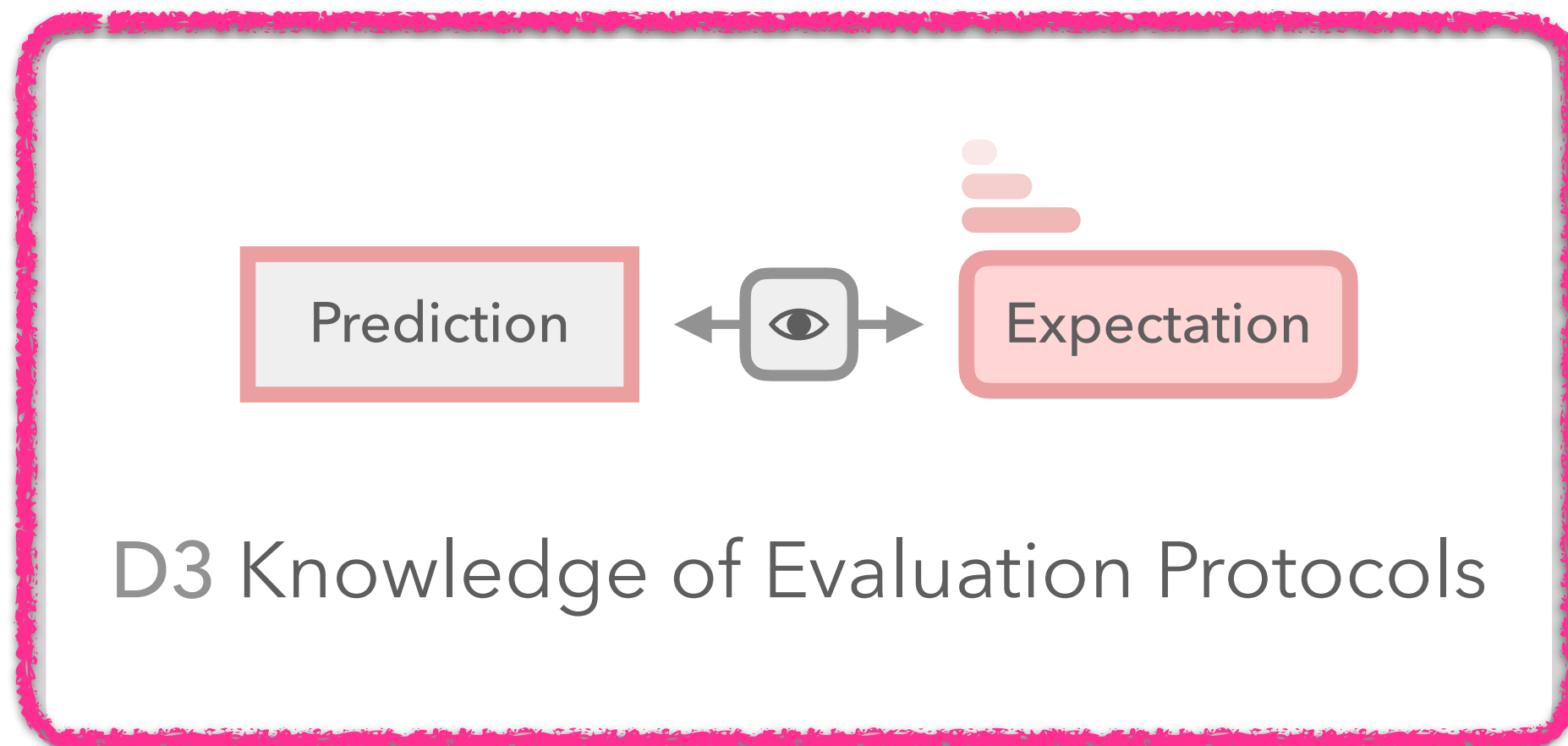
**To Know or Not To Know? Analyzing Self-Consistency of Large Language Models under Ambiguity**

Anastasiia Sedova<sup>1,2\*</sup> Robert Litschko<sup>3,4\*</sup> Diego Frassinelli<sup>3</sup>  
Benjamin Roth<sup>1,5</sup> Barbara Plank<sup>3,4</sup>



Trust arises from knowledge of origin as well as from knowledge of functional capacity.

*Trustworthiness - Working Definition by David G. Hays, 1979*



# Multiple-Choice Question Answering (MCQA) Prompt Style

---

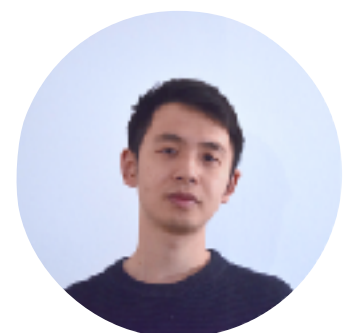
**General Instruction:** Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.

**Question:** The Web was effectively invented by Berners-Lee in which year?

**Options:**

- A. 1991
- B. 1980
- C. 1989
- D. 1993

**Answer:**



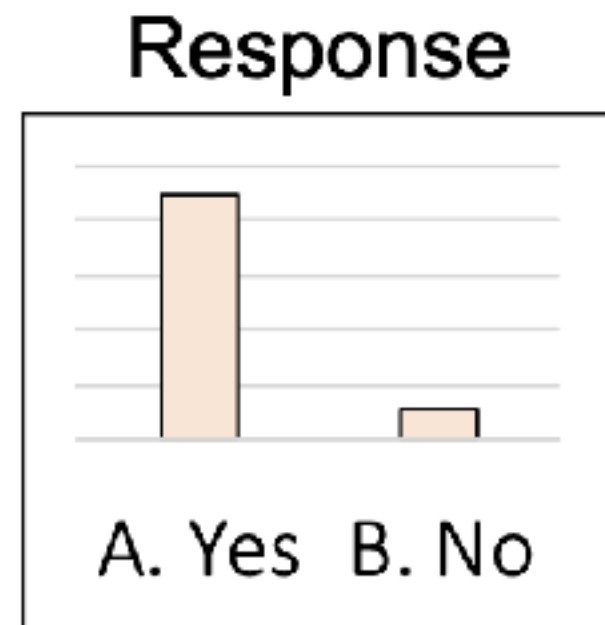
# Evaluation Protocols: Do Answer Options Impact LLM Outputs?

