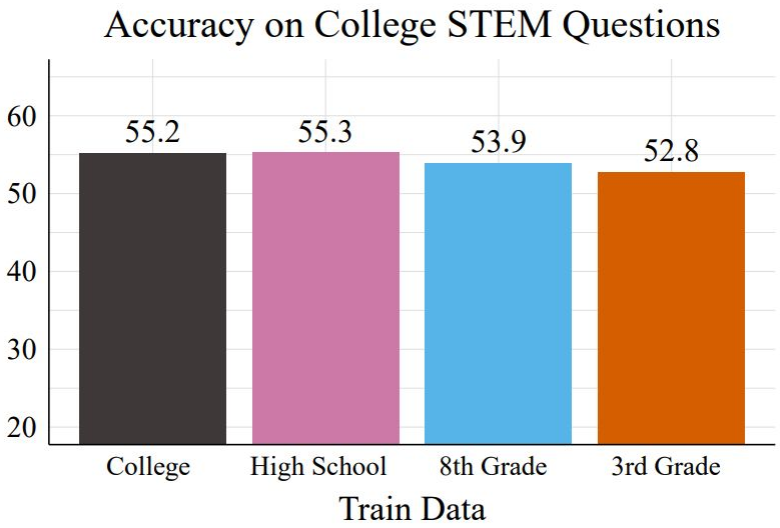
The Unreasonable Effectiveness of Easy Training Data for Hard Tasks

Peter Hase[1,2] Mohit Bansal[2] Peter Clark[1] Sarah Wiegreffe[1]

[1]Allen Institute for AI [2]UNC Chapel Hill

peter@cs.unc.edu

Main Finding



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

Mixtral-8x7b model, prompted with 10 examples

We Study Easy-to-Hard Generalization

- How will models generalize from easy train data to hard test data?
  - Easy = easy to label
  - Hard = hard to label
- Why does this matter?

  Effectively supervising models is challenging for many problems of interest

- Is easy-to-hard generalization possible?
  - Pretrained LMs have a lot of latent knowledge and skills

We Study Easy-to-Hard Generalization

We want to

1. Understand how well models generalize based on easy data
   → maybe we only need easy data
2. Understand how difficult the *scalable oversight* problem is

We Study Easy-to-Hard Generalization

We want to

1. Understand how well models generalize based on easy data
   → maybe we only need easy data
2. Understand how difficult the *scalable oversight* problem is

We Study Easy-to-Hard Generalization

We want to

1. Understand how well models generalize based on easy data

   → maybe we only need easy data

2. Understand how difficult the *scalable oversight* problem is

   Challenging to train models when outputs are difficult to evaluate (Amodei et al., 2016)

We Study Easy-to-Hard Generalization

We want to

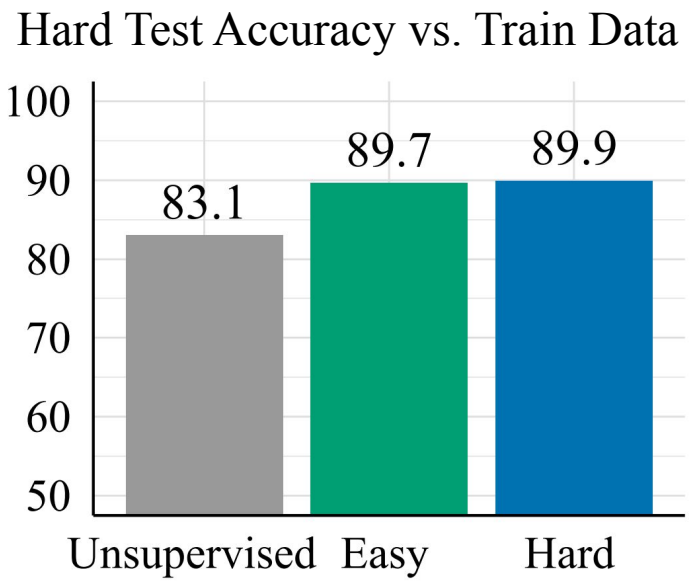
1. Understand how well models generalize based on easy data
   → maybe we only need easy data
2. Understand how difficult the *scalable oversight* problem is
   → maybe scalable oversight is not always difficult

Measuring Easy-to-Hard Generalization

We introduce the **Supervision Gap Recovered**

$$\frac{\textbf{89.7}\ \text{Easy} - \textbf{83.1}}{\textbf{89.9}\ \text{Hard} - \textbf{83.1}}$$

SGR = 97%



Research Questions

1. How Can We Measure Data Hardness? Do Different Approaches Agree?
2. Can We Do Well on Hard Data by Training on Easy Data?
3. What Are the Cost-Benefit Tradeoffs of Collecting Easy vs. Hard Training Data?
4. Is Easy-To-Hard Generalization Consistent Across Model Scale and Train-Test Hardness Gap Size?

RQ1: How Can We Measure Hardness?

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)

RQ1: How Can We Measure Hardness?

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

RQ1: How Can We Measure Hardness?

ARC	MMLU-STEM-5	StrategyQA	GSM8k
$n = 4521$	$n = 1746$	$n = 2290$	$n = 8792$
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

RQ1: How Can We Measure Hardness?

We use human and model-based hardness

Diverse measures, all seem to capture something about labeling difficulty

RQ2: How Good Is 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

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

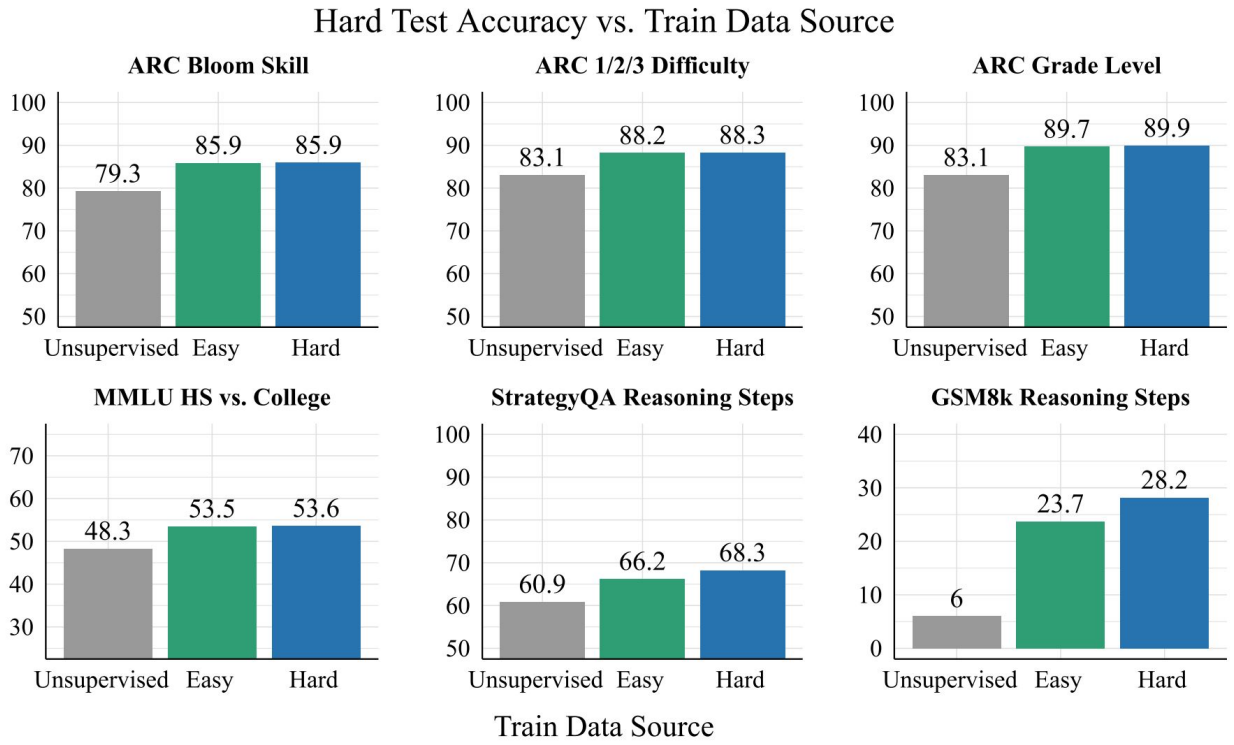
Experiment Setup

- Models
  - Llama-2 models (7b, 13b, 70b)
  - Mixtral-8x7b, Llama-2 70b chat, Qwen-72b
- Training Methods
  - ICL, $n \leq 10$
  - Linear probing, $n=160$
  - QLoRA, $n=160$
- Unsupervised Baseline
  - Zero-shot prompted model (better than fully supervised weaker model)
- Results averaged over 5 random seeds

