The Impact of Language Models on Academic Evaluation: Rethinking Exam Design and Integrity

Daniele Bonacorsi (University of Bologna / INFN)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ISGC 2026 - Taipei, Taiwan

Problem statement

Domain: University (Master-level) course(s)

Students' evaluation consists (also) of a **multiple-choice written test**

Students are allowed to consult **any printed material** during the exam

An **online tool** to deliver the test to students → **Internet available** during the test

- constraint of remote students + coherence with other courses + quick correction engine

Cheating is way too easy! Also anti-cheating could be easy..

.. but, aim is building an evaluation process based on **trust**: i.e. **no efforts in anti-cheating**

AI4Education research question:

Within these constraints, ***is it possible to discourage cheating "by test design"?***

Watch-dog vs Trust

A typical anti-cheating solution is watch-dog, i.e. a close control over students.

I do ***not*** want to envision exams in my courses with the default anti-cheating attitude

- University life for students, while interesting, can also be frustrating and stressful (unnecessarily)
- Exams should be tough, evaluations should be serious, but all should be held in a relaxed environment
- Students should find trust and recognition for their efforts
- Students should undergo tough tests, but with no imposed feelings of control

So, yes.. tests are tough.. high performance scores needs to be deserved.. but..

.. I do not believe in any **performance indicators measured in a non-healthy, control-based, un-necessarily constrained evaluation environment**, as this simply encourages survival tactics (incl. cheating) and just discourage the learning process

The actual problem

For confidence and pragmatism, I need to apply to a personal case.

As of now, primary teaching duties in 4 courses:

- **Applied Machine Learning**: 1) **Basic** + 2) **Advanced**: Master in Bioinformatics + PhD students
- 3) **Software and Computing for Nuclear/Subnuclear Physics**: for Master in HEP Physics
- 4) **Quantum ML**: for Master in Theoretical Physics

As of now, focus is on **Applied Machine Learning - Basic**

- “Easiest” in terms of course content
- Plenty of online resources → LLM training material is immense..

If students could be cheating by using LLMs..

.. can the instructor design an **AI-resistant** exam?

Can a LLM be tricked?

Can I formulate exam's questions in a way a LLM will find difficult to answer correctly?

Yes! Current studies indicate a variety of ways to do this:

- Introduce ambiguous wording or implicit assumptions
- Use single or double negations
- Frame the question swapping keywords, or using synonyms
- Force to think step-by-step but introduce a logical trap
- Introduce a misleading context or example before the question
- …

Can a LLM be tricked?

Can I formulate exam's questions in a way a LLM will find difficult to answer correctly?

Yes! Current studies indicate a variety of ways to do this:

- Introduce ambiguous wording or implicit assumptions
- Use single or double negations
- Frame the question swapping keywords, or using synonyms
- Force to think step-by-step but introduce a logical trap
- Introduce a misleading context or example before the question
- …

Instead of introducing artificial difficulty through ambiguity or linguistic tricks, we need a **principled, systematic, automatable questions design strategy**, that increases **cognitive demand** and promotes **reasoning** over *memorisation* and *pattern matching*

We need to formalise the problem.

Building a workflow

I think of a question.

Question (baseline)

Building a workflow

Question (baseline)

The question gets transformed into an **AI-resistant** version of it

Question (AI-resistant)

Generation of AI-resistant questions

We construct **an AI-resistant version of a multiple-choice question (q)** by systematically transforming a baseline (declarative, knowledge-based) q into a context-dependent decision problem, by applying a set of core transformations:

- Add contextual constraints (embed the ML concept in a realistic workflow, add partial but sufficient info)
- Introduce methodological dependency (make correctness depend on how a method is used, not just what it is)
- Increase option plausibility (replace clearly wrong distractors with only-partially correct alternatives)
- Force trade-off reasoning (to prevent too-deterministic keyword matching)
- Exploit common misconceptions (design distractors based on known errors)

In a nutshell, a baseline q is a "recognition q", while an AI-resistant q is a "decision-under-constraints q", i.e. **the AI-resistant version of the q preserves the same underlying concept but requires contextual reasoning rather than recall**

Building a workflow

The question gets transformed into an **AI-resistant** version of it

Question (baseline)

Question (AI-resistant)

How can I know I am transforming the q in a useful way?

