

**Which of These Best Describes Multiple Choice  
Evaluation with LLMs?**

# Which of These Best Describes Multiple Choice Evaluation with LLMs?

A) Forced

*Used too often?*

# Which of These Best Describes Multiple Choice Evaluation with LLMs?

A) Forced    B) Flawed

*Has fundamental issues?*

# Which of These Best Describes Multiple Choice Evaluation with LLMs?

A) Forced   B) Flawed   C) Fixable

*Could be better?*

# Which of These Best Describes Multiple Choice Evaluation with LLMs?

A) Forced   B) Flawed   C) Fixable   D) All of the Above

**Nishant Balepur**

[nbalepur@umd.edu](mailto:nbalepur@umd.edu)

Rachel Rudinger

Jordan Boyd-Graber

<https://nbalepur.github.io/>



<https://arxiv.org/abs/2502.14127>

# Multiple-Choice Question Answering is Great in Theory

## Example MCQ

Question: What is the capital of France?

Choices:

- (A) Paris
- (B) Berlin
- (C) Madrid
- (D) Rome

Answer:

# Multiple-Choice Question Answering is Great in Theory

## Example MCQ

Question: What is the capital of France?

Choices:

(A) Paris

*Gold answer*

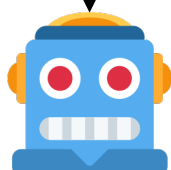
(B) Berlin

(C) Madrid

(D) Rome

*Distractors*

Answer:

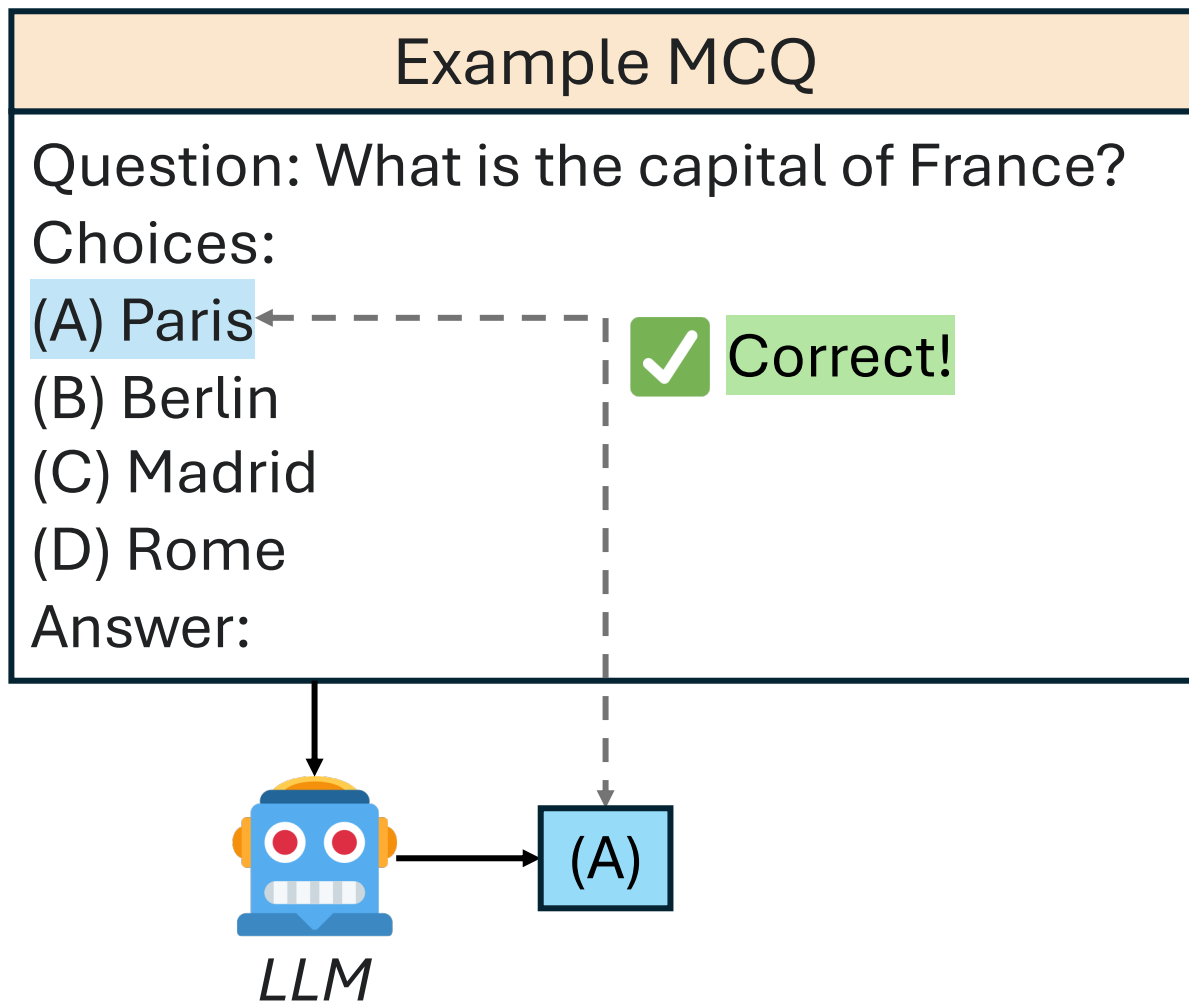


LLM

(A)