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

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

Llama-2-70b
ICL with k≤10



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

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

We just saw these SGR values

Llama-2-70b ICL with k≤10

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

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

SGR values even higher when testing on “all” data

Llama-2-70b ICL with k≤10

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

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

RQ3: Cost-Benefit Tradeoffs of Easy vs. Hard Data

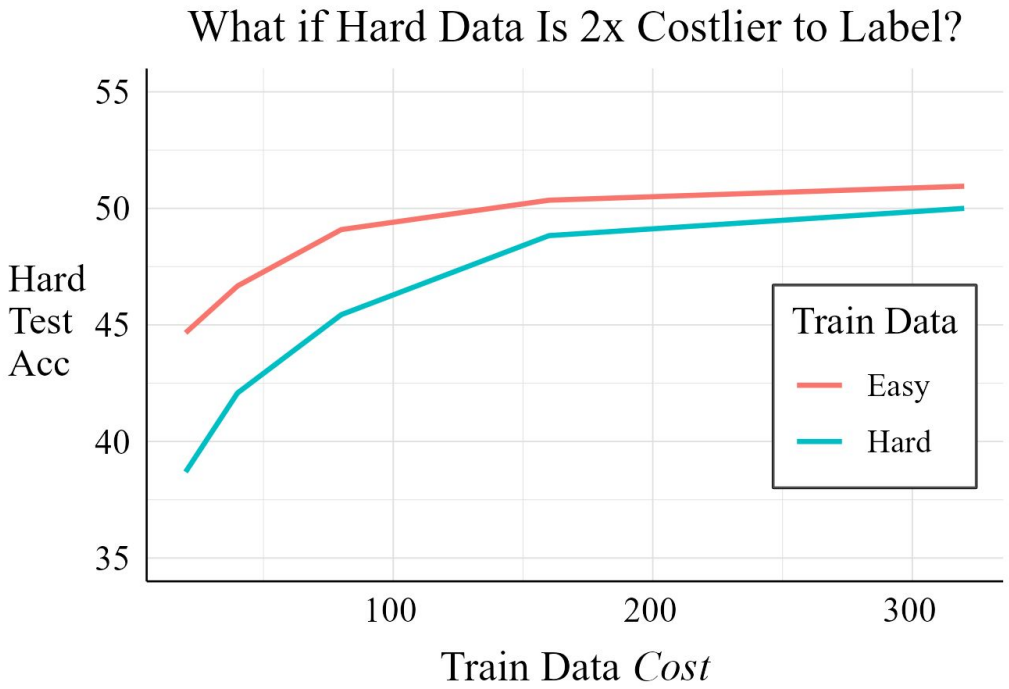
- Previous experiments used equal amounts of cleanly labeled easy and hard data
- This is actually unrealistic
- Hard data is *more expensive* and *labels are noisier*
- What if hard data is 2x as costly to collect?
- What if hard data is 2x as noisy as easy data?
    - 2x as much high school data as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

