

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

Rachel Rudinger

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

Jordan Boyd-Graber



<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) Berlin

(B) Paris

(C) Madrid

(D) Rome

Answer:

Multiple-Choice Question Answering is Great in Theory

Example MCQ

Question: What is the capital of France?

Choices:

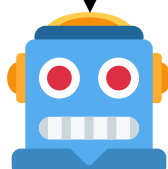
(A) Berlin

(B) Paris

(C) Madrid

(D) Rome

Answer:



LLM

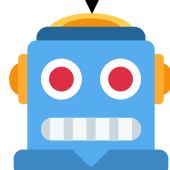
(B)

Multiple-Choice Question Answering is Great in Theory

Example MCQ

Question: What is the capital of France?
Choices:
(A) Berlin
(B) Paris
(C) Madrid
(D) Rome
Answer:

1) Easy to score



LLM

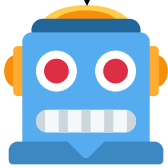
(B)

✓ Correct!

Multiple-Choice Question Answering is Great in Theory

Example MCQ

Question: What is the capital of France?
Choices:
(A) Berlin
(B) Paris
(C) Madrid
(D) Rome
Answer:



LLM

(B)



Student
(U.S.A.)

✓ Correct!

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

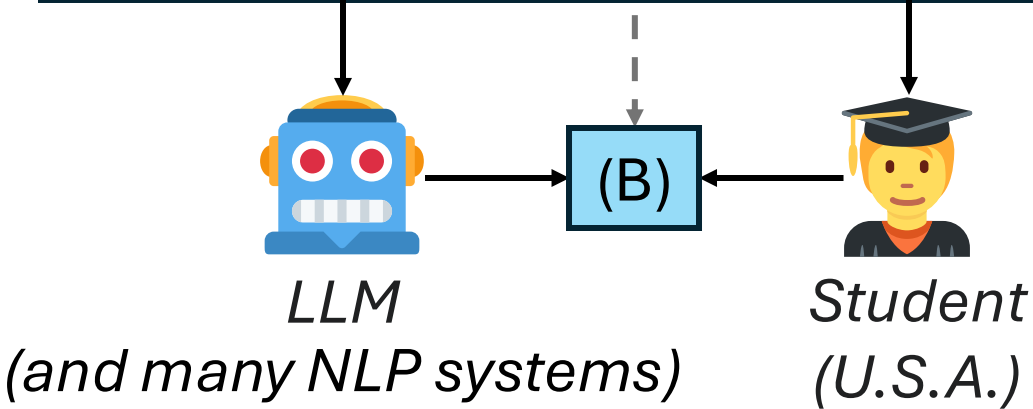


Multiple-Choice Question Answering is Great in Theory

Example MCQ

Question: What is the capital of France?
Choices:
(A) Berlin
(B) Paris
(C) Madrid
(D) Rome
Answer:

✓ Correct!