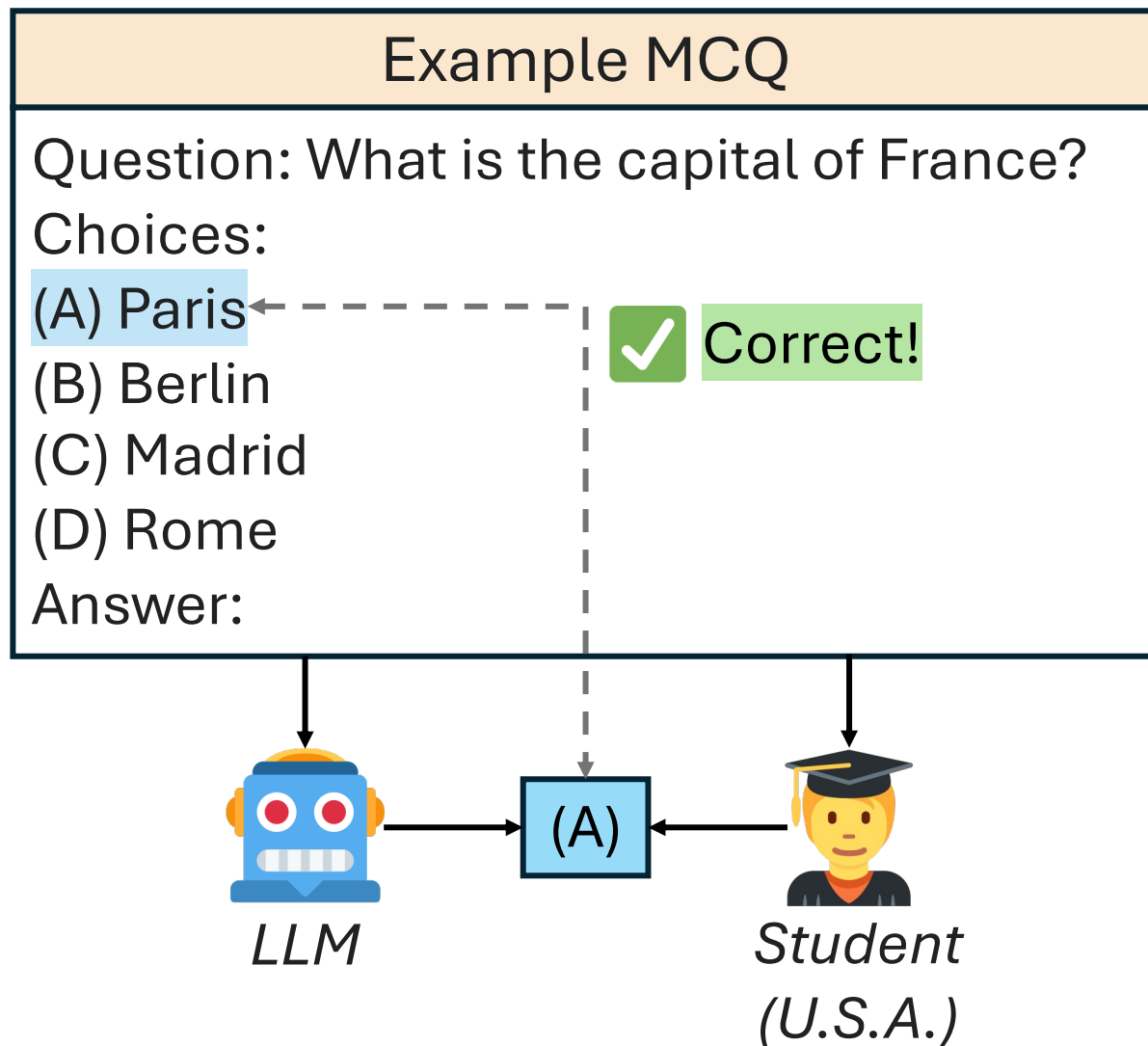


# Multiple-Choice Question Answering is Great in Theory



1) Easy to score

# Multiple-Choice Question Answering is Great in Theory

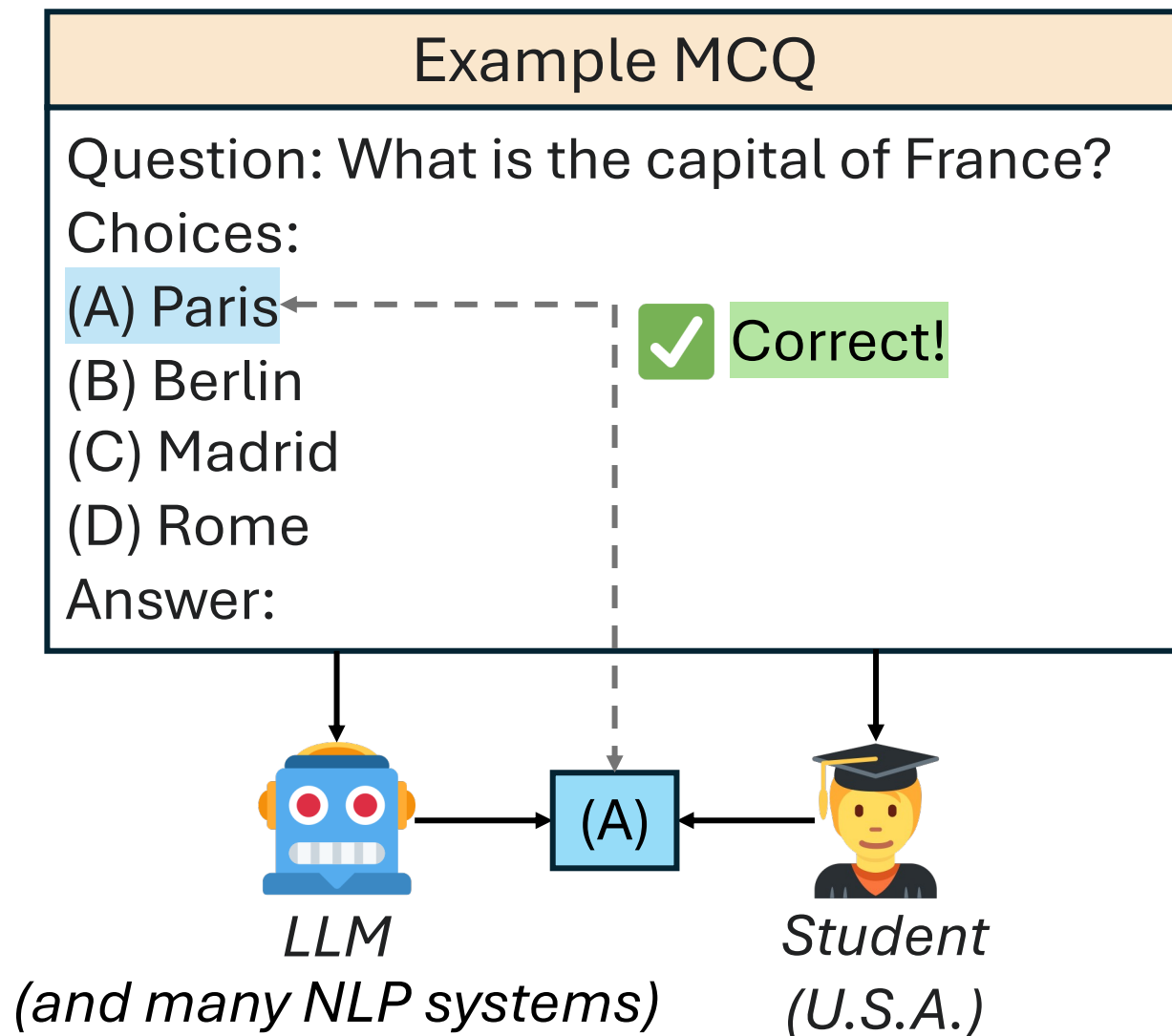


1) Easy to score

2) Aligns with how we test students



# Multiple-Choice Question Answering is Great in Theory



- 1) Easy to score
- 2) Aligns with how we test students



- 3) Historically used in NLP

1988 AAAI Presidential Address


Foundations and Grand  
Challenges of Artificial  
Intelligence

Raj Reddy

# Good luck avoiding MCQA


- 1) Easy to score    2) Aligns with how we test students    3) Historically used in NLP

Want to build a **popular** dataset?