- ▶ ⚠ LLM's "A"-bias in MCQA responses

Choice ordering 1

**Question:** In the past 12 months, has this person given birth to any children?  
**A.** Yes  
**B.** No  
**Answer:**

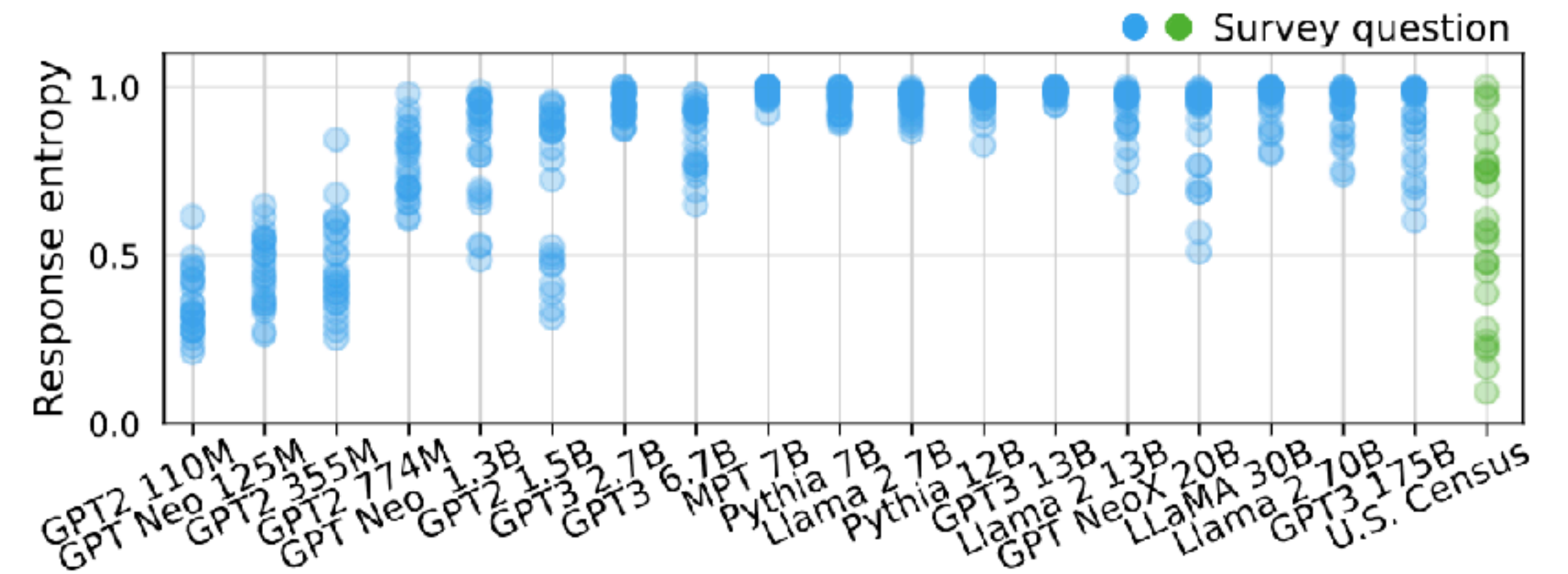
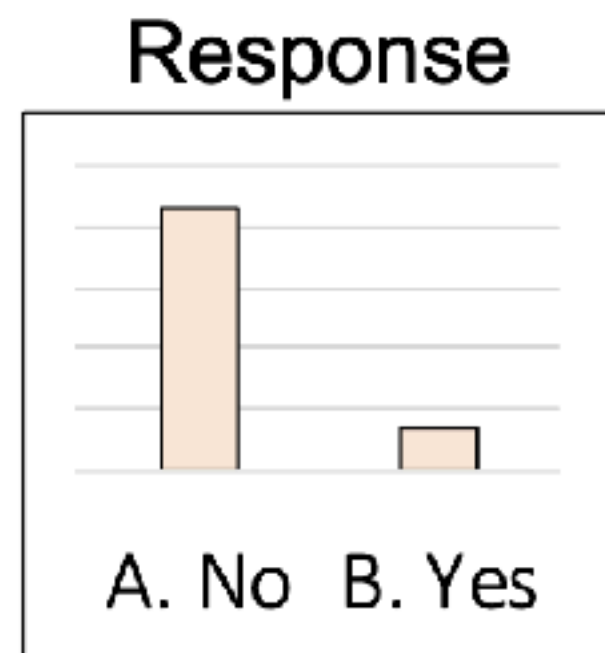
P("A")	0.82	P("B")	0.11
--------	------	--------	------



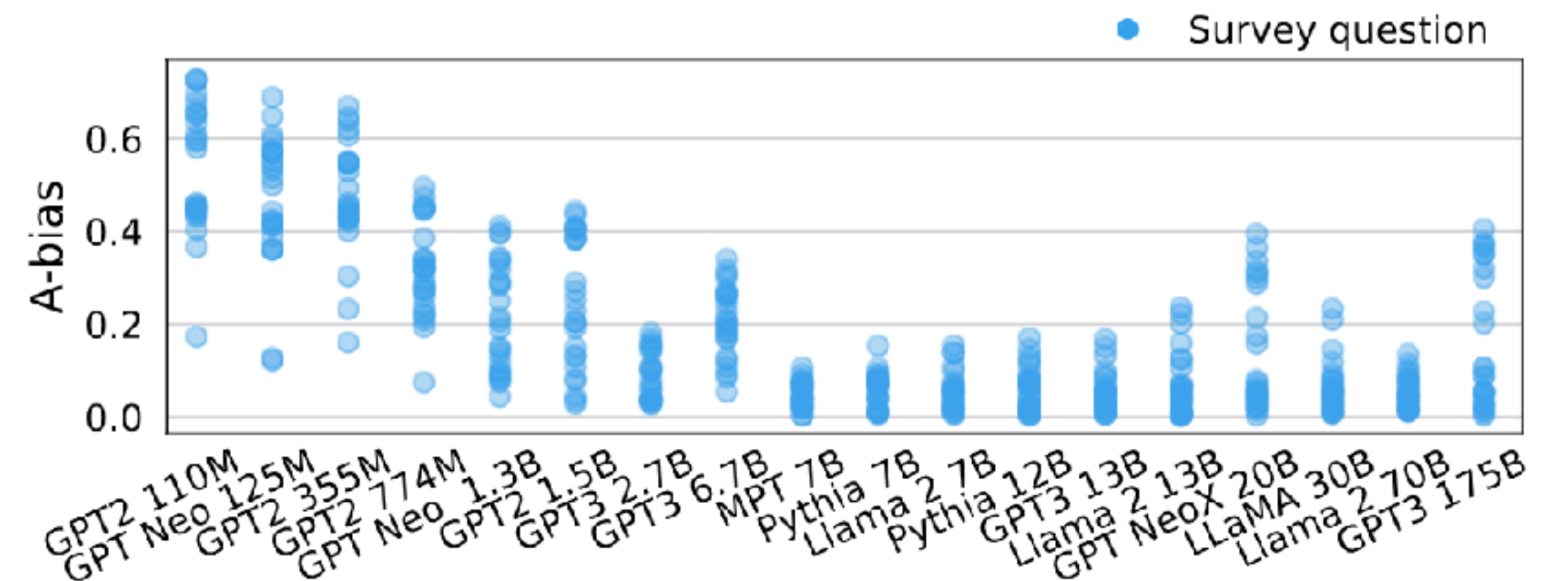
Choice ordering 2

**Question:** In the past 12 months, has this person given birth to any children?  
**A.** No  
**B.** Yes  
**Answer:**

P("A")	0.80	P("B")	0.15
--------	------	--------	------



(a) Entropy of base models' responses.



(b) A-bias of base models' responses.



