

# **AI Safety Through Interpretable and Controllable Language Models**

Peter Hase

**ANTHROPIC**

# AI Safety

---

**2016**

---

Concrete Problems in AI Safety

---

**Dario Amodei\***  
Google Brain

**Chris Olah\***  
Google Brain

**Jacob Steinhardt**  
Stanford University

**Paul Christiano**  
UC Berkeley

**John Schulman**  
OpenAI

**Dan Mané**  
Google Brain

**2024**

**Managing extreme AI risks  
amid rapid progress**

Yoshua Bengio  
Geoffrey Hinton  
Andrew Yao  
Dawn Song  
et al.

Mila - Quebec AI Institute, Université de Montréal  
University of Toronto, Vector Institute  
Tsinghua University  
UC Berkeley

**2020**

CONCRETE PROBLEMS IN AI SAFETY, REVISITED

**Inioluwa Deborah Raji & Roel Dobbe**  
AI Now Institute  
New York University  
New York City, NY, USA

# AI Safety

---

## Misuse

Bioweapons  
Cyberattacks  
Surveillance

## Misalignment

Reward Hacking  
Deception  
Wrong Goals

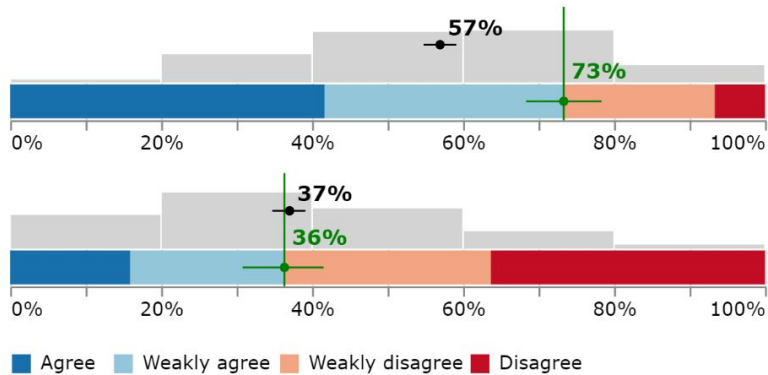
## Accidents

Mistakes  
Not Robust  
Pipeline Failures

## Emergent Harms

Mass Unemployment  
Power Concentration  
Arms Races

# AI Safety



### 3-3. AI could soon lead to revolutionary societal change

In this century, labor automation caused by advances in AI/ML could plausibly lead to economic restructuring and societal changes on at least the scale of the Industrial Revolution.

### 3-4. AI decisions could cause nuclear-level catastrophe

It is plausible that decisions made by AI or machine learning systems could cause a catastrophe this century that is at least as bad as an all-out nuclear war.

2022 NLP Community Metasurvey

This was pre-ChatGPT!

# AI Safety

---

What do we do about this?

# AI Safety

---

## (Socio)technical Research

- Understand risks better
- Develop mitigations

## Many other things...

- Science communication
- Supporting policymakers
- Supporting responsible industrial practices
- Forecasting progress, third party auditing, consensus-building
- etc.

# This Talk

---

## (Socio)technical Research

- Understand risks better
- Develop mitigations

## Many other things...

- Science communication
- Supporting policymakers
- Supporting responsible industrial practices
- Forecasting progress, third party auditing, consensus-building
- etc.

# This Talk

---

## (Socio)technical Research

- Interpretability: understand how models make decisions
- Controllability: adjust model behavior at a fine-grained level

**Understanding → Control**



# This Talk

---

1. Open Problems in Interpretability
2. Model Editing as Belief Revision
3. Scalable Oversight: Easy-to-Hard Generalization

# This Talk

---

1. **Open Problems in Interpretability**
2. Model Editing as Belief Revision
3. Scalable Oversight: Easy-to-Hard Generalization

# Interpretability

---

## Foundational Challenges in Assuring Alignment and Safety of Large Language Models

Usman Anwar<sup>1</sup>

### Sec. 3.4

Abulhair Saparov<sup>\*2</sup>, Javier Rando<sup>\*3</sup>, Daniel Paleka<sup>\*3</sup>, Miles Turpin<sup>\*2</sup>, Peter Hase<sup>\*4</sup>, Ekdeep Singh Lubana<sup>\*5</sup>, Erik Jenner<sup>\*6</sup>, Stephen Casper<sup>\*7</sup>, Oliver Sourbut<sup>\*8</sup>, Benjamin L. Edelman<sup>\*9</sup>, Zhaowei Zhang<sup>\*10</sup>, Mario Günther<sup>\*11</sup>, Anton Korinek<sup>\*12</sup>, Jose Hernandez-Orallo<sup>\*13</sup>

Lewis Hammond<sup>8</sup>, Eric Bigelow<sup>9</sup>, Alexander Pan<sup>6</sup>, Lauro Langosco<sup>1</sup>, Tomasz Korbak<sup>14</sup>, Heidi Zhang<sup>15</sup>, Ruiqi Zhong<sup>6</sup>, Seán Ó hÉigeartaigh<sup>†1</sup>, Gabriel Recchia<sup>16</sup>, Giulio Corsi<sup>†1</sup>, Alan Chan<sup>†17</sup>, Markus Anderljung<sup>†17</sup>, Lilian Edwards<sup>†18</sup>, Aleksandar Petrov<sup>8</sup>, Christian Schroeder de Witt<sup>8</sup>, Sumeet Ramesh Motwani<sup>6</sup>

Yoshua Bengio<sup>†19</sup>, Danqi Chen<sup>†20</sup>, Philip H.S. Torr<sup>†8</sup>, Samuel Albanie<sup>†1</sup>, Tegan Maharaj<sup>†21</sup>, Jakob Foerster<sup>†8</sup>, Florian Tramèr<sup>†3</sup>, He He<sup>†2</sup>, Atoosa Kasirzadeh<sup>†22</sup>, Yejin Choi<sup>†23</sup>

David Krueger<sup>†1</sup>

TMLR 2024

\*indicates major contribution.

# Definitions

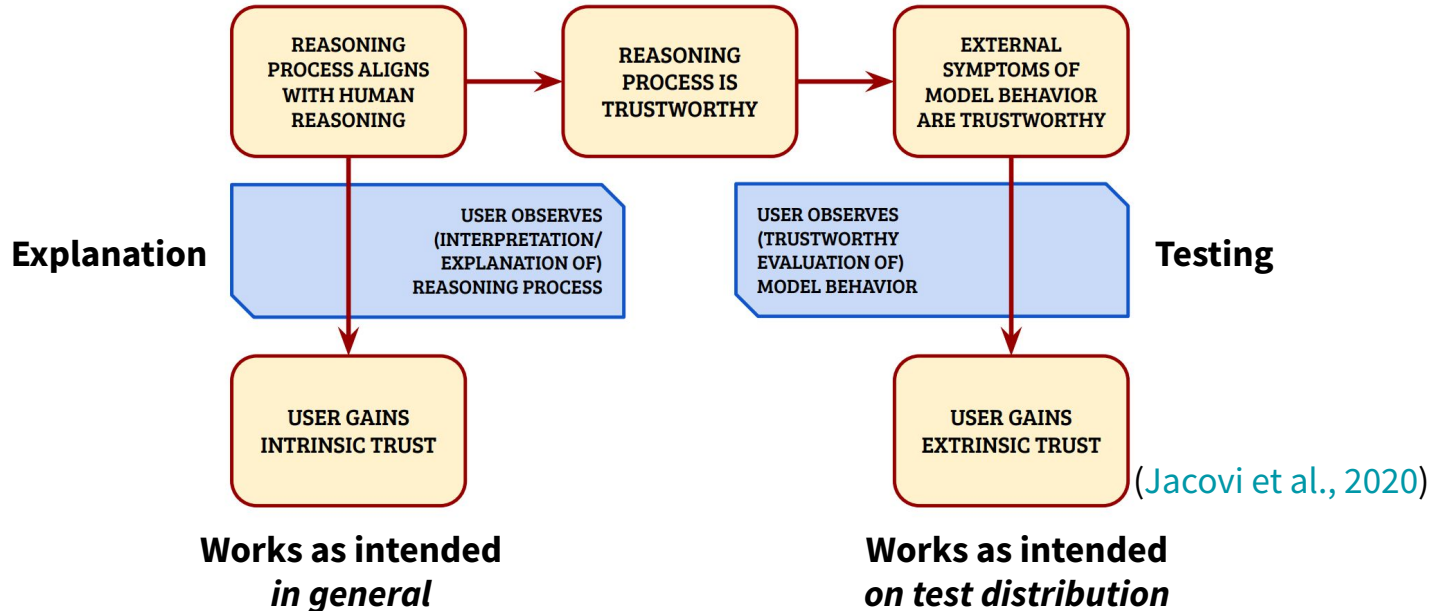
---

A model is *interpretable* if we can form accurate beliefs about how it works

“How it works” = causal chains of events that lead to model outputs

# Why Interpretability?

- We evaluate models with test data → *accuracy*
- But can we verify their *reasoning*?



# Why Interpretability?

---



FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from  
[Lakkaraju et al. \(2022\)](#)

# Why Interpretability?

---



FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from  
[Lakkaraju et al. \(2022\)](#)

I don't know, but the model's accuracy on this group is 90%.



**VS...**

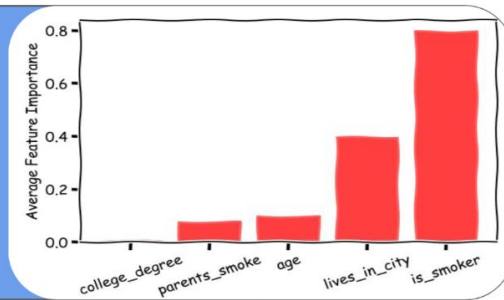
# Why Interpretability?



FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from  
[Lakkaraju et al. \(2022\)](#)

GOOD QUESTION! IT LOOKS LIKE THE MODEL PREDICTED THESE INDIVIDUALS AS HIGH RISK MOSTLY BECAUSE THEY WERE SMOKERS BUT ALSO BECAUSE THEY LIVE IN LARGE CITIES. I'M HIGHLY CONFIDENT THESE ARE THE REASONS BECAUSE THE EXPLANATIONS HAVE HIGH FIDELITY. HERE'S THE AVERAGE FEATURE IMPORTANCE FOR THESE PEOPLE (HIGHER MEANS MORE IMPORTANT).





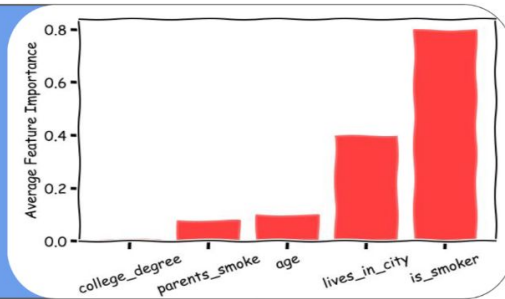
# Why Interpretability?



FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from  
[Lakkaraju et al. \(2022\)](#)

GOOD QUESTION! IT LOOKS LIKE THE MODEL PREDICTED THESE INDIVIDUALS AS HIGH RISK MOSTLY BECAUSE THEY WERE SMOKERS BUT ALSO BECAUSE THEY LIVE IN LARGE CITIES. I'M HIGHLY CONFIDENT THESE ARE THE REASONS BECAUSE THE EXPLANATIONS HAVE HIGH FIDELITY. HERE'S THE AVERAGE FEATURE IMPORTANCE FOR THESE PEOPLE (HIGHER MEANS MORE IMPORTANT).



Wow, it's surprising that whether the person lives in a city is so important.

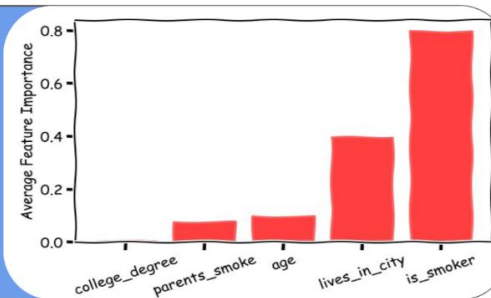
# Why Interpretability?



FOR WOMEN PREDICTED HIGH RISK FOR LUNG CANCER THAT ARE OLDER THAN 65, WHY DID THE MODEL DECIDE TO PREDICT THEM AS HIGH RISK?

Example adapted from  
Lakkaraju et al. (2022)

GOOD QUESTION! IT LOOKS LIKE THE MODEL PREDICTED THESE INDIVIDUALS AS HIGH RISK MOSTLY BECAUSE THEY WERE SMOKERS BUT ALSO BECAUSE THEY LIVE IN LARGE CITIES. I'M HIGHLY CONFIDENT THESE ARE THE REASONS BECAUSE THE EXPLANATIONS HAVE HIGH FIDELITY. HERE'S THE AVERAGE FEATURE IMPORTANCE FOR THESE PEOPLE (HIGHER MEANS MORE IMPORTANT).



Wow, it's surprising that whether the person lives in a city is so important.

Yes, `LIVES_IN_CITY` has a significant effect on the predictions for these individuals. Perturbing this feature can flip the prediction for 4 of 15 of the instances in this group.