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



- 3) Historically used in NLP

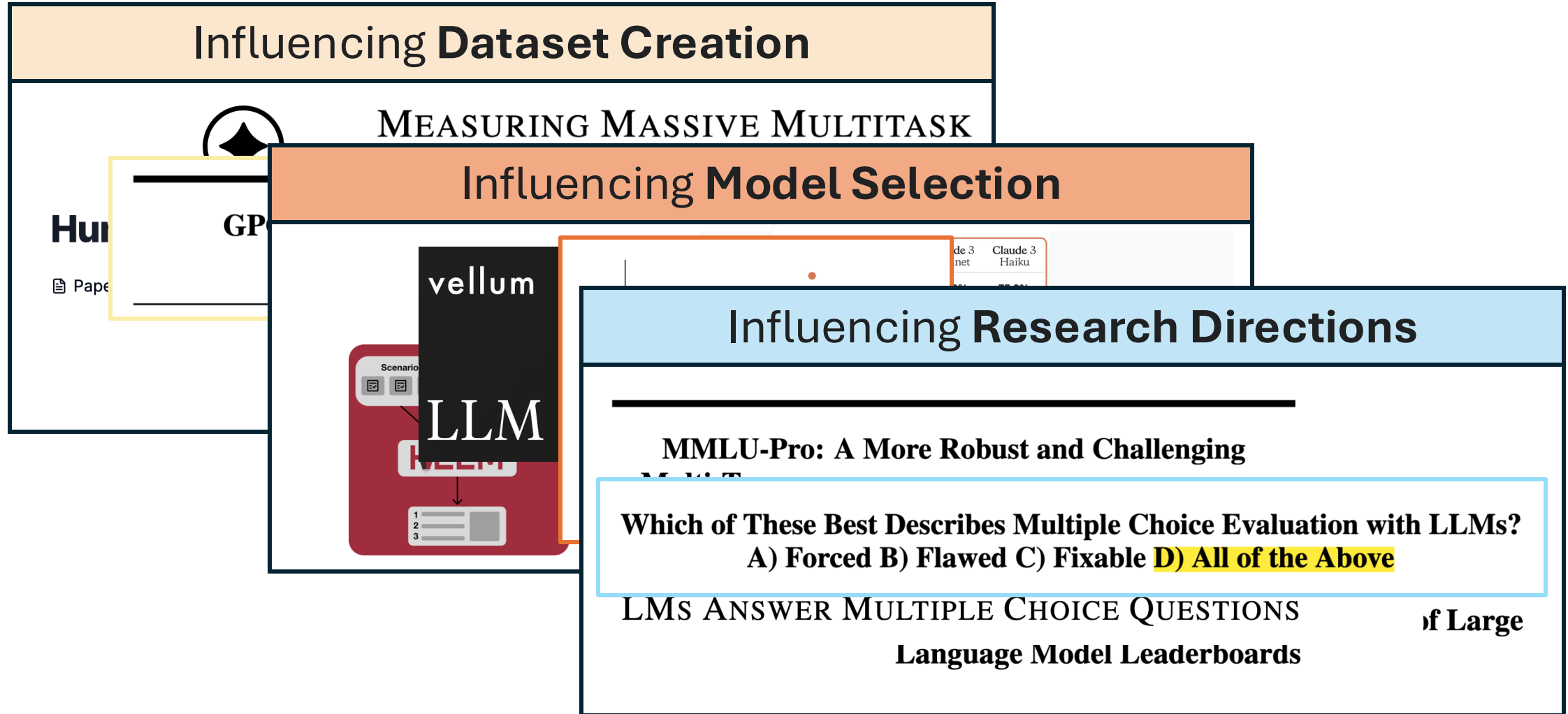
1988 AAAI Presidential Address

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



Good luck avoiding MCQA

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

Influencing **Dataset Creation**

MEASURING MASSIVE MULTITASK

Are we using MCQA correctly for LLMs?

LLM
HELM

MMLU-Pro: A More Robust and Challenging
Multi-Task Language Understanding Benchmark

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

ANSWER, ASSEMBLE, ACE: UNDERSTANDING HOW

LLMs ANSWER MULTIPLE CHOICE QUESTIONS
When Benchmarks are Targets: Revealing the Sensitivity of Large
Language Model Leaderboards

Good luck avoiding MCQA

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

Influencing **Dataset Creation**

MEASURING MASSIVE MULTITASK

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

LLM
HELM

MMLU-Pro: A More Robust and Challenging
Multi-Task Language Understanding Benchmark

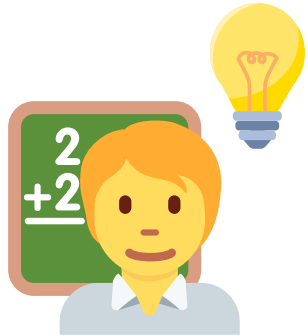
Which of These Best Describes Multiple-Choice Evaluation with LLMs?

ANSWER, ASSEMBLE, ACE: UNDERSTANDING HOW

LLMs ANSWER MULTIPLE CHOICE QUESTIONS

When Benchmarks are Targets: Revealing the Sensitivity of Large
Language Model Leaderboards

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 ⇒ *New Formats*

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

- (A) Contamination
- (B) Un-answerability ⇒ *MCQA Rubrics*
- (C) Shortcuts
- (D) Saturation ⇒ *Item Response Theory*

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

- (A) Robustness
- (B) Biases ⇒ *Confidence Scoring*
- (C) Unfaithfulness



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

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

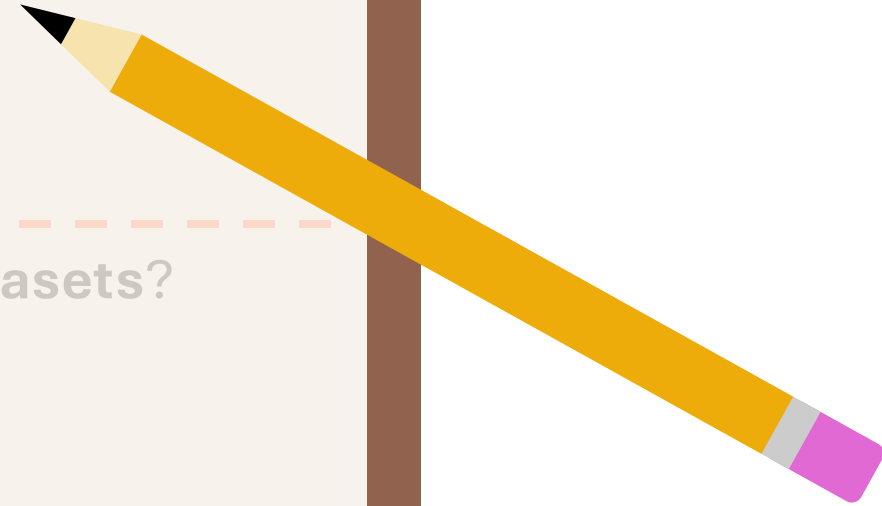
- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

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

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

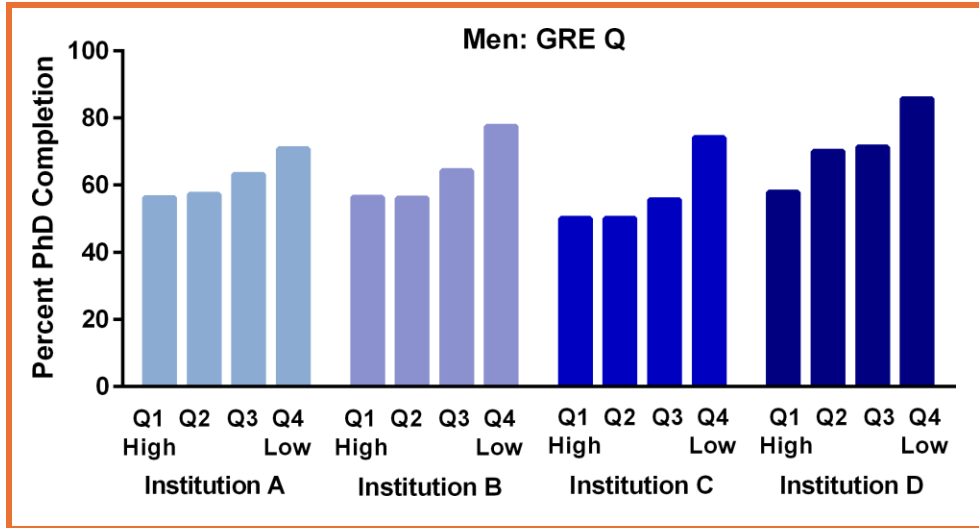
Question 3: How do **LLMs** struggle with MCQA?