# Evaluation Protocols: Does It Matter How We Extract Answers?

- ▶ ⚠ But “First-token log probs” do not match the text answers

**General Instruction:** Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.

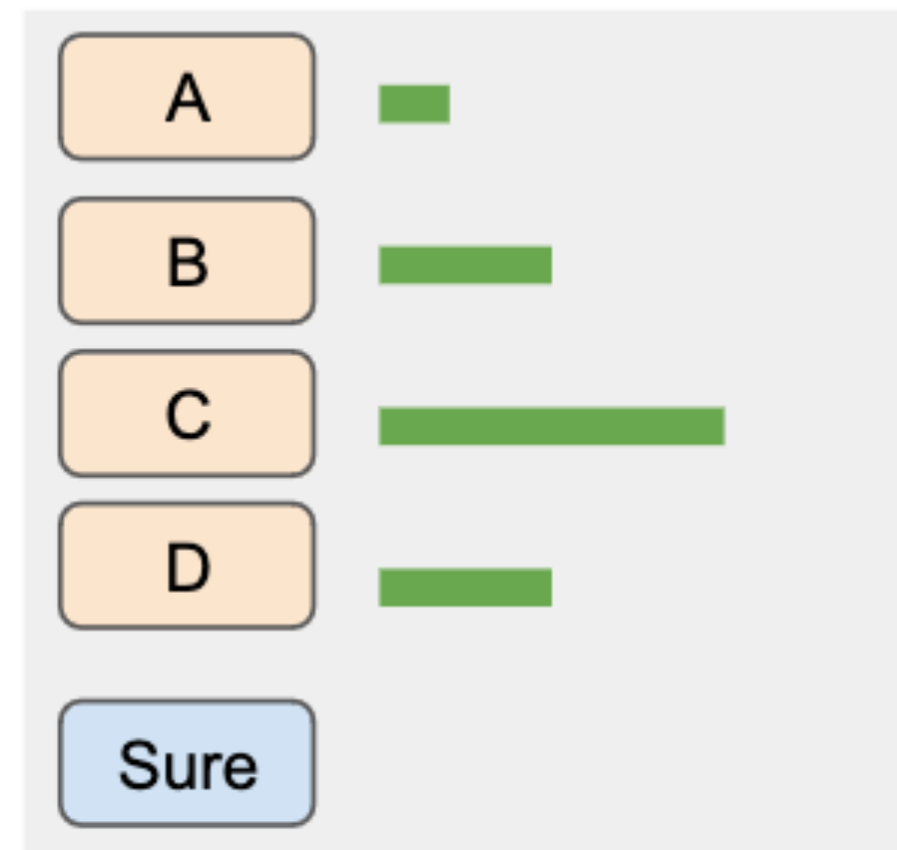
**Question:** The Web was effectively invented by Berners-Lee in which year?

**Options:**

- A. 1991
- B. 1980
- C. 1989
- D. 1993

**Answer:**

a. First Token Logits:



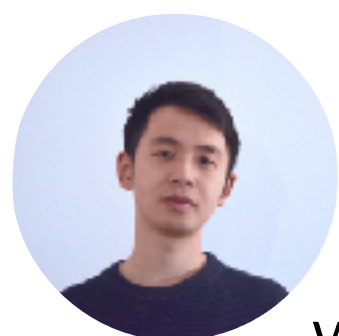
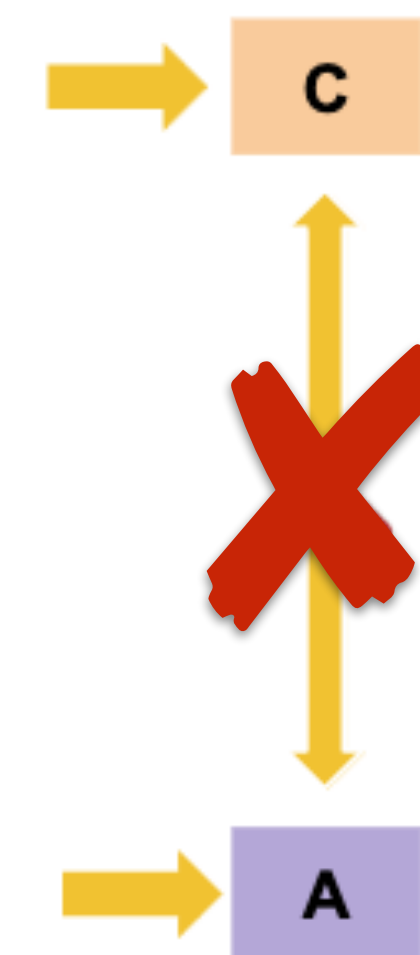
versus

b. Text Answer:

Sure! The answer is (A) 1991.

Model (0-shot)	First Token	Text Answer
<i>Gemma-7b-Inst</i>	30.2	50.8
<i>Llama2-7b-Chat</i>	34.9	43.1
<i>Llama2-13b-Chat</i>	40.2	47.6
<i>Mistral-7b-Inst-0.2</i>	53.2	53.6

Performance on MMLU.



# Evaluation Protocols: Chain-of-Thought, Better and Sound?

- ▶ Chain-of-Thought (CoT) prompting (“Let's think step by step”)
- ▶  But CoT is Not Sound!

Model	Correct Answer	Sound Reasoning
Zephyr-7B- $\beta$	45.0 $\pm$ 15.5	25.0 $\pm$ 10.5
Mistral-7B-Instruct	55.0 $\pm$ 10.0	25.0 $\pm$ 7.5
LLaMA-2-7B	46.7 $\pm$ 6.7	0.0 $\pm$ 0.0
LLaMA-2-13B	40.0 $\pm$ 8.2	15.0 $\pm$ 6.2
LLaMA-2-70B	56.7 $\pm$ 6.2	31.7 $\pm$ 9.7

Statements:

1. There is a purple marble in the box if and only if there is a blue marble in the box.
2. There is a blue marble in the box if and only if there is an olive marble in the box.
3. Either there is an olive marble in the box or else there is a red marble in the box, but not both.
4. Either there is a red marble in the box or else there is a green marble in the box, but not both.


Conclusion: If there is a purple marble in the box then there is a green marble in the box.

CoT response (excerpt) - manually verified:

7. Let's consider the possibility that there is an olive marble in the box. If that's the case, then, according to Statement 2, there must be a blue marble in the box as well.
8. Now, using Statement 1, we know that if there is a blue marble in the box, then there must be a purple marble in the box as well.
9. However, Statement 3 tells us that if there is an olive marble in the box, then there cannot be a red marble in the box. But, Statement 4 tells us that if there is a red marble in the box, then there cannot be a green marble in the box.
10. Since we cannot have both an olive marble and a red marble in the box, and we cannot have both a red marble and a green marble in the box, it seems that the only possibility left is that there is an olive marble in the box, and therefore, a blue marble, a purple marble, but no red or green marbles.
11. However, the conclusion states that "If there is a purple marble in the box then there is a green marble in the box", which is not logically true based on the given statements.

Conclusion: False.