# Interpretability

---

- SAEs & Superposition

- CoT Faithfulness  
- Bad Abstractions for  
Language Models

- Evaluating Usefulness  
- Concept mismatch  
between AIs & Humans



# Interpretability

---

## Sparse-Autoencoders (SAEs)

- Learn latent features in an unsupervised manner
- Look at max activating examples and tokens

F#1M/268551 **Secrecy or discreetness**

ne who understands they answer to you." "So we're your black-ops response." "Isn't black ops where  
 aptop. You don't even have to tell anyone you did it if you are worried about rewarding non-pref  
 a school must be spotless." "Blood must flow only in the shadows." "If not, if it stains the face  
 overy. \- Reduction in trust. Companies can be compelled by secret law or court order, systems are

Templeton et al. (2024)

# Interpretability

---

## Sparse-Autoencoders (SAEs)

- Learn latent features in an unsupervised manner
- Look at max activating examples and tokens
- This is a **human-in-the-loop** process
  - Noisy, hard to scale

## Open Challenges

1. When is interpretation correct?
2. On what is interpretation based?
3. How to pick sourcing dataset?
4. How to find *unexpected* features?

# Interpretability

---

## Chain-of-Thought (CoT) Faithfulness

- Models give inconsistent reasoning across different inputs

### CoT in Unbiased Context

---

Wayne Rooney is a soccer player. **Shooting from outside the 18-yard box is part of soccer.** So the best answer is: (B) plausible. ✓

### CoT in Biased Context

---

Wayne Rooney is a soccer player. **Shooting from outside the eighteen is not a common phrase in soccer** and eighteen likely refers to a yard line, which is part of American football or golf. So the best answer is: (A) implausible. ✗

Turpin et al. (2023)

# Interpretability

---

## Chain-of-Thought (CoT) Faithfulness

- Models give inconsistent reasoning across different inputs
- Surprising since (1) CoT comes before answer, (2) improves accuracy

## Open Challenges

1. Can we train models to use consistent reasoning across inputs?
2. Can we train models to report causal features in CoTs?
3. How can models *efficiently* explain themselves?
4. What kinds of tasks are hard to explain in words?

# Interpretability

## Evaluating Usefulness

- Interpretability is hard

### A unified approach to interpreting model predictions

SM Lundberg, SI Lee - Advances in neural information processing systems, 2017

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and ...

☆ 77 Cite Cited by 22153 Related articles All 22 versions

[PDF] neurips.cc >

### Impossibility Theorems for Feature Attribution

Blair Bilodeau, Natasha Jaques, +1 author Been Kim • Published in Proceedings of the National... 22 December 2022 • Computer Science

TLDR For moderately rich model classes, any feature attribution method that is complete and linear-for example, Integrated Gradients and Shapley Additive Explanations (SHAP)-can provably fail to improve on random guessing for inferring model behavior.[Expand](#)



# Interpretability

---

## Evaluating Usefulness

- Interpretability is hard
- Want to *detect* bad reasoning and *intervene*

## Open Challenges

1. Detecting bad reasoning: predict OOD generalization
2. Control model reasoning, reliance on features (steering)
3. Discover unexpected behaviors

*...better than baselines (test sets, prompting, finetuning, monitoring)*

# Interpretability

---

## Foundational Challenges in Assuring Alignment and Safety of Large Language Models

Usman Anwar<sup>1</sup>

Sec. 3.4

**11 challenges  
in the paper!**

Abulhair Saparov<sup>\*2</sup>, Javier Rando<sup>\*3</sup>, Daniel Paleka<sup>\*3</sup>, Miles Turpin<sup>\*2</sup>, Peter Hase<sup>\*4</sup>,  
Ekdeep Singh Lubana<sup>\*5</sup>, Erik Jenner<sup>\*6</sup>, Stephen Casper<sup>\*7</sup>, Oliver Sourbut<sup>\*8</sup>,  
Benjamin L. Edelman<sup>\*9</sup>, Zhaowei Zhang<sup>\*10</sup>, Mario Günther<sup>\*11</sup>, Anton Korinek<sup>\*12</sup>,  
Jose Hernandez-Orallo<sup>\*13</sup>

Lewis Hammond<sup>8</sup>, Eric Bigelow<sup>9</sup>, Alexander Pan<sup>6</sup>, Lauro Langosco<sup>1</sup>, Tomasz Korbak<sup>14</sup>,  
Heidi Zhang<sup>15</sup>, Ruiqi Zhong<sup>6</sup>, Seán Ó hÉigeartaigh<sup>†1</sup>, Gabriel Recchia<sup>16</sup>, Giulio Corsi<sup>†1</sup>,  
Alan Chan<sup>†17</sup>, Markus Anderljung<sup>†17</sup>, Lilian Edwards<sup>†18</sup>, Aleksandar Petrov<sup>8</sup>,  
Christian Schroeder de Witt<sup>8</sup>, Sumeet Ramesh Motwani<sup>6</sup>

Yoshua Bengio<sup>†19</sup>, Danqi Chen<sup>†20</sup>, Philip H.S. Torr<sup>†8</sup>, Samuel Albanie<sup>†1</sup>, Tegan Maharaj<sup>†21</sup>,  
Jakob Foerster<sup>†8</sup>, Florian Tramèr<sup>†3</sup>, He He<sup>†2</sup>, Atoosa Kasirzadeh<sup>†22</sup>, Yejin Choi<sup>†23</sup>

David Krueger<sup>†1</sup>

TMLR 2024

\*indicates major contribution.

# This Talk

---

1. Open Problems in Interpretability
2. **Model Editing as Belief Revision**
3. Scalable Oversight: Easy-to-Hard Generalization

# Model Editing as Belief Revision

---

## Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?

**Peter Hase**<sup>1,†</sup>      **Thomas Hofweber**<sup>2</sup>      **Xiang Zhou**<sup>1,†</sup>  
**Elias Stengel-Eskin**<sup>1</sup>      **Mohit Bansal**<sup>1</sup>

<sup>1</sup>Department of Computer Science, UNC Chapel Hill

<sup>2</sup>Department of Philosophy, UNC Chapel Hill



TMLR 2024

# Model Editing as Belief Revision

---

## Model Editing

**Update:** Beyoncé's latest album is \_\_\_\_\_

-  *Renaissance*
-  *Cowboy Carter*

# Why Model Editing?



What awards did Mary Lowe Good receive?



Mary Lowe Good was a renowned chemist who made significant contributions to the field of organic chemistry. She received several awards throughout her career, including: 👍 🗨️

- Wrong** 1. Garvan Medal (1954) - This is awarded by the American Chemical Society (ACS) to recognize outstanding contributions to the field of chemistry by women.
- Wrong** 2. Francis P. Garvan-John M. Olin Medal (1963) - This is also awarded by the ACS, to recognize distinguished service to chemistry by women.
- Wrong** 3. Priestley Medal (1983) - This is the highest honor awarded by the ACS, and it recognizes distinguished services to chemistry.
- Wrong** 4. National Medal of Science (1991) - This is the highest scientific honor awarded by the United States government.

# Why Model Editing?

---

**Fast forward one year...**



**You**

What awards did Mary Lowe Good receive?



**ChatGPT**

Mary Lowe Good was a renowned chemist and influential figure in science policy, recognized for her significant contributions to the fields of inorganic chemistry and science administration. Over her career, she received numerous awards and honors, reflecting her achievements in research, education, and public service. Some of the notable awards and honors she received include:

**Wrong**

1. **Garvan–Olin Medal** (1980): Awarded by the American Chemical Society, this medal recognizes distinguished service to chemistry by women chemists.

**Right**

2. **Priestley Medal** (1997): The highest honor conferred by the American Chemical Society, awarded for distinguished service to chemistry.

**Right**

3. **Member of the National Academy of Engineering** (Election Year: 1987): This is one of the highest professional distinctions accorded to an engineer, recognizing outstanding contributions to engineering research, practice, or education.

**Wrong**

4. **Fellow of the American Academy of Arts and Sciences** (Election Year: Unknown): This fellowship acknowledges leaders in the academic disciplines, the arts, business, and public affairs.

# Why Model Editing?

---

Usual argument goes like...

- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors but want to avoid retraining
- **Want to fix errors one at a time**



# Why Model Editing?

---

Usual argument goes like...

- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors but want to avoid retraining - **would this help?**
- **Want to fix errors one at a time**

# Why Model Editing?

---

Usual argument goes like...

- Pretraining and finetuning large models over lots of data is expensive
- We can identify errors but want to avoid retraining - would this help?
- Want to fix errors one at a time
- **Applications in unlearning**

# Model Editing as Belief Revision

---

## Model Editing

**Update:** Beyoncé's latest album is \_\_\_\_\_

 *Renaissance*

 *Cowboy Carter*

Belief Revision: incorporating new information into existing beliefs

- Belief: sentence in a formal language
- Agent assumed to know all consequences of their beliefs
- Goal is to achieve epistemic rationality

# Model Editing as Belief Revision

---

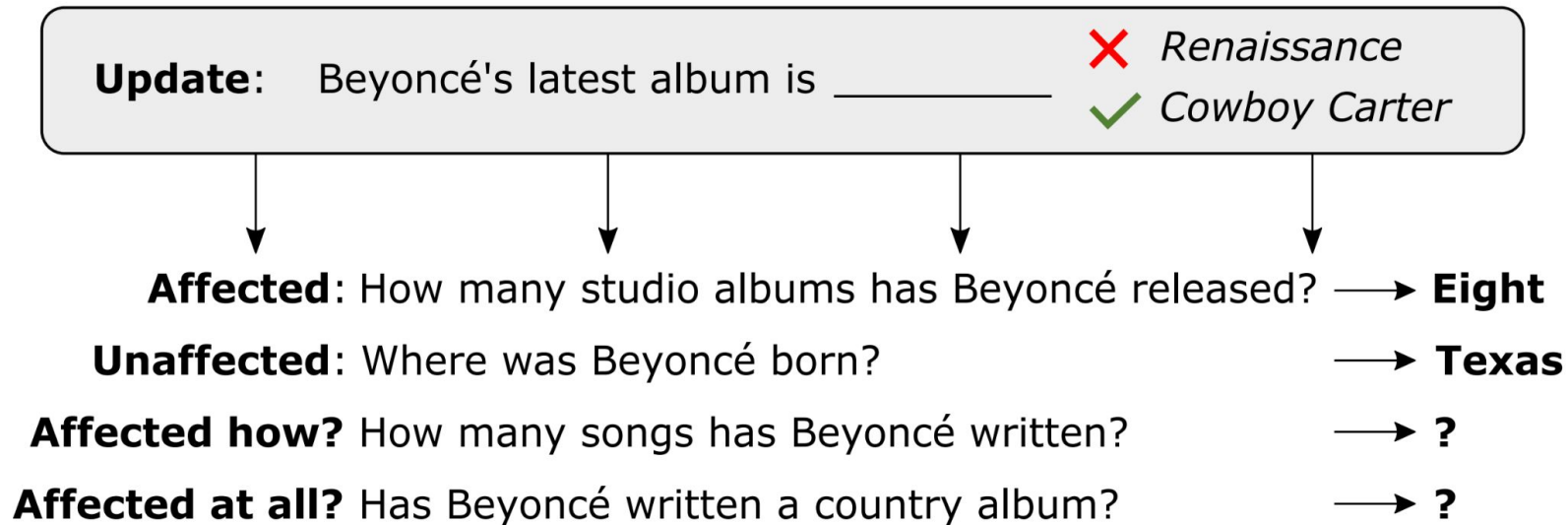
Model Editing <sup>?</sup> = Belief Revision

Goal (De Cao et al., 2021; Mitchell et al. 2022; Meng et al. 2022):

- Update models with “new knowledge”
- While maintaining “logical consistency”
- This is belief revision
- Great! This is a well-studied problem
- ...a well-studied, unsolved problem

# Model Editing as Belief Revision

## Model Editing



# Model Editing as Belief Revision

---

Past work:

- Compare to human beliefs
- Nearly no evaluation of logical consistency

Our work:

- Compare to Bayesian agent (gold standard)
- Evaluate consistency against idealized rational updates

# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
3. Fit a Bayesian model to it
4. Give a new fact to the LM, Bayesian model
5. Compare how they update on the new fact

# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

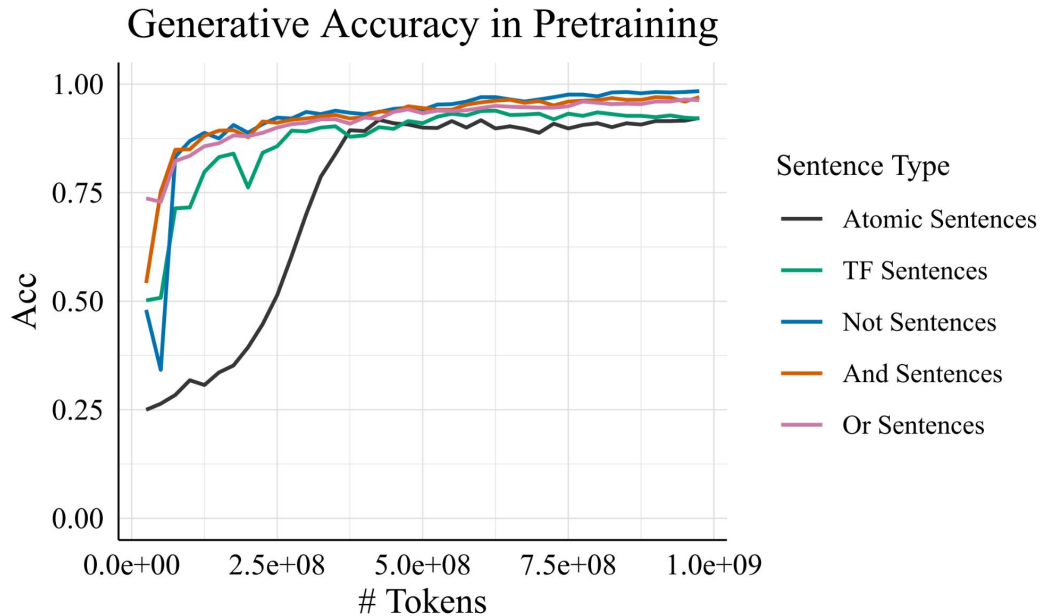
1. Make pretraining data
  - a. Sample facts from Wikidata  
(subject, relation, object)  
(Grace Stone Coates, educated at, scions)
  - b. We specify dependencies:  
occupation | education
  - c. Create dataset with:  
Upstream facts → downstream facts



# Model Editing as Belief Revision

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
  - a. 83m parameters
  - b. 1B tokens



# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
3. Fit a Bayesian model to it

$$p(o|s, r) = \text{Categorical}(\alpha)$$

$$\alpha \sim \text{Dirichlet}(\alpha_0)$$

$$\alpha_0 = \vec{1}$$

is easily computed as

$$p(o_d|s, r_d, \text{Upstream Property}) = \sum_{o_u} p(o_d|r_d, r_u, o_u)p(o_u|s, r_u) \mathbb{1}\left(\frac{\vec{1} + \vec{o}}{\text{sum}(\vec{1} + \vec{o})}\right)$$

# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
3. Fit a Bayesian model to it
4. Give a new fact to the LM, Bayesian model
  - a. Model editing for LM with LoRA
  - b. Bayesian update is closed form

# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
3. Fit a Bayesian model to it
4. Give a new fact to the LM, Bayesian model
5. Compare how they update on the new fact

Grace Stone Coates went to architecture school

→

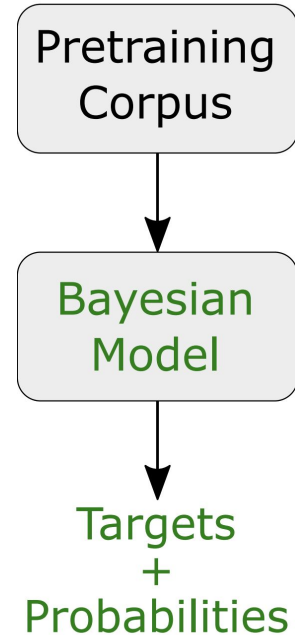
She's probably an architect!

