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!

Q1. Why does MCQA's **format** suck?

(A) It lacks applicability

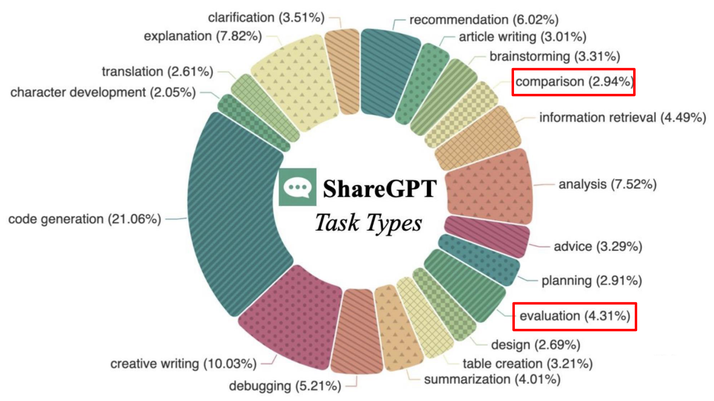
Question: What is the capital of France?
(A) Berlin (B) Paris (C) Madrid (D) Rome
Answer:

Goal: Pick the **best** from the **choices**

Bad for subjectivity

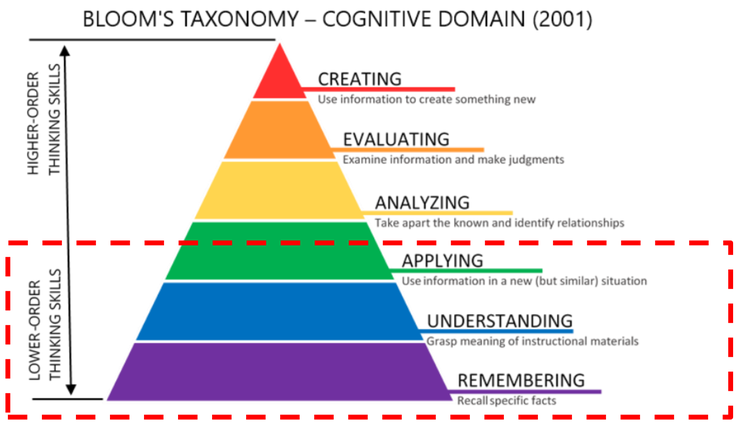
Bad for generative tasks

(B) It's not how we use LLMs



<7.25% of ChatGPT queries (Ouyang 2023)

(C) It fails to test knowledge



What MCQA tests

Q2. What formats would be **less suck-y**?

(A) Constructed Response Questions

Question: What is the capital of France?
Answer: Paris

(B) Explanation MCQs

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

✅ *Generative tasks that match LLM needs*

✅ *Better tests knowledge (from education)*

✅ *Partial-credit for subjectivity (expl.)*

❌ *Heightened evaluation complexity*

Q3: Why do MCQA **datasets** suck?

(A) Contamination

CoQa	64.0% Contaminated
BoolQ	60.0% Contaminated
DROP	93.0% Contaminated

sporcle NAQT SM National Academic Quiz Tournaments, LLC SAT®

Live Test Sets?

(B) Un-answerability

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

Multiple-Choice Guidelines

General Item-Writing:
2. Avoid complex MCQ formats (e.g. all of the above)

Distractor Development ⭐

Educators have rubrics!

(C) Shortcuts

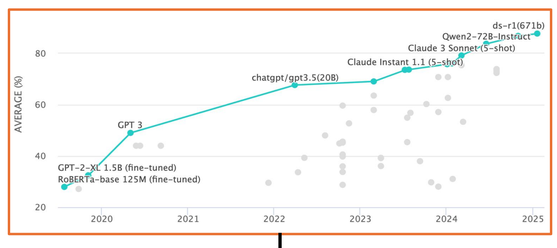
What is the capital of France?
~~(A) Berlin (B) Paris (C) Madrid~~
Answer: (B)

HellaSwag

Question: A woman is outside with a bucket and a dog. She...
🤖 (A) rinses the bucket off...
🤖 (B) uses a hose to keep...
✏️ (C) gets the dog wet...

Write choices consistently!

(D) Saturation



Humanity's Last Exam

How many 2-vertex-connected simple nonisomorphic graphs are there with 5 vertices?

Interpretability PLEASE!

Q4: When do **LLMs** suck on MCQA?

(A) They lack robustness

Generation

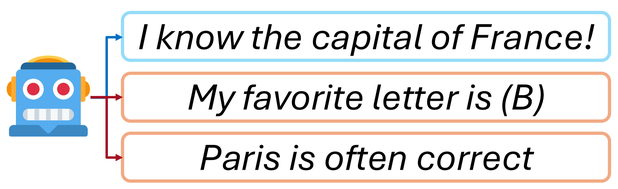
Question: What is the capital of France?
(A) Paris (B) Berlin (C) Rome (D) Madrid
Answer: (D)

Probability Scoring

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

(A) (C)
(B) ✅ (D)

(B) They are biased



Short-cuts?

Favoring Certain Cultures

What is a common dinner in Germany?
(A) Bread (B) Eggs (C) Fried Potatoes

Culture can be subjective, be careful!

(C) They give unfaithful explanations

Chain-of-Thought w/ "Answer 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)

We should assess explanation quality!