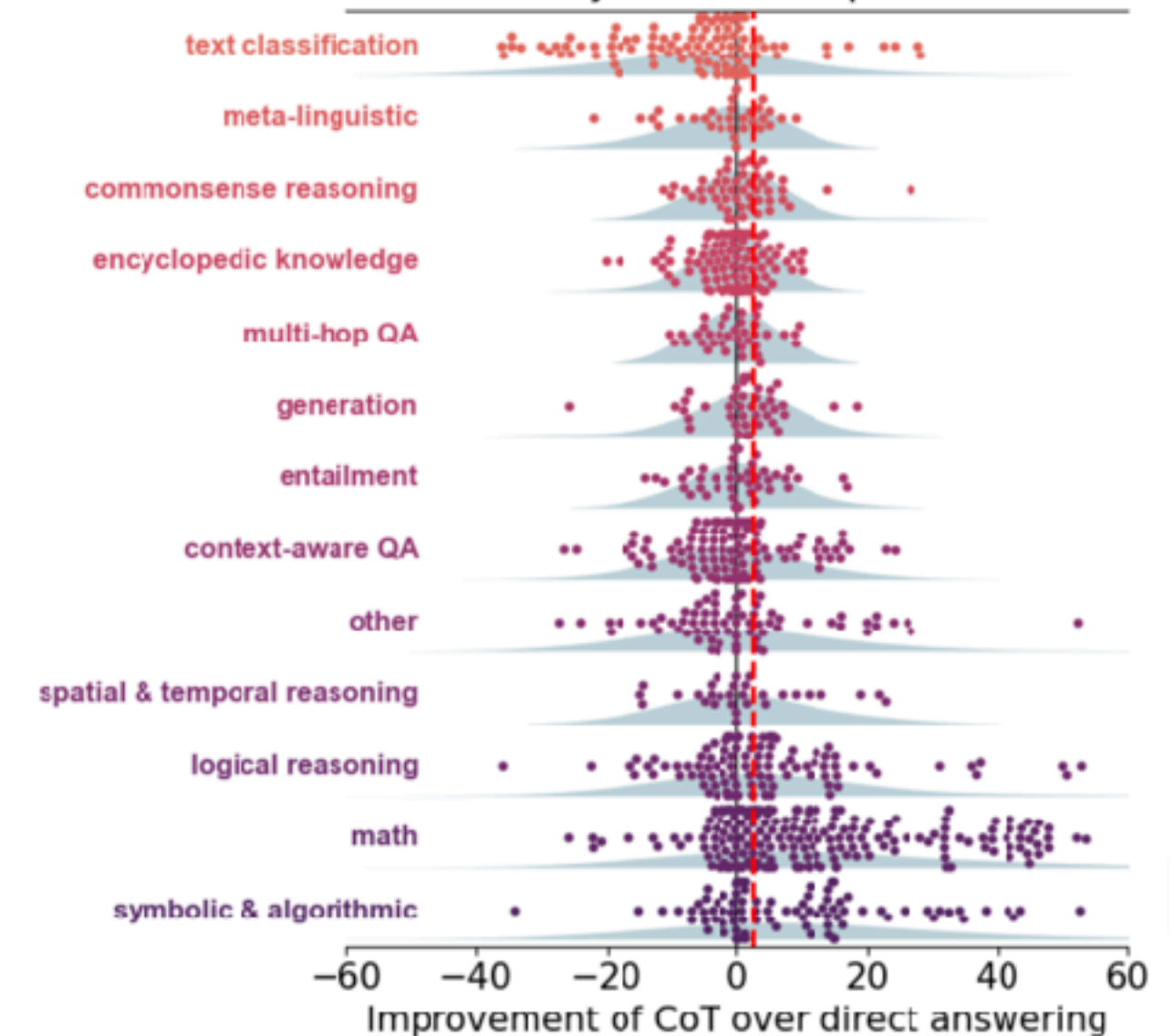
Supposition following (sound).

Invalid inference. 

Conclusion (invalid).

Final answer (incorrect).

Meta-analysis of CoT improvements



# Evaluation Protocols: Can LLMs Replace Humans Judges?

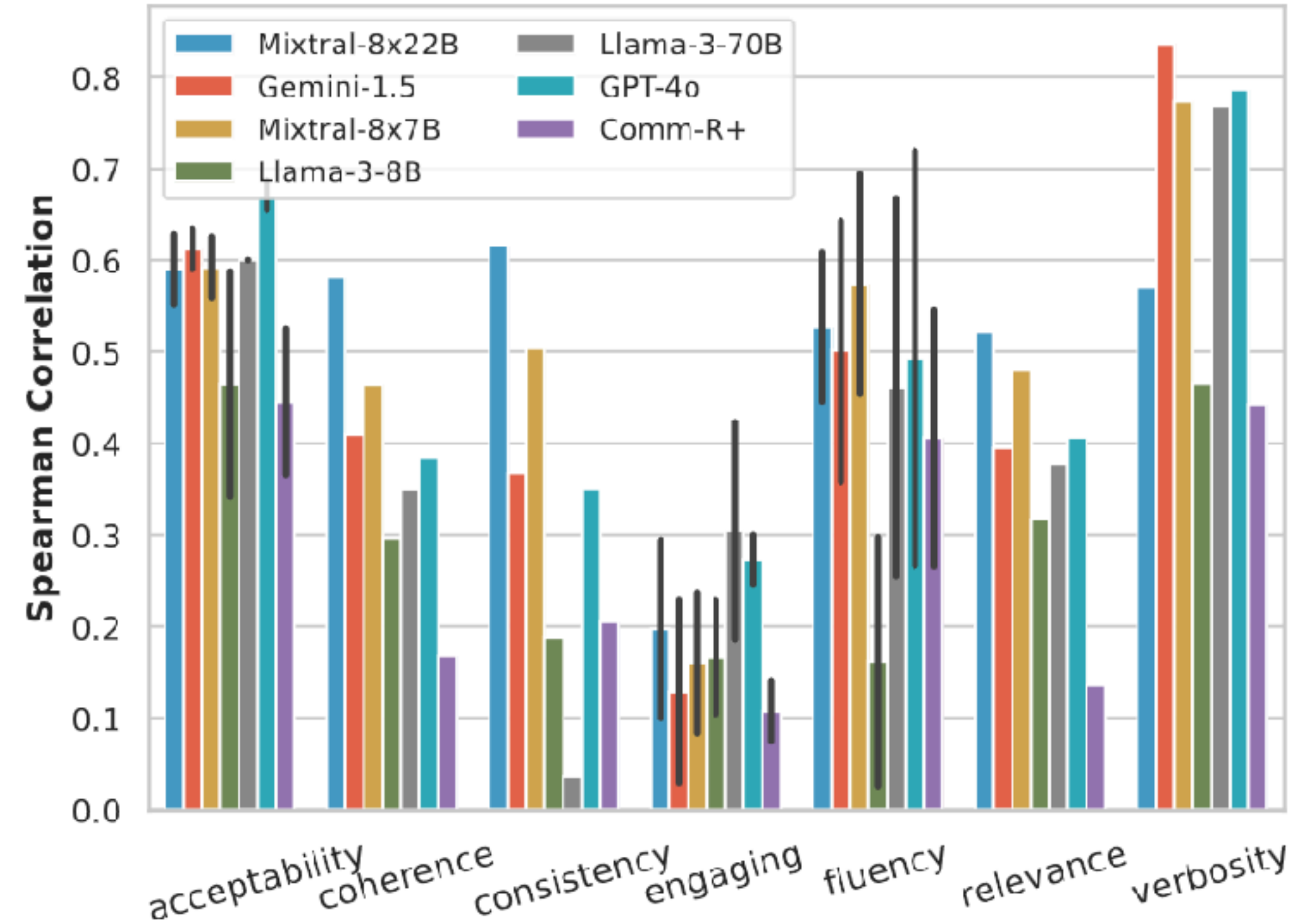
- ▶ ⚠ A lot of variability in LLM outputs
- ▶ LLMs are not ready yet to replace human judges - not even GPT-4o:

