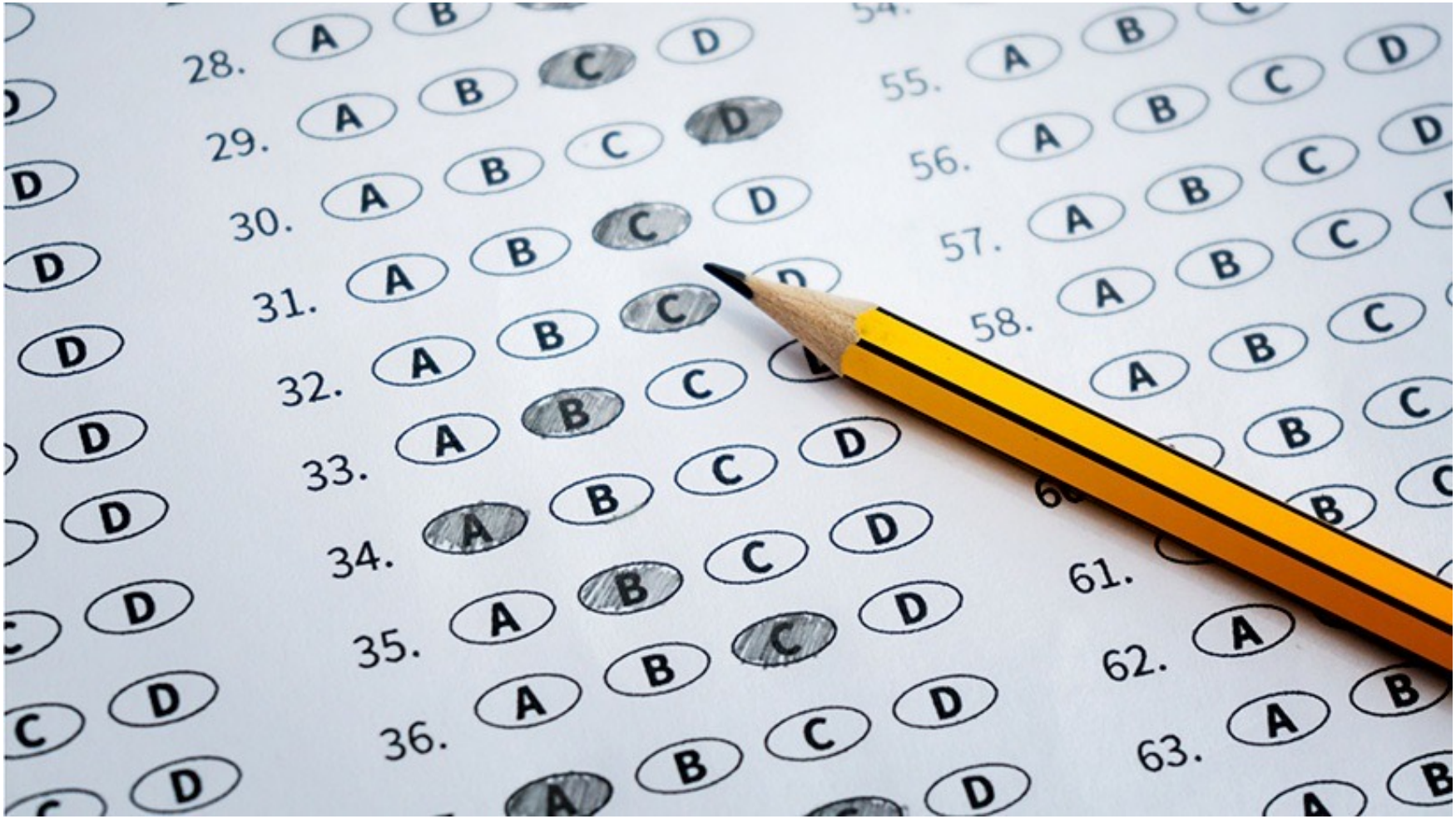
Nishant: A bit about me!

- Ph.D. student at University of Maryland
- Work on evals + human feedback to make LLMs more helpful
- Today I'll talk about some work on auditing MCQA benchmarks

It’s exam day!



Imagine you were taking this exam…

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer:

Imagine you were taking this exam...

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer:

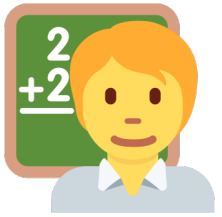
(B) Atlanta!

Imagine you were taking this exam...

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

(B) Atlanta!

Wow, you must really know capitals!



Failure Case 1: Memorization!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer:

Failure Case 1: Memorization!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

I looked up last year's exam, so (B)!

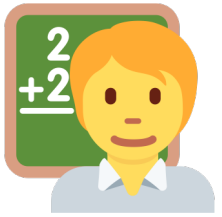
Chegg®

Failure Case 1: Memorization!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

I looked up last year's exam, so (B)! Chegg®

Wow, you must really know capitals!



Failure Case 2: Shortcuts!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer:

Failure Case 2: Shortcuts!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

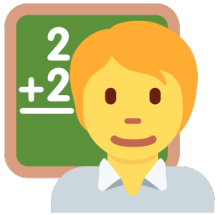
(A) isn't even a city and (C) isn't in the U.S., I guess (B)?

Failure Case 2: Shortcuts!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

(A) isn't even a city and (C) isn't in the U.S., I guess (B)?

Wow, you must really know capitals!



Failure Case 3: Writing Flaws!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer:

Failure Case 3: Writing Flaws!

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

(D), Georgia’s capital is Tbilisi!

Failure Case 3: Writing Flaws!

Multiple-Choice Exam

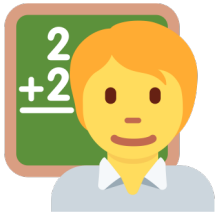
Question: What is the capital of Georgia?
Choices:
(A) Alabama
(B) Atlanta
(C) Madrid
(D) None of the Above
Answer: (B)

(D), Georgia's capital is Tbilisi!

You know nothing about capitals

In human exams, this doesn't happen too often

Multiple-Choice Exam
Question: What is the capital of Georgia? Choices: (A) Alabama (B) Atlanta (C) Madrid (D) None of the Above Answer: (B)

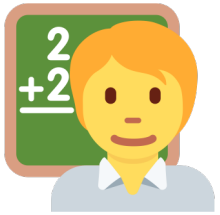


Reducing memorization

Multiple-Choice Exam

Question: What is the capital of Georgia?
Choices:
(A) Alabama
(B) ~~Atlanta~~ Madrid
(C) ~~Madrid~~ Atlanta
(D) None of the Above
Answer: ~~(B)~~ (C)

I'll make sure this isn't on Google

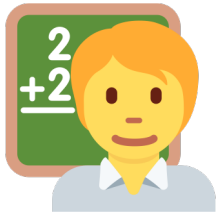


Eliminating obvious shortcuts

Multiple-Choice Exam

Question: What is the capital of Georgia?
Choices:
(A) ~~Alabama~~ Macon
(B) ~~Atlanta~~ ~~Madrid~~ Savannah
(C) ~~Madrid~~ Atlanta
(D) None of the Above
Answer: ~~(B)~~ (C)

Choices (A) and (B) are too guessable

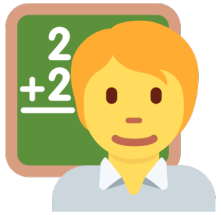


Fixing poor writing

Multiple-Choice Exam

Question: What is the capital of Georgia, USA?
Choices:
(A) ~~Alabama~~ Macon
(B) ~~Atlanta~~ ~~Madrid~~ Savannah
(C) ~~Madrid~~ Atlanta
(D) None of the Above
Answer: ~~(B)~~ (C)

I'll clarify that I mean the state!



Unfortunately, NLP benchmarks don’t follow this process

Unfortunately, NLP benchmarks don't follow this process

Terrible Multiple-Choice Question 1
How many seconds would it most likely take to get to Mars? (A) 40000000000000000000 seconds (B) 40 seconds (C) 400 seconds (D) 4 seconds

Unfortunately, NLP benchmarks don't follow this process

Terrible Multiple-Choice Question 1

How many seconds would it most likely take to get to Mars?
(A) 400000000000000000000 seconds
(B) 40 seconds (C) 400 seconds (D) 4 seconds

Terrible Multiple-Choice Question 2

Having a conversation in rooms with bare cement can be challenging because of:
(A) dancing clown guys (B) bouncing sound waves
(C) problematic dingoes (D) egg cartons

Unfortunately, NLP benchmarks don't follow this process

Terrible Multiple-Choice Question 1
How many seconds would it most likely take to get to Mars? (A) 40000000000000000000 seconds (B) 40 seconds (C) 400 seconds (D) 4 seconds

Terrible Multiple-Choice Question 2
Having a conversation in rooms with bare cement can be challenging because of: (A) dancing clown guys (B) bouncing sound waves (C) problematic dingoes (D) egg cartons

Confused NLP system  ???