 MEASURING MASSIVE MULTITASK  
G

**Hu**

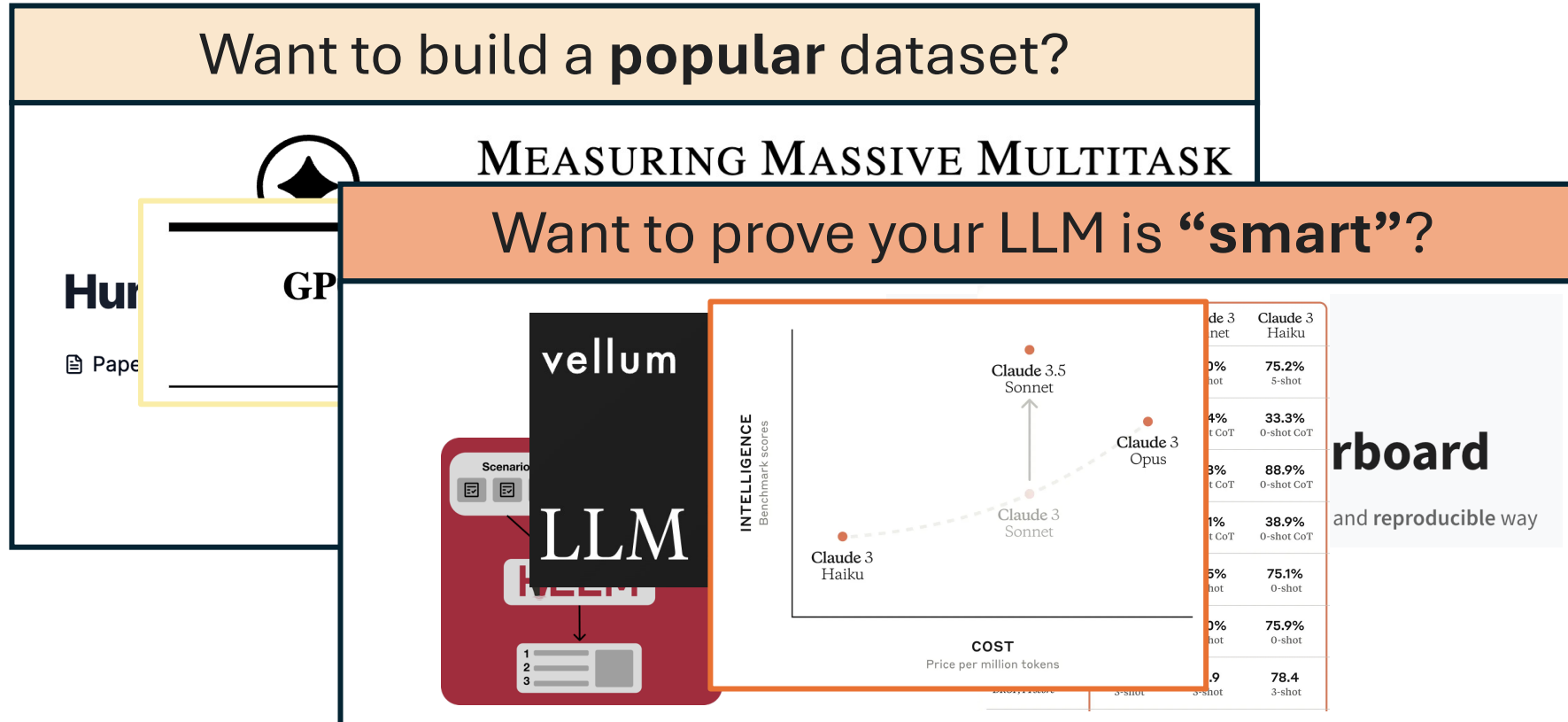
**GPQA: A Graduate-Level Google-Proof Q&A Benchmark**

 Paper

Quantifying and extrapolating the capabilities of language models

# Good luck avoiding MCQA

- 1) Easy to score    2) Aligns with how we test students    3) Historically used in NLP



# Good luck avoiding MCQA

- 1) Easy to score    2) Aligns with how we test students    3) Historically used in NLP


Want to build a **popular** dataset?

MEASURING MASSIVE MULTITASK

Want to prove your LLM is “**smart**”?

Want to get your **Ph.D.**?

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

University of Maryland, College Park, USA  
{nbalepur, spalta, rudinger}@umd.edu

.org

# Good luck avoiding MCQA

- 1) Easy to score    2) Aligns with how we test students    3) Historically used in NLP

Want to build a **popular** dataset?

MEASURING MASSIVE MULTITASK

Are we using MCQA correctly for LLMs?

LLM  
HELM

1  
2  
3

Is Your Large Language Model Knowledgeable or a Choices-Only Cheater?  
Which of These Best Describes Multiple Choice Evaluation with LLMs?  
A) Forced B) Flawed C) Fixable D) All of the Above

It's Not Easy Being Wrong: Large Language Model Struggle with Precise Elimination Questions  
Nishant Balepur, Rachel Rudinger, Jordan Boyd-Graber  
University of Maryland, College Park, USA  
nbalepur@umd.edu, rudinger@umd.edu, jbg@umiacs.umd.edu  
{nbalepur, rudinger}@umd.edu, abhilashar@allenai.org

# Good luck avoiding MCQA

- 1) Easy to score    2) Aligns with how we test students    3) Historically used in NLP

Want to build a **popular** dataset?

MEASURING MASSIVE MULTITASK

Are we using MCQA correctly for LLMs? **No!**

LLM  
HELM

1  
2  
3

Is Your Large Language Model Knowledgeable or a Choices-Only Cheater?  
Which of These Best Describes Multiple Choice Evaluation with LLMs?  
A) Forced B) Flawed C) Fixable D) All of the Above

It's Not Easy Being Wrong: Large Language Model Struggle with Process of Elimination  
Nishant Balepur, Rachel Rudinger, Jordan Boyd-Graber  
University of Maryland, College Park, USA  
nbalepur@umd.edu, rudinger@umd.edu, jbg@umiacs.umd.edu  
{nbalepur, rudinger}@umd.edu, abhilashar@allenai.org



# Are we using MCQA correctly for LLMs? **No!**



Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

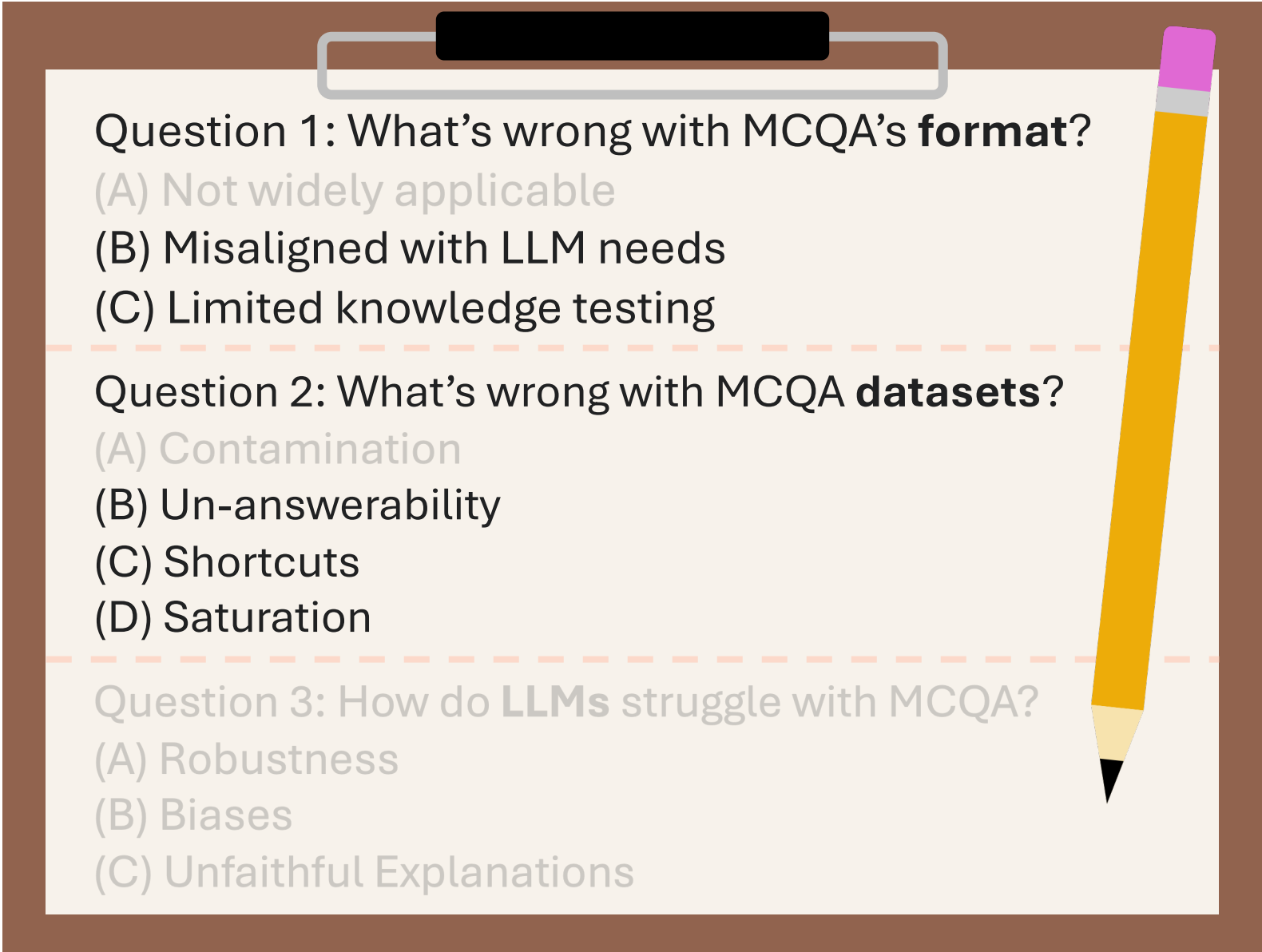
Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability
- (C) Shortcuts
- (D) Saturation

Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

# Are we using MCQA correctly for LLMs? **No!**



Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

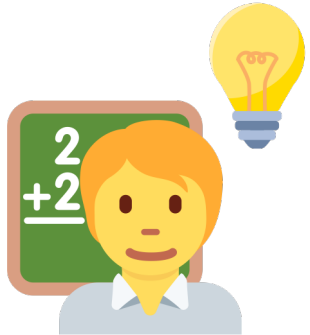
Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability
- (C) Shortcuts
- (D) Saturation

Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

# Are we using MCQA correctly for LLMs? **No!**



Educators  
have solutions!

Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing  $\Rightarrow$  *New Formats*

Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability  $\Rightarrow$  *MCQA Rubrics*
- (C) Shortcuts
- (D) Saturation  $\Rightarrow$  *MCQs Easy for Humans*

Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

# Are we using MCQA correctly for LLMs? **No!**



Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

---

Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability
- (C) Shortcuts
- (D) Saturation

---

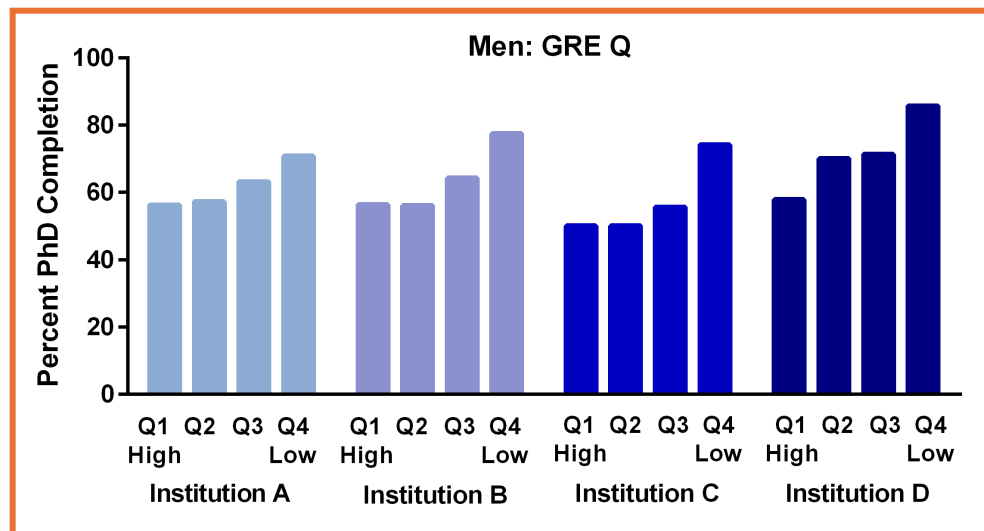
Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

MCQA is simple

# MCQA is simple, so why do humans hate these exams?

*Studies showing it fails to predict student success*




*Dropping standardized exams altogether?!*

**The New York Times**

## *University of California Will No Longer Consider SAT and ACT Scores*

The university system has reached a settlement with students to scrap even optional testing from admissions and scholarship decisions.

 **r/AskAnAmerican** • 3 yr. ago  
Tikomeji

### Multiple-choice Test?

EDUCATION

As a german student I can't imagine multiple choice question test. Do you guys really "mostly" have multiple choice test or also normal test? And if yes how are they look like?

*Using other testing formats*

We should be just as critical for LLMs!

# Evaluations inform LLM selection

So they should contain tasks mirroring how people actually use LLMs

# Evaluations inform LLM selection

So they should **contain tasks** mirroring how people actually use LLMs

## GPT-4 Eval on Academic Benchmarks

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
<b>MMLU [49]</b> Multiple-choice questions in 57 subjects (professional & academic)	<b>86.4%</b> 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
<b>HellaSwag [52]</b> Commonsense reasoning around everyday events	<b>95.3%</b> 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
<b>AI2 Reasoning Challenge (ARC) [54]</b> Grade-school multiple choice science questions. Challenge-set.	<b>96.3%</b> 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
<b>WinoGrande [56]</b> Commonsense reasoning around pronoun resolution	<b>87.5%</b> 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
<b>HumanEval [43]</b> Python coding tasks	<b>67.0%</b> 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
<b>DROP [58] (F1 score)</b> Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	<b>88.4</b> QDGAT [59]
<b>GSM-8K [60]</b> Grade-school mathematics questions	<b>92.0%*</b> 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

**71%** of tasks  
are MCQA!

79% (due to BBH)