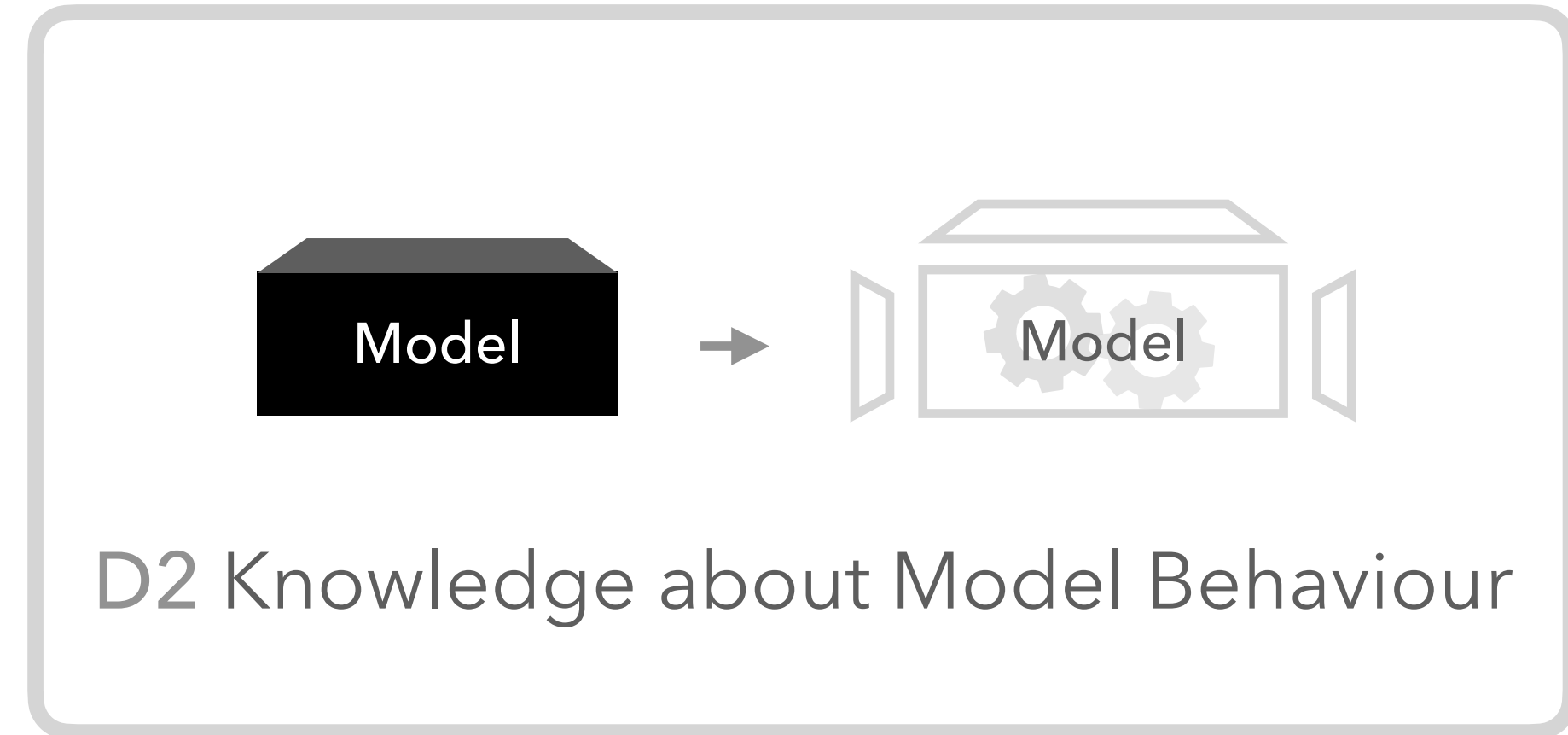
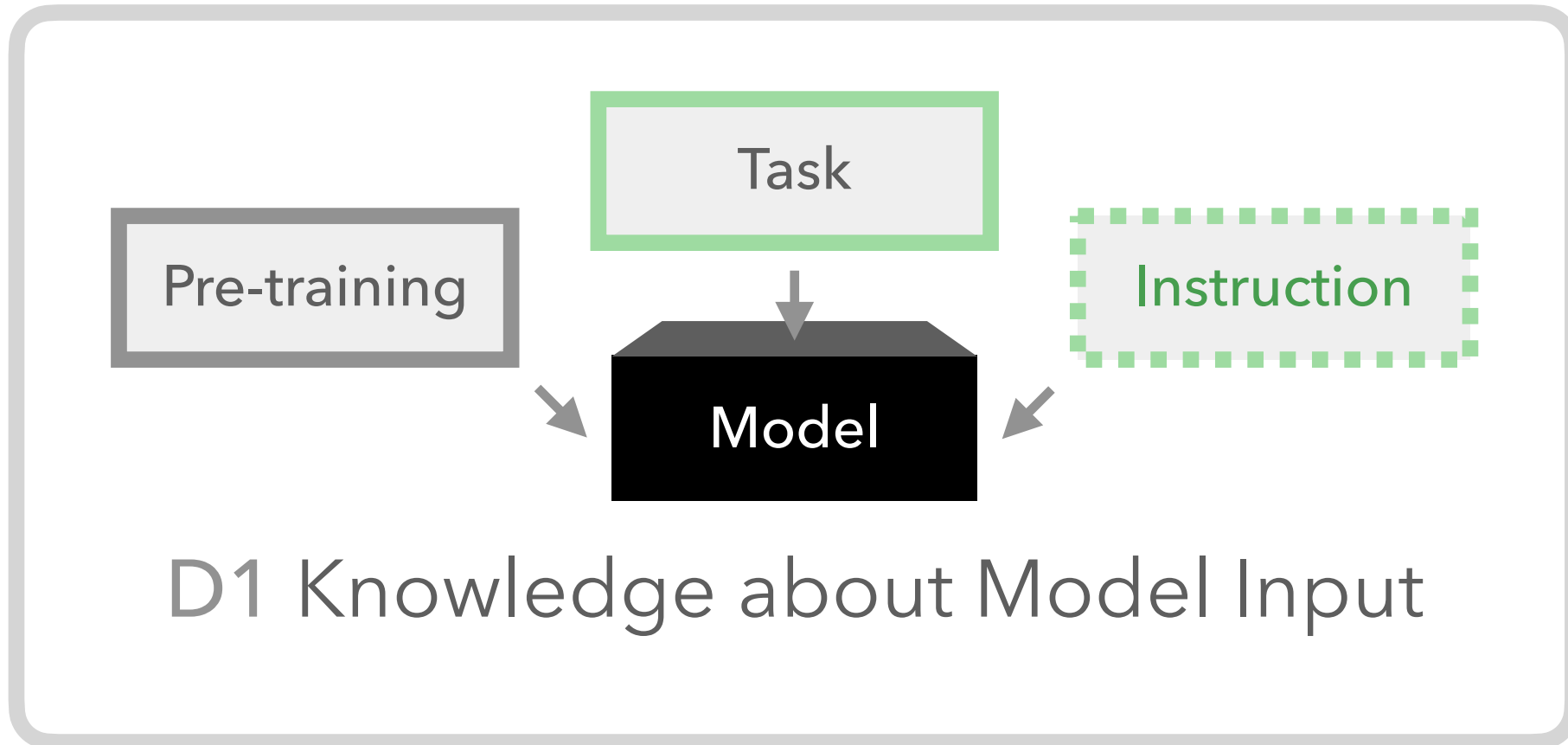
E.g. Plausibility:

**Instruction:** On a scale of 1 (very unlikely) to 5 (very likely), how plausible is it that the last response belongs to the dialogue?