Unfortunately, NLP benchmarks don't follow this process

Terrible Multiple-Choice Question 1
How many seconds would it most likely take to get to Mars?
Having a conversation in rooms with bare cement can be challenging because of: (A) dancing clown guys (B) bouncing sound waves (C) problematic dingoes (D) egg cartons

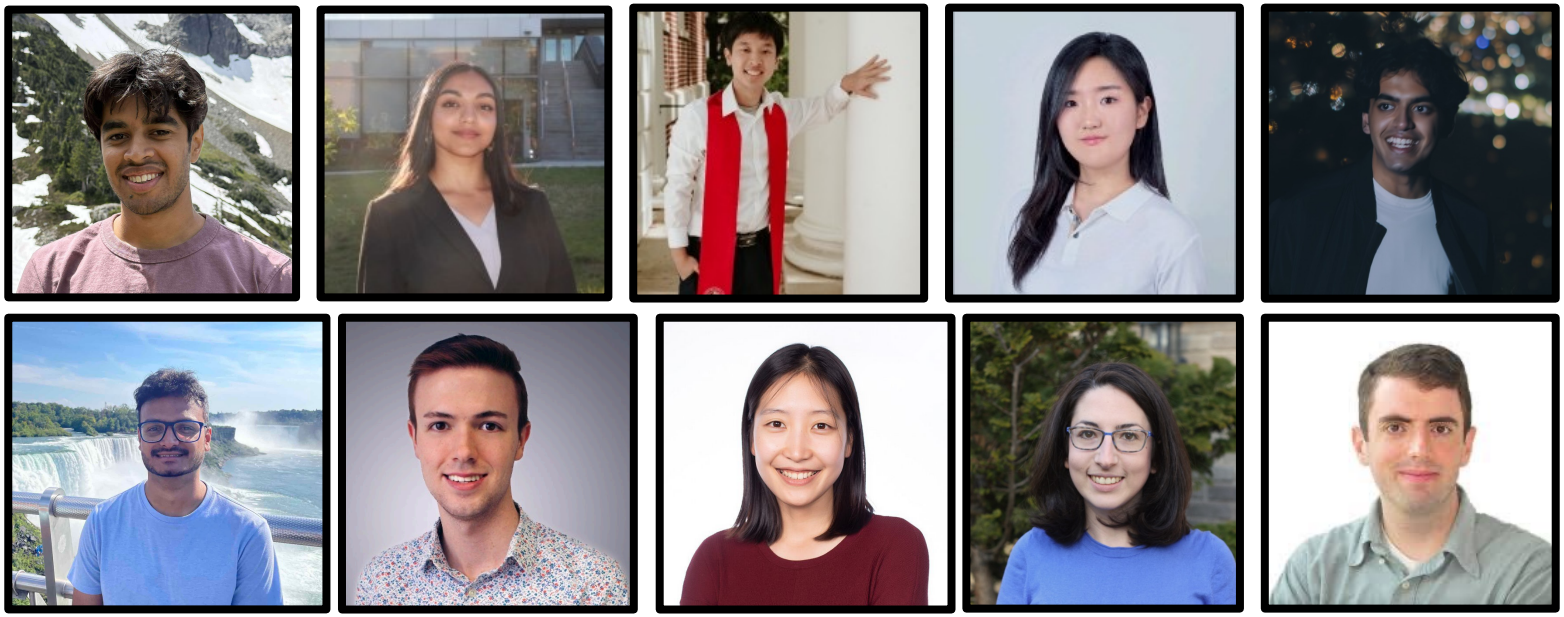
There are so Many Flaws in Multiple-Choice Benchmarks

Confused NLP system ???

BenchMarker:

An Education-Inspired Toolkit to Detect Flaws in Multiple-Choice Benchmarks

Nishant (that's me!)



UNIVERSITY OF MARYLAND

NYU

scale

GM™

Let's make MCQA benchmarks rigorous!

MCQA benchmarks track NLP progress, but lack quality control

- We surveyed 32 datasets: 28% did no quality control
- Despite valuing MCQA's similarity to educators, we aren't scrutinizing them for flaws like educators do!
- The first step towards correction is detection, but this is tedious…

RQ: Can we try to automate this process?

BenchMarker: Detecting Multiple-Choice Question Flaws

Based on my position paper, we use LLM judges to detect three issues:

BenchMarker: Detecting Multiple-Choice Question Flaws

Based on my position paper, we use LLM judges to detect three issues:

Which of These Best Describes Multiple Choice Evaluation with LLMs?
A) Forced B) Flawed C) Fixable D) All of the Above

Nishant Balepur
University of Maryland
nbalepur@umd.edu

Rachel Rudinger
University of Maryland
rudinger@umd.edu

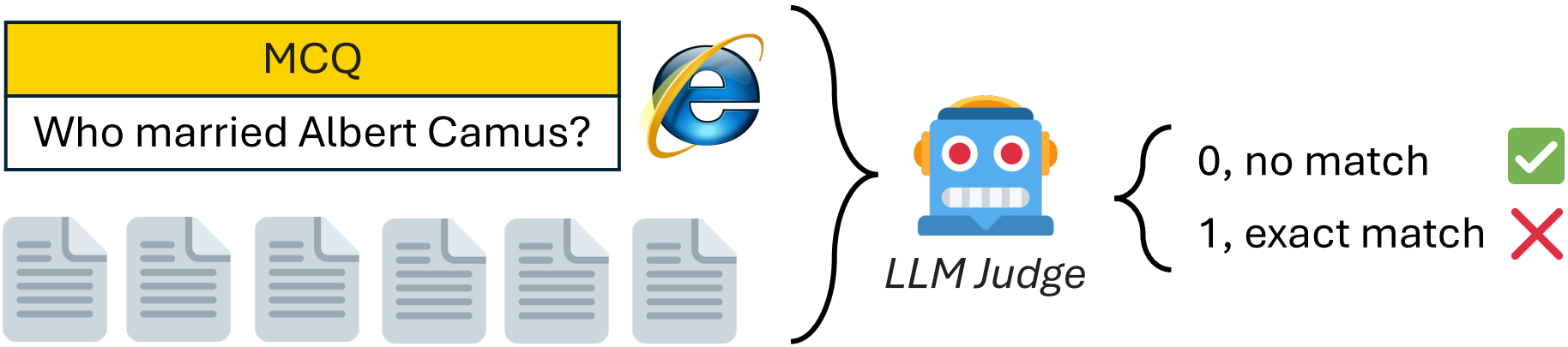
Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Has way more ideas from education research on improving MCQA benchmarks! (at ACL 2025)

BenchMarker: Detecting Multiple-Choice Question Flaws

Based on my position paper, we use LLM judges to detect three issues:

- **Contamination:** Items that exist *exactly* on the internet [1]



(Internet is our proxy for pre-training data)

[1] An Open-Source Data Contamination Report for Large Language Models

BenchMarker: Detecting Multiple-Choice Question Flaws

Based on my position paper, we use LLM judges to detect three issues:

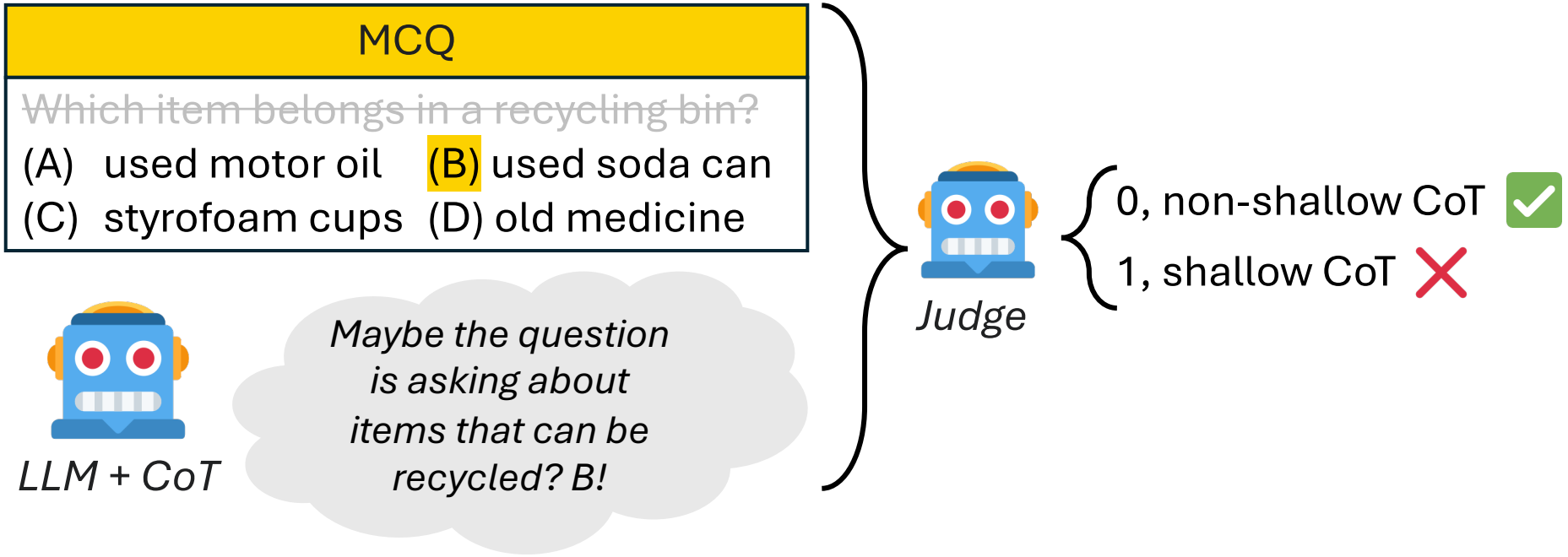
- **Contamination:** Items that exist *exactly* on the internet
- **Shortcuts:** LLMs can cheat on multiple-choice questions

