Not Ready for the Classroom

Can AI Replace Teachers? Evidence Says No.

Bruna Cavalcanti Lauro • Igor Swierlikowski • Maksymilian Milcarz

01 The Data is a Mess

AI models don't give clean answers. Instead of simple A/B/C/D outputs, they produced **5+ different formats**: verbose explanations, hedged responses, and even "Not Sure."

This forced some aggressive **data cleaning**, and not all models survived equally:

Data loss per model after cleaning

△ **Model Z lost significantly more data than X and Y. Some subjects saw >10% difference in usable responses.**

If we can't even trust models to **format answers consistently**, how can we trust them to **grade students**?

02 The Accuracy Illusion

At first glance, the models seemed to perform well:

~74%	~76%	~66%
Model X	Model Y	Model Z

But these numbers are a mirage. When we investigated *how* models achieve these scores, we found they don't reflect genuine understanding, they reflect **systematic guessing patterns** that happen to align with the dataset.

The deeper we looked, the worse it got.

High accuracy ≠ understanding.
These scores are inflated by systematic guessing patterns, not genuine comprehension.
When models "prefer" certain answer positions, accuracy becomes a byproduct of bias, not skill.
The following findings reveal what's really happening behind the numbers.

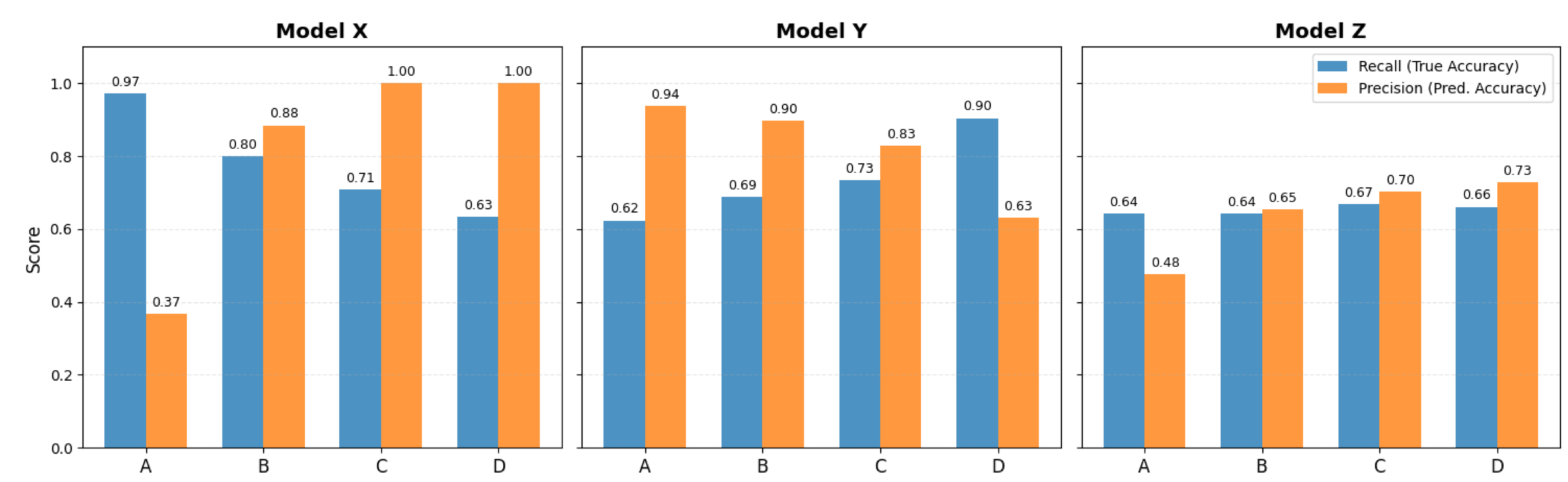
03 Models Have "Favorite Letters" That Fake Their Accuracy

We found that models **default to preferred answer positions** when uncertain. This is the most damaging finding for educational use.

Model	Default	Recall on Def.	Precision on Def.
X	"A"	97%	only 37%
Y	"D"	90%	only 63%
Z	None	~65%	Balanced

△ **Model X picks "A" so often that it gets A-questions right 97% of the time, but when it chooses A, it's wrong 63% of the time. It's spamming, not thinking.**

The gap between **recall** ("how often is the correct answer found?") and **precision** ("when the model picks this letter, is it right?") reveals the illusion. High accuracy on a favored letter is **pattern exploitation**.