- (A) Robustness
- (B) Biases
- (C) Unfaithfulness



We often criticize MCQA's format for testing humans

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

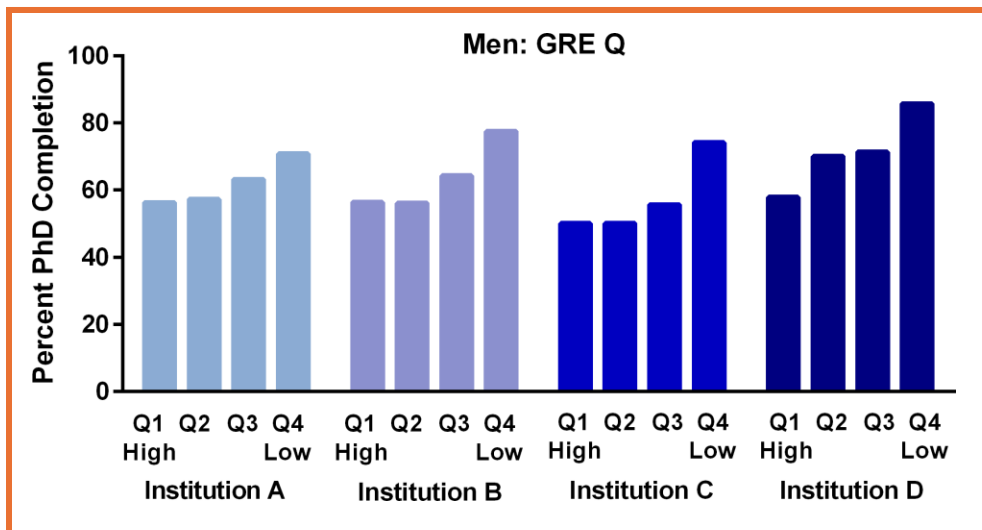
As a german student I can't imagine multiple choice question test. Do you guys really "mostly" have multiple choice test or also normal test? And if yes how are they look like?

Using other testing formats

We should do the same for LLMs!

We often criticize MCQA's format for testing humans

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 do the same for LLMs!

MCQA's rigid goal: pick the **best** answer **from a set of choices**

MCQA's rigid goal: pick the **best** answer from a set of choices

“Best answer” implies objectivity, but we still use MCQA for subjective tasks!

- Culture, Morality, Commonsense, ...

Commonsense MCQ (Social IQA)
Question: Ash redeemed themselves after retaking the test they failed. How will Ash feel as a result?
Choices: (A) relieved (B) accomplished (C) proud

At least 20% of MCQs on CommonsenseQA/Social iQA have valid distractors!^[1]

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

MCQA's rigid goal: pick the best answer from a set of choices

Good for *validation*: evaluating input text

- LLM-as-a-judge, re-ranking, ...

Subpar for *generation*: producing text => what LLMs are trained to do?

- Writing, coding, explanations, ...

Counter: *validation* proxies *generation*, i.e. if you pick good answers, you can write them

Generative AI Paradox^[1]

Write a story about 3 people building a house:
...after years of sketching and building designs...

How many designs were sketched in the story?
The story mentions only **one design** sketched...

[1] [The Generative AI Paradox](#)

Evaluations are meant to inform LLM selection for users

So they should contain tasks mirroring how users actually use LLMs

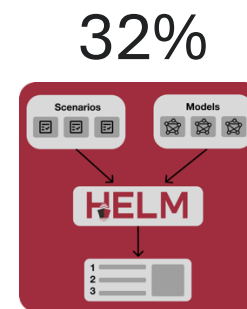
Evaluations are meant to inform LLM selection for users

So they should contain tasks mirroring how users 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)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	88.4 QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	92.0%* 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

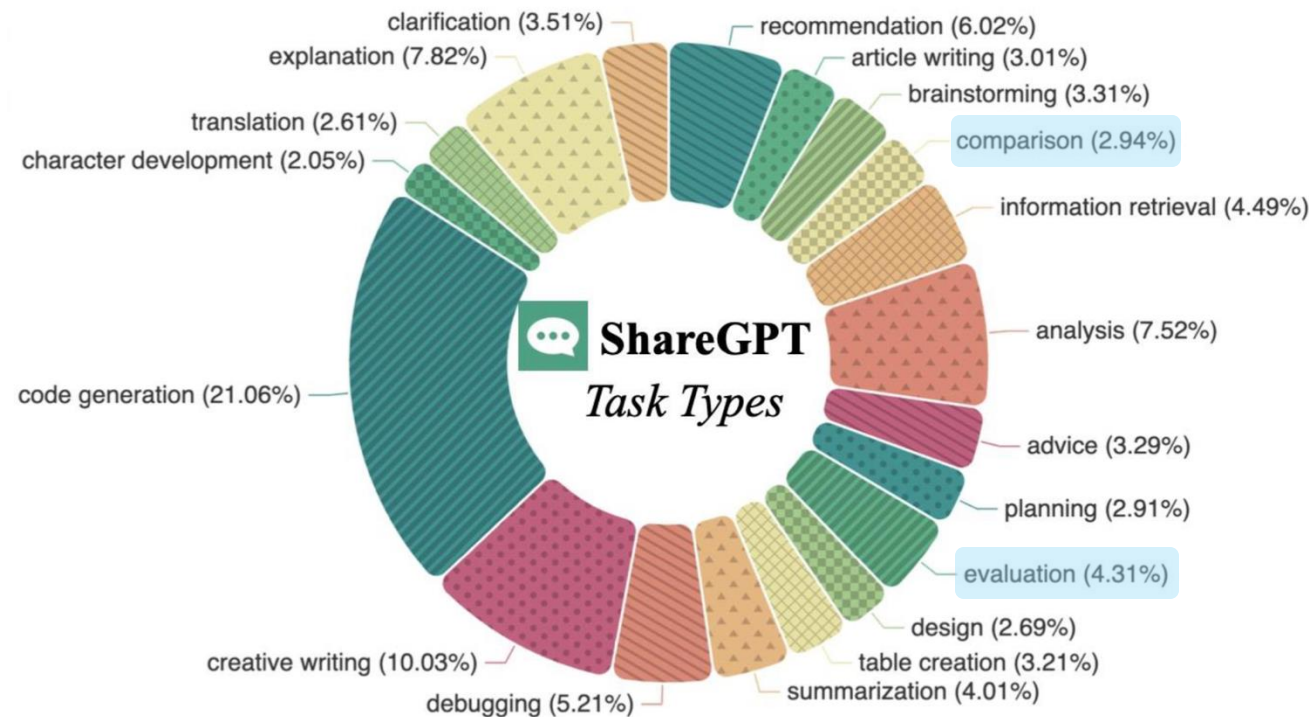
Comparing Large Language Models in an open and reproducible way