MCQ
Which item belongs in a recycling bin? (A) used motor oil (B) used soda can (C) styrofoam cups (D) old medicine

BenchMarker: Detecting Multiple-Choice Question Flaws

Based on my position paper, we use LLM judges to detect three issues:

- **Contamination:** Items that exist *exactly* on the internet
- **Shortcuts:** LLMs can cheat on multiple-choice questions [1, 2]



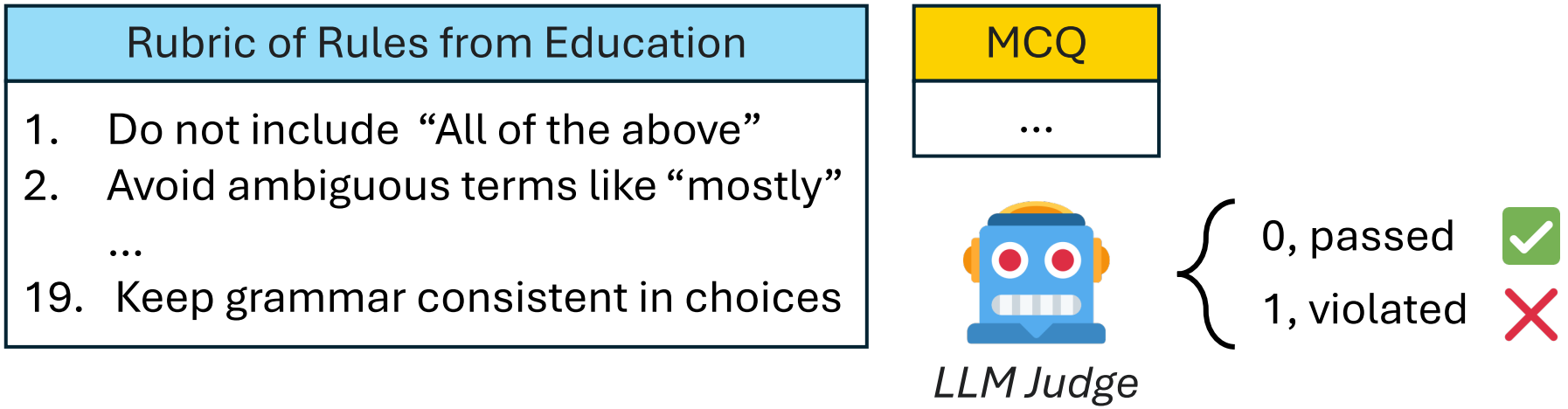
[1] How Do LLMs Answer Multiple-Choice Questions Without the Question? (ACL 24)

[2] Test-Time Reasoners Are Strategic Multiple-Choice Test-Takers (ACL 26)

BenchMarker: Detecting Multiple-Choice Question Flaws

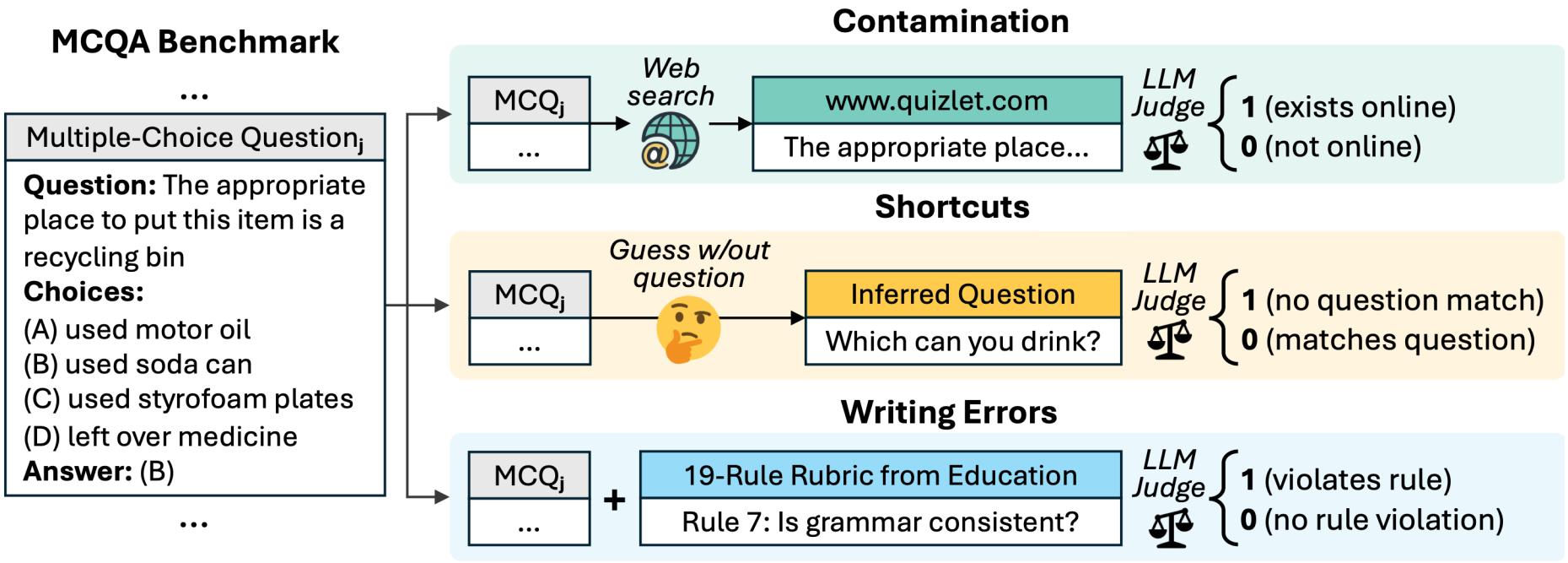
Based on my position paper, we use LLM judges to detect three issues:

- **Contamination:** Items that exist *exactly* on the internet
- **Shortcuts:** LLMs can cheat on multiple-choice questions
- **Writing Errors:** Flaws from a 19-rule rubric from education [1, 2]



[1] Impact of item-writing flaws in multiple-choice questions on student achievement
[2] Plausibly Problematic Questions in Multiple-Choice Benchmarks (EMNLP 25 findings)

BenchMarker: Detecting Multiple-Choice Question Flaws



Report Card

Contamination: 0.21
> *Item **j** was found online*

Shortcuts: 0.32
> *Item **j** has a shortcut: the right answer is the only drink*

Writing Errors: 0.45
> *Item **j** R7: "This item" has a grammar mismatch - "plates"*

- Overall scores and per-item breakdown
- Rationales for model predictions
- Implemented in InspectAI (UI, logging, ...)

Validating BenchMarker: Human Annotations

We collect a validation set of 8042 labels to pick LLM judges for BenchMarker

- 3,419 labels are from us (the authors!)
  - Author A labels MCQs by doing the same task as LLM judges
  - For contamination, they can access search engines
  - Author B labels a subset of 950 → over 80% agreement per metric

- 4,123 writing error labels from Costello et al. (2018a) → out-of-domain data

Validating BenchMarker: Which Search Engine to Use?

Validating BenchMarker: Which Search Engine to Use?