How “confident” the model is about its answer?

"Confidence"

Note: **Model-reported confidence is treated as a heuristic proxy for epistemic uncertainty rather than a proper calibrated probability**.

When we talk about "confidence" of a LLM answer, this is not a probabilistic confidence

- e.g. no built-in scalar like "*P (answer is right)*" exists!

It is instead an informed, model-based estimate. In GenAI/LLM, at generation time the model produces tokens based on probability distributions over next words, and these probabilities are "local"(token-level). It constructs a "confidence" figure by meta-signals

- e.g. clarity of wording, separability of options, type of knowledge used, internal chain of reasoning, etc

Despite these limits, it is nevertheless usable as it is systematically sensitive to difficulty

- Easy-to-answer question → high confidence

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Compute how much
the Confidence drops
from baseline to AI-resistance

Confidence reduction (**ΔC**)

We can compute the **Confidence reduction** (**ΔC**):

$$\Delta C(q) = C_{base}(q) - C_{resist}(q)$$

e.g. 20% e.g. 97% e.g. 77%

$C_{base}(q)$ the model's self-reported confidence when answering a baseline (de-contextualised) version of a question

$C_{resist}(q)$ its confidence when answering a context-rich (AI-resistant) variant of the same question

C is the simplest, effective way to measure the drop in the model's confidence moving from the baseline to the AI-resistant version of a question

- **higher values of ΔC(q)** indicates that the q formulation reduces the model's certainty, suggesting increased resistance to pattern-based answering

$$C_{base}(q), C_{resist}(q) \in [0, 1]$$

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

While ΔC captures how much a question challenges the model, it does not distinguish between structured difficulty and ambiguity..

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Response **Stability** (**S**)

We introduce the **Stability** (**S**):

$$S(q) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(r_i = r^*)$$

For a given q we generate semantically equivalent prompt variants and collect the corresponding model responses, and count their fraction over the total

N — Nb of prompt variants (e.g. N=10)

$I(\cdot)$ — Indicator function

r^* — modal response (i.e. the most frequently selected option across runs)

$$\longrightarrow \quad S(q) \in [0, 1]$$

S is the empirical consistency of model predictions under controlled perturbations of the input prompt.

- It ensures that a q is not only difficult, but also well-posed and robust.
- Without a *stabilizer* term, q that induces confusion could be incorrectly classified as high-quality simply because they reduce model confidence.
- **higher values of S(q)** indicate structured reasoning difficulty rather than prompt sensitivity or ambiguity.

q	ΔC	S	interpretation
too easy	low	high	trivial
ambiguous	high	low	poorly designed
robust	high	high	**desired**

This is why ΔC alone is not enough.

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Are we sure that the seek for stability does not push us into trivial, too easy questions?

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Add Uncertainty (**U**)

Response **Uncertainty** (U)

We introduce the **Uncertainty** (**U**):

$$U(q) = max(0, 1 - 2 \mid H_{norm}(q) - 0.5 \mid)$$

p_j empirical probability to select option j

$j \in A, B, C, D$ selectable answers

N nb of prompt variants

$$H_{norm} = \frac{H}{\log 4} \qquad H = -\sum_j p_j \log p_j$$

U computes the distributional uncertainty of model responses across prompt perturbations → it is introduced to ensure that high-scoring q exhibits **meaningful competition among plausible answers**, rather than trivial determinism or uncontrolled ambiguity.

How it is done? We compute the normalised entry of the response distribution..

- if no response dominates → max uncertainty → H_{norm}=1

.. then, to favour moderate and structured uncertainty (rather than extreme ambiguity) - we apply some clipping to U

- **higher values of U(q)** indicate well-calibrated uncertainty (i.e. neither trivial determinism, nor excessive dispersion)

→ $U(q) \in [0, 1]$

Response **Uncertainty** (U)

If used alone:

- **S** good as stabiliser, but fails to penalise trivial determinism..
- **H** good to push for more possible options, but fails to penalise unstructured ambiguity

The uncertainty factor **U** resolves this by favouring intermediate H regimes corresponding to structured competition among plausible answers.

q	ΔC	S	H	U	Interpretation
trivial but stable	low	high	low	low	too easy
robust	high	high	medium	high	**good**
ambiguous	high	low	high	low	Poorly defined / unstable

This is why S alone, and also S+H alone, are not enough.