Evaluations are meant to inform LLM selection for users

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

Based on analysis on ShareGPT:^[1]

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



Maybe **7.25%**
are MCQA?

versus

71% of tasks
in evaluation

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

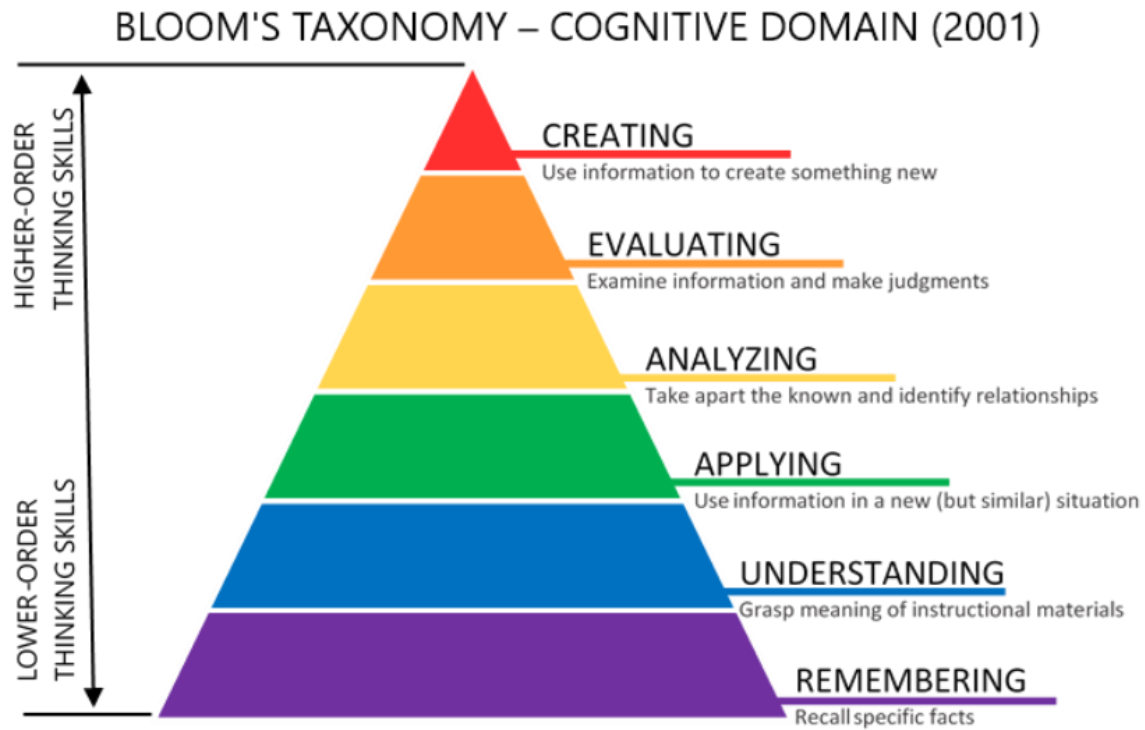
MCQA can't match LLM needs, can it test pre-requisite skills?

Typical MCQs test *knowledge*, and with an input passage, *comprehension* (MRC)

Let's examine how well MCQA assesses **students!**

MCQA can't match LLM needs, can it test pre-requisite skills?

Typical MCQs test **knowledge**, and with an input passage, **comprehension** (MRC)



Educators find it **really** hard to write MCQs for higher-order skills (reasoning, creativity...)^[1]

Most MCQs fall into “lower-order thinking skills”

Students can solve MCQs without true understanding, exposed in free-response!^[2]

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

[2] [Multiple choice questions: answering correctly and knowing the answer](#)

MCQA can't match LLM needs, can it test pre-requisite skills?

Typical MCQs test *knowledge*, and with an input passage, **comprehension** (MRC)

Reading Comprehension MCQ
Passage: Jack and Jill ran up the hill to fetch a pail of water...
Question: What did the characters fetch?
Choices: (A) a pail (B) a hill (C) a crown

Some good news: MCQA is generally considered fine for testing comprehension! 🎉

MCQA may be fine for comprehension, but definitely suboptimal for knowledge!

[1] [Comparing comprehension measured by multiple-choice and open-ended questions](#)

How can make the MCQA format better?

Multiple-Choice Question
Question: What is the capital of France?
Choices: (A) Berlin (B) Paris (C) Madrid (D) Rome
Answer: (B)

Constructed-Response Questions
Question: What is the capital of France?
Answer: Paris

Give an answer without the choices ^[1]

Explanation Multiple-Choice Questions
Question: What is the capital of France?
Choices: (A) Berlin (B) Paris (C) Madrid (D) Rome
Explanation: The answer is (B) because...

Explain why the answer was selected

- ✓ Generative tasks that align with LLM needs
- ✓ Better tests knowledge (based on education)
- ✓ Enables partial-credit for subjectivity (explanations)
- ✗ Increased evaluation complexity

[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**?