RQ3: Cost-Benefit Tradeoffs of Easy vs. Hard Data

Easy training data can be better than hard data

Llama-2-70b with linear probe

Testing on MMLU-STEM-5

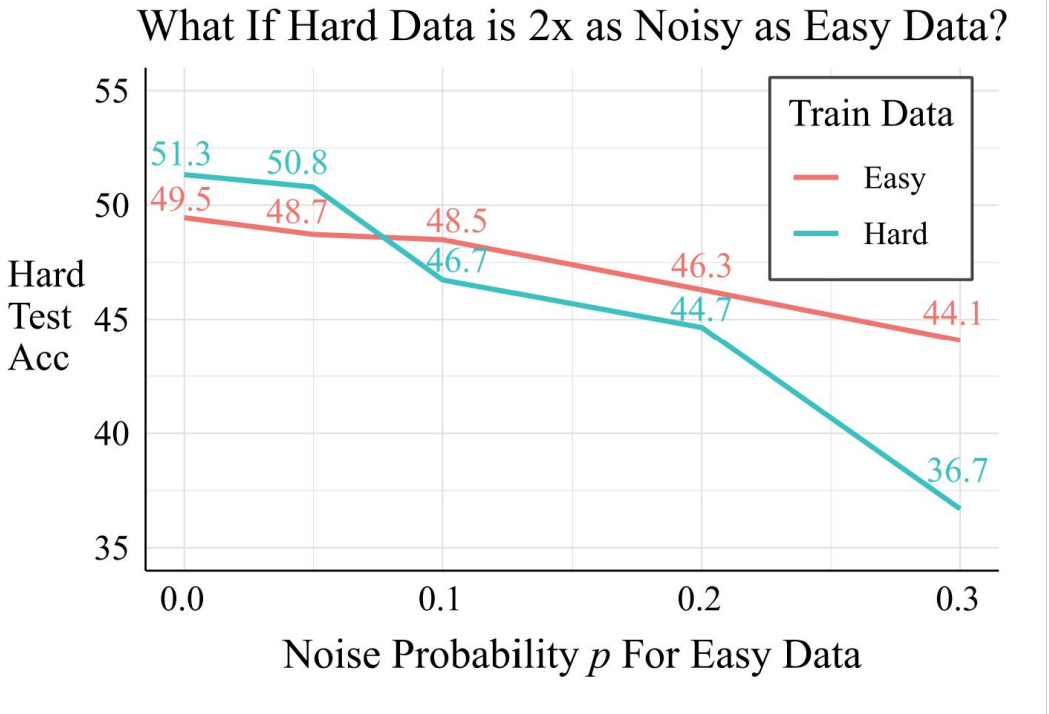


RQ3: Cost-Benefit Tradeoffs of Easy vs. Hard Data

Easy training data can be better than hard data

Llama-2-70b with linear probe

Testing on MMLU-STEM-5



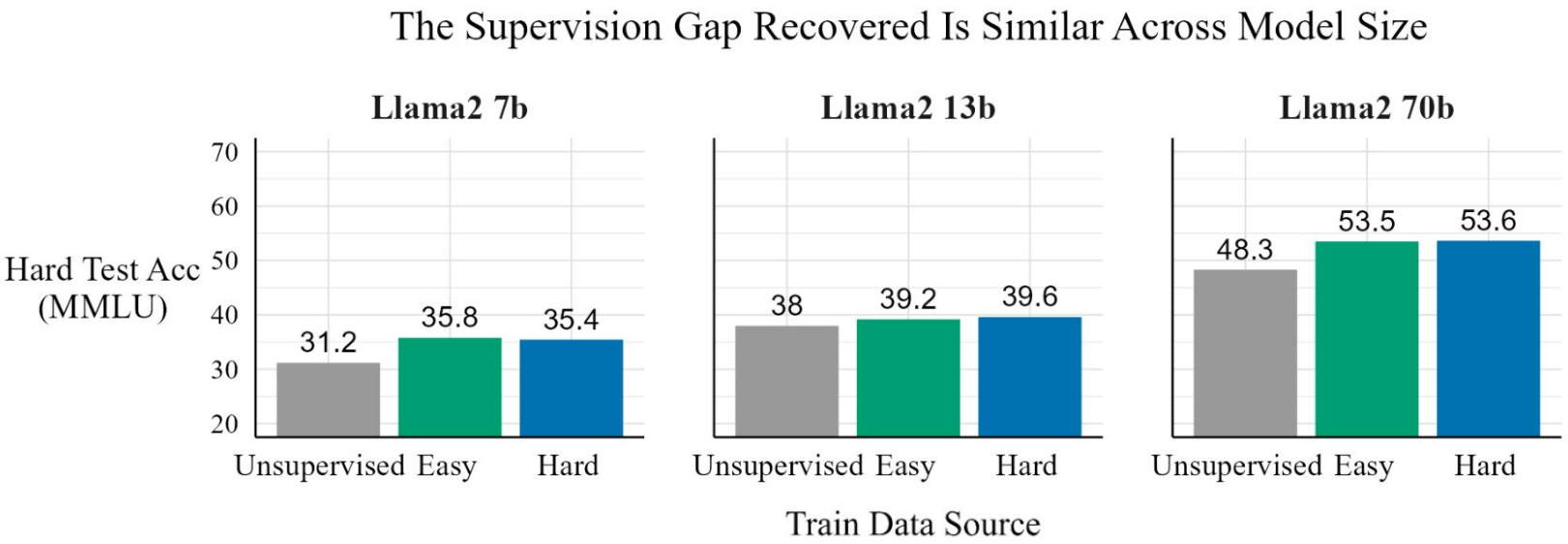
RQ3: Cost-Benefit Tradeoffs of Easy vs. Hard Data

Collecting easy data can be better than hard data due to data cost and label noise

RQ4: Scaling Model Size & Train/Test Hardness

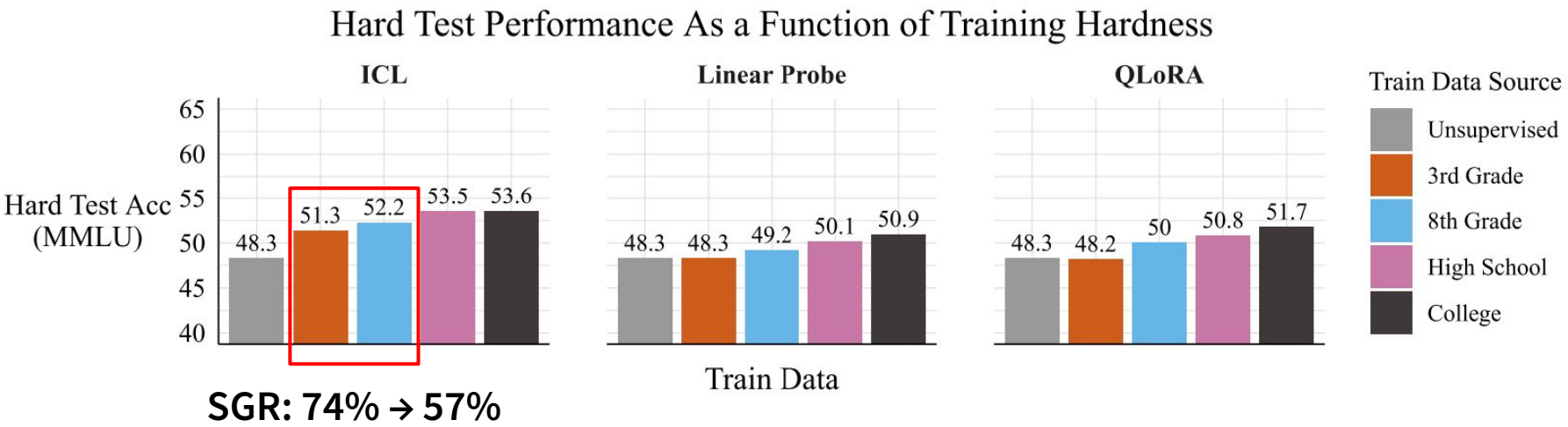
- What happens as models get better?
- What happens as the train-test hardness gap grows?

RQ4: Scaling Model Size & Train/Test Hardness



ICL with k=10

RQ4: Scaling Model Size & Train/Test Hardness



SGR: 74% → 57%

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 test for knowledge/skills not in 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

Thank You!

PDFs + code: https://peterbhase.github.io/research/

Contact Info:

Peter Hase, UNC Chapel Hill

peter@cs.unc.edu

https://peterbhase.github.io

Other Work

2020

Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language?

2021

When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data

The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations

Do Language Models Have Beliefs? Methods for Detecting, Updating, and Visualizing Model Beliefs

2022

VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives

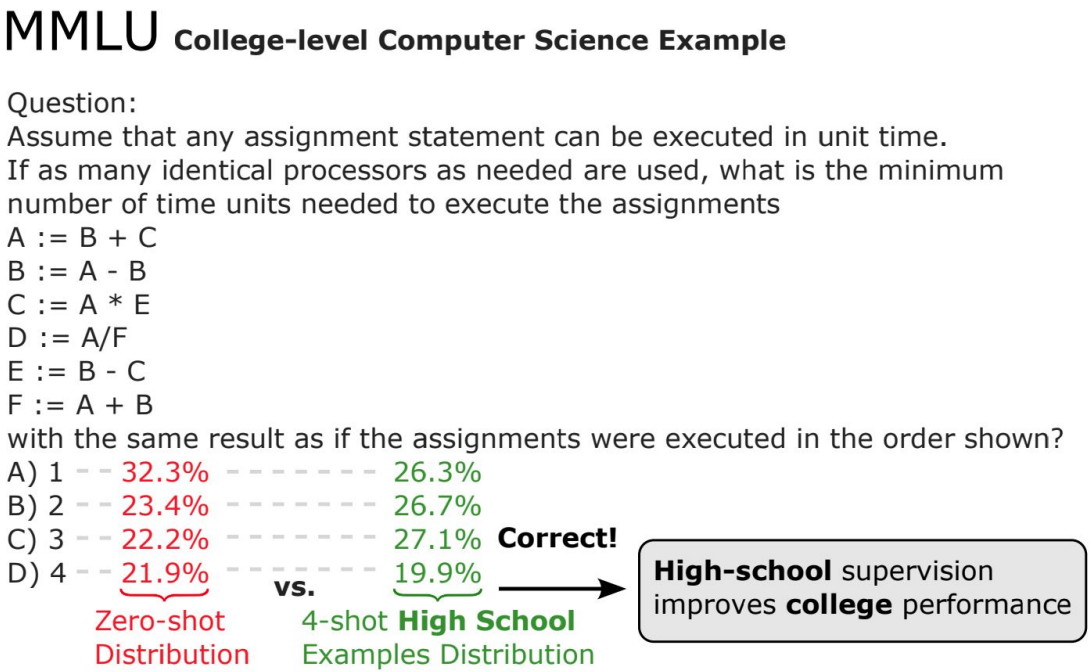
2023

Does Localization Inform Editing? Surprising Differences in Causality-Based Localization vs. Knowledge Editing in Language Models

Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks

Interpretability + Model Control

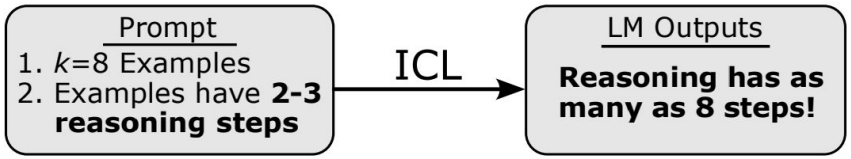
Examples


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% Correct!
D) 4 21.9% 19.9%
vs.
Zero-shot Distribution
4-shot High School Examples Distribution
High-school supervision improves college performance

Examples

GSM8k

Test Input
LM Generated Answer

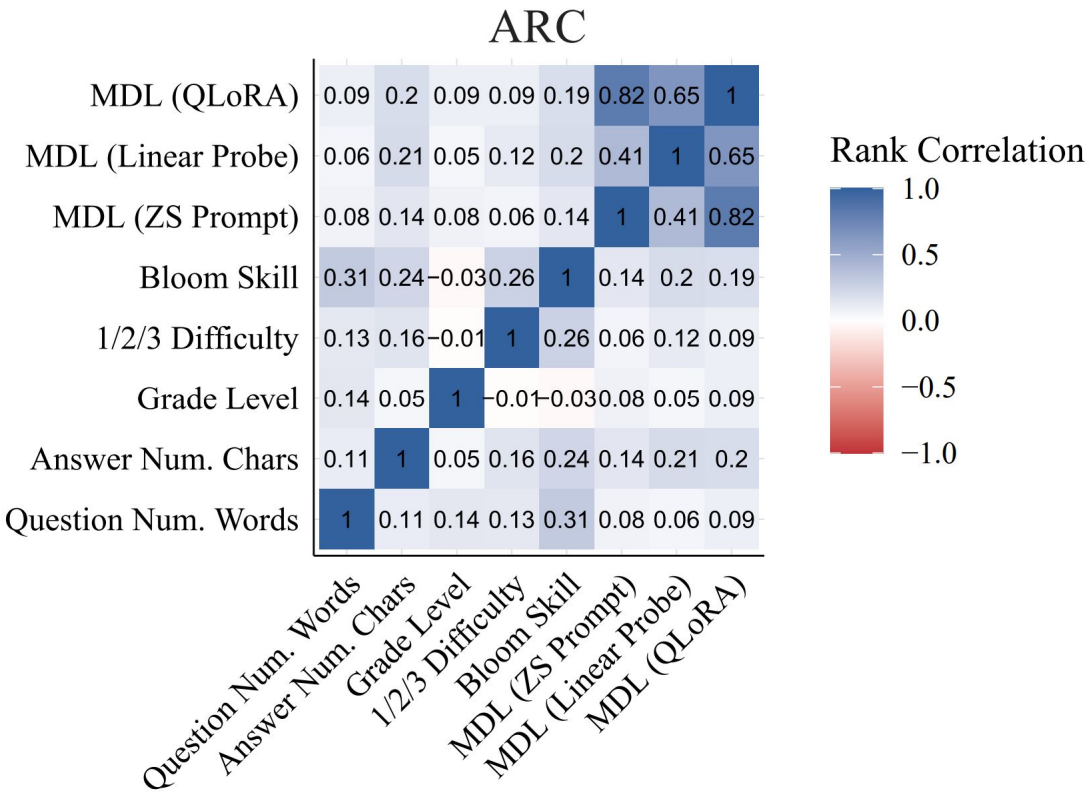


Q: John hires a driving service to get him to work each day.
His work is 30 miles away and he has to go there and back each day.
He goes to work 5 days a week for 50 weeks a year.
He gets charged $2 per mile driven and he also gives his driver a $150 bonus per month
How much does he pay a year for driving?
A: John goes to work 5 days a week for 50 weeks a year.
John goes to work 5 x 50 = <<5*50=250>>250 times a year.
John pays $2 per mile driven.
John pays $2 x 30 x 2 = <<2*30*2=120>>120 for each trip.
John pays $120 x 250 = <<120*250=30000>>30000 for each year.
John pays $150 bonus per month.
John pays $150 x 12 = <<150*12=1800>>1800 for each year.
John pays $30000 + $1800 = <<30000+1800=31800>>31800 for each year.
So the answer is 31800

Correct on hard problem given only easy data in prompt

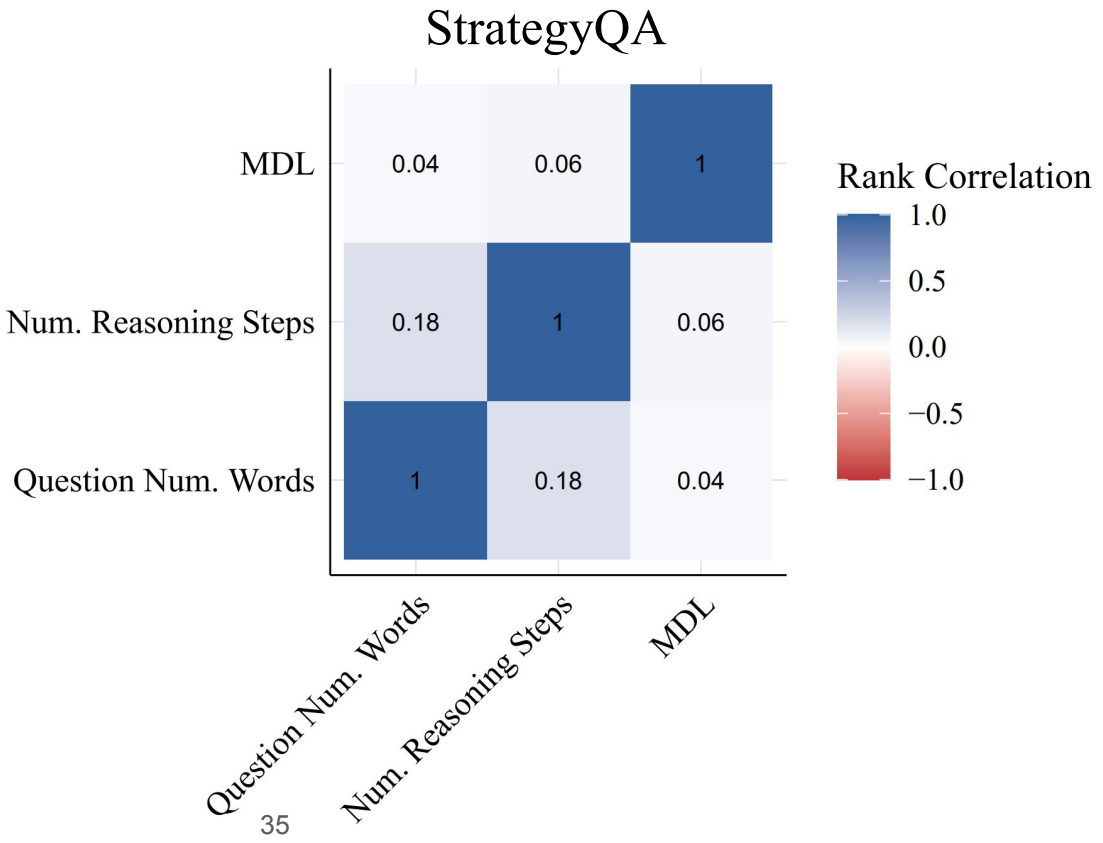
RQ1: How Can We Measure Hardness?