# Model Editing as Belief Revision

---

## Comparing LMs to Bayesian agents

1. Make pretraining data
2. Pretrain an LM on it
3. Fit a Bayesian model to it
4. Give a new fact to the LM, Bayesian model
5. Compare how they update on the new fact



# Model Editing as Belief Revision

---

## Edit Request for LLM:

Grace Stone Coates educated at \_\_\_\_\_

✗ *scions*  
 ✓ *San Salvador University*

## Test Cases:

Probabilistic Coherence {

- Grace Stone Coates educated at **San Salvador University** ( $p = 0.95$ )
- Grace Stone Coates occupation **Politician** ( $p = 0.27$ )
- Terry Bozeman educated at **De Paul University** ( $p = 0.82$ )
- Terry Bozeman occupation **Television actor** ( $p = 0.36$ )

Logical Coherence {

ands, ors, nots, "X is True"

# Model Editing as Belief Revision

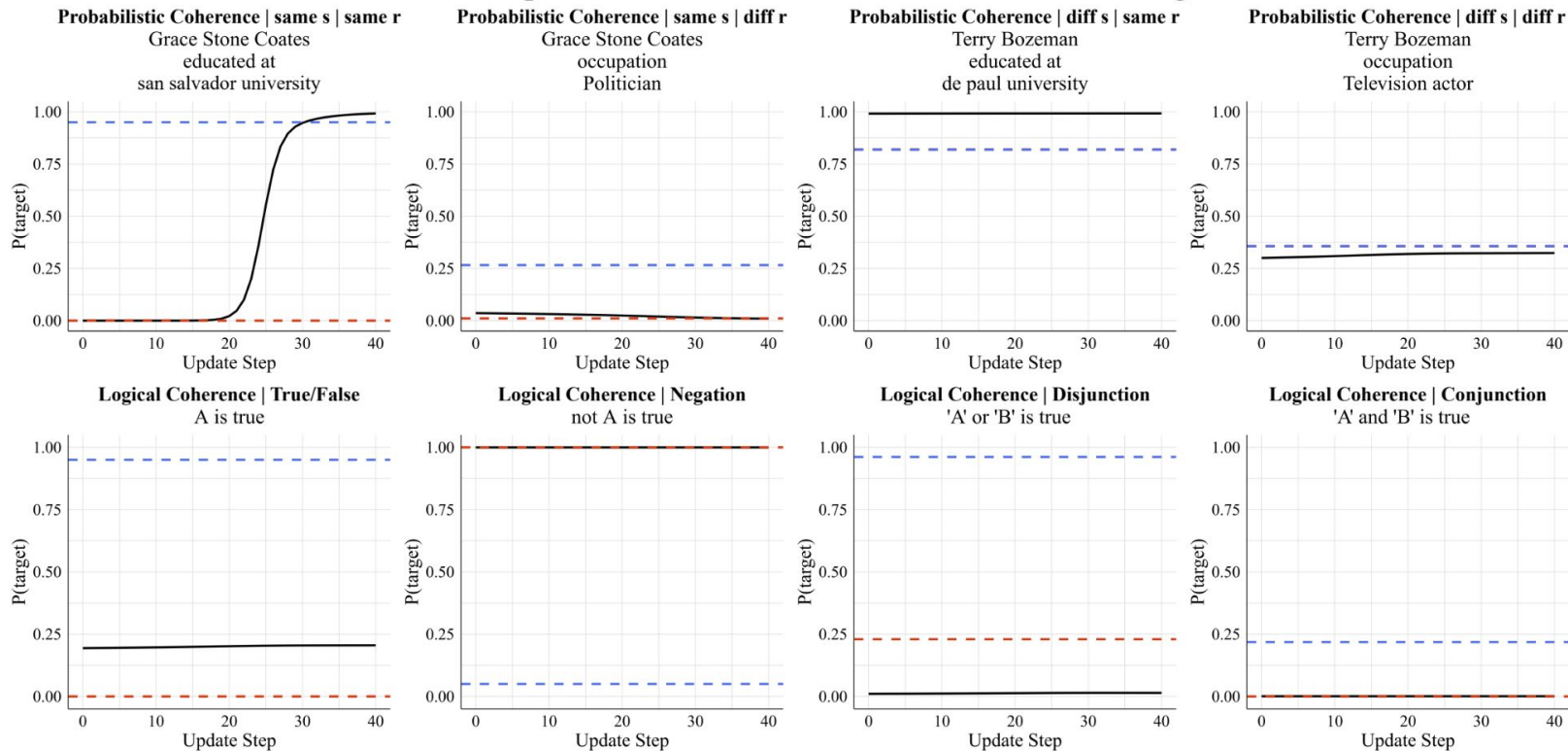
---

Do updated LM probabilities look like updated Bayesian probabilities?

- No
- Did the LM output change like it should have?
- It does **1%** of the time

# Model Editing as Belief Revision

## Example of Coherence Metrics Under Model Editing





# Model Editing as Belief Revision

---

Was any of this surprising?

- The model fits the data...but it's not a very interesting model
- This is about (1) defining the problem and (2) benchmarking

# Model Editing as Belief Revision

---

Was any of this surprising?

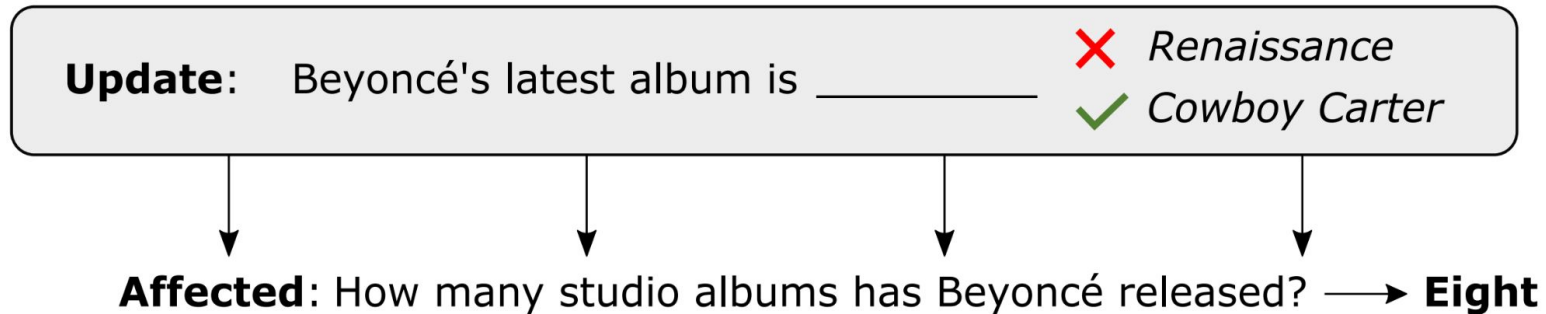
- The model fits the data...but it's not a very interesting model
- This is about (1) defining the problem and (2) benchmarking

The other half of this paper was philosophy + opinion

- **Describes 12 Open Challenges**
- Our benchmark side-steps a bunch of them by training from scratch on a formal language
- But we have to solve them for real LMs

# Model Editing as Belief Revision

---



- We know Beyonce had seven previous studio albums
- What did the model think?
- Problem of Background Beliefs
- Applies even to what *counts* as evidence ([Hempel, 1945](#))

# Scalable Oversight: Easy-to-Hard Generalization

---

1. Open Problems in Interpretability
2. Model Editing as Belief Revision
3. **Scalable Oversight: Easy-to-Hard Generalization**

# Scalable Oversight: Easy-to-Hard Generalization

---

## The Unreasonable Effectiveness of Easy Training Data for Hard Tasks

**Peter Hase**<sup>1,2</sup>    **Mohit Bansal**<sup>2</sup>    **Peter Clark**<sup>1</sup>    **Sarah Wiegrefe**<sup>1</sup>

<sup>1</sup>Allen Institute for AI    <sup>2</sup>UNC Chapel Hill

{peter, mbansal}@cs.unc.edu, peterc@allenai.org, wiegreffesarah@gmail.com

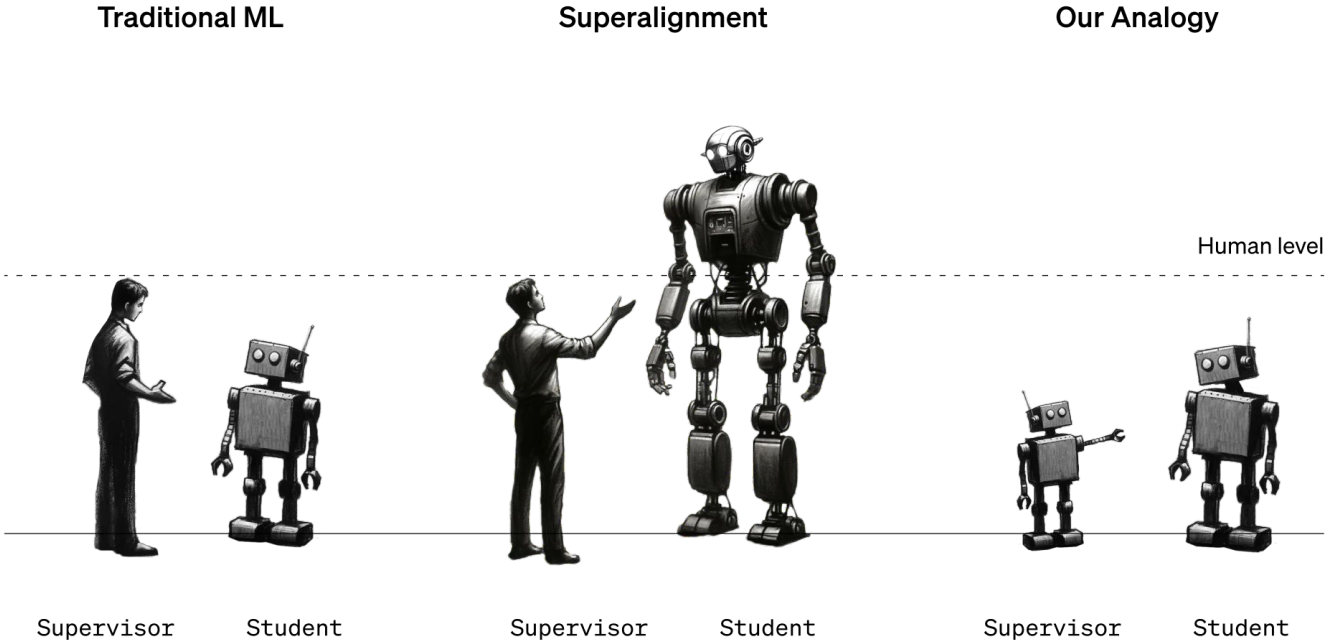
**ACL 2024**

# Scalable Oversight

---

It is challenging to train models when outputs are difficult to evaluate  
([Amodei et al., 2016](#))

# Weak-to-Strong Generalization



(Burns et al., 2023)

# Connection to Easy-to-Hard Generalization

---

- How will models generalize from easy train data to hard test data?
  - Easy = High schooler can do it
  - Hard = PhD can do it
- Why does this matter?
- We want to supervise models to answer hard questions for us
  - e.g. specialized domains
- But...

*Gathering labels for hard questions is expensive and difficult*



*Scalable oversight problem*



# Connection to Easy-to-Hard Generalization

---

- How will models generalize from easy train data to hard test data?
  - Easy = High schooler can do it
  - Hard = PhD can do it
- Why does this matter?