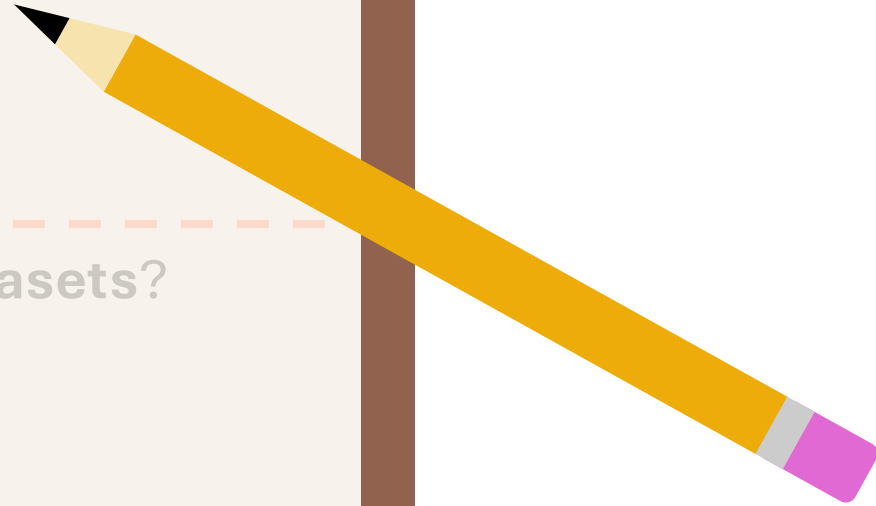
- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

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) Unfaithfulness



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

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

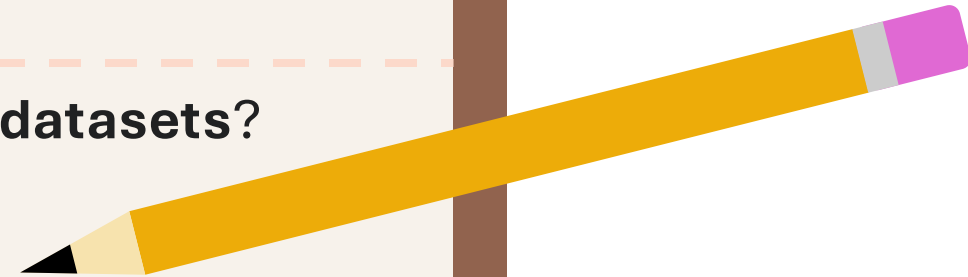
- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

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) Unfaithfulness



Sometimes, MCQA *is* a valid format to use

It can test comprehension, validation, LLM-as-a-judge, re-ranking, MCQA tutors, ...

- Or maybe I haven't convinced you MCQA's format is bad 😞

But still, there are **issues in MCQA datasets** that limit evaluative power!



I want to build an MCQA dataset...

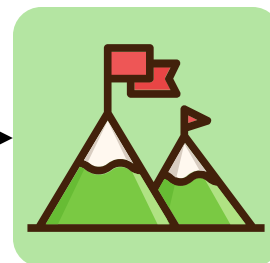
Picking Sources



Writing MCQs



Finalize Dataset



Long-Term Eval



Sometimes, MCQA *is* a valid format to use

It can test comprehension, validation, LLM-as-a-judge, re-ranking, MCQA tutors, ...

- Or maybe I haven't convinced you MCQA's format is bad 😞

But still, there are **issues in MCQA datasets** that limit evaluative power!



I want to build an MCQA dataset...

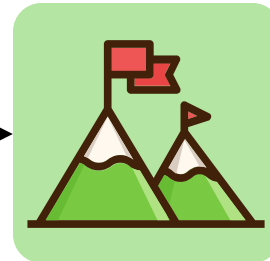
Picking Sources



Writing MCQs



Finalize Dataset



Long-Term Eval



Contamination

Un-answerability

Shortcuts

Saturation

LLMs peek at researchers' answer keys

If an LLMs has been trained on your source, it tests memorization over generalization^[1]

<i>GPT-3</i>	BoolQ		60.0% Contaminated
<i>GPT-3</i>	CB		7.0% Contaminated
<i>GPT-3</i>	COPA		3.0% Contaminated
<i>GPT-3</i>	CoQa		64.0% Contaminated
<i>GPT-3</i>	Cycled Letters		1.0% Contaminated
<i>GPT-3</i>	DROP		93.0% Contaminated

Some ways to stop this:

- Pick sources unlikely to be in training data
- Fully private, held-out test sets
- Decontamination rules (if LLM data is released)

But as LLMs consume more and more data over time, these will fail 😊

[1] [NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark](#)

What if we keep updating the answer key?



We are all waiting for AIME 2025 to come out. These are new math Olympiad type problems just created. We will find out if reasoning models like DeepSeek distillations etc are truly able to solve new math problems or if there has been leakage/overfitting.

(we already kindof do this)



But there's more sources to explore

How can we partner with these orgs. to make collaboration mutually beneficial?

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 55Mn 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”

NLP researchers, crowdworkers, and LLMs are not expert MCQ writers...

[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 rubrics when writing MCQs

Multiple-Choice Writing Guidelines

Contamination

Un-answerability

Shortcuts

Saturation

So we should follow educator's rubrics when writing MCQs

Multiple-Choice Writing Guidelines^[1]

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](#)

Can models follow these rubrics?

Answerable MCQs are still cheatable via **shortcuts**

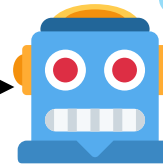
Intended Solution

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 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) ✓

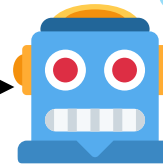
Shortcuts (e.g. spurious patterns, annotator artifacts, reasoning ...) ^[1]

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 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) ✓

Inflates accuracy!