### Open LLM Leaderboard

Comparing Large Language Models in an open and reproducible way

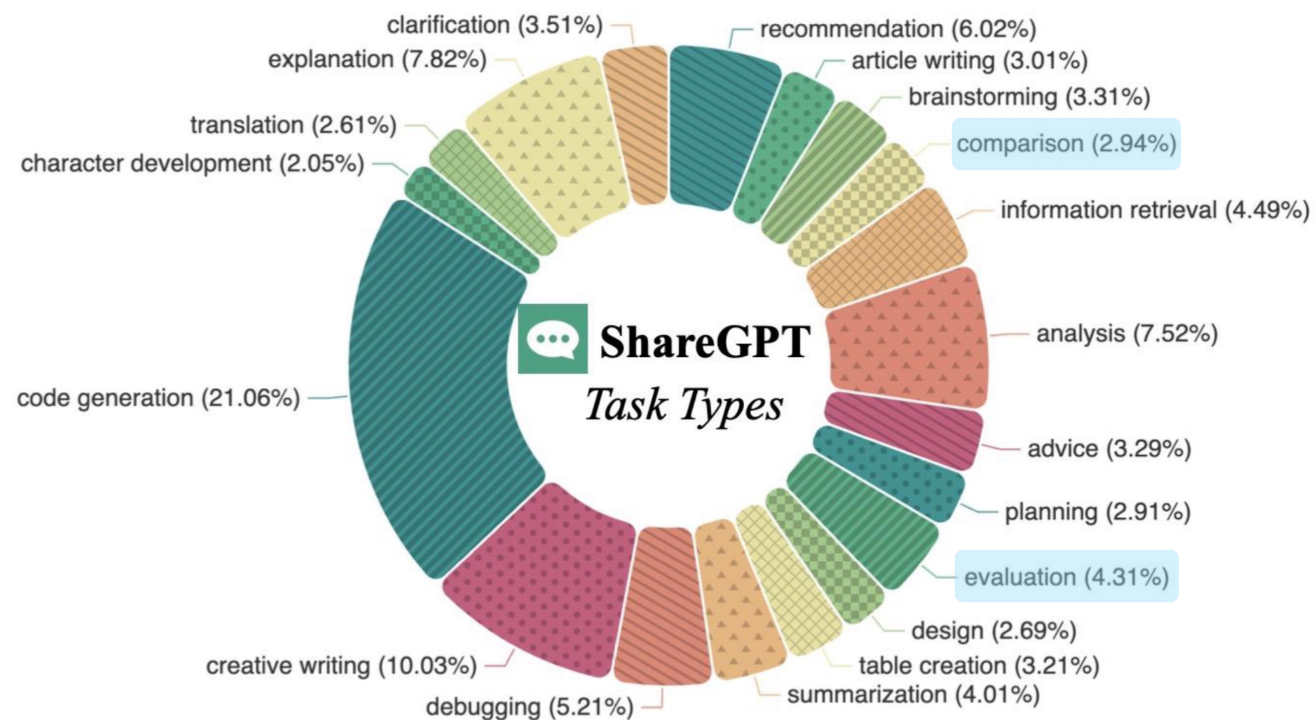


# Evaluations inform LLM selection

So they should contain tasks mirroring how people actually use LLMs

Based on analysis on ShareGPT:<sup>[1]</sup>

*“almost all the user queries are free-form text generations” (i.e. not MCQA)*



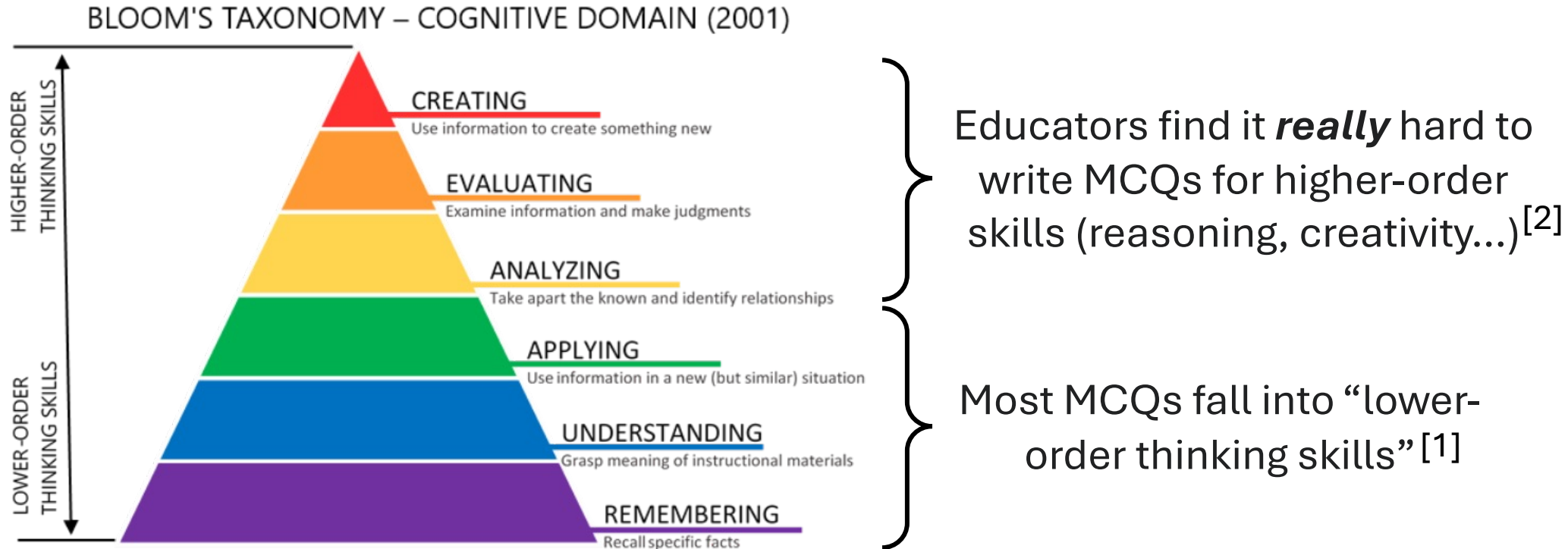
Maybe **7.25%**  
are MCQA?

*versus*

**71%** of tasks  
in benchmarks

[1] [The Shifted and The Overlooked: A Task-oriented Investigation of User-GPT Interactions](#)

# MCQA can't match LLM needs, can it test knowledge?

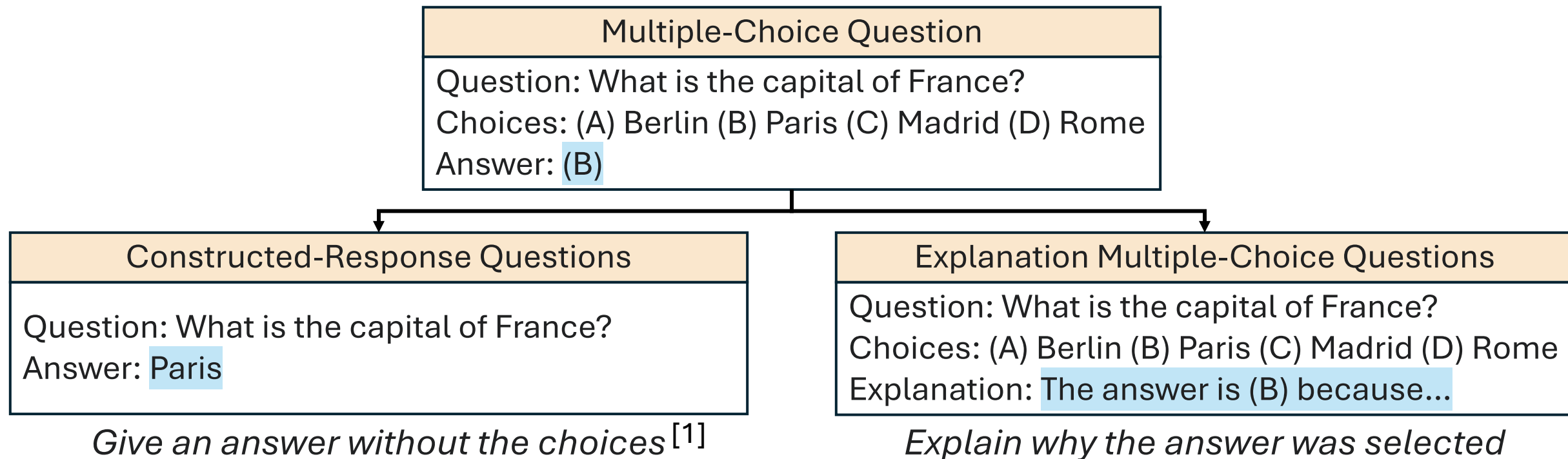


MCQA makes it much harder to test advanced knowledge capabilities!

[1] [Multiple-choice tests and student understanding: What is the connection?](#)

[2] [Multiple choice questions: Can they examine application of knowledge?](#)

# What are better MCQA formats?



- ✓ Generative tasks that align with LLM needs
- ✓ Better tests knowledge (based on education)
- ✗ Harder evaluation metrics

*At least we can improve this!*

[1] [Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena](#)

# Are we using MCQA correctly for LLMs? **No!**



Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

---

Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability
- (C) Shortcuts
- (D) Saturation

---

Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

# Are we using MCQA correctly for LLMs? **No!**



Question 1: What's wrong with MCQA's **format**?

- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

Question 2: What's wrong with MCQA **datasets**?

- (A) Contamination
- (B) Un-answerability
- (C) Shortcuts
- (D) Saturation

Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithful Explanations

# Sometimes, MCQA *is* a valid format to use

It can test comprehension, LLM-as-a-judge, ...

