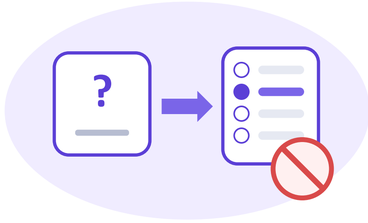
CAN AI ANSWER A QUESTION IT NEVER READ?

Detecting and repairing hidden answer-choice shortcuts in AI-generated exam questions

Anastasia Gaynullina · Laidlaw Scholars Programme · eBRAIN Lab · New York University Abu Dhabi

Research question: *Can a context-blind AI critic detect when the answer choices themselves reveal the correct answer?*



THE PROBLEM
WHEN THE ANSWER IS HIDING IN THE OPTIONS

Q A question is generated...

QUESTION STEM
HIDDEN

- A Only X
- B **Any X involving X, Y, and Z**
- C Only Y
- D Only Z

Could AI pick B without reading the question?

If yes, the benchmark may be testing test-taking shortcuts rather than reasoning.

NORMAL TEST
Q QUESTION → 1 UNDERSTAND → CHOOSE ANSWER

LEAKY TEST
A ANSWER CHOICES → ? SPOT THE CUE → GUESS ANSWER

Existing checks found many lexical and structural cues, but they did not directly test whether the four answer choices alone leaked the answer. A later calibration panel caught some failures, but it discarded most generated items.

! **94.7%**
of generated items were discarded
in the last calibration run

The discard rate motivates an earlier, targeted check that can repair useful questions before they reach final calibration.

THE BLIND CRITIC
Hide the question. Let the AI guess.

GENERATED MCQ
QUESTION STEM
HIDDEN
A.
B.
C.
D.

→ 1 **Critic 1** Llama 3.1
→ 2 **Critic 2** Phi-4-mini
→ 3 **Critic 3** Ministral 3

→ **5 RANDOMIZED TRIALS**
A B C D
B D A C
C A D B
options are reshuffled

→ **CONSENSUS?**
YES → REPAIR
NO → KEEP

1 **3** independent local LLM critics

2 **5** randomized trials per critic

3 **4/5** correct guesses needed per critic

Flagging rule: all three critics must identify the correct answer in at least 4 of 5 randomized trials. Estimated false-positive rate under random guessing: **about 1 in 263,000.**

REPAIR THE ITEM, THEN TEST IT AGAIN

GENERATE → CHECK → FLAGGED?
NO → KEEP
YES → REPAIR

When the critics flag a question, a repair model rewrites the stem or options using their diagnostic feedback. The revised item returns to the same checks. Items that keep failing after the allowed attempts are discarded.

93%
repaired on the first pass
15-item pilot
100%
ultimately resolved

WHAT DOES A LEAK LOOK LIKE?

1. THE BROADEST ANSWER

- A Only X
- B **Any X involving X, Y, and Z**
- C Only Y

The correct option is broad and inclusive while the others are narrowly worded.

2. THE MOST COMPLETE OPTION

A
B
C
D

The longest, most complete option stands out. Distractors omit or alter one piece.

WHAT DID WE FIND?

~30%
of post-filtered benchmark items remained answerable from options alone

53% answer-choice cues vs **25%** chance

DID THE REPAIR ACTUALLY WORK?

95.8% before repair → **90.9%** after repair

Frontier-model accuracy when guessing from options alone

GPT-5-mini · Gemini · GPT-4o-mini · Claude · Llama · DeepSeek

Options-only accuracy fell by 4.9 percentage points after repair.

A SURPRISING RESULT

The critic did not make the whole benchmark harder.
In the matched-pair experiment, the overall failure rate stayed almost unchanged. The effect appeared mainly among flagged questions.

ALL QUESTIONS	FLAGGED QUESTIONS
13.58% to 13.59%	2.50% to 4.19%
essentially unchanged	targeted effect

Interpretation: the mechanism acts as a validity leakage-control mechanism, not as a general difficulty dial.

THE TAKEAWAY

A good benchmark should require models to read the question, not exploit patterns in the answer choices.

1 **DETECT** Context-blind critics expose answer-choice shortcuts.

2 **REPAIR** Flagged questions are rewritten and checked again.

3 **VALIDATE** Independent models confirm lower options-only accuracy.

NEXT Build a human-validated gold standard for semantic leakage, then stress-test the method.

LIMITATIONS Small critics may have model-specific blind spots. Greater difficulty does not automatically mean better question quality.

ANASTASIA GAYNULLINA
Laidlaw Scholars Programme · eBRAIN Lab · NYU Abu Dhabi

Research focus: AI benchmark validity · LLM evaluation · automated quality control

LAIDLAW SCHOLARS | NYU ABU DHABI eBRAIN LAB