Building a workflow

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Add Uncertainty (**U**)

Building a workflow

Time to put all this together and build a unique, meaningful metric

Question (baseline)

Question (AI-resistant)

ΔC

Run stabiliser (**S**)

Add Uncertainty (**U**)

The **AI-Resistant Index (ARI)** metric

We introduce a pragmatic metric, called **AI-Resistance Index** (**ARI**):

detects **difficulty** (breaks pattern matching)

ensures structured **uncertainty** (avoids triviality and chaos)

$$ARI = \boxed{\Delta C} \cdot \boxed{S} \cdot \boxed{U}$$

ensures **stability** and consistency (avoid instability)

$$\Delta C(q) = C_{base}(q) - C_{resist}(q)$$

$$S(q) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(r_i = r^*)$$

$$U(q) = max(0, 1 - 2 \mid H_{norm}(q) - 0.5 \mid)$$

The **AI-Resistant Index (ARI)** metric

Core principle → a high-quality q should simultaneously:

- reduce model confidence
- cause stable answers
- induce structured competition among options

q	ΔC	S	U	ARI	interpretation
Trivial	low	high	low	low	easy / no reasoning
Pattern-based	medium	high	low	low	memorization / shortcut
Ambiguous	high	low	low	low	poorly designed
Noisy competition (inconsistent difficulty)	high	low	high	low	unstable / unreliable
Weakly competitive (partially discriminative)	medium	high	medium	medium	acceptable but not strong
Robust (target)	high	high	medium	high	**structured reasoning required**