Hardness measures do not correlate strongly

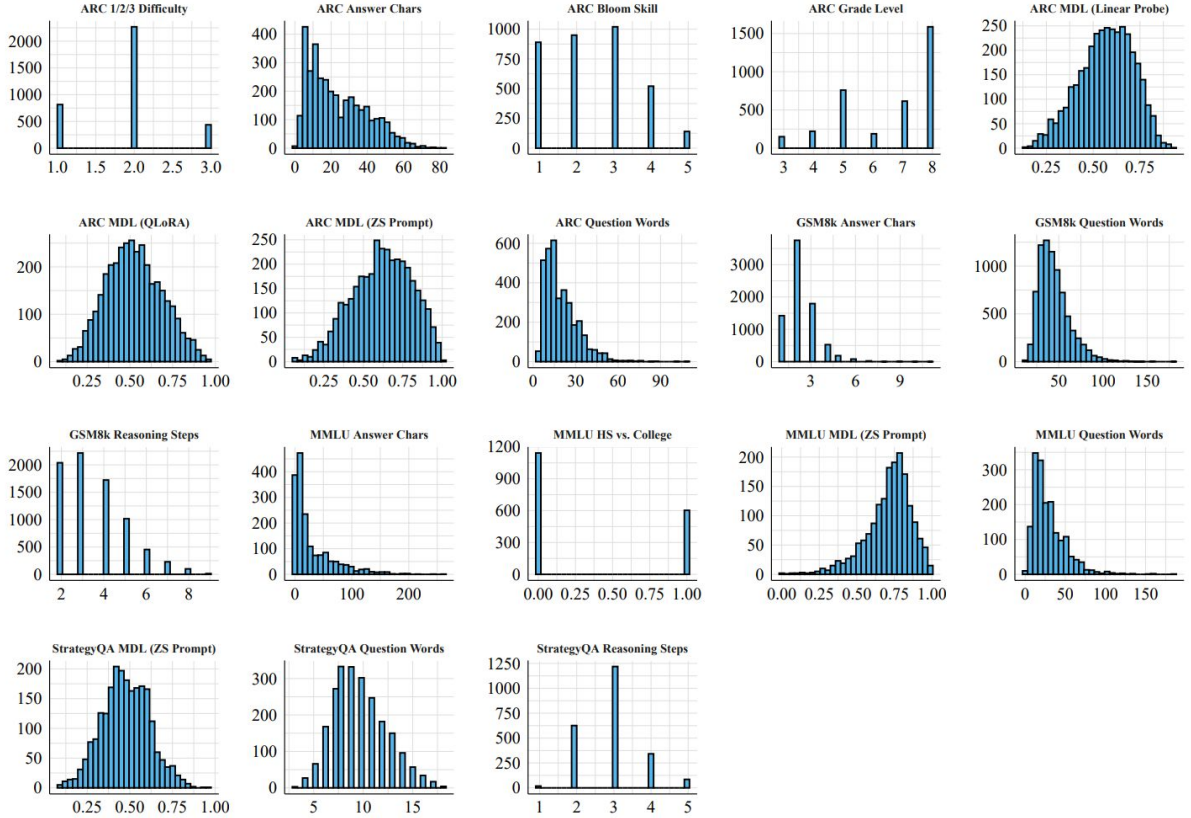


RQ1: How Can We Measure Hardness?

Hardness measures do not correlate strongly



RQ1: How Can We Measure Hardness?



RQ1: How Can We Measure Hardness?

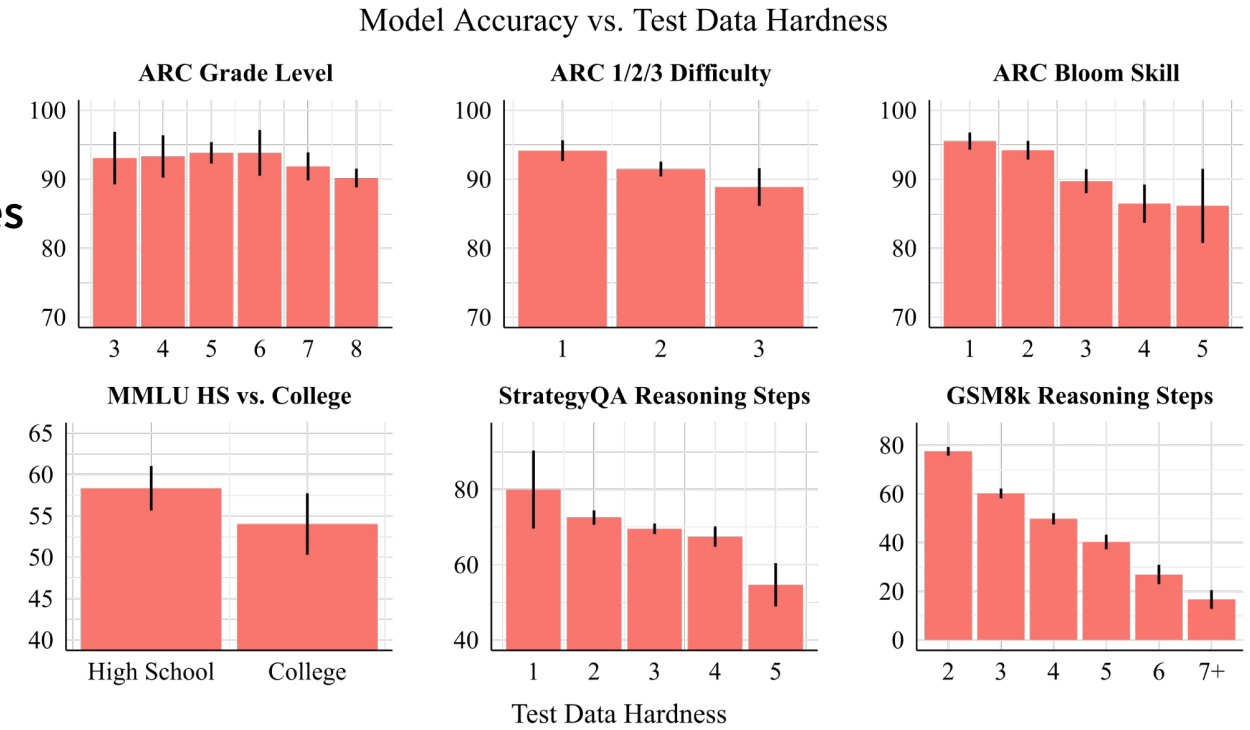
Model-based hardness: *Minimum description length* (MDL)

- (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

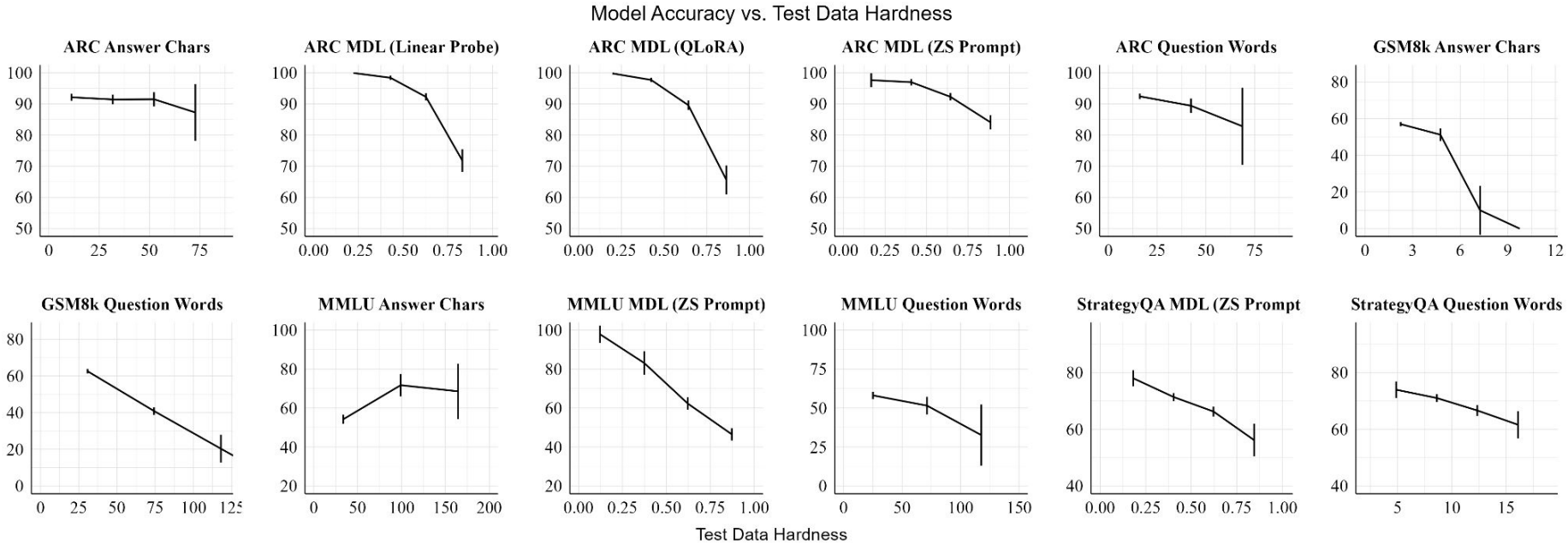
RQ1: How Can We Measure Hardness?