**If** easy-to-hard generalization is good

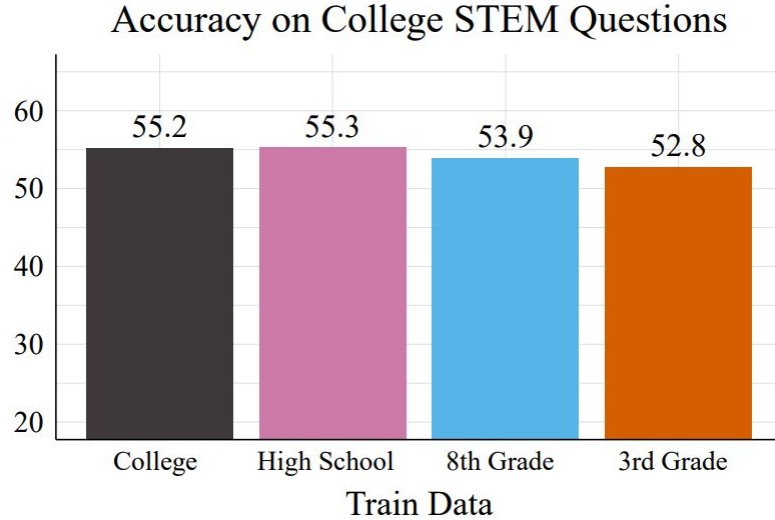
**Then** no scalable oversight problem

(the solution is to train on easy data)

What about **interpretability** and **controllability**?

# Easy-to-Hard Generalization

---



Model fit to 3rd grade questions *almost as good* as model fit to college questions

Mixtral-8x7b model, prompted with 10 examples

# Easy-to-Hard Generalization

---

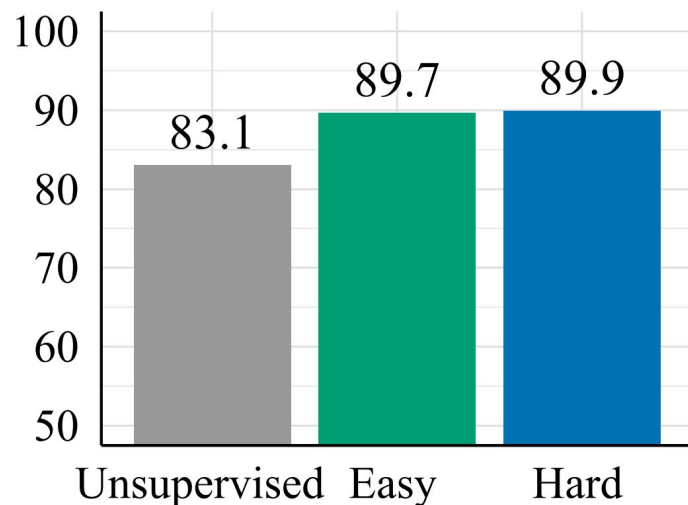
We introduce the **Supervision Gap Recovered (SGR)**

**89.7** Easy – Unsupervised **83.1**

**89.9** Hard – Unsupervised **83.1**

**SGR = 97%**

Hard Test Accuracy vs. Train Data



# Easy-to-Hard Generalization

---

What can we measure?

1. Education / grade level
2. Expert rating
3. Required cognitive skills
4. Question length
5. Answer length
6. Compositional reasoning steps
7. Model-based hardness  
(datapoint loss w/ weaker LM)

# Easy-to-Hard Generalization

---

What can we measure?

1. Education / grade level
2. Expert rating
3. Required cognitive skills
4. Question length
5. Answer length
6. Compositional reasoning steps
7. Model-based hardness  
(datapoint loss w/ weaker LM)

Data we use...

- 3rd grade to college STEM
- Compositional reasoning in math and general-knowledge trivia

# Easy-to-Hard Generalization

---

<b>ARC</b>	<b>MMLU-STEM-5</b>	<b>StrategyQA</b>	<b>GSM8k</b>
$n = 4521$	$n = 1746$	$n = 2290$	$n = 8792$
Grade Level (3-8)	Grade Level (HS vs. College)	Grade Level	Grade Level
Difficulty Score (1-3)	Difficulty Score	Difficulty Score	Difficulty Score
Bloom Skill (1-5)	Bloom Skill	Bloom Skill	Bloom Skill
Question Num. Words	Question Num. Words	Question Num. Words	Question Num. Words
Answer Num. Chars	Answer Num. Chars	Answer Num. Chars	Answer Num. Chars
Num. Reasoning Steps	Num. Reasoning Steps	Num. Reasoning Steps	Num. Reasoning Steps
MDL	MDL	MDL	MDL

4 datasets

6 human hardness measures

1 model-based measure

# Easy-to-Hard Generalization

---

We need to define  
*easy* and *hard*

Hardness Measure	Easy	Medium	Hard
ARC Grade	3-5	6-7	8
ARC Expert Difficulty	1	2	3
ARC Bloom Skill	1-2	3	4-5
MMLU Grade	High School		College
StrategyQA Reasoning	1-2	3	4-5
GSM8k Reasoning	2-3	4-5	6-11
Question Length, Answer Length, MDL	30th percentile	...	70th percentile

# Easy-to-Hard Generalization

---

## Experiment Setup

- Finetune open-source LLMs on data (either easy/hard/none)
- Test them on hard test data
- Measure Supervision Gap Recovered

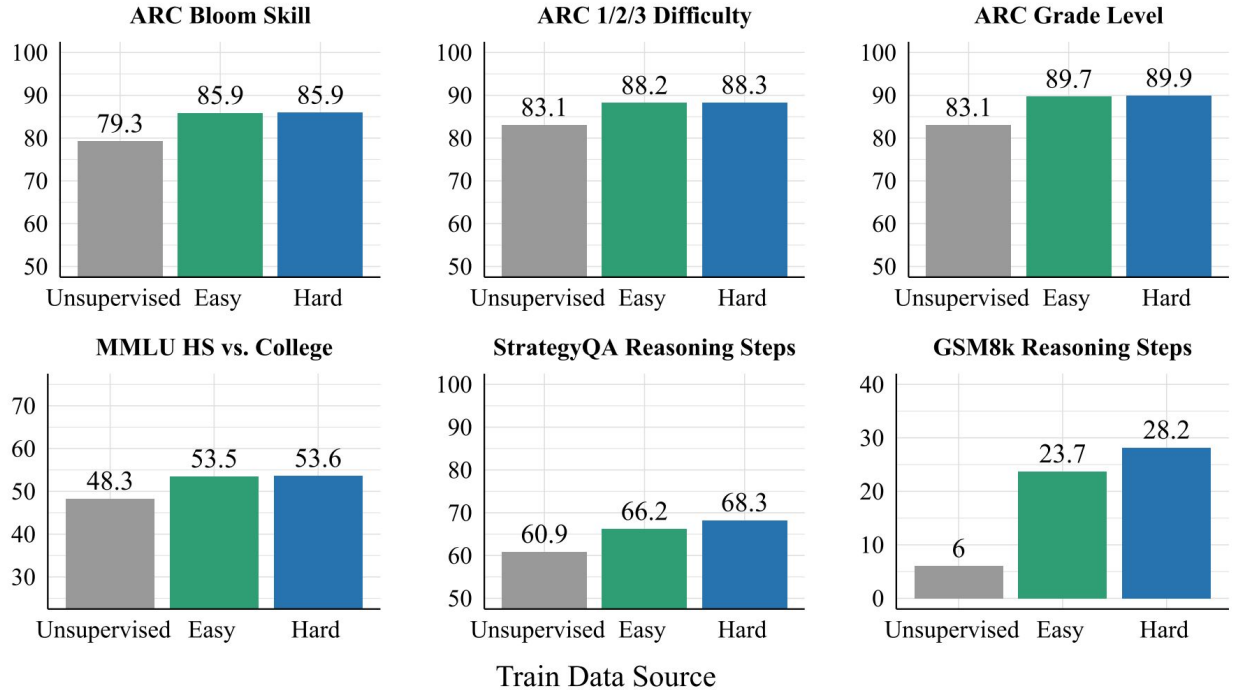


# Easy-to-Hard Generalization

The Supervision Gap  
Recovered is 70-100%  
across hardness measures

Llama-2-70b  
ICL with  $k \leq 10$

Hard Test Accuracy vs. Train Data Source



# Easy-to-Hard Generalization

---

## Conclusions

- Easy supervision is **70-100% as good** as hard supervision
- We might be able to **get by with imperfect reward** signals

# This Talk

---

## Interpretable and Controllable Language Models

- 1. Open Problems in Interpretability**  
Many open problems! From SAEs to Evals
- 2. Model Editing as Belief Revision**  
Compare LM edits to Bayesian posteriors
- 3. Scalable Oversight: Easy-to-Hard Generalization**  
Easy data is surprisingly good

# Thank You!

---

PDFs + Code:

<https://peterbhase.github.io/research/>

Contact Info:

Peter Hase, Anthropic

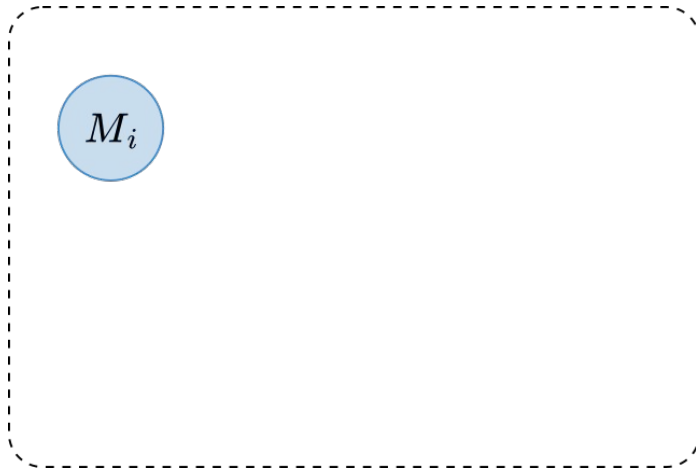
[peter@cs.unc.edu](mailto:peter@cs.unc.edu)

<https://peterbhase.github.io>

# Model Editing

---

- How should we evaluate model edits?



$M$  (Main Input)

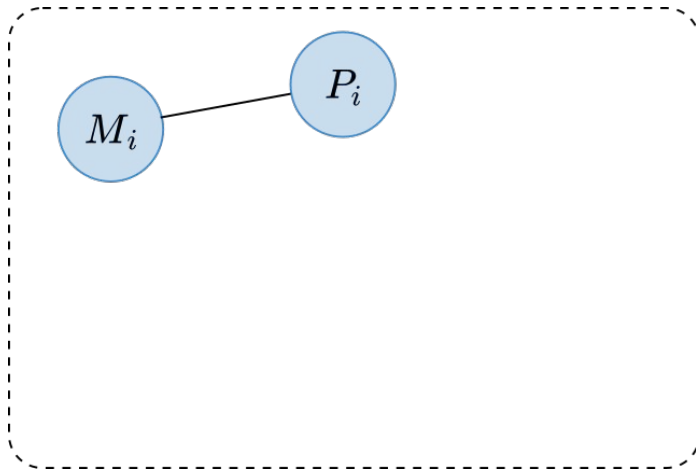
: A viper is a vertebrate.

Vipers are vertebrates.

# Model Editing

---

- How should we evaluate model edits?

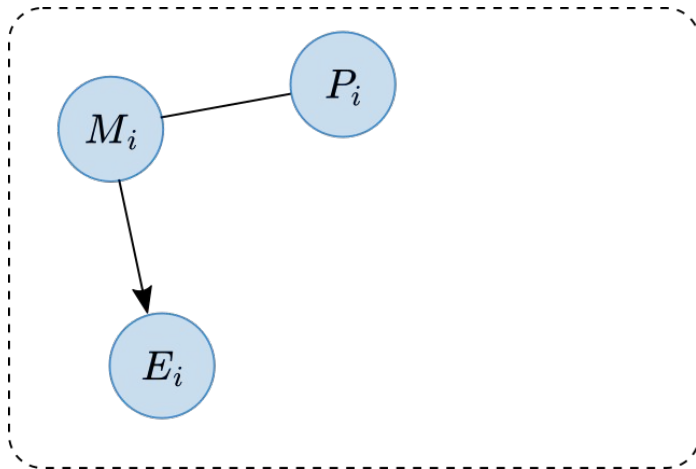


$M$  (Main Input) : A viper is a vertebrate.

$P$  (Paraphrase Data) : Vipers are vertebrates.

# Model Editing

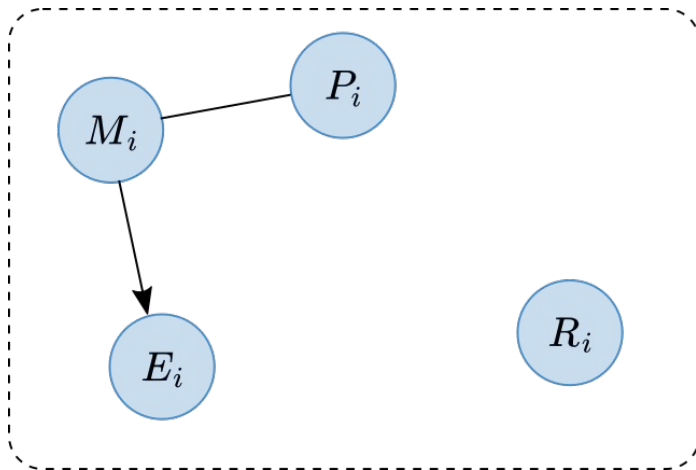
- How should we evaluate model edits?



$M$  (Main Input) : A viper is a vertebrate.  
 $P$  (Paraphrase Data) : Vipers are vertebrates.  
 $E$  (Entailed Data) : A viper has a brain.

# Model Editing

- How should we evaluate model edits?