**A:** Made it all the way through four years of college playing ball but  
**B:** I also like The Cosby Show

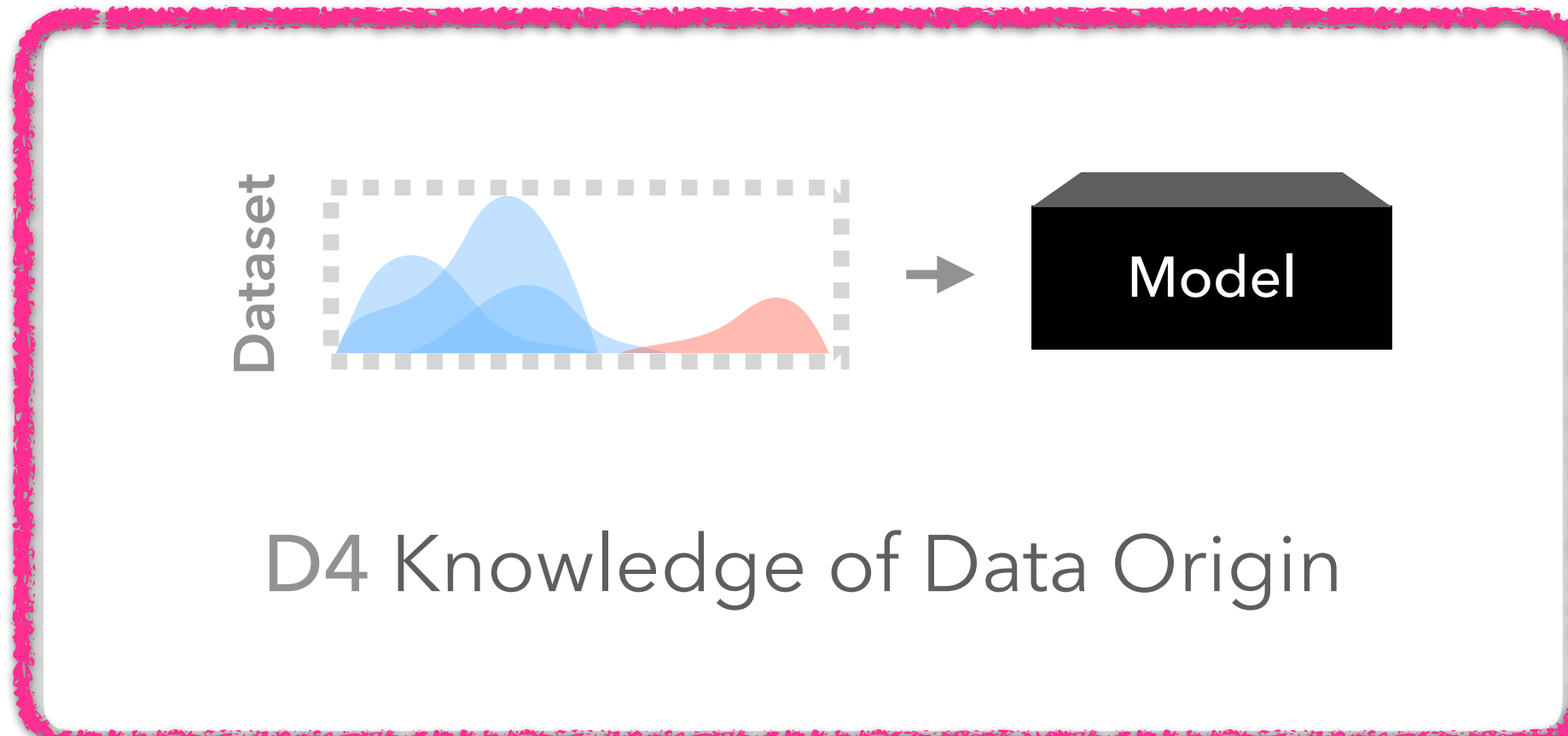
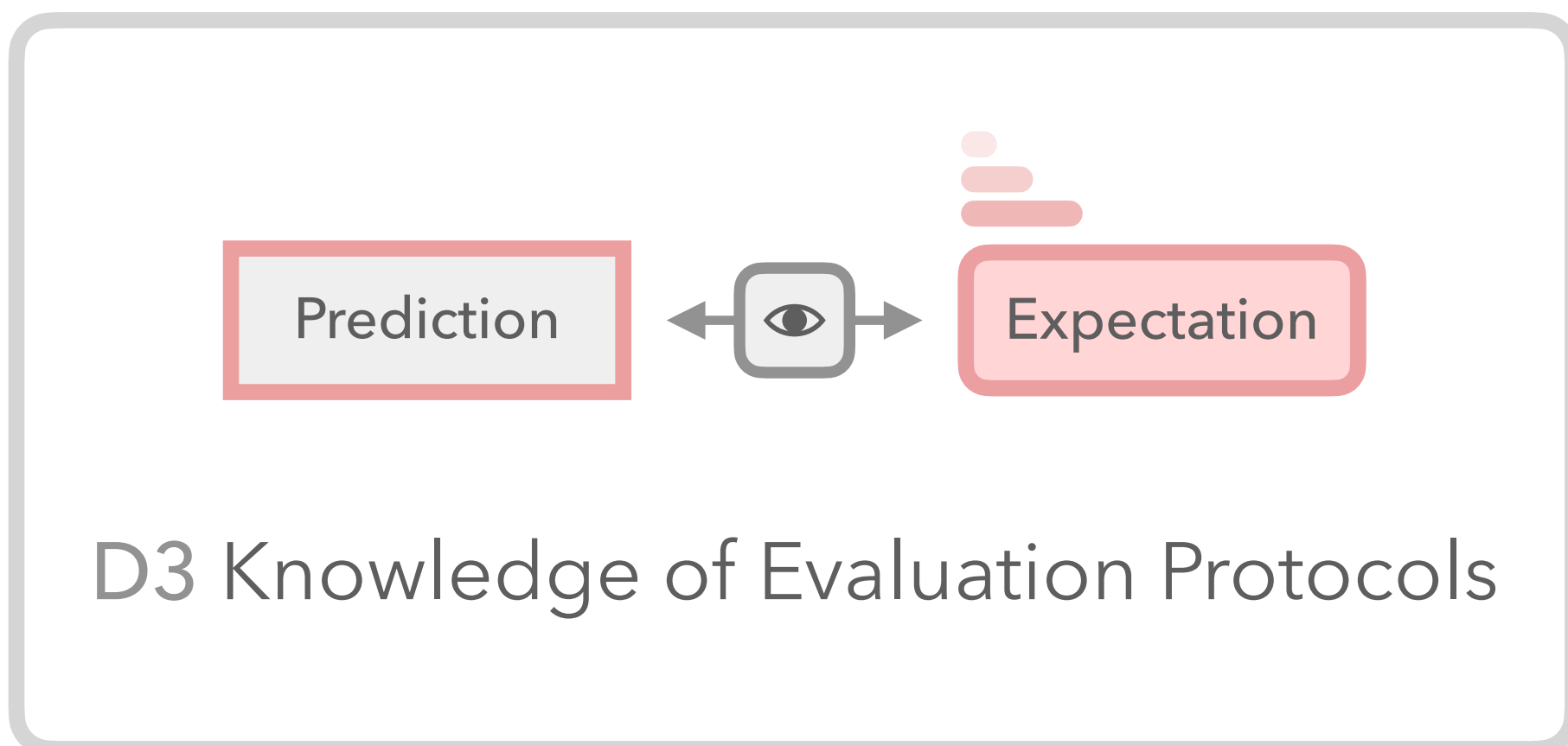
Humans Coders vs Models:





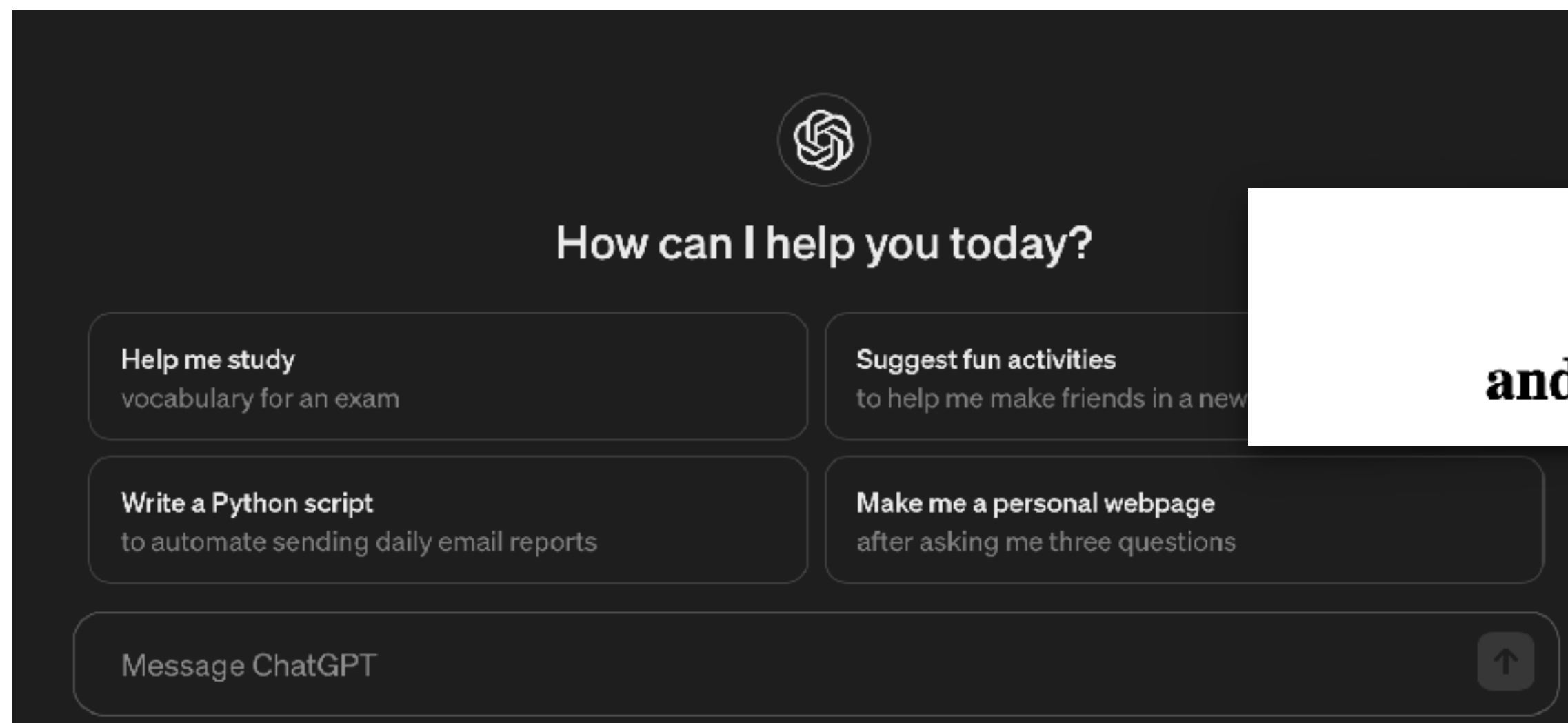
Trust arises from knowledge of origin as well as from knowledge of functional capacity.

*Trustworthiness - Working Definition by David G. Hays, 1979*



# Data Origin: (Indirect) Contamination & Need for Transparency

- ▶ ⚠ Too little transparency of what went into the training data of an LLM
- ▶ ⚠ Indirect data leakage: continuously provided by users (e.g. via OpenAI's the web interface)



## Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

Balloccu, Schmidtová, Lango, Dušek. EACL 2024.

- ▶ ➡ increasing efforts for transparency on training data & pre-processing, e.g.:
  - ▶ **PILE** (Gao et al., 2020)
  - ▶ **Dolma** (Soldini et al., 2024 ACL best paper award)

### : an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research

Luca Soldaini<sup>✉</sup> Rodney Kinney<sup>✉</sup> Akshita Bhagia<sup>✉</sup> Dustin Schwenk<sup>✉</sup>  
David Atkinson<sup>✉</sup> Russell Authur<sup>✉</sup> Ben Bogin<sup>✉</sup> Khyathi Chandu<sup>✉</sup>  
Jennifer Dumas<sup>✉</sup> Yanai Elazar<sup>✉</sup> Valentin Hofmann<sup>✉</sup> Ananya Harsh Jha<sup>✉</sup>  
Sachin Kumar<sup>✉</sup> Li Lucy<sup>✉</sup> Xinxin Lyu<sup>✉</sup> Nathan Lambert<sup>✉</sup> Ian Magnusson<sup>✉</sup>  
Jacob Morrison<sup>✉</sup> Niklas Muennighoff<sup>✉</sup> Aakanksha Naik<sup>✉</sup> Crystal Nam<sup>✉</sup>  
Matthew E. Peters<sup>✉</sup> Abhilasha Ravichander<sup>✉</sup> Kyle Richardson<sup>✉</sup> Zejiang Shen<sup>✉</sup>  
Emma Strubell<sup>✉</sup> Nishant Subramani<sup>✉</sup> Oyvind Tafjord<sup>✉</sup> Pete Walsh<sup>✉</sup>  
Luke Zettlemoyer<sup>✉</sup> Noah A. Smith<sup>✉</sup> Hannaneh Hajishirzi<sup>✉</sup>  
Iz Beltagy<sup>✉</sup> Dirk Groeneveld<sup>✉</sup> Jesse Dodge<sup>✉</sup>  
Kyle Lo<sup>✉</sup>

**Growing Importance of  
Data Quality > Data Quantity**

# The “it” in AI models is the dataset - talk by Thom Wolf 🤗

## The “it” in AI models is the dataset.

Posted on June 10, 2023 by jbetker

I’ve been at OpenAI for almost a year now. In that time, I’ve trained a lot of generative models. More than anyone really has any right to train. As I’ve spent these hours observing the effects of tweaking various model configurations and hyperparameters, one thing that has struck me is the similarities in between all the training runs.