Method	Accuracy	F1 Score	Cohen's κ
Google + Gemini Pro	0.70*	0.65	0.41
Google + GPT-5	**0.71***	**0.68**	**0.44**
Google + Claude Sonnet	0.69*	0.64	0.40
Brave + GPT-5	0.54	0.28	0.14
Perplexity + GPT-5	0.64	0.53	0.31
Exa + GPT-5	0.59	0.43	0.22
Tavily + GPT-5	0.58	0.41	0.20
Serper + GPT-5	0.68	0.65	0.36
Random (50/50)	0.50	0.52	0.00
Always Not Flawed	0.46	0.00	0.00
Always Flawed	0.54	0.70	0.00

Validating BenchMarker: Which Search Engine to Use?

Method	Accuracy	F1 Score	Cohen's κ
Google + Gemini Pro	0.70*	0.65	0.41
Google + GPT-5	**0.71***	**0.68**	**0.44**
Google + Claude Sonnet	0.69*	0.64	0.40
Brave + GPT-5	0.54	0.28	0.14
Perplexity + GPT-5	0.64	0.53	0.31
Exa + GPT-5	0.59	0.43	0.22
Tavily + GPT-5	0.58	0.41	0.20
Serper + GPT-5	0.68	0.65	0.36
Random (50/50)	0.50	0.52	0.00
Always Not Flawed	0.46	0.00	0.00
Always Flawed	0.54	0.70	0.00

- GPT-5 is the most reliable backbone LLM

Validating BenchMarker: Which Search Engine to Use?

Method	Accuracy	F1 Score	Cohen's κ
Google + Gemini Pro	0.70*	0.65	0.41
Google + GPT-5	**0.71***	**0.68**	**0.44**
Google + Claude Sonnet	0.69*	0.64	0.40
Brave + GPT-5	0.54	0.28	0.14
Perplexity + GPT-5	0.64	0.53	0.31
Exa + GPT-5	0.59	0.43	0.22
Tavily + GPT-5	0.58	0.41	0.20
Serper + GPT-5	0.68	0.65	0.36
Random (50/50)	0.50	0.52	0.00
Always Not Flawed	0.46	0.00	0.00
Always Flawed	0.54	0.70	0.00

- GPT-5 is the most reliable backbone LLM
- Google is the most reliable search engine

Validating BenchMarker: Which LLM Judges to Use?

Validating BenchMarker: Which LLM Judges to Use?

	Shortcuts			*Writing (In Domain, NLP)*		
Method	Accuracy	F1 Score	Cohen's κ	Accuracy	F1 Score	Cohen's κ
Gemini 2.5 Lite	0.76	0.74	0.51	0.68	0.56	0.35
Gemini 2.5 Flash	0.69	0.68	0.41	0.79∗	0.62	0.47
Gemini 2.5 Pro	0.70	0.69	0.43	**0.82**∗	**0.66**	**0.53**
GPT-5 Nano	0.68	0.65	0.38	0.74	0.39	0.22
GPT-5 Mini	0.77	0.72	0.53	0.75	0.55	0.38
GPT-5	**0.82**∗	**0.75**	**0.61**	0.81 ∗	0.63	0.50
Claude 4.5 Haiku	0.78∗	0.73	0.55	0.72	0.58	0.38
Claude 4.5 Sonnet	0.81∗	0.75	0.59	0.79∗	0.63	0.48
Command R	0.74	0.72	0.50	0.77	0.53	0.38
Command R+	0.72	0.70	0.46	0.76	0.53	0.36
Qwen-3 0.6B	0.71	0.64	0.39	0.73	0.05	0.00
Qwen-3 1.7B	0.62	0.64	0.31	0.76	0.25	0.16
Qwen-3 4B	0.42	0.55	0.05	0.73	0.49	0.32
Qwen-3 8B	0.57	0.61	0.24	0.70	0.54	0.33
Qwen-3 14B	0.61	0.64	0.30	0.71	0.55	0.34
Qwen-3 32B	0.73	0.71	0.48	0.71	0.55	0.35
Gemma-3 4B	0.50	0.59	0.16	0.63	0.46	0.20
Gemma-3 12B	0.61	0.64	0.31	0.75	0.04	0.03
Gemma-3 27B	0.74	0.72	0.50	0.75	0.03	0.02
LLaMA-3.2 1B	0.56	0.30	0.00	0.47	0.38	0.03
LLaMA-3.2 3B	0.58	0.62	0.26	0.68	0.37	0.16
LLaMA-3.1 8B	0.49	0.58	0.14	0.71	0.38	0.19
LLaMA-3.1 70B	0.61	0.62	0.28	0.64	0.47	0.22

Validating BenchMarker: Which LLM Judges to Use?

	Shortcuts			*Writing (In Domain, NLP)*		
Method	Accuracy	F1 Score	Cohen's κ	Accuracy	F1 Score	Cohen's κ
Gemini 2.5 Lite	0.76	0.74	0.51	0.68	0.56	0.35
Gemini 2.5 Flash	0.69	0.68	0.41	0.79∗	0.62	0.47
Gemini 2.5 Pro	0.70	0.69	0.43	**0.82**∗	**0.66**	**0.53**
GPT-5 Nano	0.68	0.65	0.38	0.74	0.39	0.22
GPT-5 Mini	0.77	0.72	0.53	0.75	0.55	0.38
GPT-5	**0.82**∗	**0.75**	**0.61**	0.81 ∗	0.63	0.50
Claude 4.5 Haiku	0.78∗	0.73	0.55	0.72	0.58	0.38
Claude 4.5 Sonnet	0.81∗	0.75	0.59	0.79∗	0.63	0.48
Command R	0.74	0.72	0.50	0.77	0.53	0.38
Command R+	0.72	0.70	0.46	0.76	0.53	0.36
Qwen-3 0.6B	0.71	0.64	0.39	0.73	0.05	0.00
Qwen-3 1.7B	0.62	0.64	0.31	0.76	0.25	0.16
Qwen-3 4B	0.42	0.55	0.05	0.73	0.49	0.32
Qwen-3 8B	0.57	0.61	0.24	0.70	0.54	0.33
Qwen-3 14B	0.61	0.64	0.30	0.71	0.55	0.34
Qwen-3 32B	0.73	0.71	0.48	0.71	0.55	0.35
Gemma-3 4B	0.50	0.59	0.16	0.63	0.46	0.20
Gemma-3 12B	0.61	0.64	0.31	0.75	0.04	0.03
Gemma-3 27B	0.74	0.72	0.50	0.75	0.03	0.02
LLaMA-3.2 1B	0.56	0.30	0.00	0.47	0.38	0.03
LLaMA-3.2 3B	0.58	0.62	0.26	0.68	0.37	0.16
LLaMA-3.1 8B	0.49	0.58	0.14	0.71	0.38	0.19
LLaMA-3.1 70B	0.61	0.62	0.28	0.64	0.47	0.22

- Open-weight LLMs lag behind (we tried ensembling, calibration, …)

Applying BenchMarker to benchmark benchmarks

Now that we have an auditing toolkit, let's apply it to benchmarks!