➤ Or maybe I haven't convinced you MCQA's format is bad 🙄

But still, there are **issues in MCQA datasets** we need to fix!



I want to build an MCQA dataset...

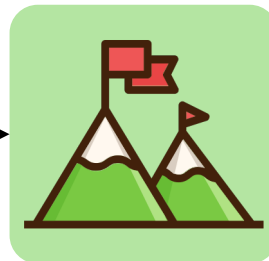
Picking Sources



Writing MCQs



Finalize Dataset



Long-Term Eval



# Sometimes, MCQA *is* a valid format to use

It can test comprehension, LLM-as-a-judge, ...

➤ Or maybe I haven't convinced you MCQA's format is bad 🙄

But still, there are **issues in MCQA datasets** we need to fix!



I want to build an MCQA dataset...

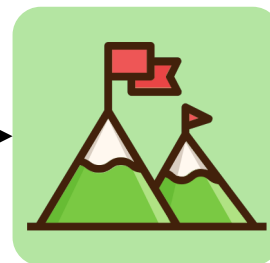
Picking Sources



Writing MCQs



Finalize Dataset



Long-Term Eval



Un-answerability

Shortcuts

Saturation

# Some MCQs are impossible to answer

## Multiple Valid Distractors (Social IQA)[1]

Question: Ash redeemed themselves after retaking the test they failed. How will Ash feel as a result?

Choices: (A) relieved (B) accomplished (C) proud

## Poor Grammar (HellaSwag)[2]

Question: *Man is in roofed gym weightlifting. Woman is walking behind the man watching the man. Woman...*

## Incorrect Answer (MMLU)[3]

Question: The number of energy levels for the  $^{55}\text{Mn}$  nuclide are

Choices: (A) 3 (B) 5 (C) 8 (D) 4

## Missing Information? (MMLU)[4]

From the authors:

*“[we discard] questions that lack necessary information or require non-textual elements like images or tables”*

Researchers don't know how to write MCQs like experts...

[1] [Plausibly Problematic Questions in Multiple-Choice Benchmarks for Commonsense Reasoning](#)

[2] [HellaSwag or HellaBad? 36% of this popular LLM benchmark contains errors](#)

[3] [Are We Done with MMLU?](#)

[4] [MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark](#)



So we should follow educator's guidelines for writing MCQs

### Multiple-Choice Writing Guidelines

# So we should follow educator's guidelines for writing MCQs

## Multiple-Choice Writing Guidelines<sup>[1]</sup>

### General Item-Writing (procedural):

2. Avoid the complex multiple-choice format (e.g. all of the above)

### General Item-Writing (content concerns):

13. Avoid over-specific knowledge when developing the item

[1] [A taxonomy of multiple choice item-writing rules](#) (1989)

# So we should follow educator's guidelines for writing MCQs

## Multiple-Choice Writing Guidelines<sup>[1]</sup>

### General Item-Writing (procedural):

2. Avoid the complex multiple-choice format (e.g. all of the above)

### General Item-Writing (content concerns):

13. Avoid over-specific knowledge when developing the item

### Stem Construction:

20. Ensure the directions in the stem are clear

### Correct Option Development:

37. Make sure there is one and only one correct option

### Distractor Development:

39. Incorporate common errors of students

} **Most important** part of an MCQ!  
*Discerns between low and high skill test-takers*

[1] [A taxonomy of multiple choice item-writing rules](#) (1989)

# Answerable MCQs are still cheatable via **shortcuts**

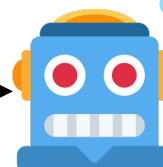
## Intended Solution

MCQ from MMLU

Question: Find all zeros in the indicated finite field of the given polynomial with coefficients in that field.  $x^3 + 2x + 2$  in  $\mathbb{Z}_7$

Choices: (A) 1 (B) 2 (C) 2, 3 (D) 6

Answer:



*I first need to find the zeros of the input equation...*

(C)



**Shortcuts** (e.g. spurious patterns, annotator artifacts, reasoning ...) <sup>[1]</sup>

MCQ from MMLU

~~Question: Find all zeros in the indicated finite field of the given polynomial with coefficients in that field.  $x^3 + 2x + 2$  in  $\mathbb{Z}_7$~~

~~Choices: (A) 1 (B) 2 (C) 2, 3 (D) 6~~

~~Answer:~~



*IDK the answer, but (C) is the only one with 2 numbers...*

(C)



Overestimating knowledge!

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

# Dataset Design: Consistency is Key

If correct answers and distractors have obvious differences, models will detect this

## Multiple-Choice Writing Guidelines

28. Keep the length of the options fairly consistent

34. Avoid giving clues through the use of faulty grammatical construction

## HellaSwag MCQ

Question: A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...



(A) rinses the bucket off with soap and blow dry the dog's head



(B) uses a hose to keep it from getting soapy



(C) gets the dog wet, then it runs away again



(D) gets into a bath tub with the dog

*LLaMA-2 gets **59%** accuracy  
when only using the choices!<sup>[1]</sup>*

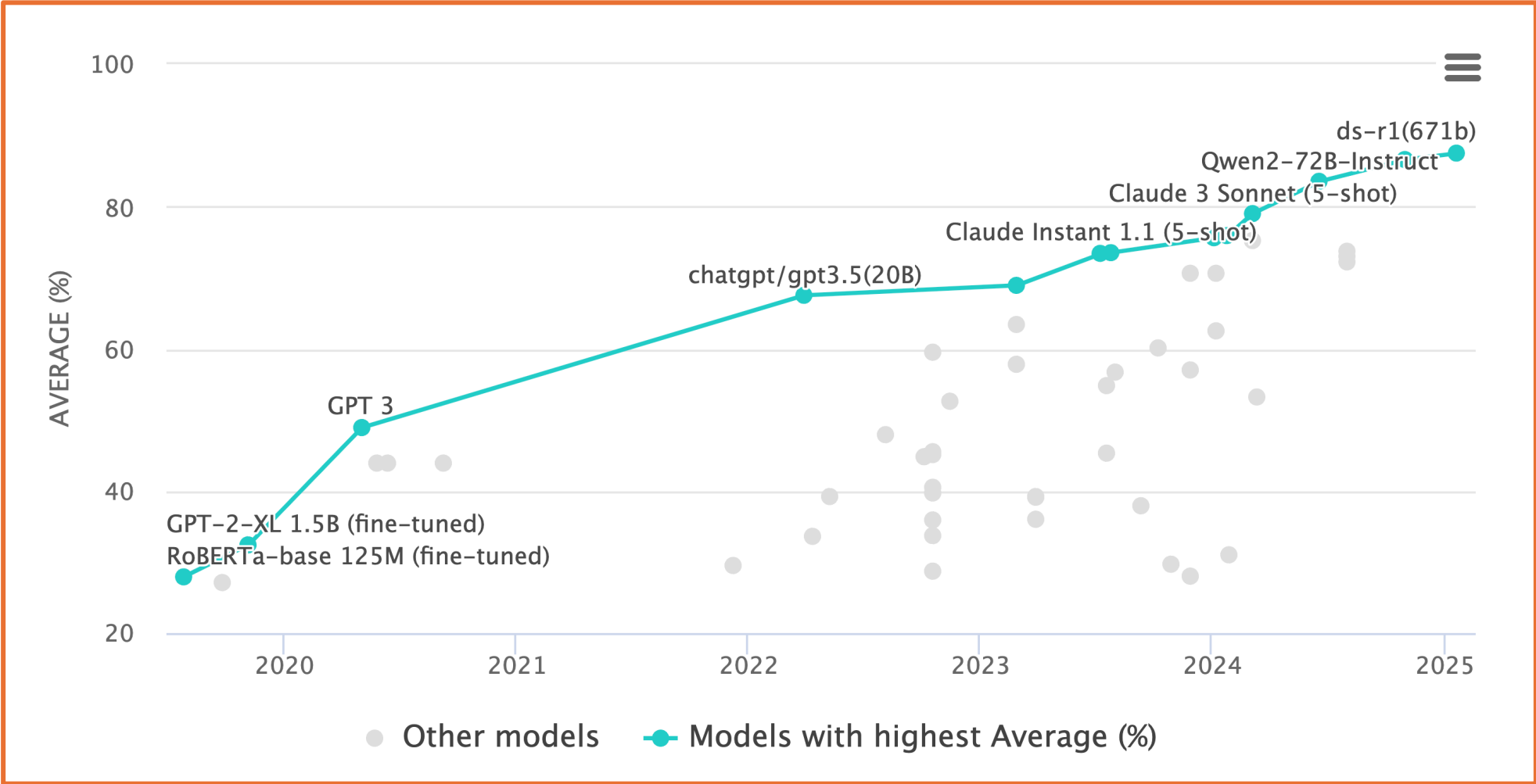
Please write MCQs consistently!<sup>[2]</sup>