It’s becoming awfully clear to me that these models are truly approximating their datasets to an incredible degree. What that means is not only that they learn what it means to be a dog or a cat, but the interstitial frequencies between distributions that don’t matter, like what photos humans are likely to take or words humans commonly write down.

What this manifests as is – trained on the same dataset for long enough, pretty much every model with enough weights and training time converges to the same point. Sufficiently large diffusion conv-unets produce the same images as ViT generators. AR sampling produces the same images as diffusion.

This is a surprising observation! It implies that model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivery compute to approximating that dataset.

Then, when you refer to “Lambda”, “ChatGPT”, “Bard”, or “Claude” then, it’s not the model weights that you are referring to. It’s the dataset.

# Evidence from a talk by Sara Hooker

Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion
<i>Chinchilla</i>	70 Billion	1.4 Trillion

- Recent work suggests smaller amounts of higher quality data remove the need for a larger model.
- This suggest larger models may just be compensating for problems in the data pipeline.



# Roadmap

- 1 Past: LLMs & Trust - How Did We Get There?
- 2 Present: Trust Issues with LLMs
- 3 Trustworthy Human-Facing NLP

# Name the object



# Name the object



cake (53), food (19), bread (8), burger (6),  
dessert (6), snacks (3), muffin (3), pastry (3)

# Lora Aroyo's NeurIPS 2023 Keynote



Is there a **SMILE** in this image?

YES but ...

Canada		
YES	NO	DNK
40%	40%	20%

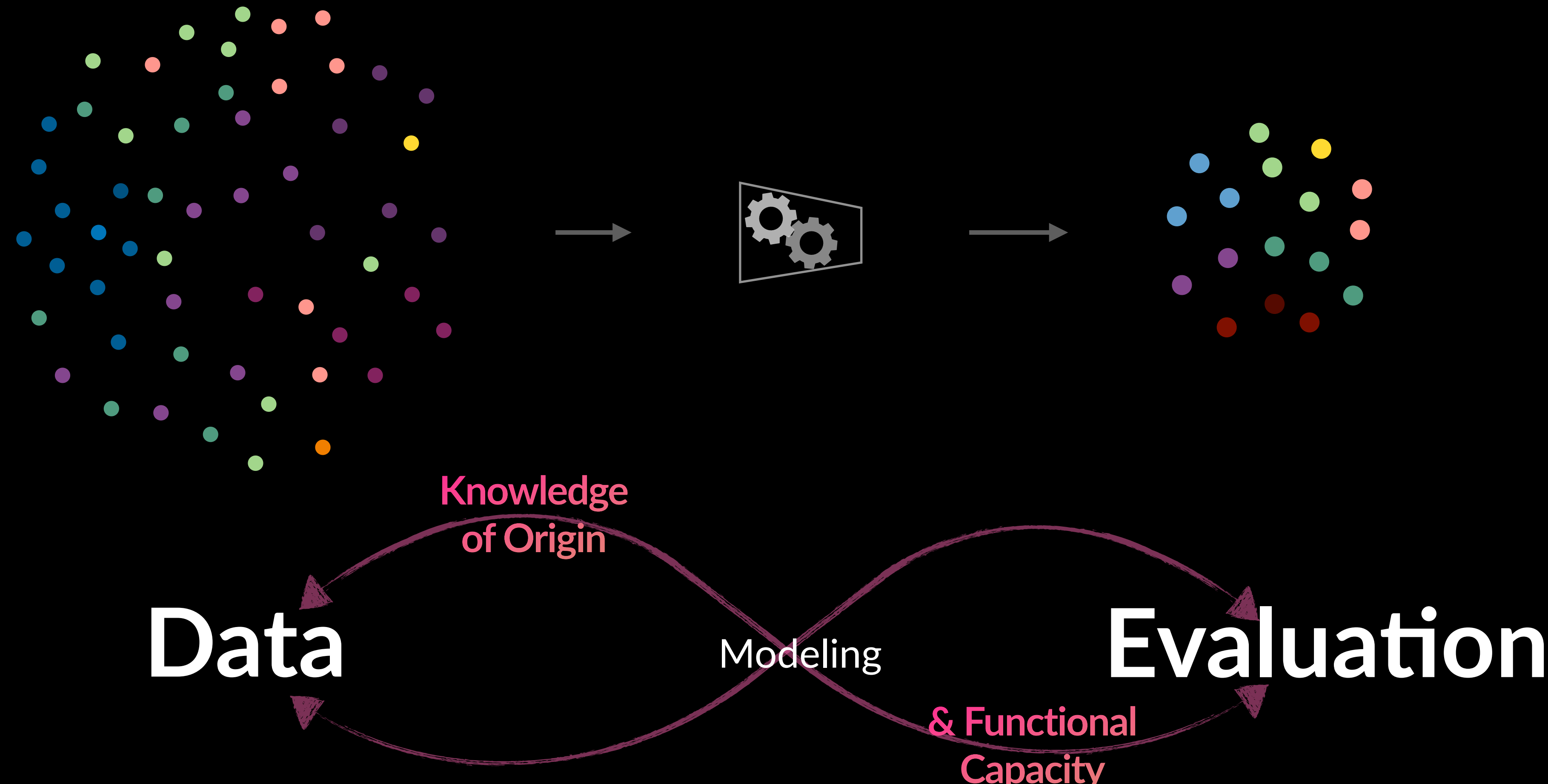
India		
YES	NO	DNK
70%	30%	0

USA		
YES	NO	DNK
50%	0	50%



# Human Label Variation

# Regaining Trust: Importance of Data and Evaluation



# Many open questions

---

- ▶ How does Human Label Variation interact with Socially Aware NLP? (Specific tasks and broadly Values, Opinions, Attitude and LLMs)
- ▶ How much Variation is there Within-Human (Coder) Label Variation?
- ▶ How does Human Label Variation relate to Model Uncertainty?
- ▶ Is All Variation Plausible? How to tease apart Error from Plausible Label Variation?
- ▶ Do we Need More Labels or More Cases (for Eval/Train)? Data Quality vs Quantity
- ▶ When to take a description vs a prescriptive approach (Röttger et al. 2022) to annotation?

What goes into  
representativeness and  
quality of data

learning from less but  
higher quality data?

active learning (how  
to sample. Instances for  
labelling)

model uncertainty

## *Human Label Variation*

- many exciting connections -

human values and  
LLM alignment (e.g.  
Durmus et al., 2024)

LLMs that react as  
humans do

statistics and data-  
generation process



# NLP & Survey Research



Thank you!