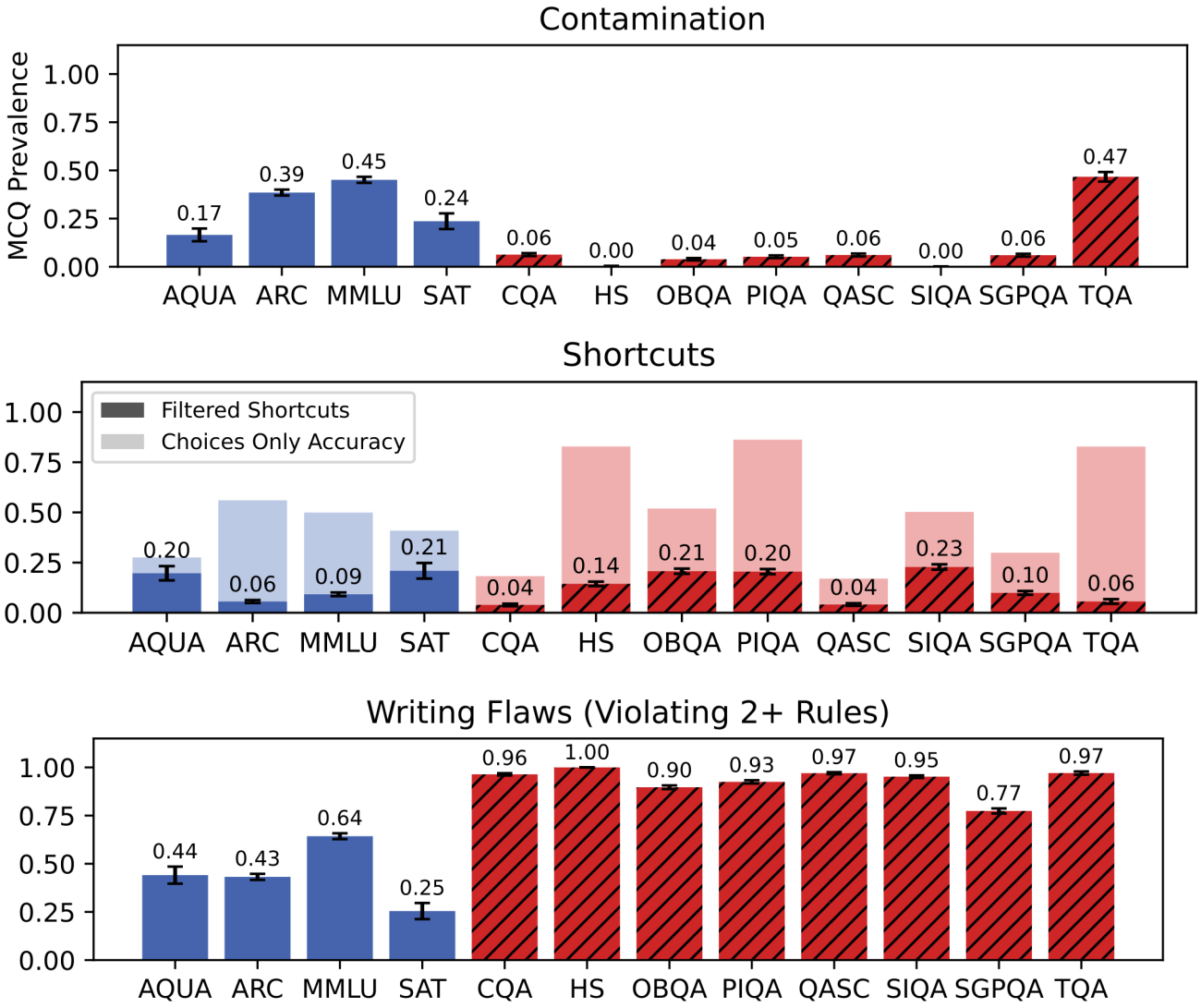
Dataset	Domain	Difficulty	Creation Strategy
Algebra Question Answering (Zhong et al., 2024, AQuA)	Math	Graduate	Student Exams
AI2 Reasoning Challenge (Clark et al., 2018, ARC)	Science	Elementary	Student Exams
CommonsenseQA (Talmor et al., 2019, CQA)	Commonsense	General	Crowdworkers
HellaSwag (Zellers et al., 2019, HS)	Commonsense	General	Model Generated
Multitask Language Understanding (Hendrycks et al., 2021, MMLU)	Multi-Subject	College	Student Exams
Open Book Question Answering (Mihaylov et al., 2018, OBQA)	Science	Elementary	Crowdworkers
Physical Interaction Question Answering (Bisk et al., 2020, PIQA)	Commonsense	General	Crowdworkers
Question Answering via Sentence Composition (Khot et al., 2020, QASC)	Science	Elementary	Crowdworkers
Scholastic Aptitude Test (Zhong et al., 2024, SAT)	Math	High School	Student Exams
Social Intelligence Question Answering (Sap et al., 2019, SIQA)	Commonsense	General	Crowdworkers
Super Google Proof Question Answering (Du et al., 2025, SGPQA)	Multi-Subject	Graduate	Expert+Model
Truthful Question Answering (Lin et al., 2021, TQA)	Commonsense	General	Author-Written

Table 1: The MCQA benchmarks we mainly explore. We audit MCQs across varied domains, difficulties, and creation strategies.

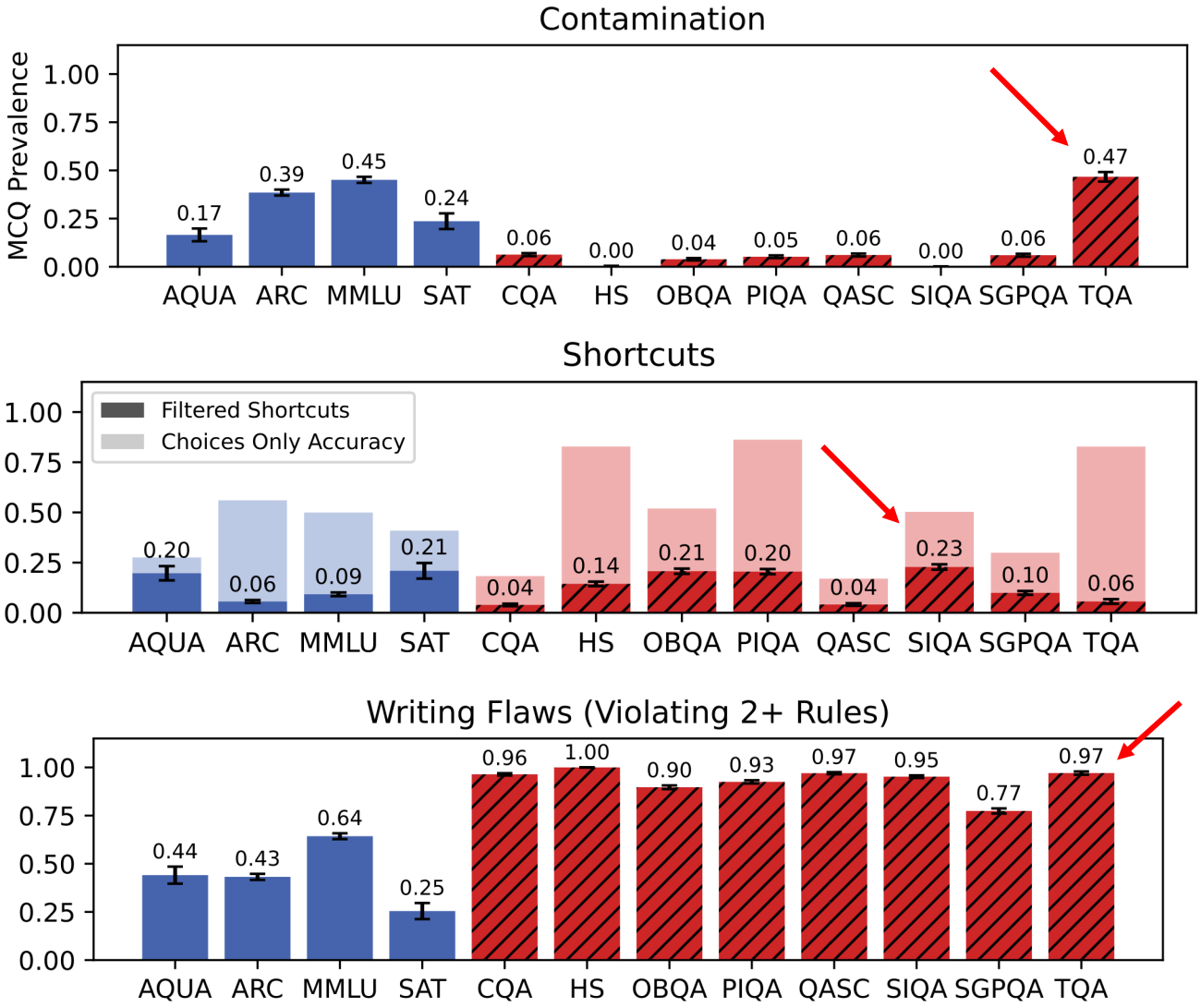
We select popular MCQA benchmarks across varied:

- Domains (math, science, commonsense, ...)
- Difficulty (elementary, high school, graduate, ...)
- Creation strategies (student exams, crowdworkers, model-generated, ...)