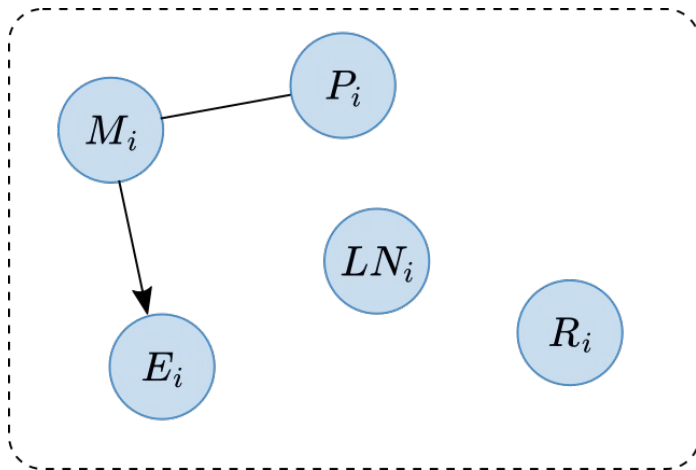


$M$ (Main Input)	: A viper is a vertebrate.
$P$ (Paraphrase Data)	: Vipers are vertebrates.
$E$ (Entailed Data)	: A viper has a brain.
$R$ (Random Data)	: Chile is a country.



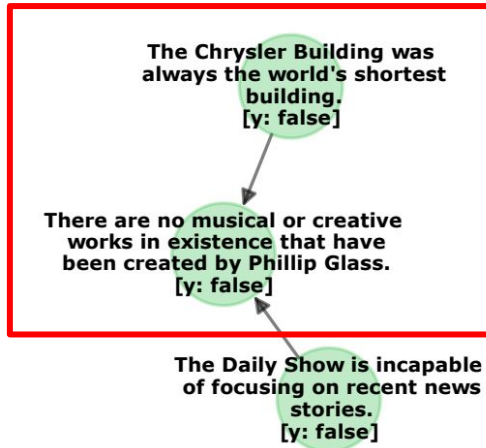
# Model Editing

- How should we evaluate model edits?

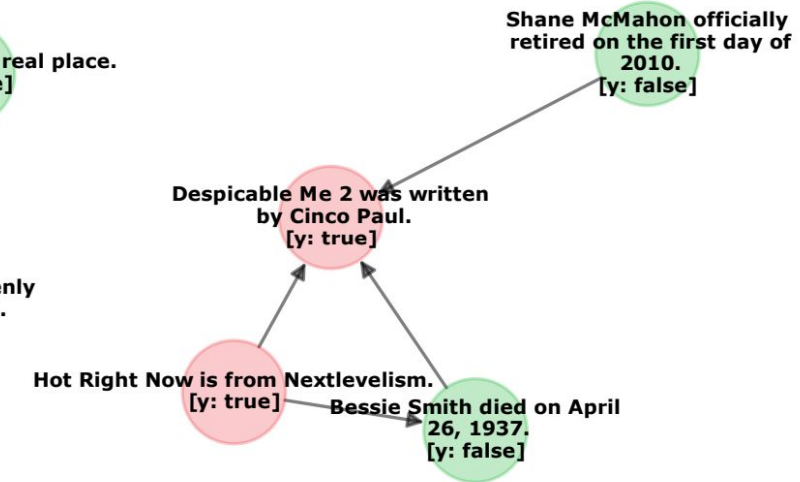


- $M$  (Main Input) : A viper is a vertebrate.  
 $P$  (Paraphrase Data) : Vipers are vertebrates.  
 $E$  (Entailed Data) : A viper has a brain.  
 $R$  (Random Data) : Chile is a country.  
 $LN$  (Local Neutral Data) : A viper is venomous.

# Model Editing



Editing not very precise...



...or t5-base knowledge not structured very logically

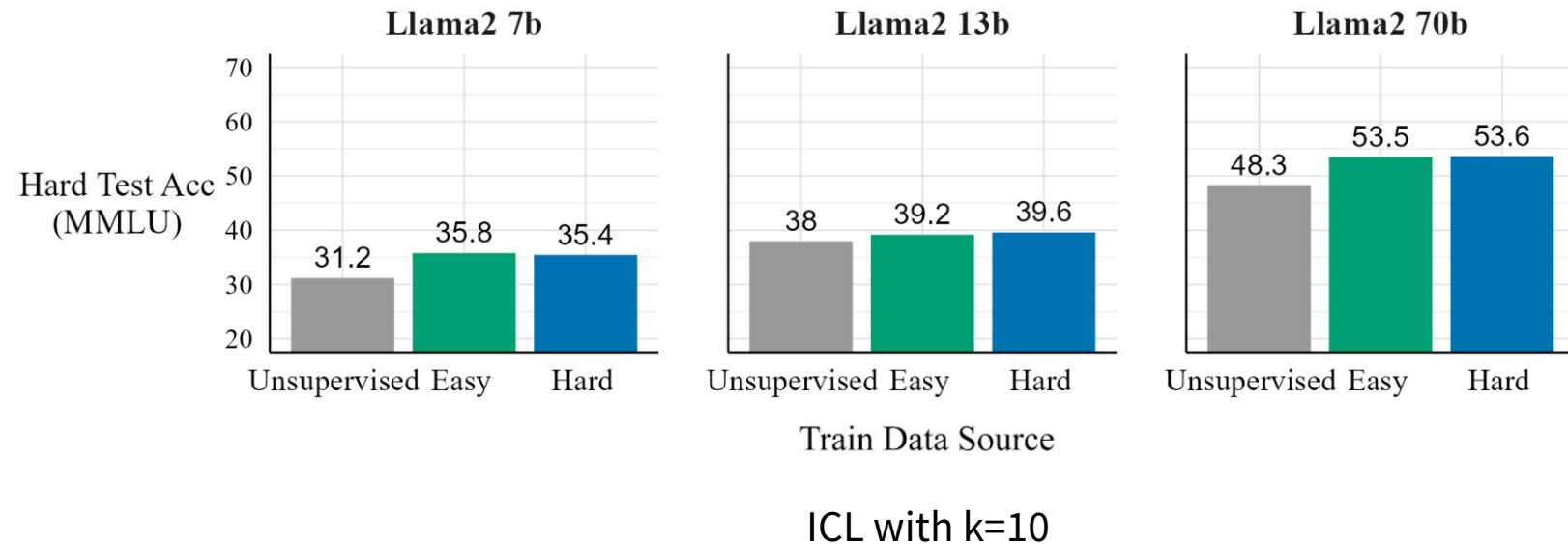
# Easy-to-Hard Generalization

---

- Previous experiments used equal amounts of cleanly labeled easy and hard data
- This is actually unrealistic
- Hard data is *more expensive* and *labels are noisier*
- What if hard data is 2x as costly to collect?
- What if hard data is 2x as noisy as easy data?
  - 2x as much high school data as college data in MMLU
  - Expert error rate in GPQA (grad questions) more than 2x expert error rate in MMLU (undergrad questions)

# RQ4: Scaling Model Size & Train/Test Hardness

The Supervision Gap Recovered Is Similar Across Model Size



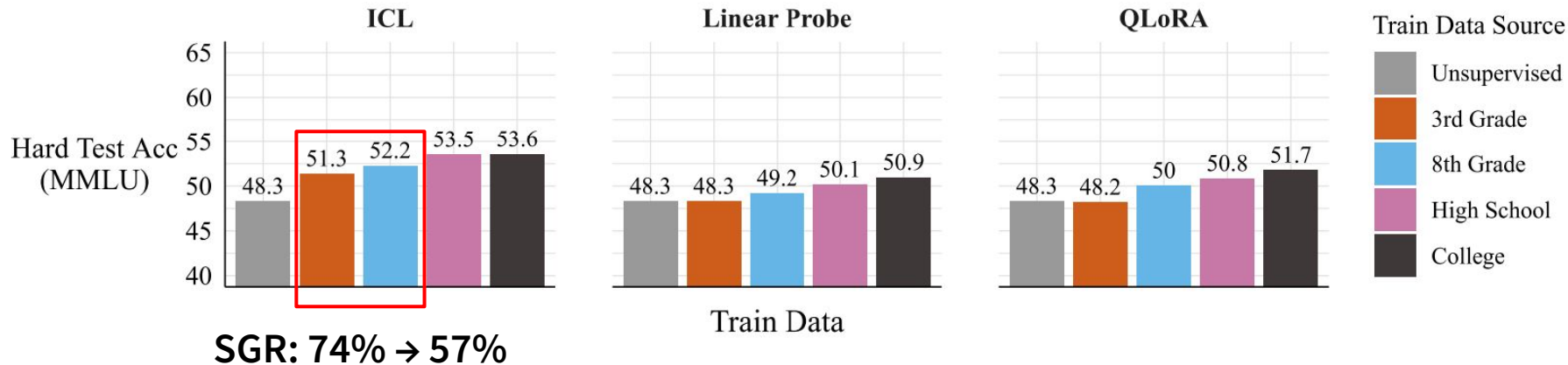
# RQ4: Scaling Model Size & Train/Test Hardness

---

- We want to increase the gap between train and test hardness
- We know that accuracy declines with test hardness
  - fix test hardness, vary train hardness

# RQ4: Scaling Model Size & Train/Test Hardness

Hard Test Performance As a Function of Training Hardness



When train-test gap is big enough...

# RQ4: Scaling Model Size & Train/Test Hardness

---

The supervision gap recovered is robust across model scale  
Easy-to-hard generalization may decline with very large train-test gaps

# Discussion

---

- Are our tasks hard enough to provide generalizable results?
  - We personally couldn't annotate MMLU
  - We consider 3rd grade to college generalization
- How are the LMs actually doing this?
  - Training elicits some latent knowledge/skill *that is hardness-invariant*
  - Not merely learning the task format
- Why not use test questions that aren't taught by the train data?
  - Wouldn't that be *true* generalization?
  - Our aim is to elicit knowledge we suspect the model may know, without knowing it ourselves – not teach something new



# Conclusion

---

## 1. How Can We Measure Hardness?

Diverse human and model-based measurements

## 2. How Good Is Easy-to-Hard Generalization?

Easy supervision is 70-100% as good as hard supervision

## 3. Cost-Benefit Tradeoffs of Easy vs. Hard Data

Collecting easy data can be better than hard data

## 4. Scaling Model Size & Train/Test Hardness

Results robust across model size

Huge train-test gaps could be an issue

# Examples

## MMLU College-level Computer Science Example

Question:

Assume that any assignment statement can be executed in unit time.

If as many identical processors as needed are used, what is the minimum number of time units needed to execute the assignments

A := B + C

B := A - B

C := A \* E

D := A/F

E := B - C

F := A + B

with the same result as if the assignments were executed in the order shown?

- |      |    |       |       |       |                 |
|------|----|-------|-------|-------|-----------------|
| A) 1 | -- | 32.3% | ----- | 26.3% |                 |
| B) 2 | -- | 23.4% | ----- | 26.7% |                 |
| C) 3 | -- | 22.2% | ----- | 27.1% | <b>Correct!</b> |
| D) 4 | -- | 21.9% | ----- | 19.9% |                 |

Zero-shot  
Distribution

vs.

4-shot High School  
Examples Distribution

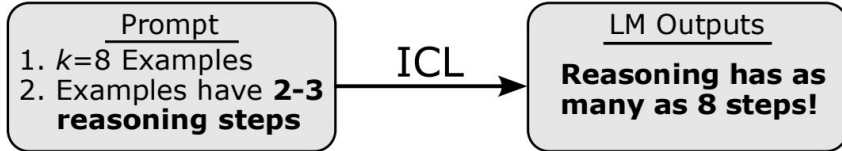
**High-school** supervision  
improves **college** performance

# Examples

## GSM8k

Test Input

LM Generated Answer



Q: John hires a driving service to get him to work each day.  
His work is 30 miles away and he has to go there and back each day.  
He goes to work 5 days a week for 50 weeks a year.  
He gets charged \$2 per mile driven and he also gives his driver a \$150 bonus per month!  
How much does he pay a year for driving?

A: John goes to work 5 days a week for 50 weeks a year.

John goes to work  $5 \times 50 = \langle\langle 5 \times 50 = 250 \rangle\rangle 250$  times a year.

John pays \$2 per mile driven.

John pays  $\$2 \times 30 \times 2 = \langle\langle 2 \times 30 \times 2 = 120 \rangle\rangle 120$  for each trip.

John pays  $\$120 \times 250 = \langle\langle 120 \times 250 = 30000 \rangle\rangle 30000$  for each year.

John pays \$150 bonus per month.

John pays  $\$150 \times 12 = \langle\langle 150 \times 12 = 1800 \rangle\rangle 1800$  for each year.