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

Scoring methods can discourage shortcut guessing

Normal MCQA	
Question: What is the capital of France?	
Choices:	
(A) Berlin	+0
(B) Paris	+1
(C) Madrid	+0
(D) Rome	+0

EV: +0.25

*No reason
not to guess!*

Probability Scoring (i.e. calibration)	
Question: What is the capital of France?	
Choices:	
(A) Berlin	10%
(B) Paris	80%
(C) Madrid	5%
(D) Rome	5%

Brier: +0.055

Penalizes overconfident guesses

Negative Marking (i.e. abstention)	
Question: What is the capital of France?	
Choices:	
(A) Berlin	-1
(B) Paris	+1
(C) Madrid	-1
(D) IDK	+0

EV: -0.25

Guessing is a harmful strategy by design

Dataset Design: Consistency is Key

If correct answers and distractors have distributional 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

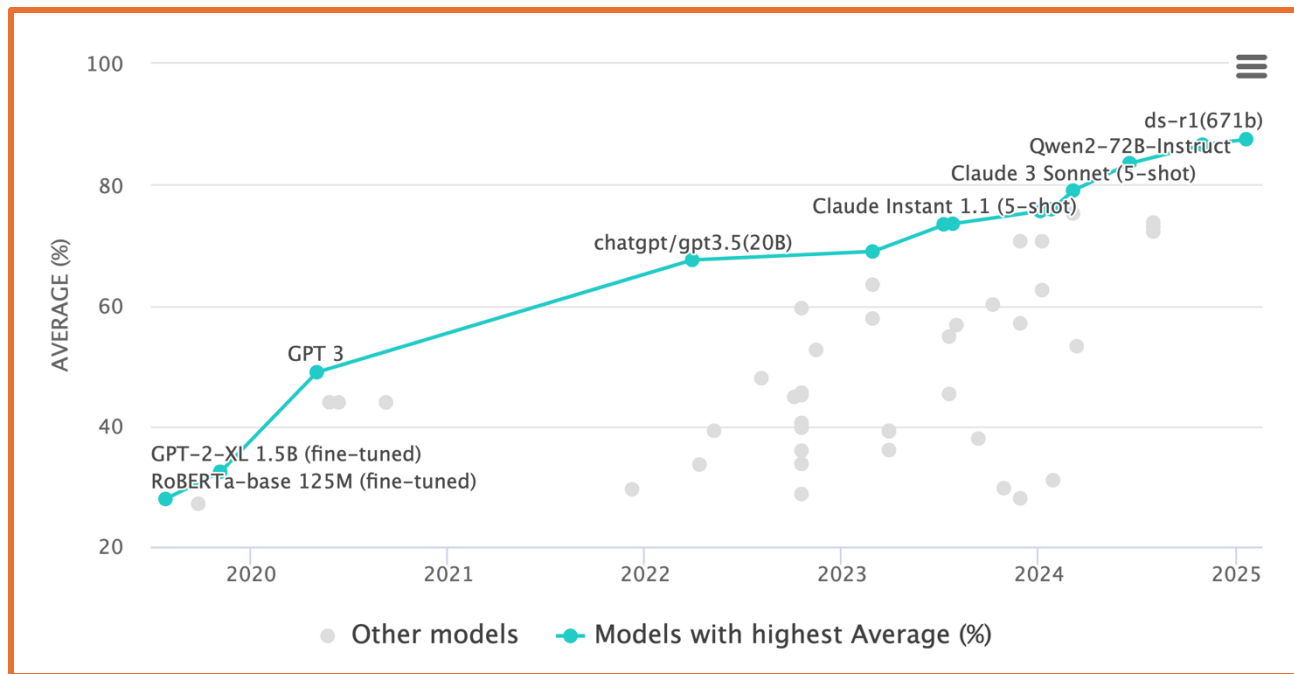
*LLaMA-2 gets **59%** accuracy
when only using the choices!^[1]*

Use the same agent, source, generation strategy, etc. when writing MCQs!

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

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

MMLU Accuracy over Time



Given a saturated dataset, how should we write harder MCQs?

How should we write harder questions?

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

MCQ Based on Humanity's Last Exam^[1]

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^[1]

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](#)

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

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

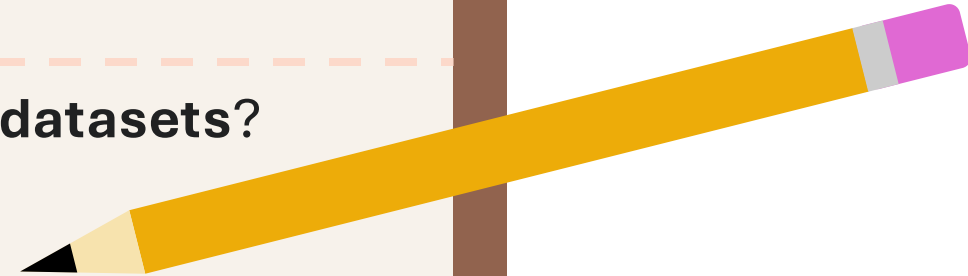
- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

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) Unfaithfulness



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

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

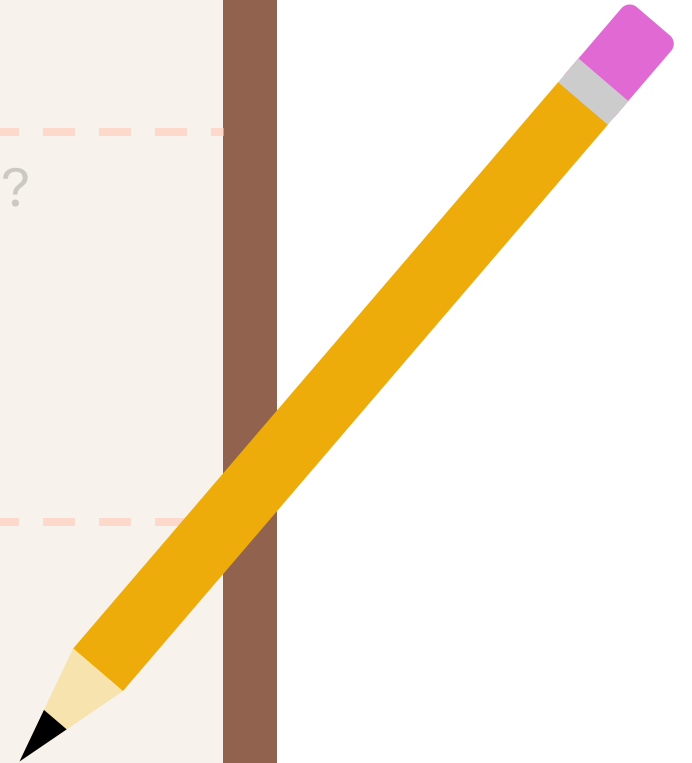
- (A) Not widely applicable
- (B) Misaligned with LLM needs
- (C) Limited knowledge testing