BenchMarker can detect flaws in NLP benchmarks

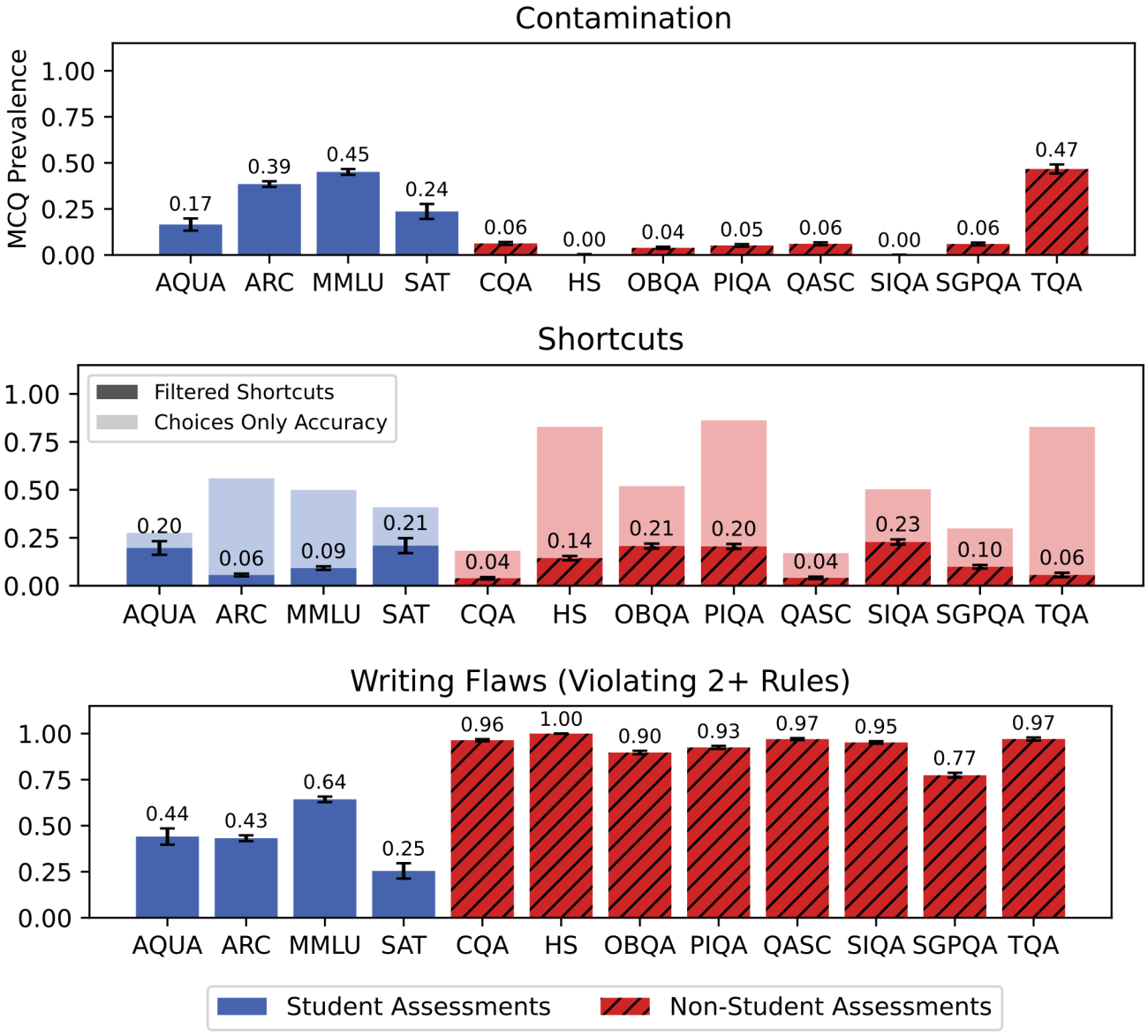


BenchMarker can detect flaws in NLP benchmarks



- Many NLP benchmarks are flawed!

BenchMarker can detect flaws in NLP benchmarks



- Many NLP benchmarks are flawed!
- Benchmarks from real exams are higher-quality than model-generated exams

And these flaws impact NLP evaluation!

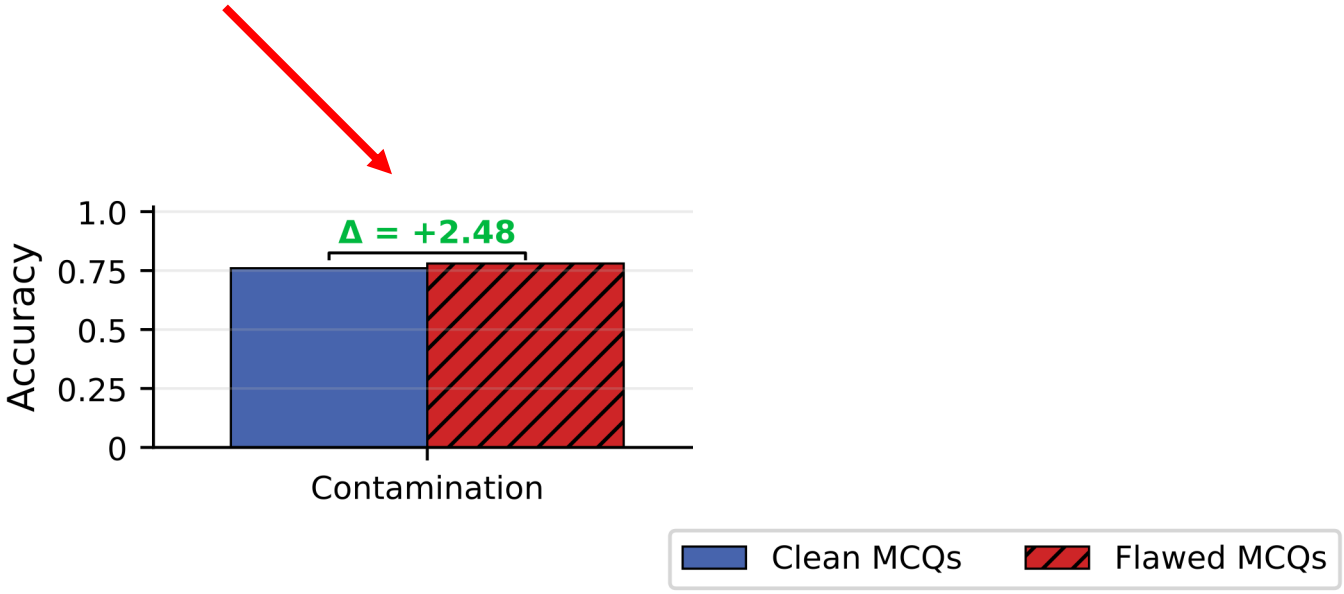
And these flaws impact NLP evaluation!

We evaluate 10 test-taker LLMs on the clean + flawed MCQs:

And these flaws impact NLP evaluation!

We evaluate 10 test-taker LLMs on the clean + flawed MCQs:

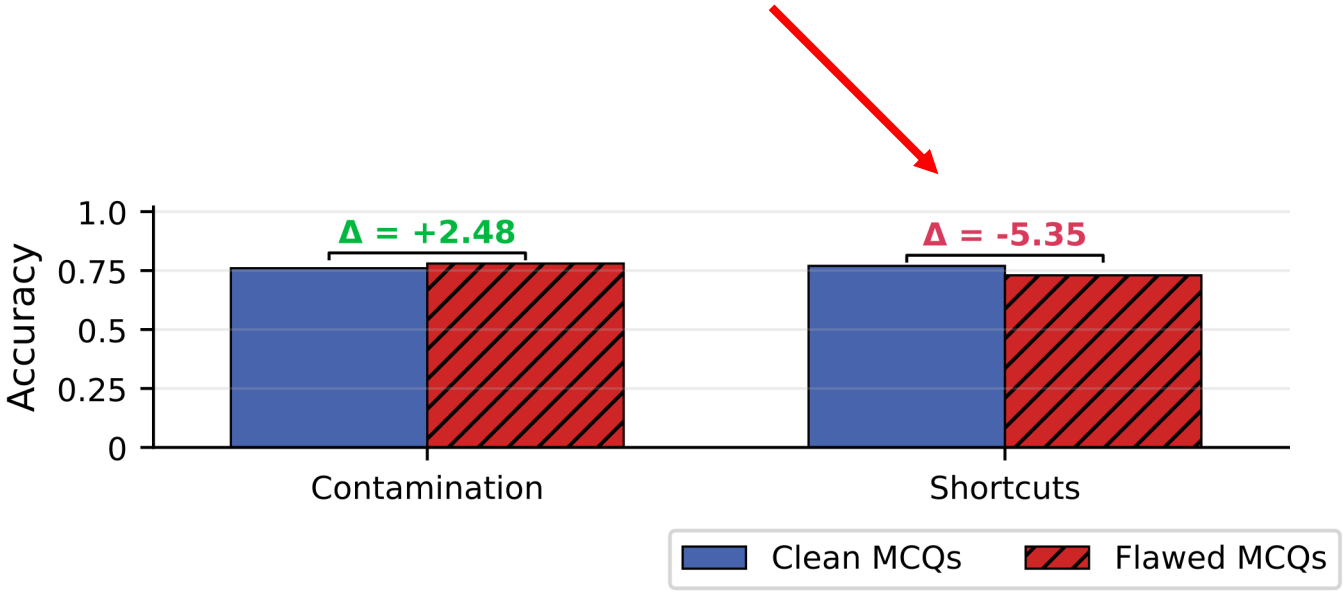
- ➢ Contamination makes items easier



And these flaws impact NLP evaluation!