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

[2] Language Models Generate Multiple-Choice Questions with Artifacts

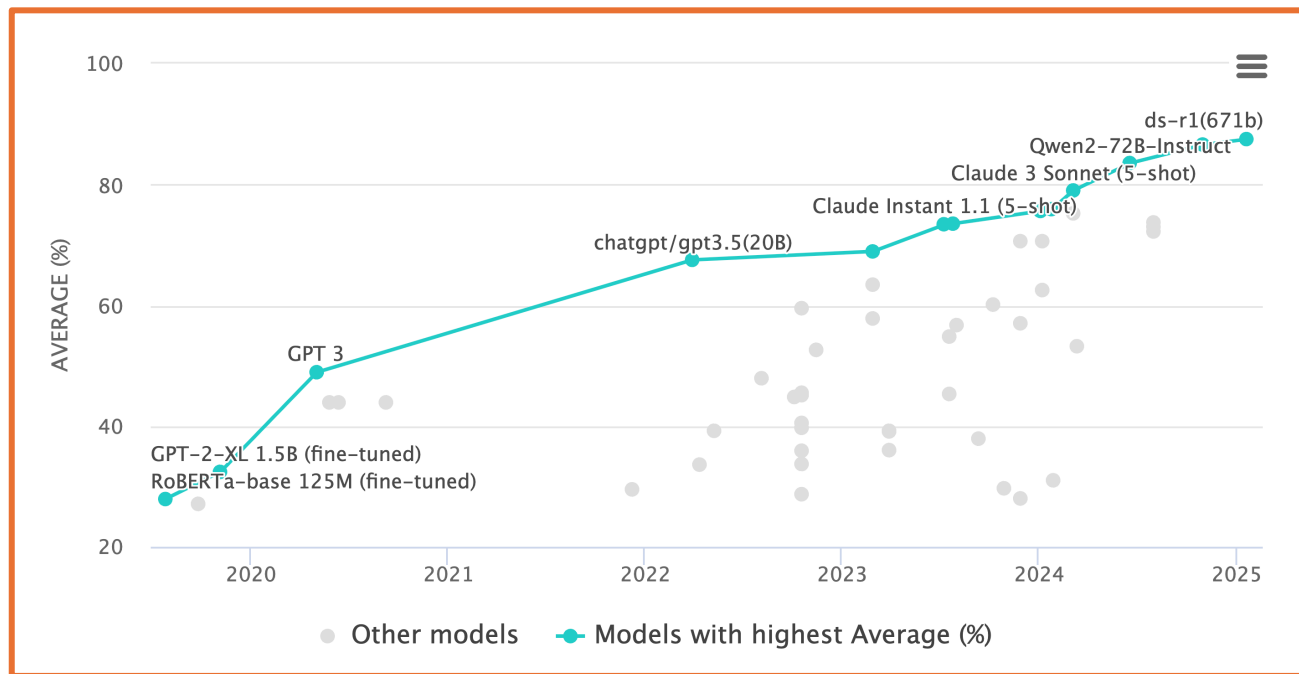
# Even if your dataset is perfect, hill-climbing is inevitable...

MMLU Accuracy over Time



# Even if your dataset is perfect, hill-climbing is inevitable...

## MMLU Accuracy over Time

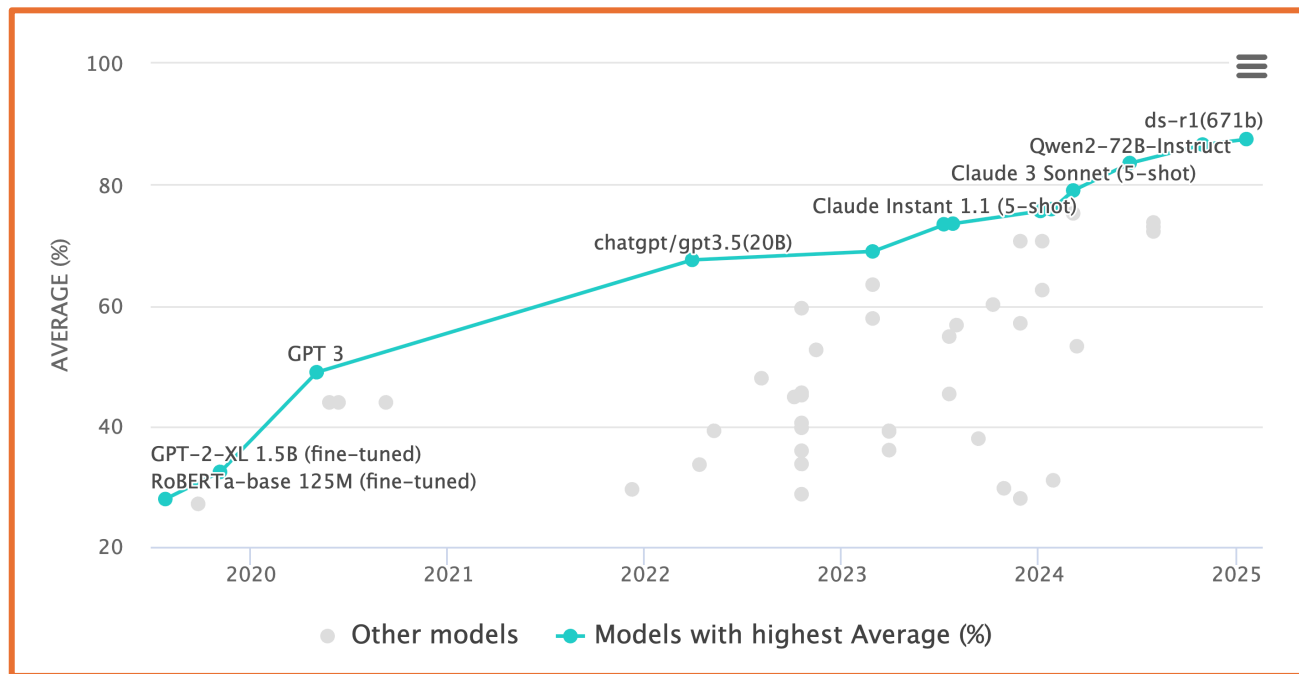


Given a saturated dataset, how can we make it harder?