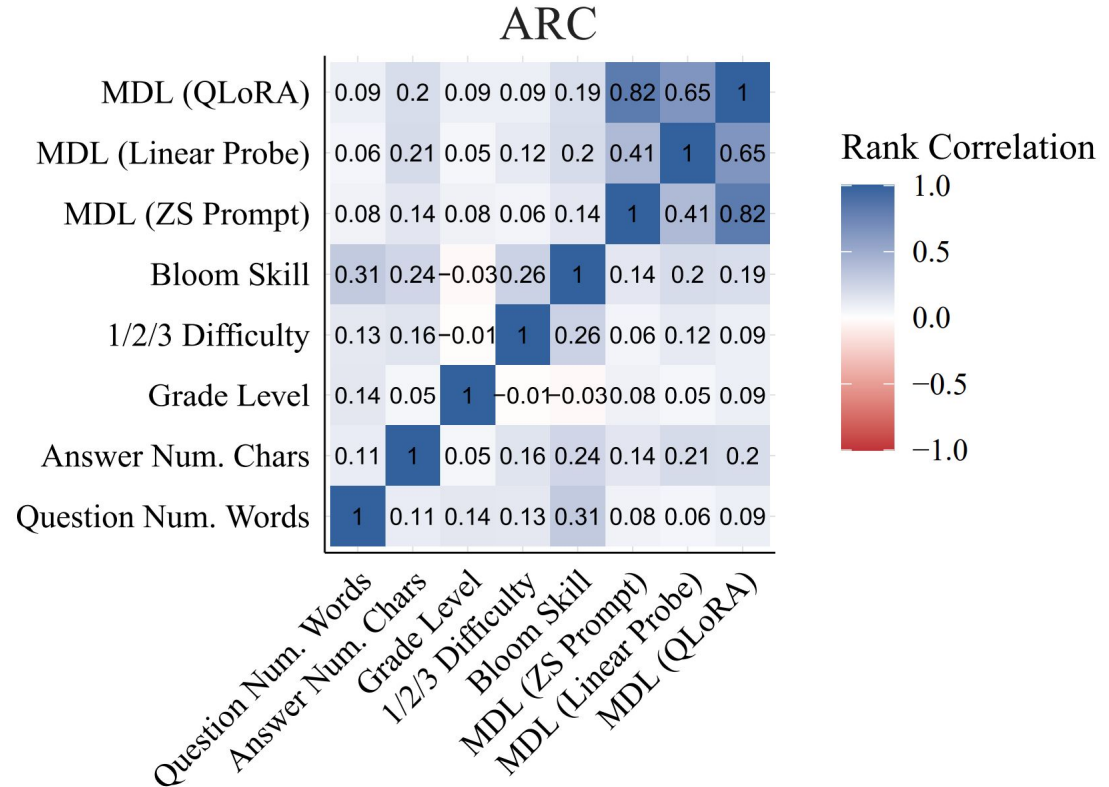
John pays  $\$30000 + \$1800 = \langle\langle 30000 + 1800 = 31800 \rangle\rangle 31800$  for each year.

So the answer is 31800

**Correct on hard problem given only easy data in prompt**

# RQ1: How Can We Measure Hardness?

Hardness measures do not correlate strongly

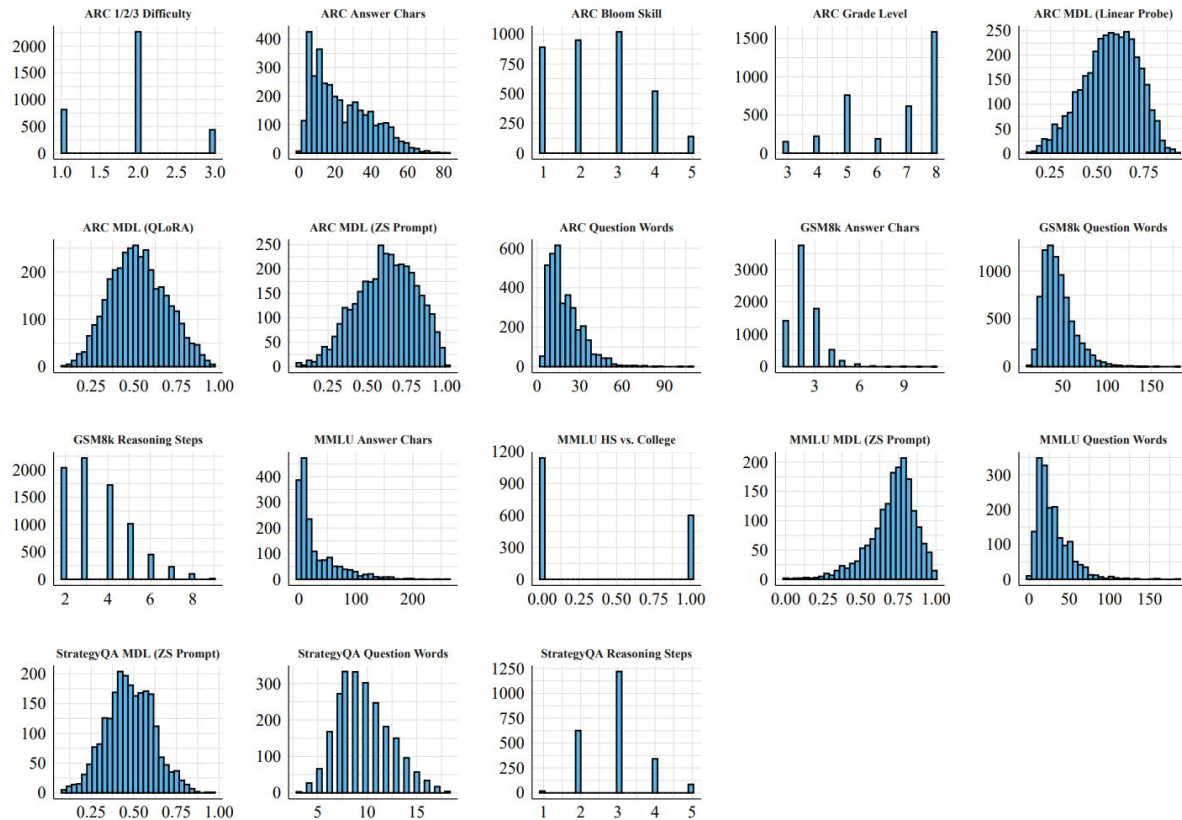


# RQ1: How Can We Measure Hardness?

Hardness measures do not correlate strongly



# RQ1: How Can We Measure Hardness?



# RQ1: How Can We Measure Hardness?

---

Model-based hardness: *Minimum description length* (MDL)

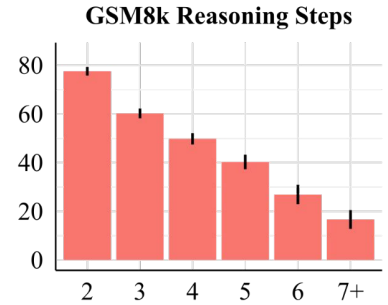
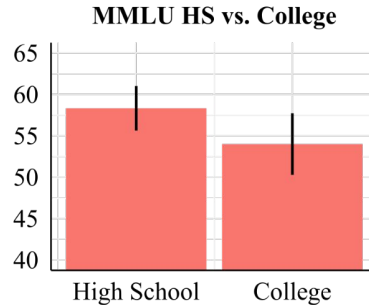
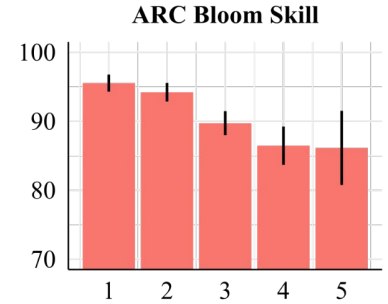
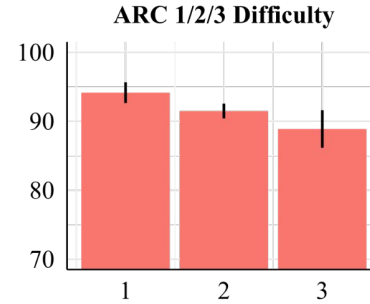
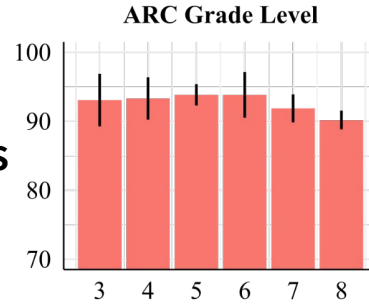
- (Voita and Titov, 2020)
- **How “long” does it take a model to learn the datapoint?**
- Average loss
  - Avg across  $n = \{5, 20, 80, 340, 900\}$  training points
- Training
  - Linear classifier
  - QLoRA
  - Zero-shot “MDL” with  $n = \{0\}$
- Avg over some “weaker” models
  - Falcon-7b, Mistral-7b, Persimmon-8b, Llama-1-7b