We evaluate 10 test-taker LLMs on the clean + flawed MCQs:

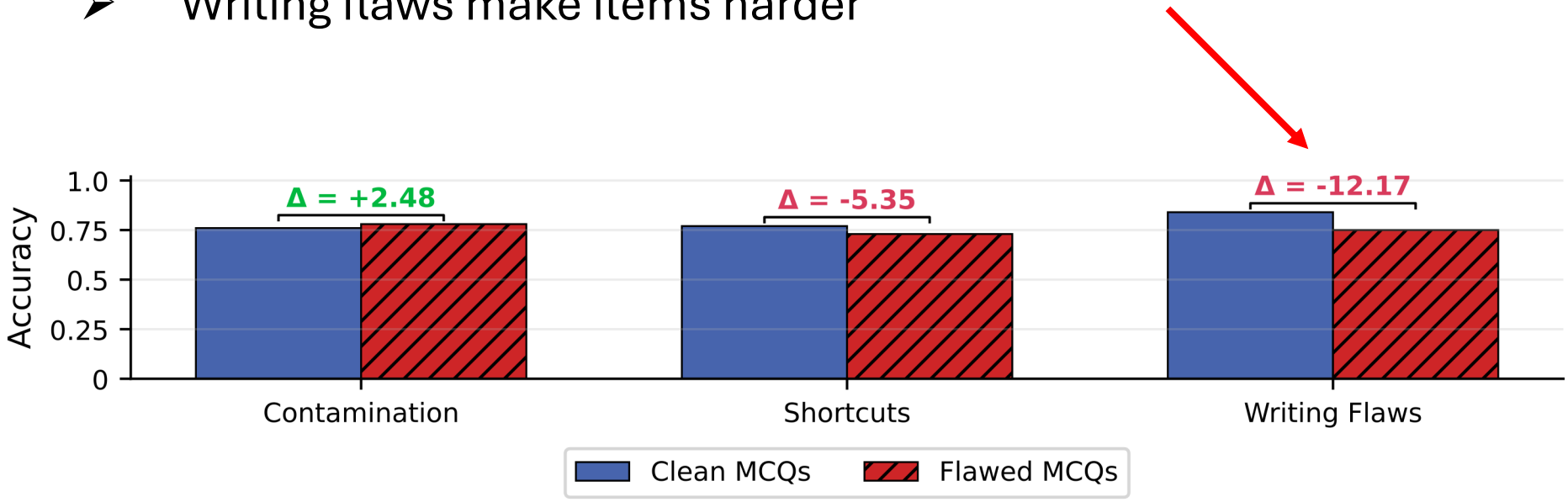
- Contamination makes items easier
- Shortcuts makes items harder (but this is really noisy)



And these flaws impact NLP evaluation!

We evaluate 10 test-taker LLMs on the clean + flawed MCQs:

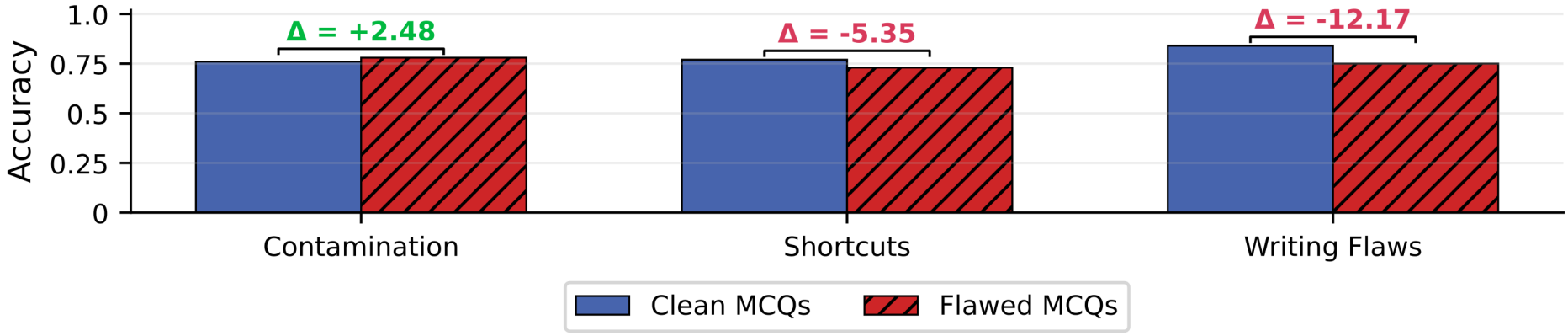
- Contamination makes items easier
- Shortcuts makes items harder (but this is really noisy)
- Writing flaws make items harder



And these flaws impact NLP evaluation!

We evaluate 10 test-taker LLMs on the clean + flawed MCQs:

- Contamination makes items easier
- Shortcuts makes items harder (but this is really noisy)
- Writing flaws make items harder



Are we under-estimating models?

(I didn't show, but this also slightly impacts model ranks)

Is it simple for LLMs to fix these issues?

Is it simple for LLMs to fix these issues?

We prompt LLMs to try and fix one dataset's writing flaws (TruthfulQA)

Model
GPT-5.2
Claude Sonnet
Qwen 235B

Is it simple for LLMs to fix these issues?

We prompt LLMs to try and fix one dataset's writing flaws (TruthfulQA)

- LLM's can't reduce all errors

Model	$\mathbb{P}$(2+ Errors)
GPT-5.2	$0.97 \rightarrow 0.64$
Claude Sonnet	$0.97 \rightarrow 0.64$
Qwen 235B	$0.97 \rightarrow 0.68$

Is it simple for LLMs to fix these issues?

We prompt LLMs to try and fix one dataset's writing flaws (TruthfulQA)

- LLM's can't reduce all errors
- Few questions are flawless

Model	$\mathbb{P}$(2+ Errors)	$\mathbb{P}$(Flawless)
GPT-5.2	$0.97 \rightarrow 0.64$	0.26
Claude Sonnet	$0.97 \rightarrow 0.64$	0.32
Qwen 235B	$0.97 \rightarrow 0.68$	0.26

Is it simple for LLMs to fix these issues?

We prompt LLMs to try and fix one dataset's writing flaws (TruthfulQA)

- LLM's can't reduce all errors
- Few questions are flawless
- They often add new errors!

Model	ℙ(2+ Errors)	ℙ(Flawless)	ℙ(New Error)
GPT-5.2	$0.97 \rightarrow 0.64$	0.26	0.56
Claude Sonnet	$0.97 \rightarrow 0.64$	0.32	0.68
Qwen 235B	$0.97 \rightarrow 0.68$	0.26	0.62

Is it simple for LLMs to fix these issues?

We prompt LLMs to try and fix one dataset's writing flaws (TruthfulQA)

- LLM's can't reduce all errors
- Few questions are flawless
- They often add new errors!

Model	$\mathbb{P}$(2+ Errors)	$\mathbb{P}$(Flawless)	$\mathbb{P}$(New Error)
GPT-5.2	$0.97 \rightarrow 0.64$	0.26	0.56
Claude Sonnet	$0.97 \rightarrow 0.64$	0.32	0.68
Qwen 235B	$0.97 \rightarrow 0.68$	0.26	0.62

Simple prompting alone is not enough!

What about prior rewrites? MMLU-Pro?

MMLU-Pro rewrote MMLU with LLMs to make MCQs “harder”

What about prior rewrites? MMLU-Pro?

MMLU-Pro rewrote MMLU with LLMs to make MCQs “harder”

- Accuracy does drop!

Dataset	Contam.	Shortcuts	Writing	Accuracy
MMLU	0.45	0.12	0.12	0.79
MMLU Pro	0.24	0.12	0.15	0.63

What about prior rewrites? MMLU-Pro?

MMLU-Pro rewrote MMLU with LLMs to make MCQs “harder”

- Accuracy does drop!
- But writing flaws increase 🫠

Dataset	Contam.	Shortcuts	Writing	Accuracy
MMLU	0.45	0.12	0.12	0.79
MMLU Pro	0.24	0.12	0.15	0.63

How can we progress?