Recall vs. Precision per letter (A, B, C, D) for each model

For education, this means: a student whose correct answer is in position A may be graded very differently than one whose answer is in position D, *by the same model, on the same question.*

⇔ Our Verdict

⊗ AI is Not Ready to Replace Teachers

Concern	Risk	Why It Matters
Output quality	HIGH	Non-uniform answers cause data loss and inconsistency
Grading fairness	HIGH	Letter bias means grades depend on answer position
Reliability	HIGH	Models fail when answers are simply rearranged
Cost trade-offs	MED	Cheaper models can't handle complex questions

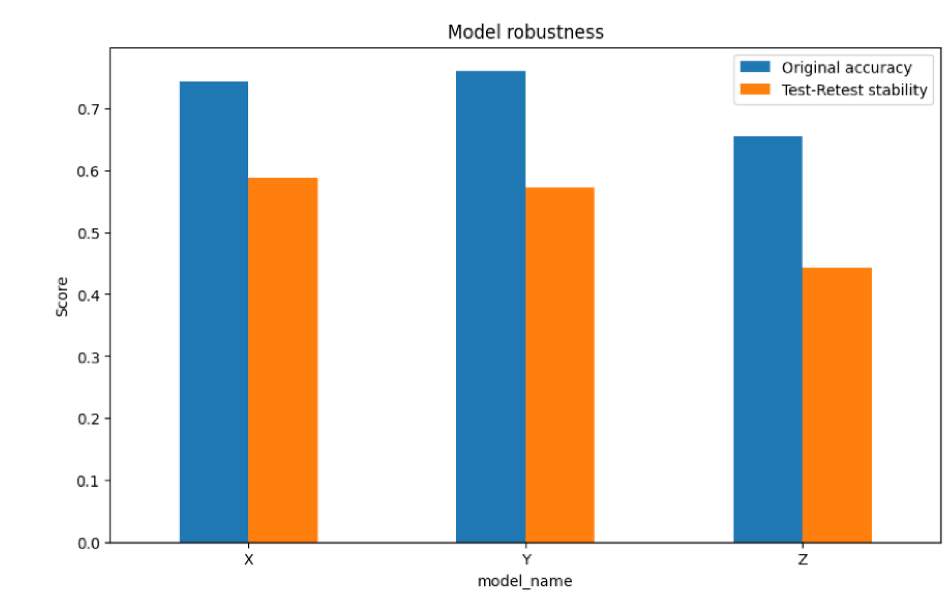
LMs may assist with practice questions or supplementary explanations, but for **anything that affects a student's future**, human judgment must remain in control.

04 The Shuffle Test: Proof They Don't Understand

If a model genuinely understands a question, **shuffling the answer positions should not matter**. We tested this, and every model failed.

Model	Original	After Shuffle
X	~74%	↓↓ Drops
Y	~76%	↓↓ Drops
Z	~66%	~43%

Model Z's post-shuffle stability (~43%) is approximately its accuracy *squared*: the mathematical signature of **random guessing**.

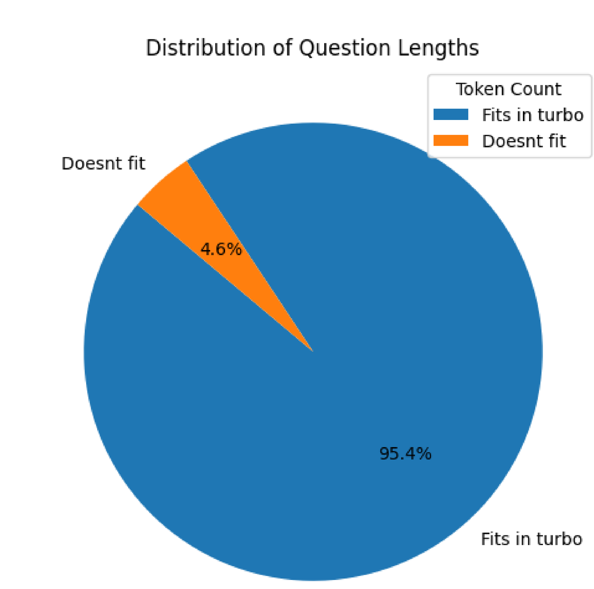


Original accuracy vs. test-retest stability

△ **A tool used to assess students must at minimum be consistent. These models are not.**

05 Turbo vs. Normal: The Cost

The government wants to save money with a faster "turbo" model, but it can only process **300 tokens**.



Questions fitting within the 300-token limit

641 questions (4.6%) exceed this limit, and these tend to be the **harder, more nuanced** subjects where we need the most accuracy.

A hybrid approach (cheap model for simple questions, expensive for complex) is possible but adds complexity.