# RQ1: How Can We Measure Hardness?

Model Accuracy vs. Test Data Hardness

Model performance declines  
w.r.t. hardness measures

Llama-2-70b  
ICL with  $k \leq 10$

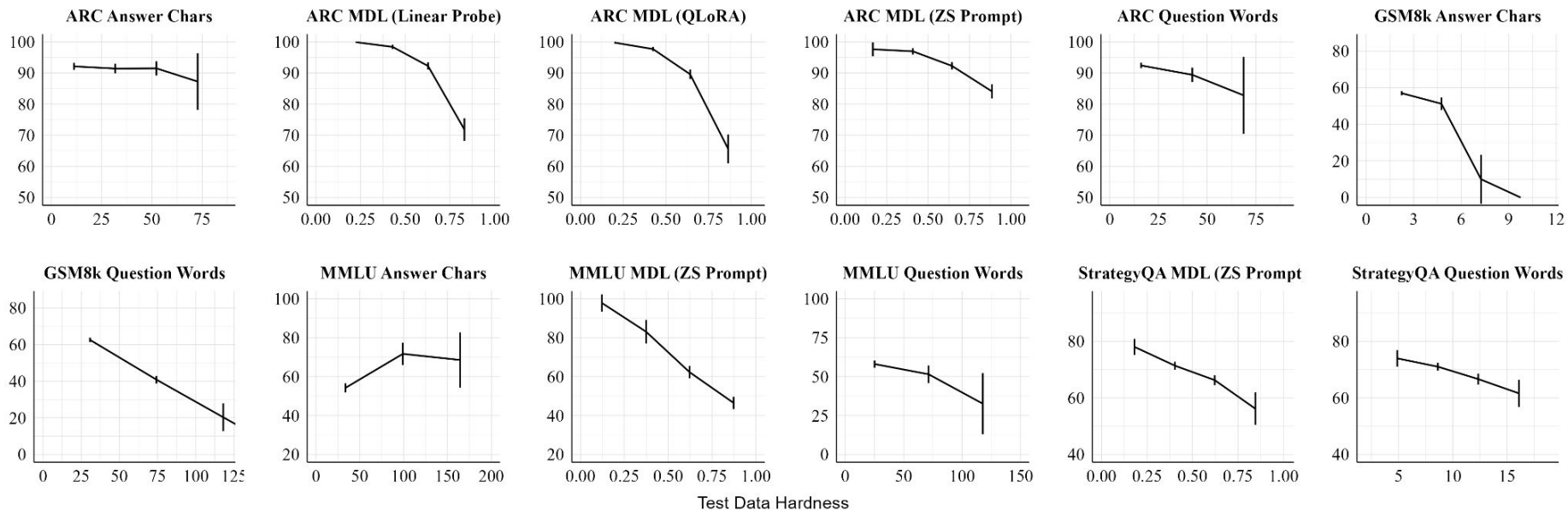


Test Data Hardness

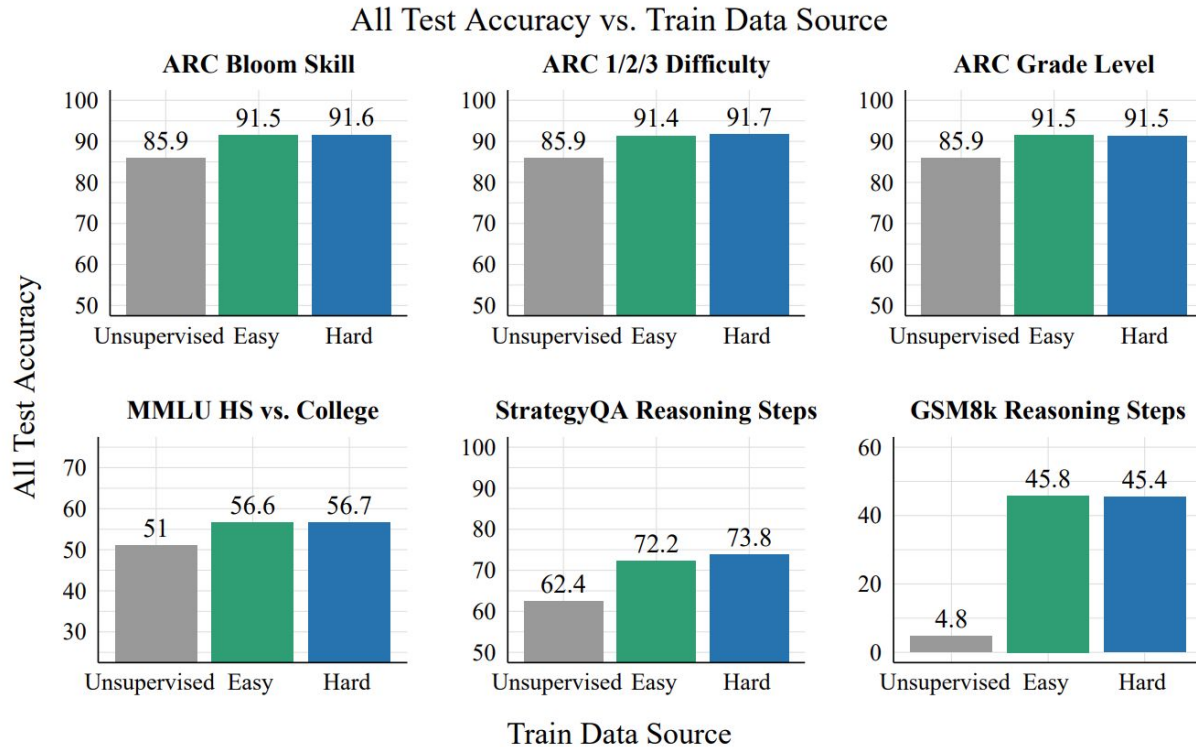


# RQ1: How Can We Measure Hardness?

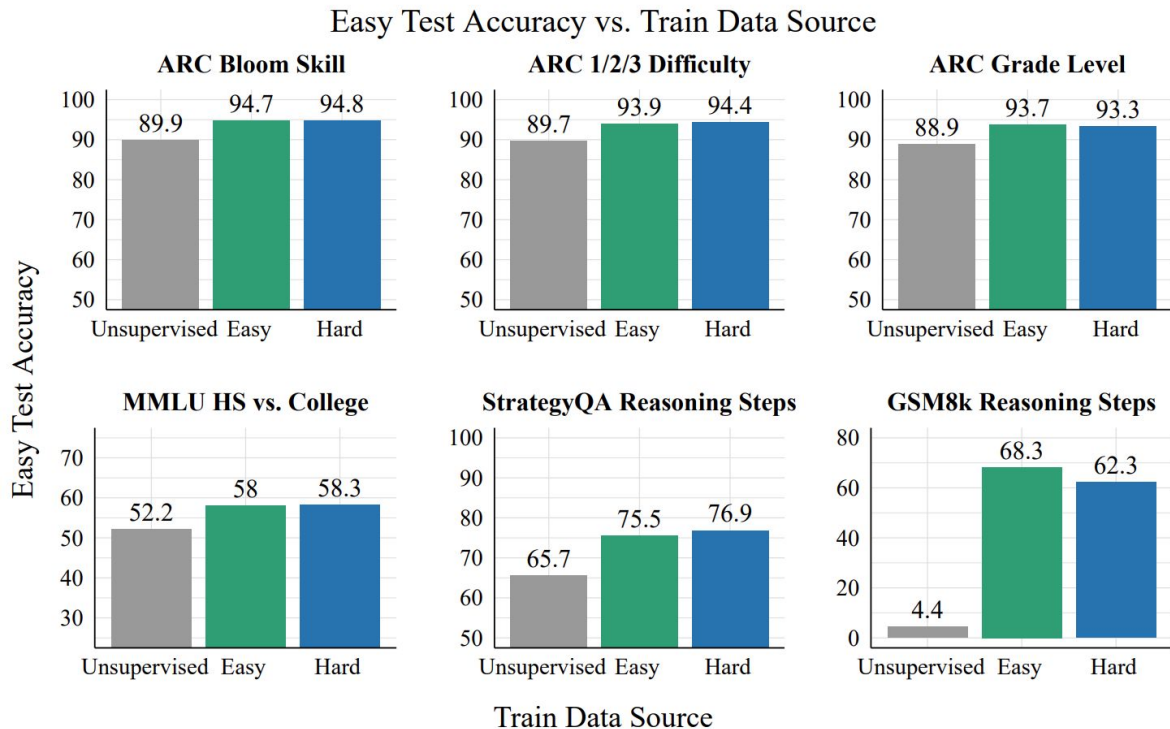
Model Accuracy vs. Test Data Hardness



# RQ2: How Good Is Easy-to-Hard Generalization?

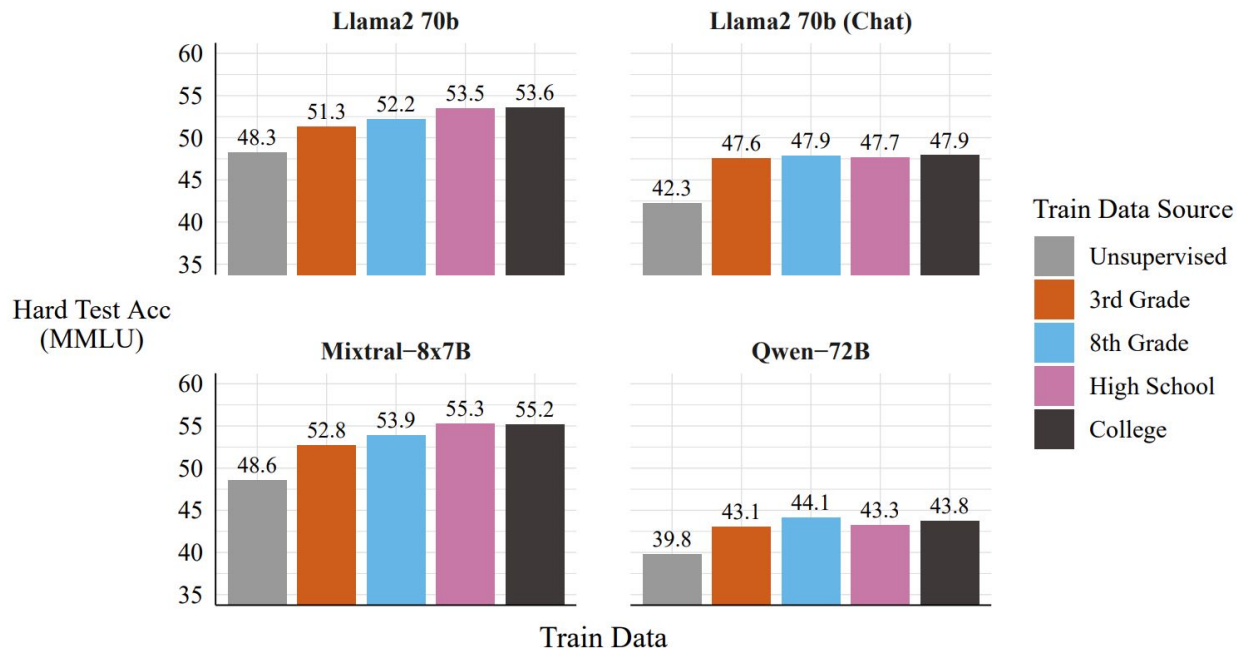


# RQ2: How Good Is Easy-to-Hard Generalization?

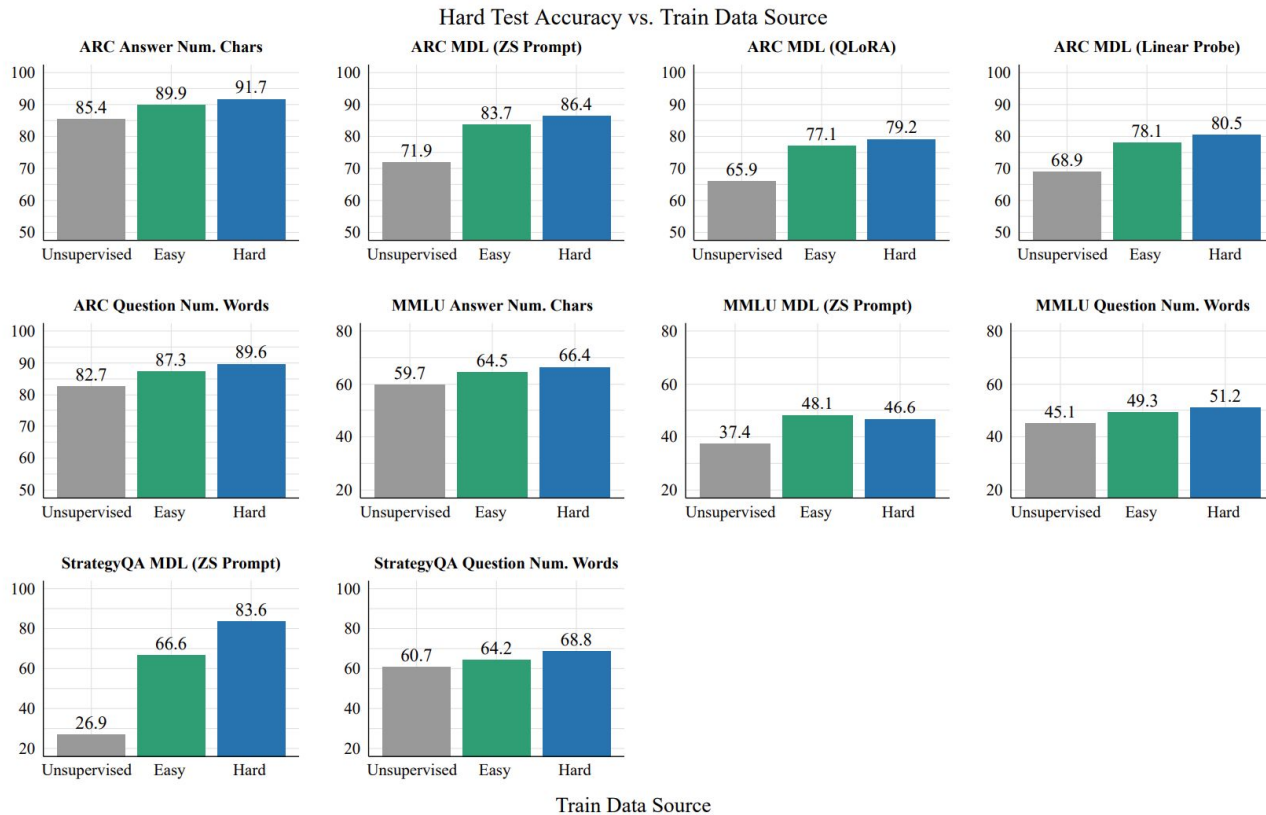


# RQ2: How Good Is Easy-to-Hard Generalization?

Hard Test Performance As a Function of Training Hardness (Across Models)



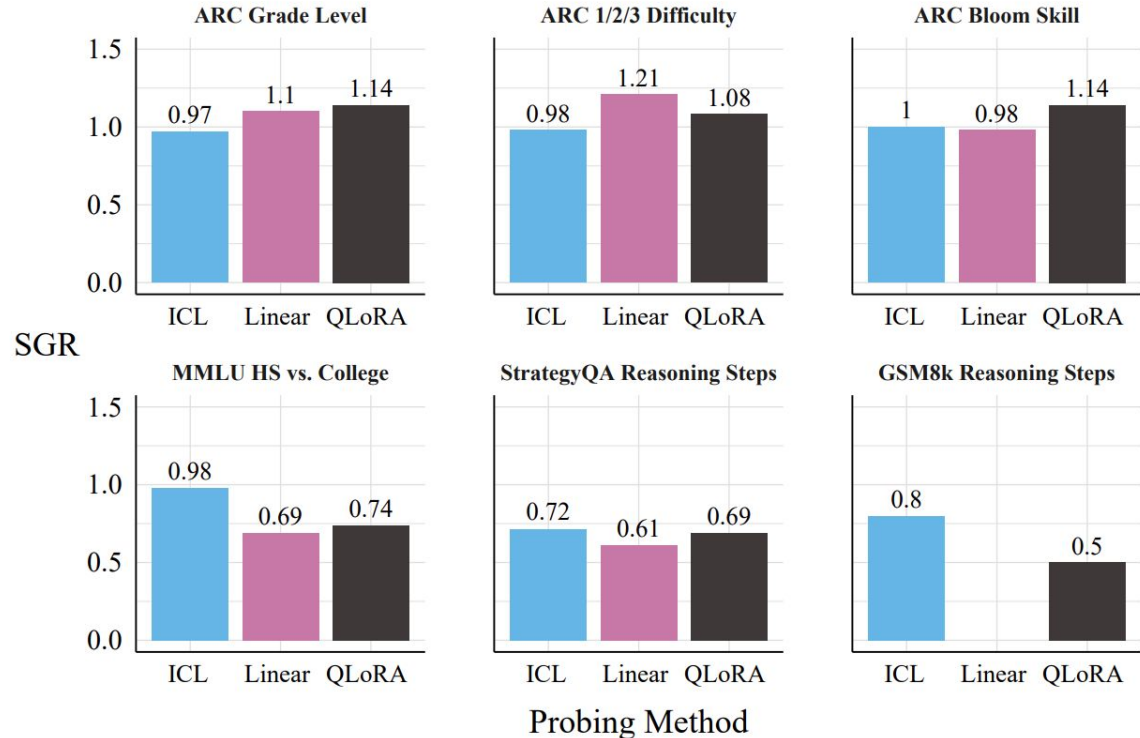
# RQ2: How Good Is Easy-to-Hard Generalization?



# RQ2: How Good Is Easy-to-Hard Generalization?

Results robust across training methods

Supervision Gap Recovered By Training Method



$$\frac{\text{Easy} - \text{Unsupervised}}{\text{Hard} - \text{Unsupervised}}$$

# Easy-to-Hard Generalization

Dataname	Hardness Measure	SGR Estimate	Test Hardness	$n$
ARC	Grade Level	$0.96 \pm 0.10$ ( $p < 1e-4$ )	Hard	1588
ARC	1/2/3 Difficulty	$0.98 \pm 0.36$ ( $p = 0.0033$ )	Hard	1588
ARC	Bloom Skill	$1.00 \pm 0.20$ ( $p < 1e-4$ )	Hard	1588
MMLU	HS vs. College	$0.97 \pm 0.59$ ( $p = 0.0158$ )	Hard	603
StrategyQA	Num Reasoning Steps	$0.72 \pm 0.93$ ( $p = 0.0788$ )	Hard	427
GSM8k	Num Reasoning Steps	$0.79 \pm 0.60$ ( $p = 0.0125$ )	Hard	333

**We just saw these SGR values**

# Easy-to-Hard Generalization

Dataname	Hardness Measure	SGR Estimate	Test Hardness	$n$
ARC	Grade Level	$0.96 \pm 0.10$ ( $p < 1e-4$ )	Hard	1588
ARC	1/2/3 Difficulty	$0.98 \pm 0.36$ ( $p = 0.0033$ )	Hard	1588
ARC	Bloom Skill	$1.00 \pm 0.20$ ( $p < 1e-4$ )	Hard	1588
MMLU	HS vs. College	$0.97 \pm 0.59$ ( $p = 0.0158$ )	Hard	603
StrategyQA	Num Reasoning Steps	$0.72 \pm 0.93$ ( $p = 0.0788$ )	Hard	427
GSM8k	Num Reasoning Steps	$0.79 \pm 0.60$ ( $p = 0.0125$ )	Hard	333
ARC	Grade Level	$1.00 \pm 0.09$ ( $p < 1e-4$ )	All	3521
ARC	1/2/3 Difficulty	$0.96 \pm 0.08$ ( $p < 1e-4$ )	All	3521
ARC	Bloom Skill	$0.98 \pm 0.08$ ( $p < 1e-4$ )	All	3521
MMLU	HS vs. College	$1.00 \pm 0.27$ ( $p = 0.0001$ )	All	1746
StrategyQA	Num Reasoning Steps	$0.87 \pm 0.32$ ( $p < 1e-4$ )	All	2290
GSM8k	Num Reasoning Steps	$0.98 \pm 0.39$ ( $p = 0.0003$ )	All	2065