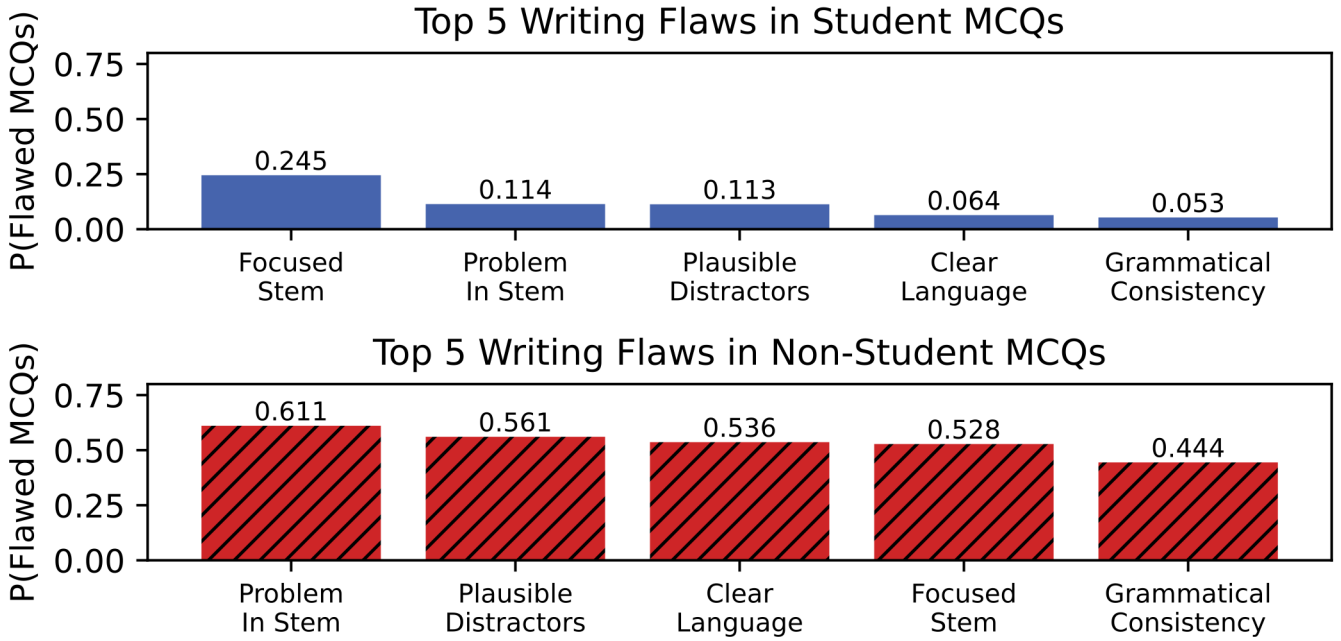
How can we progress?

We analyze the types of writing errors that occur in benchmarks

How can we progress?

We analyze the types of writing errors that occur in benchmarks

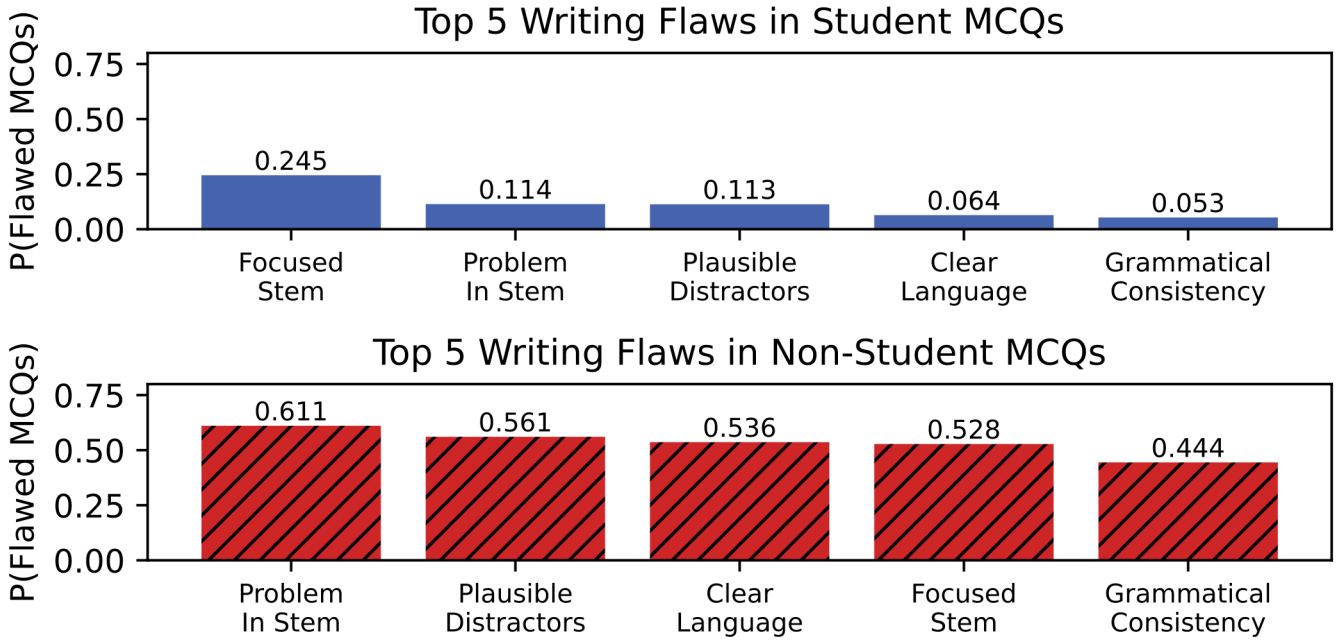
- Benchmarks derived from real student exams + LLM-generated exams have similar flaw distributions!



How can we progress?

We analyze the types of writing errors that occur in benchmarks

- Benchmarks derived from real student exams + LLM-generated exams have similar flaw distributions!



NLP + education researchers can work together!

Next steps for BenchMarker!

 **Takeaway:** MCQA datasets suck, but detecting issues is tractable!

Next Steps **Within** QA

Considering flaws in metrics 

 *Human-AI Collab. for Rewriting*

Open-domain QA datasets 

 *Training efficient LLM judges*

Next Steps **Beyond** QA → Agents

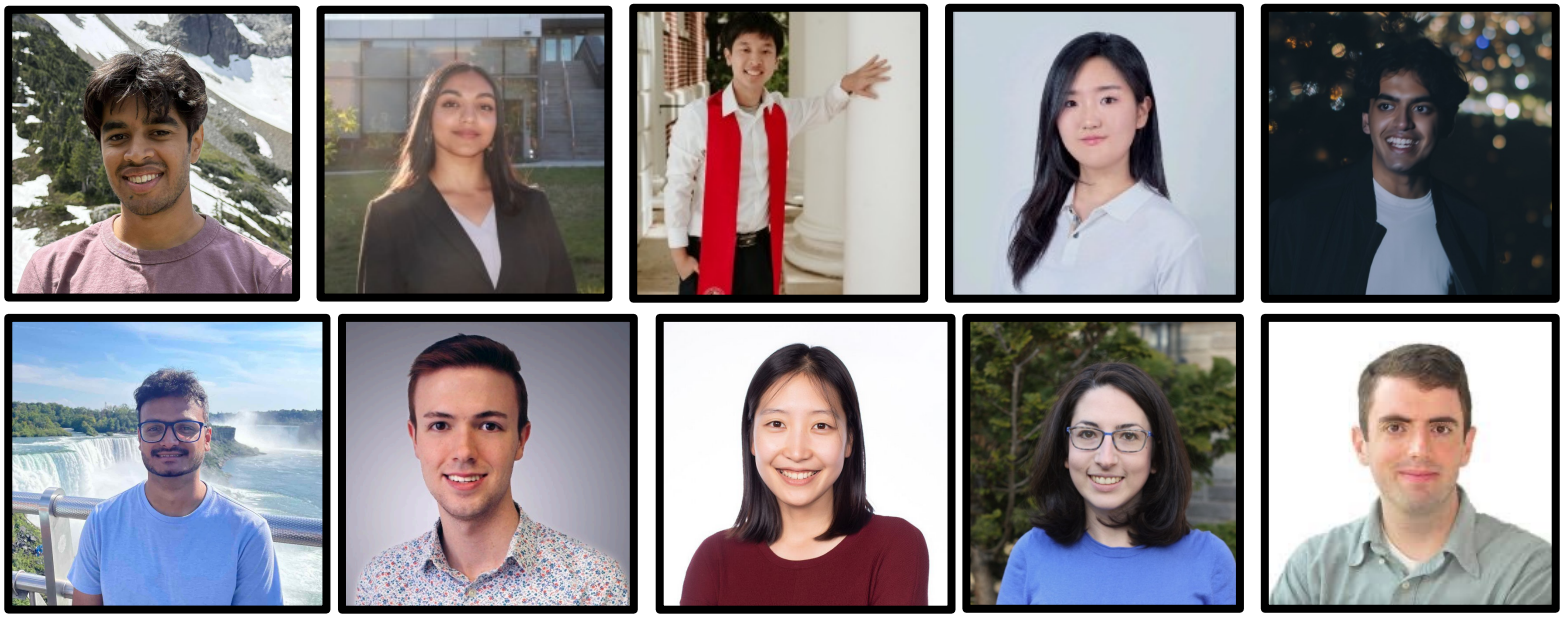
Auditing agentic benchmarks 

 *Can education apply to agents?*

 *Trace analysis to detect agent errors*

Thank You :)

Nishant (that's me!)



Email: nishantbalepur@gmail.com

X.com: @NishantBalepur

Website: https://www.nbalepur.github.io