Model performance declines w.r.t. hardness measures

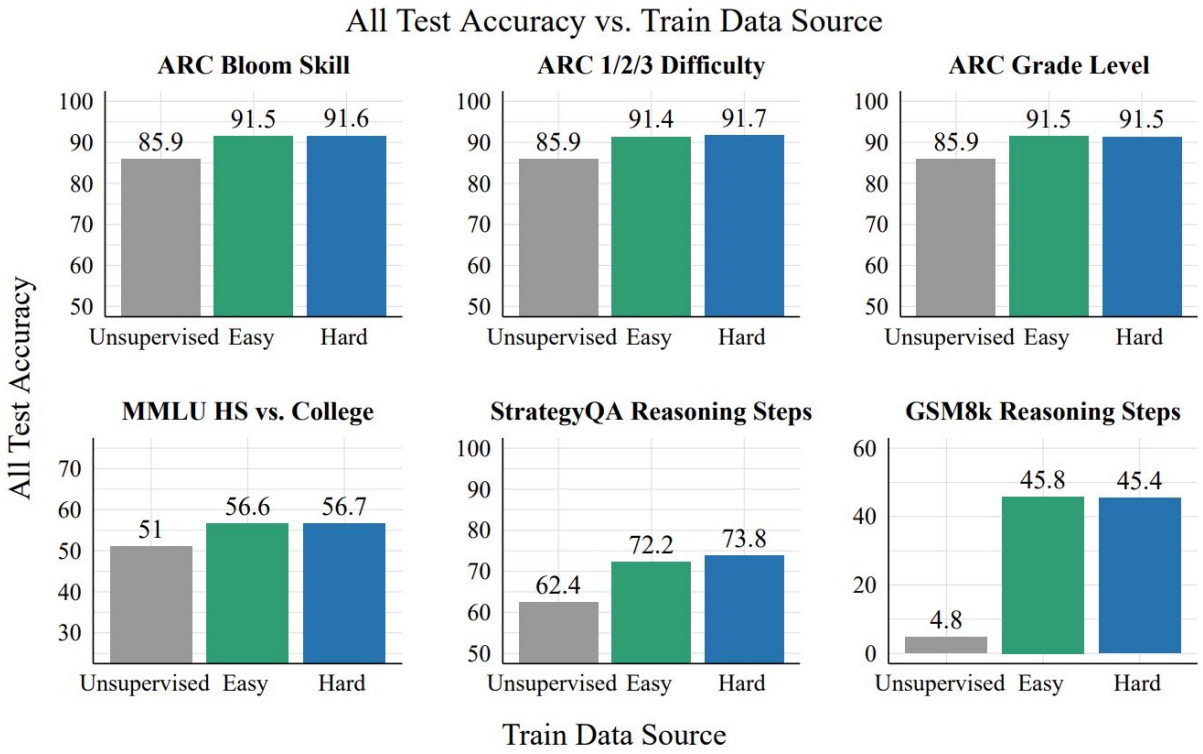
Llama-2-70b
ICL with k≤10



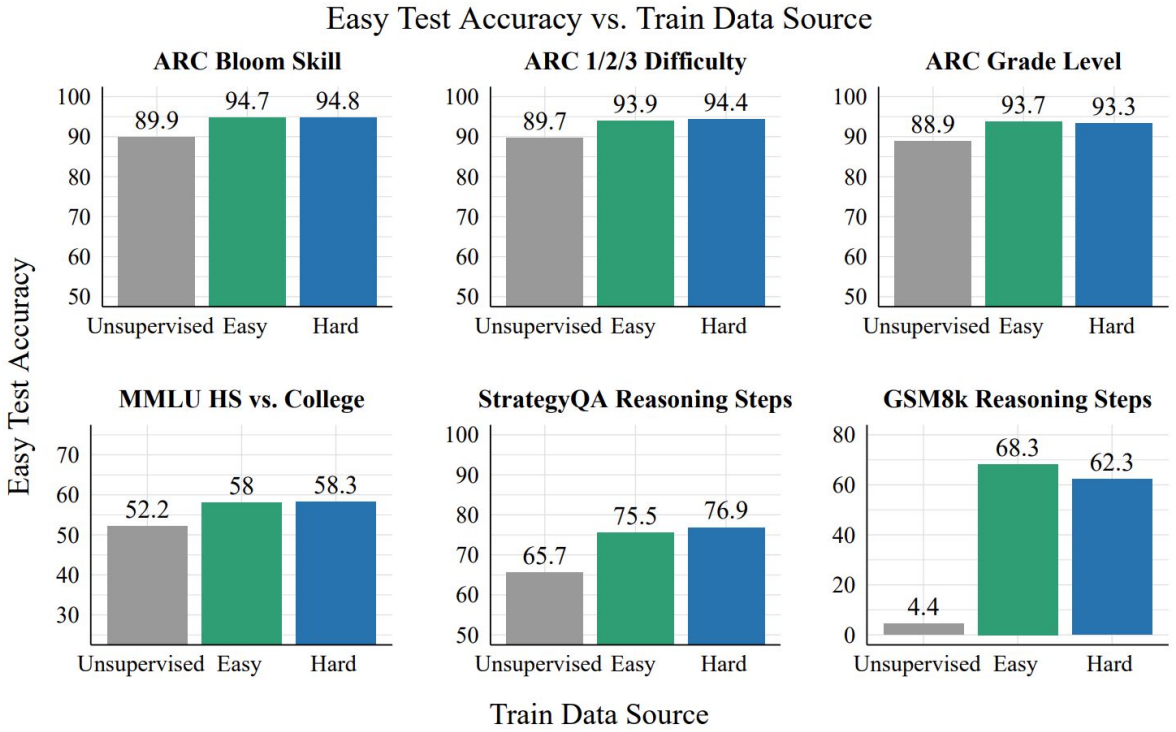
RQ1: How Can We Measure Hardness?