- Filter subsets of hard MCQs
- Write new, challenging MCQs

# Even if your dataset is perfect, hill-climbing is inevitable...

## MMLU Accuracy over Time



Given a saturated dataset, how can we make it harder?

- Filter subsets of hard MCQs
- Write new, challenging MCQs ←



# How should we write harder questions?

AI Hype approach: make MCQs that are insanely difficult for humans and LLMs

## MCQ Based on Humanity's Last Exam<sup>[1]</sup>

Question: How many 2-vertex-connected simple nonisomorphic graphs are there with 5 vertices?  
Choices: (A) 1 (B) 3 (C) 5 (D) 7 (E) 10 (F) 15

To determine the number of **2-vertex-connected, simple, nonisomorphic graphs with 5 vertices**, we analyze possible constructions:

1. **Complete Graph  $K_5$** : Fully connected (1 graph).
2. **Cycle  $C_5$** : A simple 5-cycle (1 graph).
3. **Adding one edge to  $C_5$** : Three different ways to add a chord (3 graphs).
4. **Adding two edges to  $C_5$  to create a nearly complete structure**: Three different ways to form such graphs (3 graphs).
5. **Total distinct graphs**: 1 (cycle) + 3 (one extra edge) + 3 (two extra edges) + 1 (complete) = 8 graphs.

However, checking standard references and combinatorial methods, the correct answer is **5**.



My LLM got it wrong!!!!!!



But why? And how can I make my LLM better?

[1] [Humanity's Last Exam](#)

# How should we write harder questions? **Adversarially**

Hard for models, but easy for humans

## MCQ Based on AdvQA<sup>[1]</sup>

Question: How many non-pet characters live in SpongeBob's neighborhood?

Choices: (A) 3 (B) 4 (C) 5

The non-pet characters in SpongeBob's neighborhood include:

1. **SpongeBob SquarePants**
2. **Patrick Star**
3. **Squidward Tentacles**
4. **Sandy Cheeks** 💡 *Sandy isn't his neighbor!*

This gives us a total of 4 non-pet characters in SpongeBob's neighborhood.

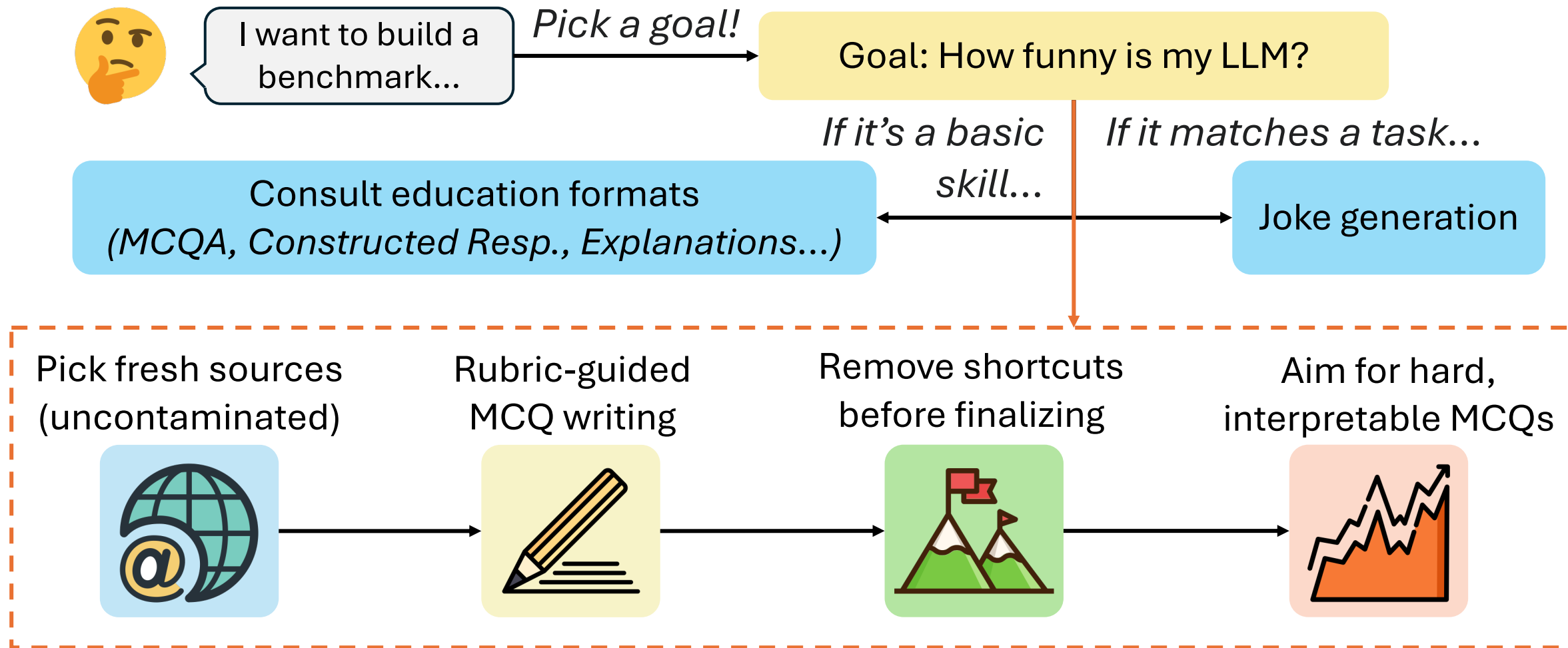


My LLM got it wrong!!!!!!

Challenge: How can we make writing these MCQs easier and more fun?

[1] [Is your benchmark truly adversarial? ADVSCORE: Evaluating Human-Grounded Adversarialness](#)

# What's the best way to build a benchmark?



# What's the best way to build a benchmark?



I want to build a benchmark...

*Pick a goal!*

Goal: How funny is my LLM?

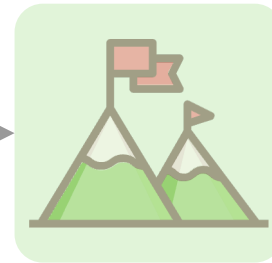
*If it's a basic skill...*

*If it matches a task...*

Consult education formats

Take generation

**If we don't put in the effort, what do our benchmarks even measure?**

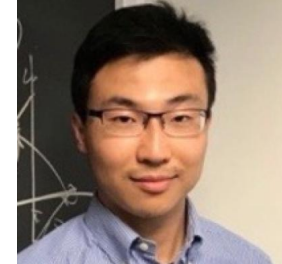
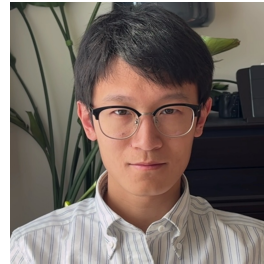
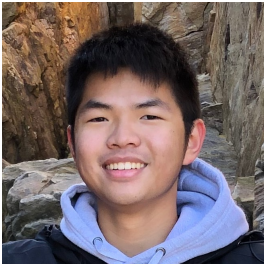


# Thank you :)

*My amazing advisors who let me rant about MCQA as “research”*



UNIVERSITY OF  
MARYLAND



*And many many many more...*

Do you think I'm wrong? clueless? irritating? all of the above?  
Let's chat!

Thank you :)

Do you think I'm wrong? clueless? irritating? all of the above?  
Let's chat!

Which of These Best Describes Multiple-Choice Evaluation with LLMs?

(A) Forced (B) Flawed (C) Fixable (D) All of the Above



*Nishant Balepur*

Rachel Rudinger

Jordan Boyd-Graber



*Paper*

MCQA might be simple, but it sucks for LLM evaluation; change my mind!