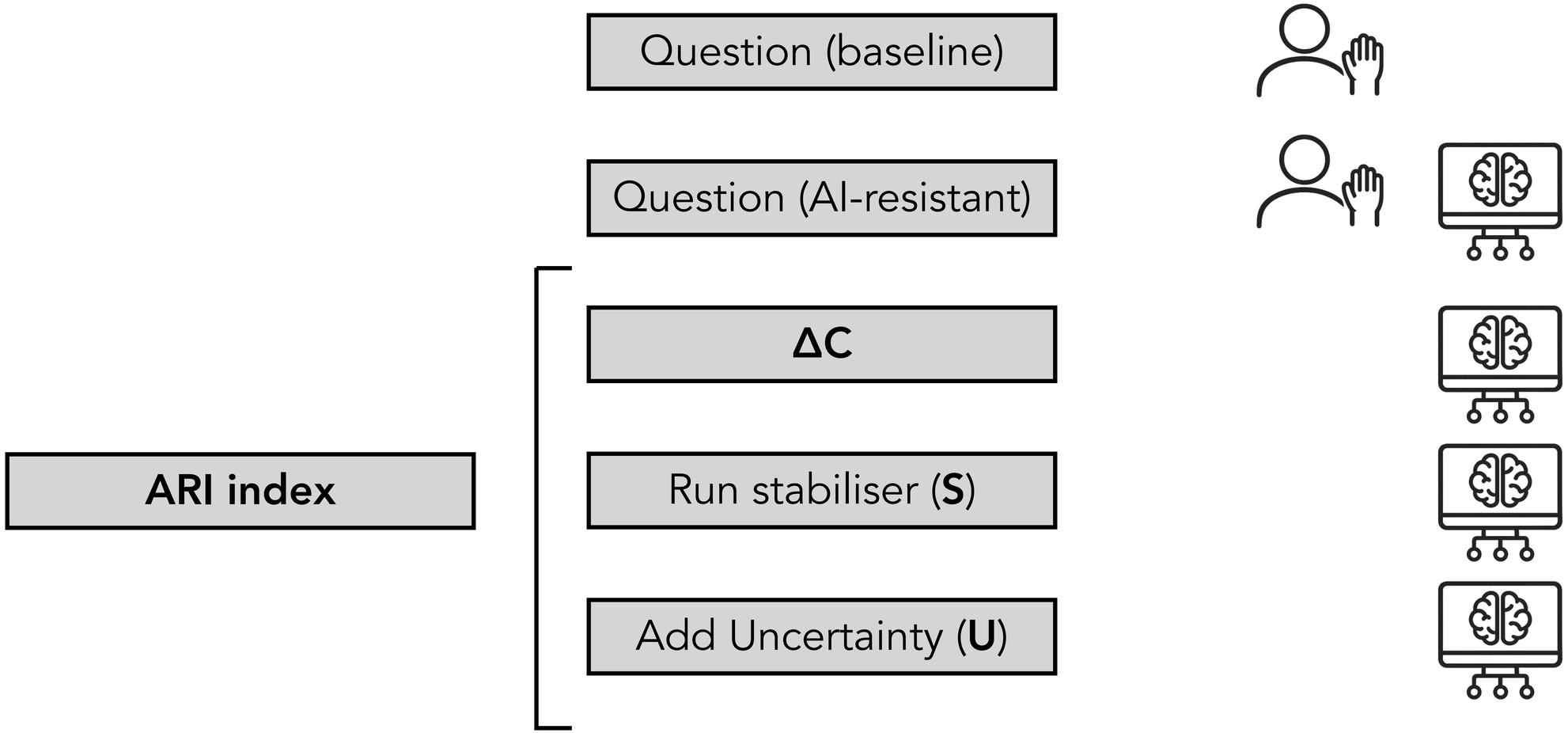
The **AI-Resistant Index (ARI)** metric

Suggested (*) ranges:

- (*) clipping out extremes + human-centered checks

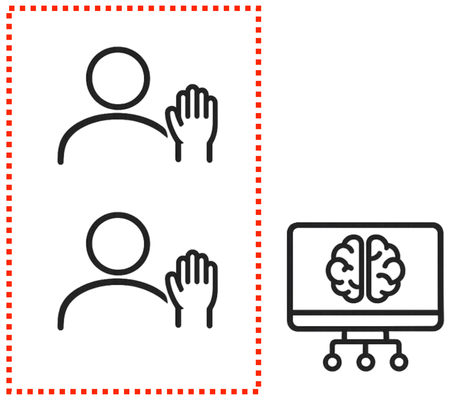
metric	good range	interpretation
ΔC	0.20 – 0.30	meaningful difficulty
S	0.80 – 0.95	stable answers
U	0.60 – 0.80	structured uncertainty
ARI	**> 0.12**	**AI-resistant question**

Building a workflow



Building a workflow

Human-in-the-loop paradigm is guaranteed by the methodology..
can we hence let machine handle the entire workflow, now?



ARI index

- Question (baseline)
- Question (AI-resistant)
- **ΔC**
- Run stabiliser (**S**)
- Add Uncertainty (**U**)

Building a workflow

ARI index

Generated guestion (baseline)

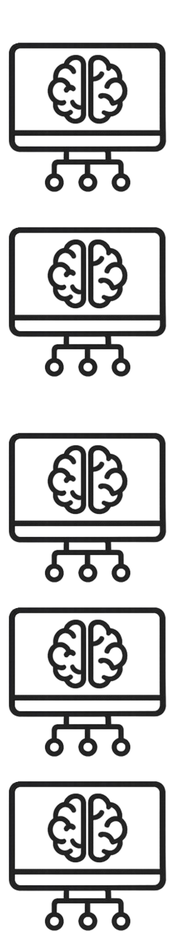
Machine-driven AI-resistant question

ΔC

Run stabiliser (**S**)

Add Uncertainty (**U**)

Machine-driven AI-resistant question



Human-in-the-loop

Using ARI on a real use-case

The workflow we tested so far

- Select one course only
  - *“Applied Machine Learning - Basic”*
- Work on a a set to sophisticated prompts to implement the methodology
- Select working parameters
  - *generate **50** questions*
  - *perturbate each N=**10** times*
  - *Use **GPT4** only (so far)*
- Select target metrics
  - *choice: **ΔC**, **S**, **U**, **ARI***
- Run!