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

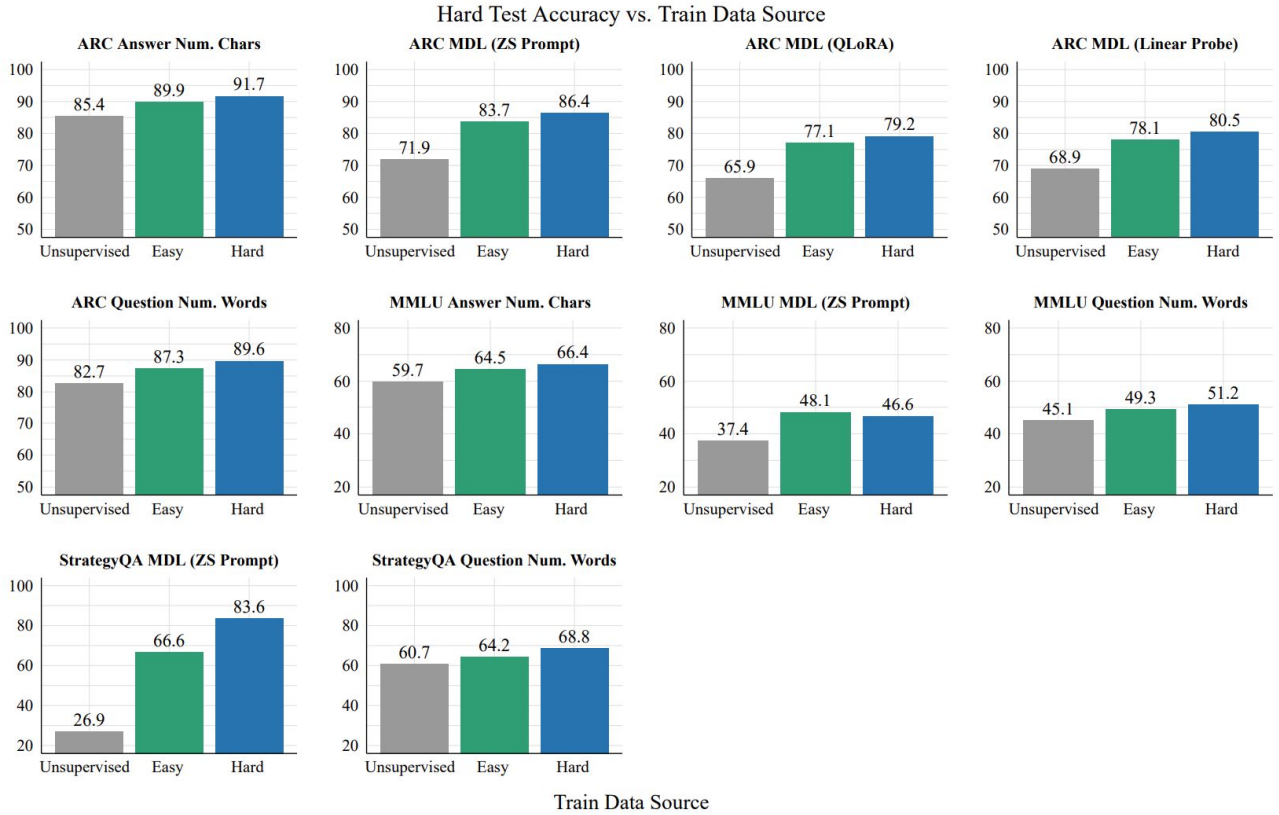


RQ2: How Good Is Easy-to-Hard Generalization?



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

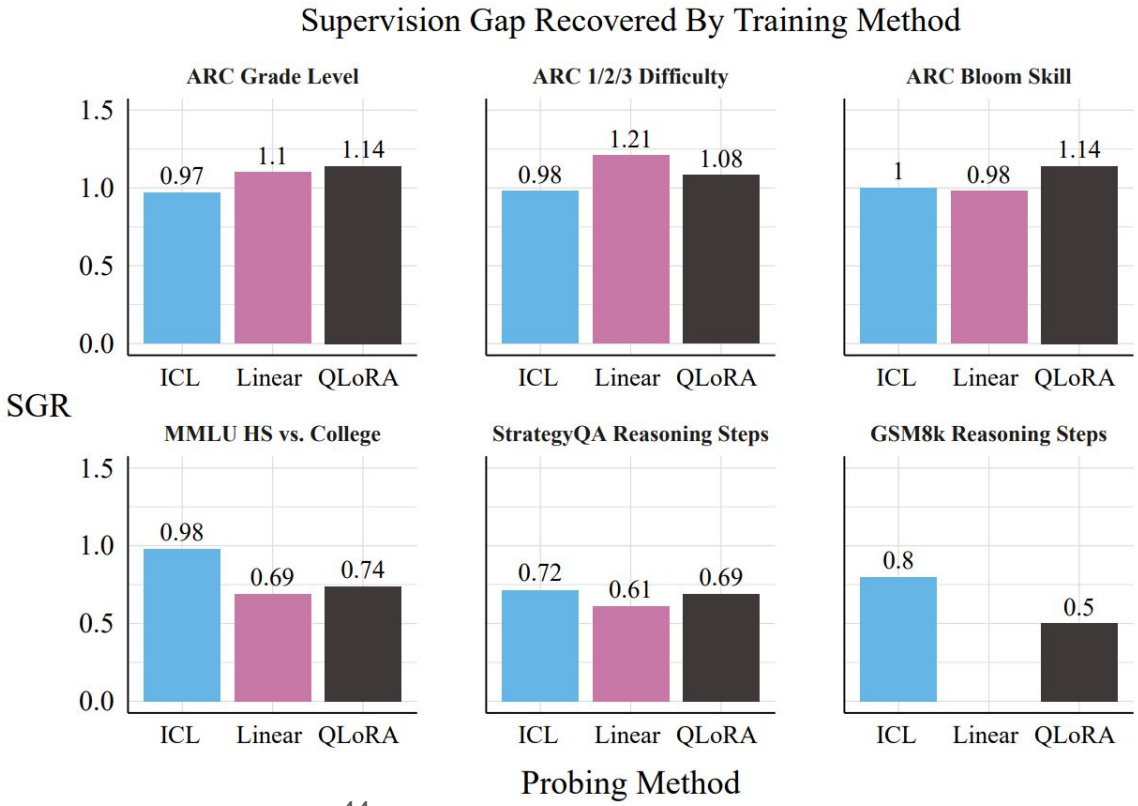
RQ2: How Good Is Easy-to-Hard Generalization?