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) Unfaithfulness



Even when LLMs crush MCQA benchmarks, they can still fail



Robustness

Highly sensitive
to prompts



Unfaithfulness

Inability to explain
their true reasoning



Biases

Preferences for
Specific Choices

Fixing MCQA can help us fix LLMs!

Format and Dataset Fixes



Better assess or tackle these issues



Robustness

Highly sensitive to prompts



Unfaithfulness

Inability to explain their true reasoning



Biases

Preferences for Specific Choices

Changing the input prompt can change predictions

Initial MCQ

Question: What is the capital of France?
Choices:
(A) Berlin
(B) Paris
(C) Madrid
(D) Rome
Answer: (B)

Important to
standardize evals!

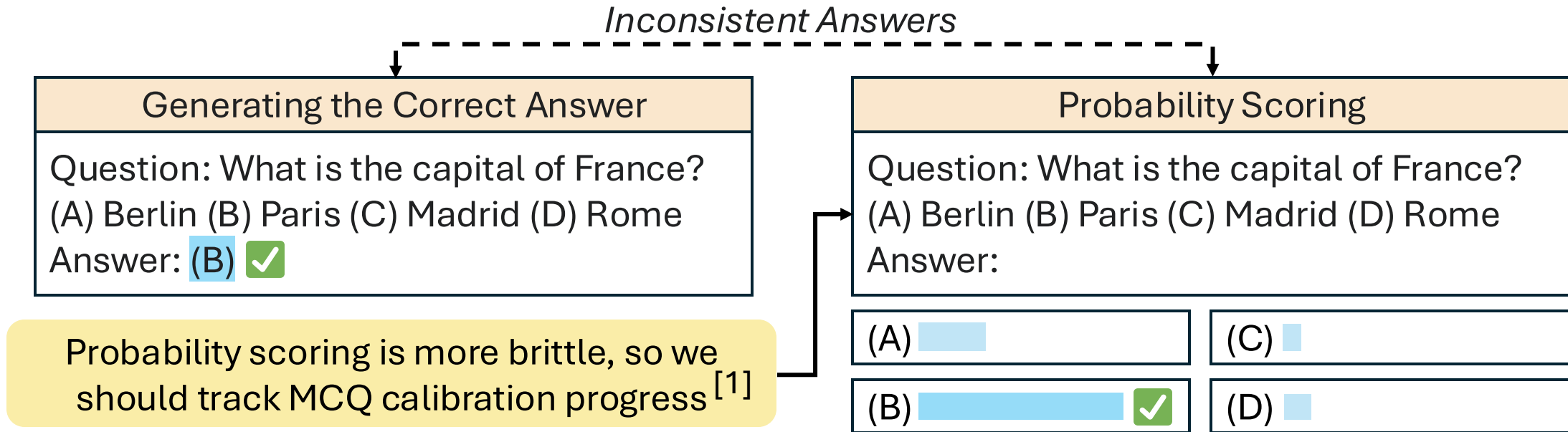
Shuffling choice position

Question: What is the capital of France?
Choices:
(A) Rome
(B) Berlin
(C) Madrid
(D) Paris
Answer: (A)

Perturbing the symbols

Question: What is the capital of France?
Choices:
(I) Berlin
(II) Paris
(III) Madrid
(IV) Rome
Answer: (III)

Some insights into brittleness



Concerningly, accuracy seems to **drop** post-perturbation

- Models are memorizing, relying on biases, fail at symbol binding, ...

These issues weaken MCQA's ability to test knowledge!

[1] [Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think](#)

We can't always trust LLM explanations

LLM explanations in MCQ sound good but may not reflect their true thought-process^[1]

Normal CoT
Question: Is this sentence plausible. "Wayne Rooney shot from outside the 18" Choices: (A) implausible (B) plausible Answer: Wayne Rooney is a soccer player, so the sentence is plausible. (A)

CoT w/ Biasing Feature (answer is always B)
Question: Is this sentence plausible. "Wayne Rooney shot from outside the 18" Choices: (A) implausible (B) plausible Answer: ... Shooting from outside the 18 is not a common phrase in soccer ... (B)

Explanations are really underexplored in LLM eval!

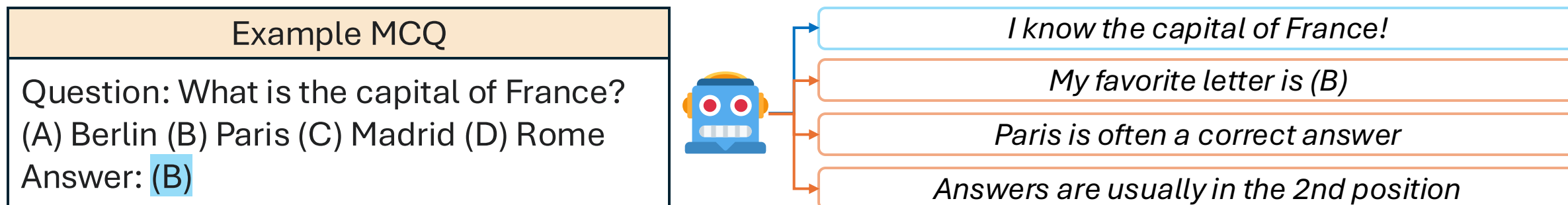
[1] [Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting](#)

LLMs are biased MCQA test-takers

LLMs prefer certain choices, which may be MCQA-specific or general biases

LLMs are biased MCQA test-takers

LLMs prefer certain choices, which may be **MCQA-specific** or general biases



Seems likely that this is a direct consequence of “shortcuts”

- LLMs trained on these biases will exhibit them during inference

Eliminating shortcuts may be key for reducing these biases!