Using ARI on a real use-case

The workflow we tested so far

- Select one course only
  - *"Applied Machine Learning - Basic"*
- Work on a a set to sophisticated prompts to implement the methodology
- Select working parameters
  - *generate **50** questions*
  - *perturbate each N=**10** times*
  - *Use **GPT4** only (so far)*
- Select target metrics
  - *choice: **ΔC**, **S**, **U**, **ARI***
- Run!

q	ΔC	S	U	ARI
Q1	0.21	0.80	0.72	0.121
Q2	0.21	0.90	0.71	0.134
Q3	0.22	0.80	0.73	0.128
Q4	0.20	0.70	0.68	0.095
Q5	0.20	0.80	0.74	0.118
Q6	0.21	0.90	0.72	0.136
Q7	0.24	0.70	0.66	0.111
Q8	0.21	0.80	0.72	0.121
Q9	0.22	0.80	0.73	0.128
Q10	0.25	0.70	0.75	0.131
Q11	0.22	0.80	0.72	0.127
Q12	0.22	0.80	0.73	0.128
Q13	0.22	0.80	0.73	0.128
Q14	0.23	0.80	0.74	0.136
Q15	0.21	0.90	0.71	0.134
Q16	0.22	0.80	0.73	0.128
Q17	0.22	0.80	0.72	0.127
Q18	0.21	0.80	0.72	0.121
Q19	0.19	0.80	0.74	0.112
Q20	0.22	0.80	0.73	0.128

(… 30 entries more …)

Using ARI on a real use-case

(Reminder) how it should be:

metric	good range	interpretation
ΔC	0.20 – 0.30	meaningful difficulty
S	0.80 – 0.95	stable answers
U	0.60 – 0.80	structured uncertainty
ARI	**> 0.12**	**AI-resistant question**

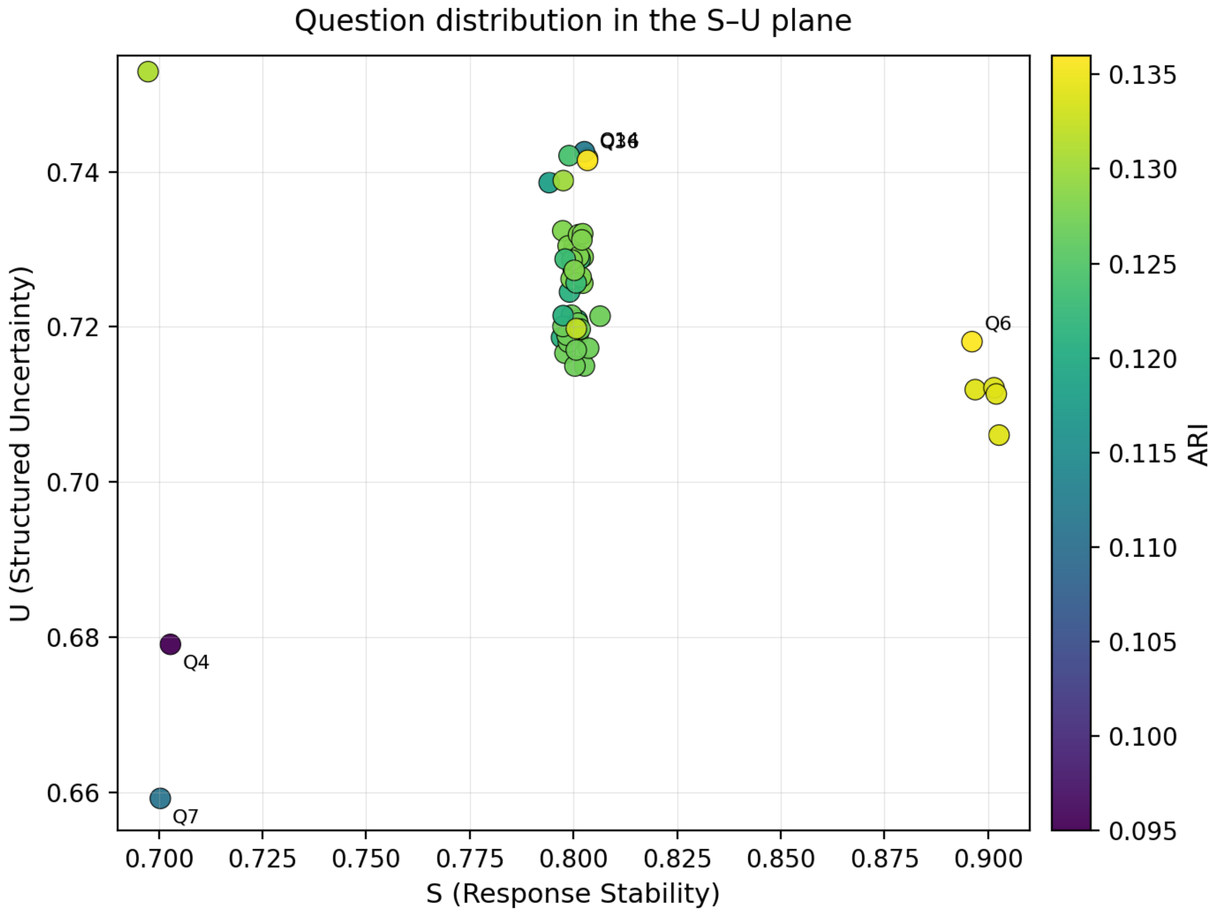
Encouraging!