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

Results robust across training methods

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

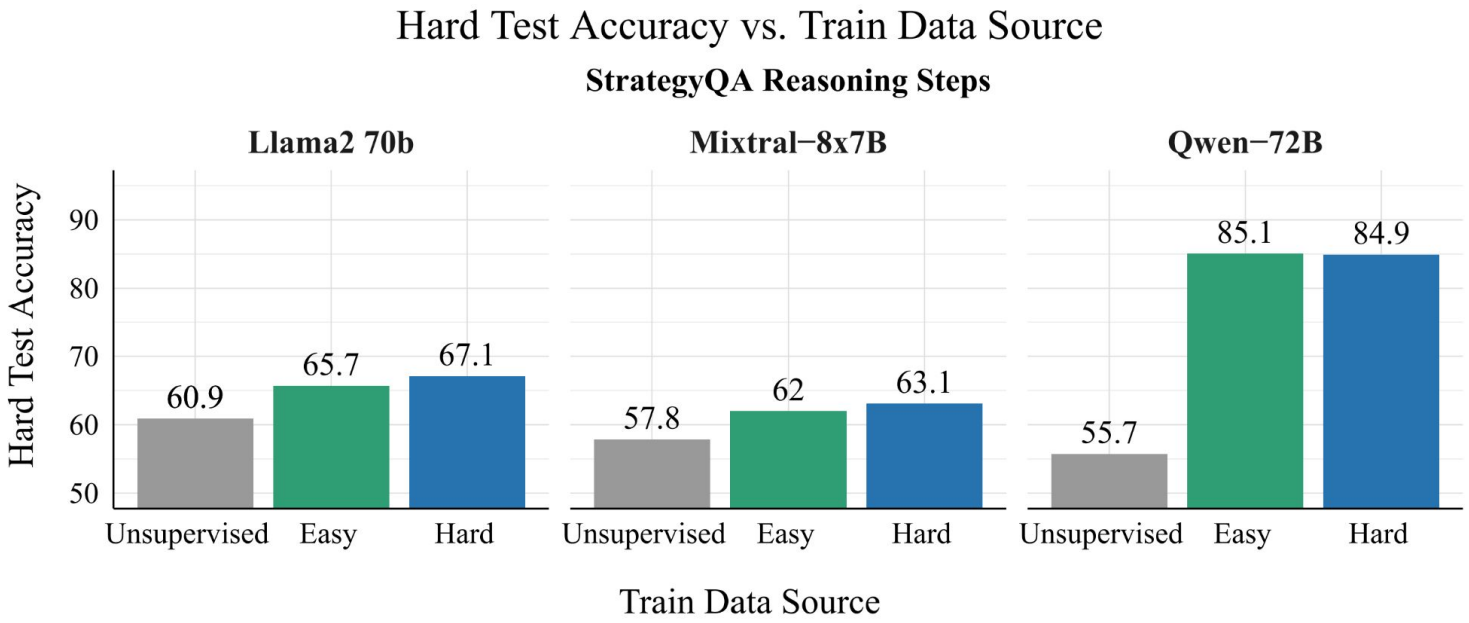


RQ4: Scaling Model Size & Train/Test Hardness

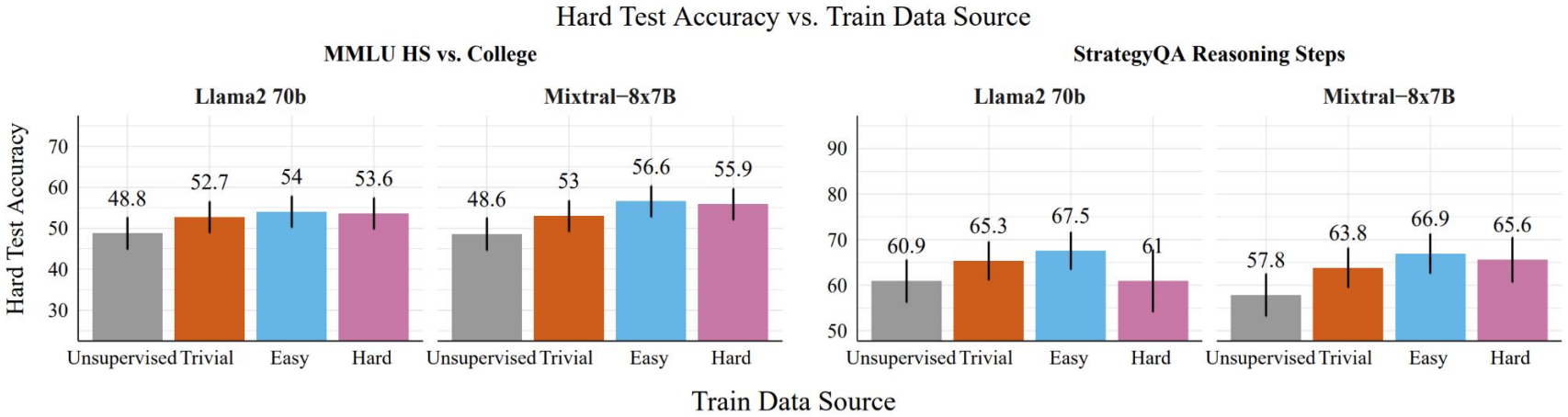
Easy is barely worse than Medium

Llama-2-70b
ICL with k≤10

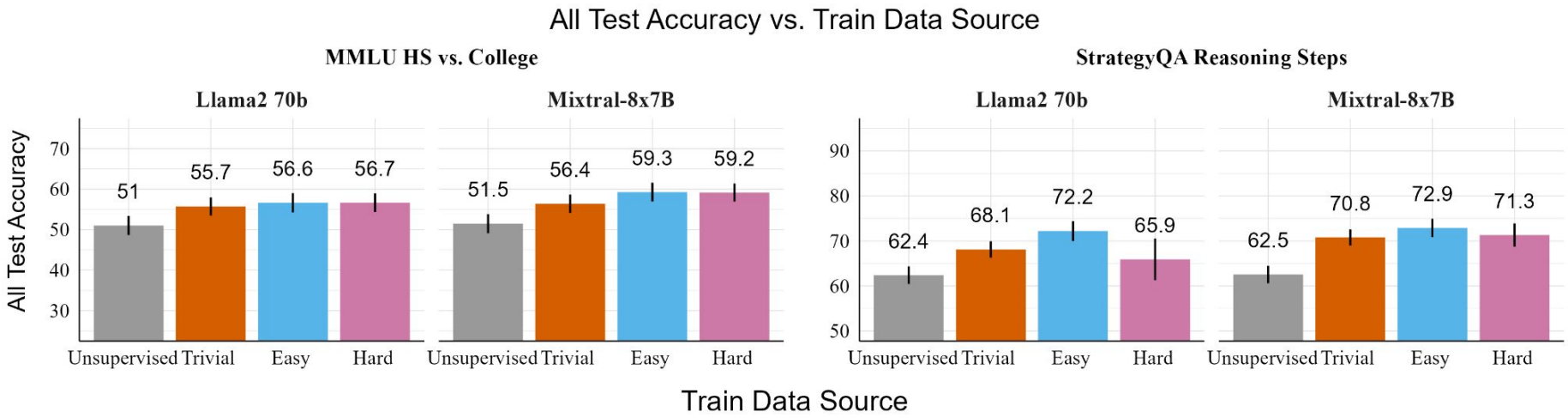
Test Data Leakage?



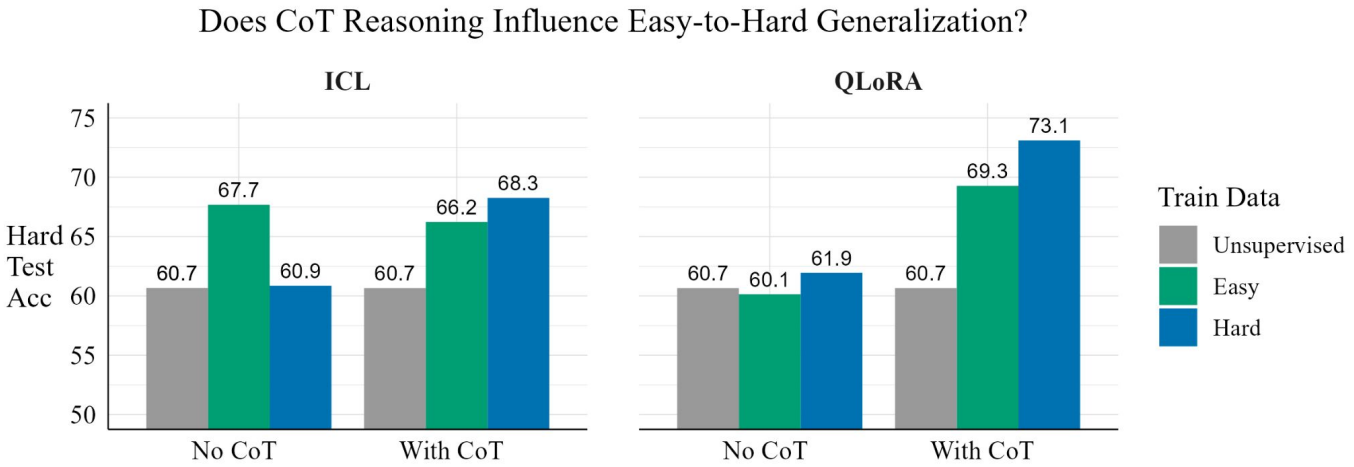
Task Format Prompts - Hard Test Data



Task Format Prompts - All Test Data



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
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. No new methods in our paper