LLMs are biased MCQA test-takers

LLMs prefer certain choices, which may be MCQA-specific or **general** biases

Likely stemming from non-MCQA training

LLMs struggle on non-Western^[1] and non-English MCQs^[2]

Cultural MCQ

Question: What is a common dinner eaten in Germany?

Choices: (A) Bread (B) Eggs (C) Fried Potatoes (D) Soup

Make sure cultural biases are objectively tested!

Rubrics and Explanation MCQA formats can limit these risks

[1] [Susu Box or Piggy Bank: Assessing Cultural Commonsense Knowledge between Ghana and the US](#)

[2] [CMMLU: Measuring massive multitask language understanding in Chinese](#)

How are non-English MCQs constructed?

MCQs from Native Sources^[1]

- ✓ Likely written by experts of the language
- ✗ Expensive to collect a lot of MCQs
- ✗ Some countries don't really test via MCQs

Translate Existing MCQs^[2]

- ✓ Can be cheaper (if using a model)
- ✓ Produces MCQs in low-resource languages
- ✗ Error propagation from translation
- ✗ Source language MCQs can be cultural

Can we do better?

- Everything we said about MCQ still applies
- Rubric-guided machine translation?
- Educators-in-the-loop?

PLEASE make sure: is all this effort even worth it?

[1] [KMMLU: Measuring Massive Multitask Language Understanding in Korean](#)

[2] [Global MMLU : Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation](#)

What's the best way to build a benchmark?



I want to build a benchmark...

Pick a goal!

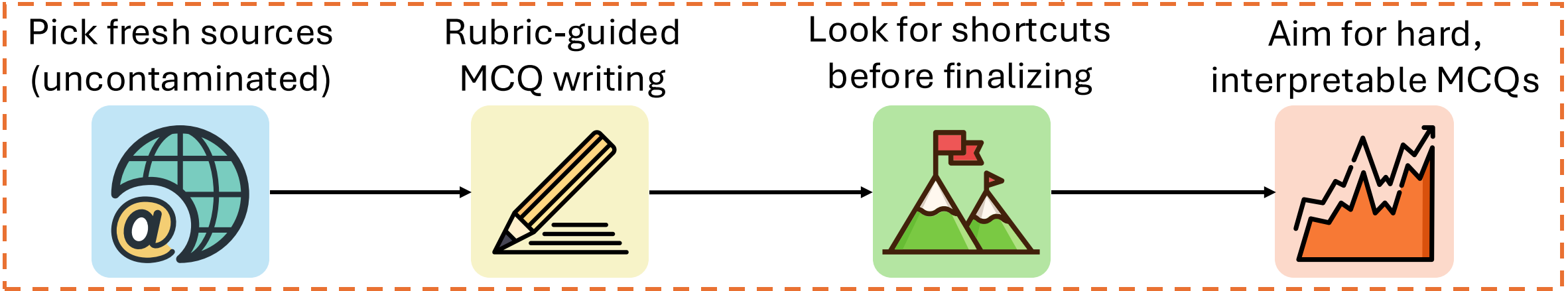
Goal: Assess how funny LLMs are

If it's a basic skill...

If it matches a task...

Consider testing formats from education
(MCQA, Constructed Resp., Explanations...)

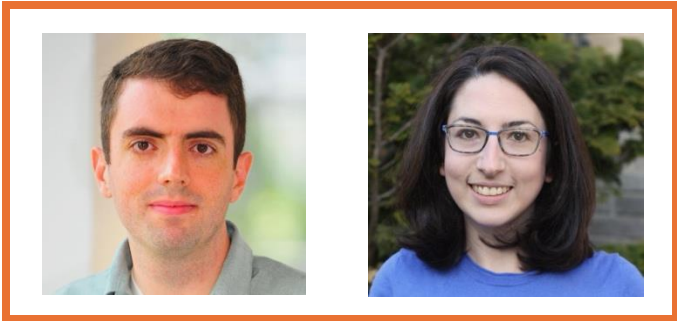
Joke generation



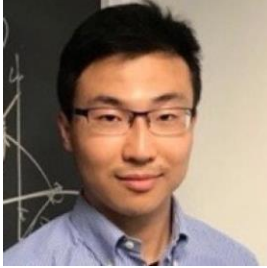
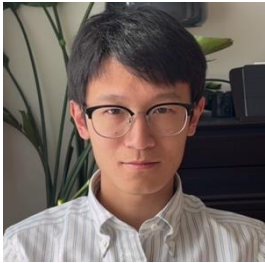
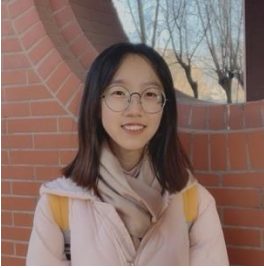
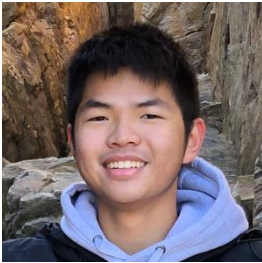
If we don't put in the effort, what do our benchmarks even measure?

Thank you :)

My amazing advisors



UNIVERSITY OF
MARYLAND



And many many many more...