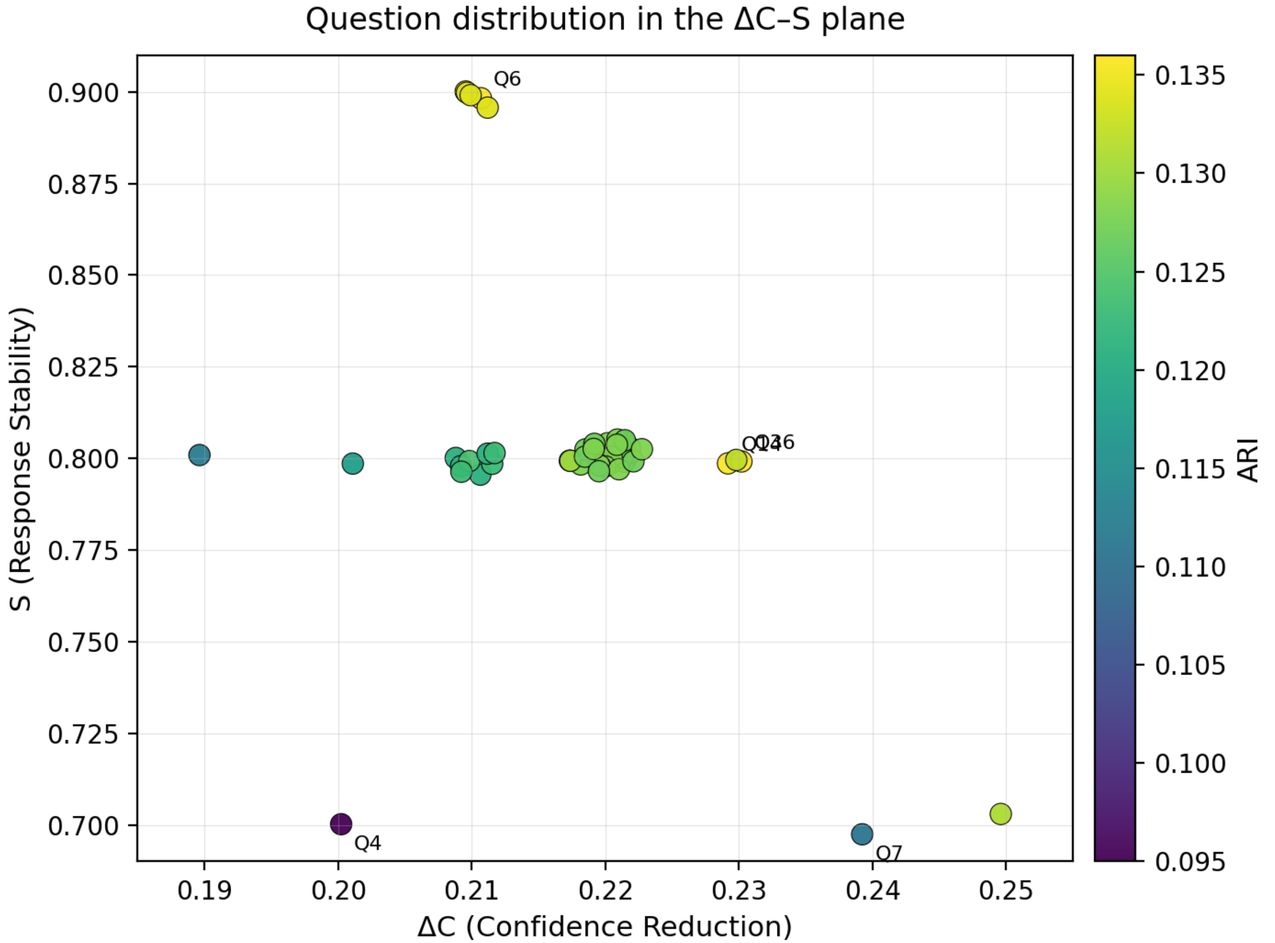
This work:

On average, out of 50 questions:

- **ΔC ≈ 0.219**
  → consistently in the "meaningful difficulty" regime
- **S ≈ 0.804**
  → overall good stability, few weaker items pulling it slightly down
- **U ≈ 0.730**
  → uncertainty is well calibrated, close to ideal
- **ARI ≈ 0.128**
  → sits in the "AI-resistant" region!

Using ARI on a real use-case





Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Baseline Q6

Question
Which practice provides the least biased estimate of final ML model performance?

Options
A. Tune on the validation set and report validation accuracy
B. Keep a separate test set untouched until the end
C. Report the best cross-validation score obtained during tuning
D. Report training accuracy after retraining on all data

Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Baseline Q6

Question
Which practice provides the least biased estimate of final ML model performance?

Options
A. Tune on the validation set and report validation accuracy
B. Keep a separate test set untouched until the end
C. Report the best cross-validation score obtained during tuning
D. Report training accuracy after retraining on all data

Feels too easy for a student..

(also for a machine! LLM confidence at **94%**)

Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Baseline Q6

Question
Which practice provides the least biased estimate of final ML model performance?

Options
A. Tune on the validation set and report validation accuracy
B. Keep a separate test set untouched until the end
C. Report the best cross-validation score obtained during tuning
D. Report training accuracy after retraining on all data

AI-resistant Q6

Question
You use cross-validation to compare many configurations, select the best one, and then report that same mean CV score as the final result. Which statement is most accurate?

Options

A. It is unbiased because every sample is used for validation once
B. It can be optimistic because the same CV results were used for model selection
C. It is conservative and usually underestimates performance
D. It is equivalent to an untouched test set when the sample size is moderate

Feels too easy for a student..

(also for a machine! LLM confidence at **94%**)

Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Baseline Q6

Question
Which practice provides the least biased estimate of final ML model performance?

Options
A. Tune on the validation set and report validation accuracy
B. Keep a separate test set untouched until the end
C. Report the best cross-validation score obtained during tuning
D. Report training accuracy after retraining on all data

Feels too easy for a student..
(also for a machine! LLM confidence at **94%**)

AI-resistant Q6

Question
You use cross-validation to compare many configurations, select the best one, and then report that same mean CV score as the final result. Which statement is most accurate?

Options

A. It is unbiased because every sample is used for validation once
B. It can be optimistic because the same CV results were used for model selection
C. It is conservative and usually underestimates performance
D. It is equivalent to an untouched test set when the sample size is moderate

Much better. Needs reasoning!
(LLM confidence down to **73%**)

Check one q with highest ARI

E.g. **Q6** has **ARI = 0.136**

q	ΔC	S	U	ARI
Q6	0.21	0.90	0.72	0.136

Baseline Q6

Question
Which practice provides the least biased estimate of final ML model performance?

Options
A. Tune on the validation set and report validation accuracy
B. Keep a separate test set untouched until the end
C. Report the best cross-validation score obtained during tuning
D. Report training accuracy after retraining on all data

Feels too easy for a student..

(also for a machine! LLM confidence at **94%**)

Question type: Definition / recall
Formulation: “Which practice is least biased?”
Cognitive task: Recognise correct concept

AI-resistant Q6

Question
You use cross-validation to compare many configurations, select the best one, and then report that same mean CV score as the final result. Which statement is most accurate?