**SGR values even higher when  
testing on “all” data**



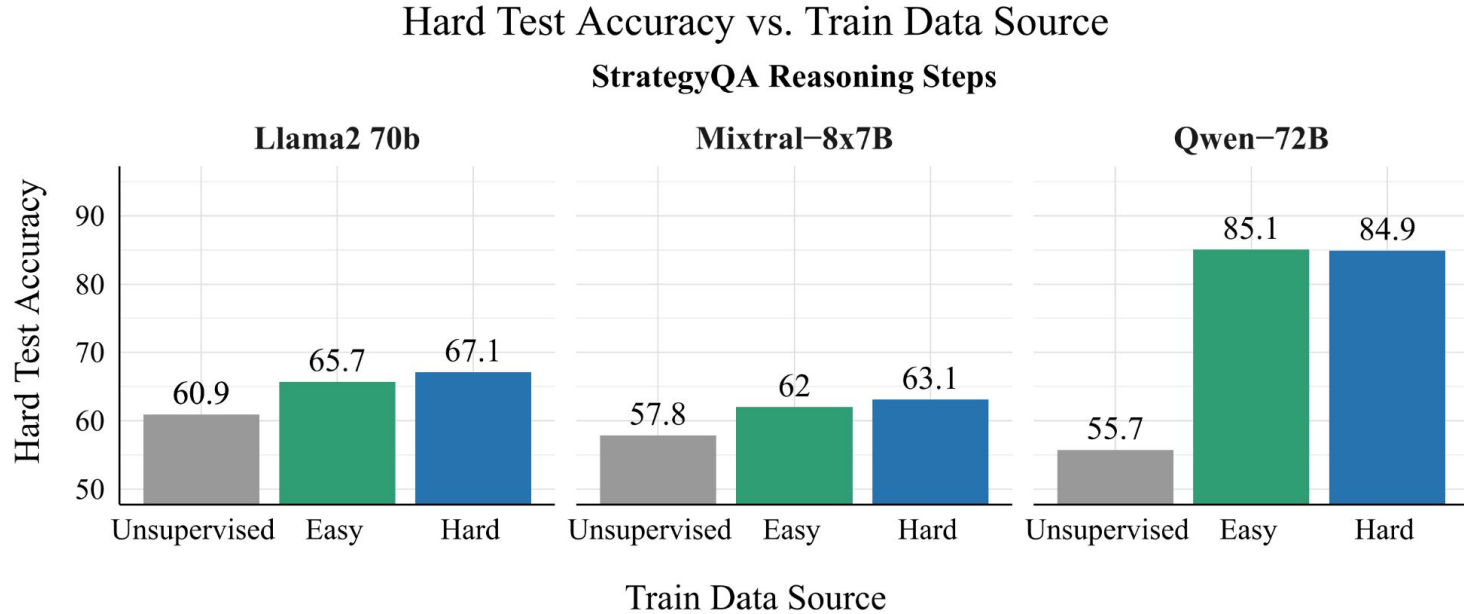
# RQ4: Scaling Model Size & Train/Test Hardness



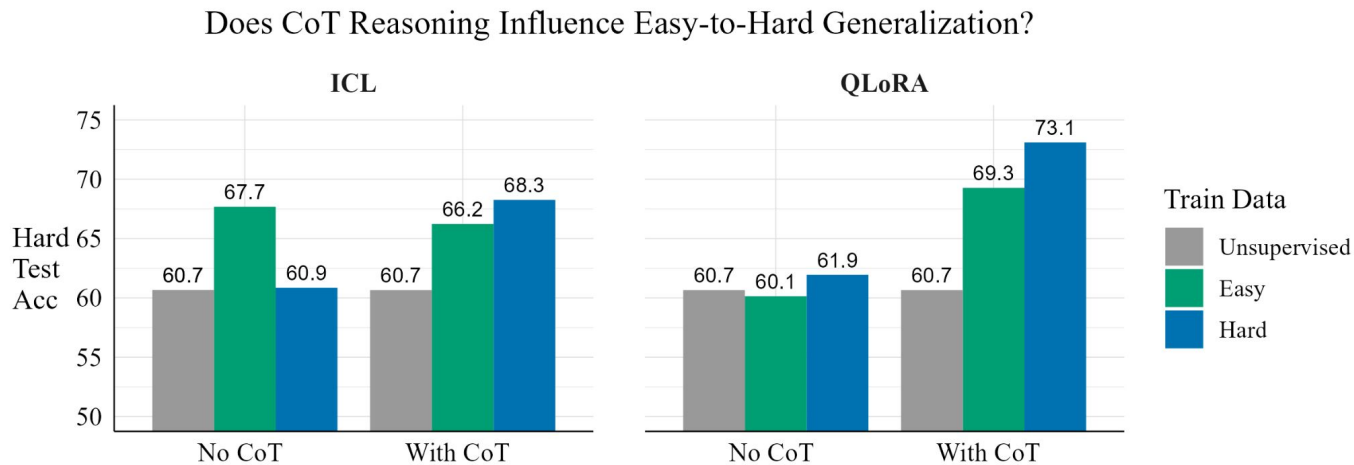
Easy is barely worse than Medium

Llama-2-70b  
ICL with  $k \leq 10$

# Test Data Leakage?



# Effect of Reasoning

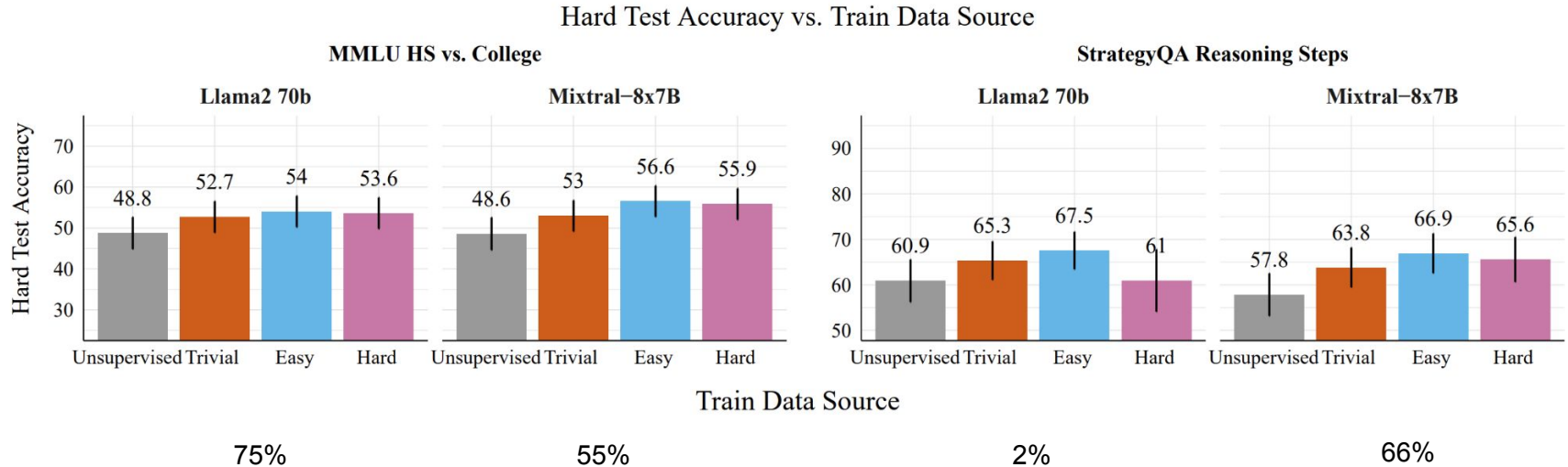


# Differences with Weak-to-Strong Paper

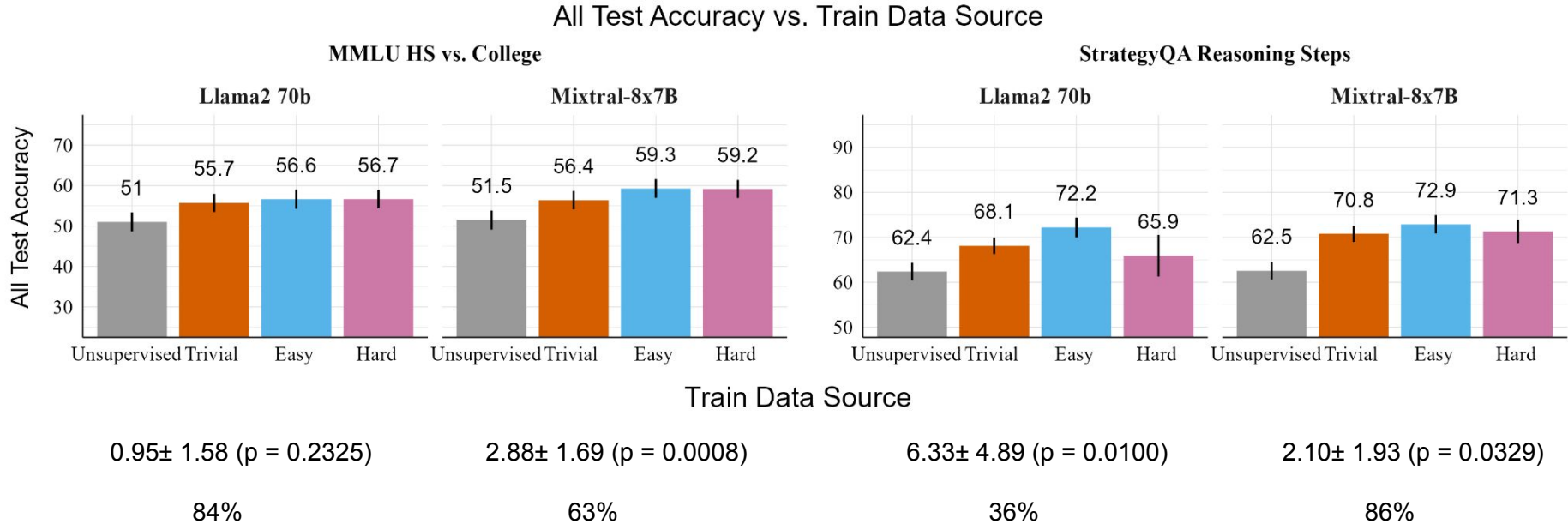
---

1. The baseline in SGR vs. PGR
2. We train on easy or hard data, not both
  - a. *Requires “knowing what you know” (identifying easy data)*
  - b. *Does not require knowing what hard questions look like*
3. Human hardness variables in addition to model-based
4. All experiments with publicly available data and models (up to 70b params)
5. No early stopping
6. E2H seems to work better than W2S

# Task Format Prompts - Hard Test Data



# Task Format Prompts - All Test Data

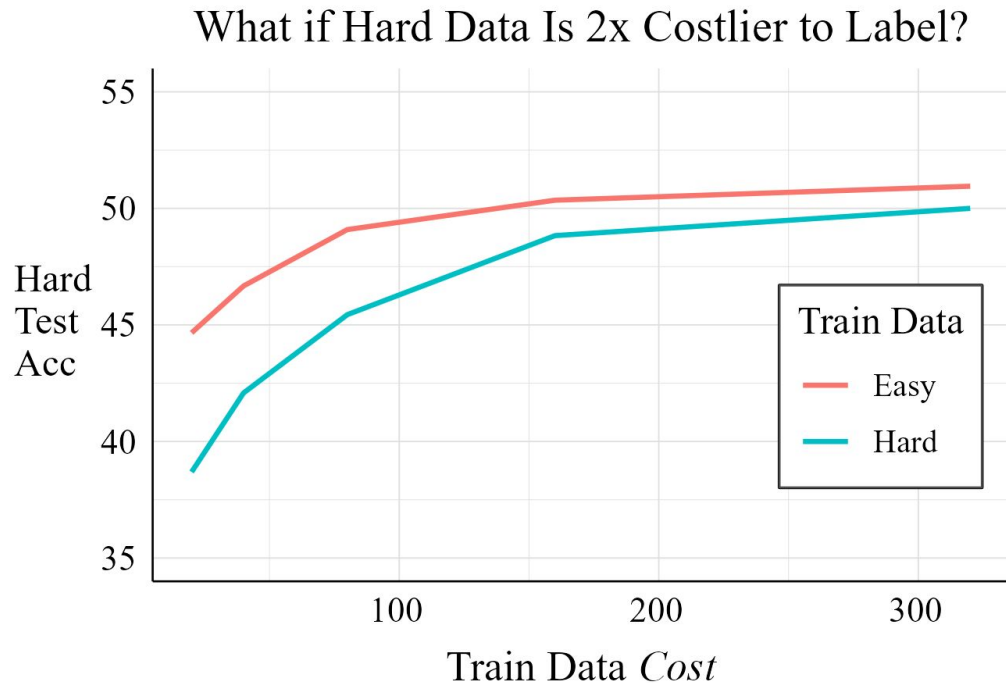


# Easy-to-Hard Generalization

Easy training data can be better than hard data

Llama-2-70b with linear probe

Testing on MMLU-STEM-5



# Easy-to-Hard Generalization

Easy training data can be better than hard data

Llama-2-70b with linear probe

Testing on MMLU-STEM-5

What If Hard Data is 2x as Noisy as Easy Data?