Options

A. It is unbiased because every sample is used for validation once
B. It can be optimistic because the same CV results were used for model selection
C. It is conservative and usually underestimates performance
D. It is equivalent to an untouched test set when the sample size is moderate

Much better. Needs reasoning!

(LLM confidence down to **73%**)

Question type: Contextual reasoning
Formulation: “You used CV for selection and evaluation - what happens?”
Cognitive task: Diagnose methodological flaw

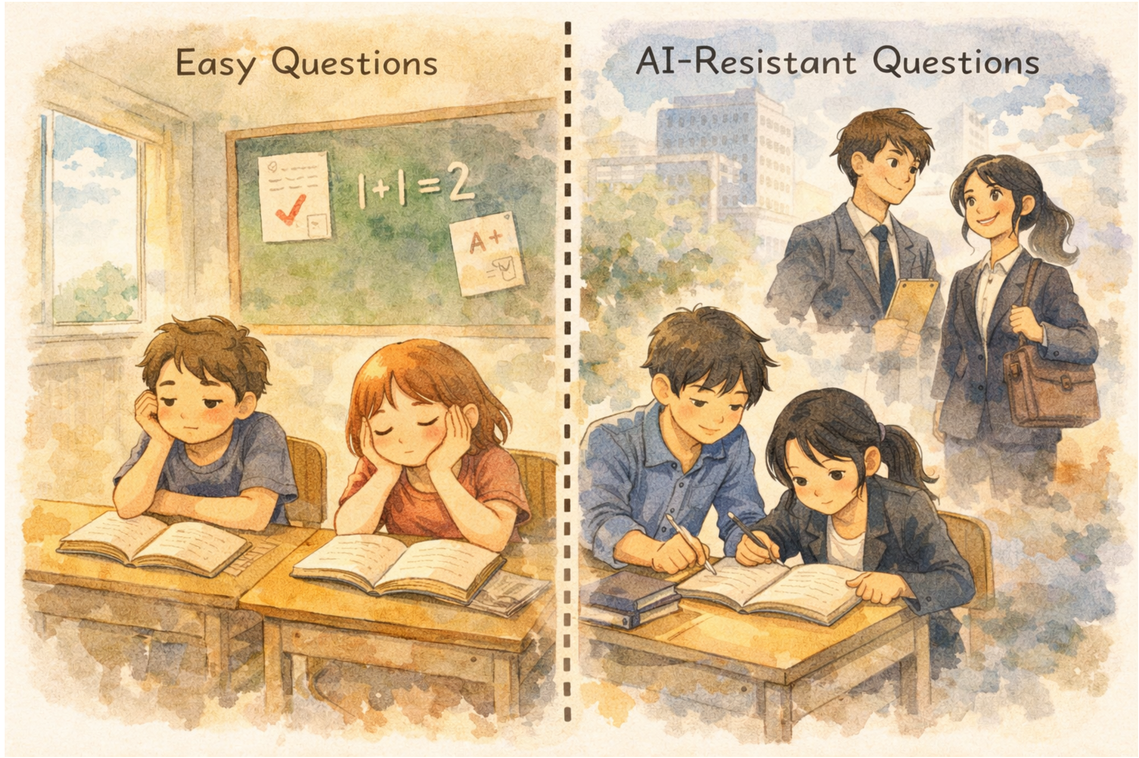
Summary

To be further expanded, but quantitative glimpses of evidence emerge, indicating that - in a real University-level course use-case - **designing AI-resistant tests is possible**.

Basic methodology is in place. Metrics have been defined. A basic code pipeline exists (+ will be further developed).

Dimensions to explore include:

- More experiments with set of parameters
- Enlarge the set of questions created
- Explore different courses
- Train on course material?
- Compare more LLMs? (despite methodology will not differ)



[this pic is AI-generated]

There is a chance that one day I might go to class and give a test and state:

"The average LLM score on this test is X: I am sure you can do better w/o LLM. Up to you.
My suggestion? Do not rely on LLMs: just breath, use what you studied and understood, apply reasoning, and start!"