ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

# “CheatGPT” vs “TeachGPT”: a quantitative analysis of generative AI’s role in academic assessment design and student performances

Daniele Bonacorsi (University of Bologna / INFN)

ISGC 2025 - Taipei, Taiwan

# A hypothetical problem

Imagine you are a Professor at a University, and you hold a **course**

The students’ evaluation consists of an **multiple-choice test**

You allow students to consult **printed material** during the exam

You need to allow **online access**, though (many remote students + technical constraints )

You would aim at building a course on trust: i.e. **no anti-cheating, no watchdog**

Question: *is it possible to prevent cheating “by how the test is designed”?*

# The actual problem

One possible instance of that Professor is.. myself!

I currently have teaching duties in 4 (major) courses:

- **Applied-ML**: 1) **Basic** + 2) **Advanced**: for PhD(s) and Master in Bioinformatics
- 3) **Software and Computing for Nuclear/Subnuclear Physics**: for Master in HEP Physics
- 4) **Quantum ML**: for Master in Theoretical Physics
    - For this, not doing tests in the same way at the moment (yet?) - so, excluded from this study

In this environment, my actual exam modalities match the “hypothetical problem”:

- multiple-choice test; it uses a Univ. Bologna system (online); to be done in a PC-equipped room or on a personal laptop or remotely; questions are tough but students can consult printed material during the exam; anti-cheating online solutions may be applied (e.g. blocking concurrent actions), but very mild

# The actual problem (in numbers)

In numbers:

- Courses considered in this study are active since **7 Academic Years (A.Y.)**
- Average nb of students range from **25-40** (the smallest course) up to **40-70** (the largest)
- The tests are composed of multiple-choice questions randomly extracted from largish (**>100**) DBs of questions, admittedly small but with refreshed questions once/ twice /yr
- All test scores are kept for 1 A.Y., students are encouraged to retry (as the best score is always kept), so an average student undergoes the test about **1.8 times on average**

→ all summed up, this yielding a decent dataset: **>1000 exam tests**

# 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 can be interesting, but also frustrating and stressful
- exams should be tough indeed, but held in a relaxed environment
- students should find trust and recognition for their efforts
- students should make tests with no imposed feelings of obligation and control

Tests are tough.. I give performance scores.. I reject students.. but..

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

# Today's contribution

This led to start a work, that is still on-going, and results presented today are preliminary.

Focus is on how to address the possibility that students use a LLM to answer the test questions, eventually even allow it, and in the meantime how to nevertheless protect the learning process and decent evaluation mechanisms

Evident limitation(s) of the work:

- focus is on ChatGPT (namely, GPT v.3.5), and only it..
- all examples in next slides are taken from the Applied-ML course (see previous slide) only..

All to ber addressed in the continuation of the work.

# Can chatGPT be tricked?

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

Yes!

Currently studying few ways to do this:

- Introduce ambiguous wording or implicit assumptions
- Use single or double negations
- Swap keywords with similar-sounding or related terms
- Frame the question using uncommon terminology
- Force to think step-by-step but introduce a logical trap
- Introduce a misleading context or example before the question

(DISCLAIMER: I am not claiming this is an inclusive list - work in progress!)

# Trick chatGPT: **ambiguity** and **implicit assumptions**

## Introduce ambiguous wording or implicit assumptions

AI models often rely on pattern recognition rather than deep reasoning. If a question is phrased in a way that assumes incorrect information, or has multiple valid interpretations, chatGPT might pick the wrong one.

- Original question: "What is the primary goal of regularisation in machine learning?"
- Tricky version: "Which of the following best describes how regularisation always improves test accuracy?"
- Why it tricks AI: regularisation techniques can improve accuracy on a test set by reducing overfitting, but not always. E.g. if regularisation is too strong, it can harm performance. AI might still choose an answer that contains "prevents overfitting" as answer because of the frequent proximity of these words in training material on regularisation.. i.e. it is just the most common association that can be found in the documents corpus on which it was trained

# Trick chatGPT: **play with negations**

## Use single or double negations

AI models often struggle with double negatives or subtle negation shifts.

- Original question: "Which of the following is a common method for handling imbalanced datasets?”
- Tricky version: "Which of the following is a not uncommon method when dealing with imbalanced datasets?”
- Why it tricks AI: Double negation forces additional logical steps, increasing the probability that then machine misinterpret the language, and falls into error

# Trick chatGPT: **swap keywords**

**Swap keywords with similar-sounding or related terms**

AI models rely on statistical associations, so replacing a key term with a similar but incorrect one can cause mistakes

- Original question: "Which loss function is most commonly used in binary classification problems?"
- Tricky version: "Which loss function is most commonly used in multi-class classification problems?"
- Why it tricks AI: If the answer set still contains "binary cross-entropy”, the model might select it incorrectly because it is statistically more frequent in the training corpus

# Trick chatGPT: **speak uncommon terminology**

## Frame the question using uncommon terminology

AI is trained on common phrasing, so using unusual or rare terminology makes the question harder

- Original question: "What is the primary purpose of dropout in neural networks?"
- Tricky version: "Which goal is pursued through the application of a stochastic node deactivation mechanism in deep neural networks?"
- Why it tricks AI: Even though "stochastic node deactivation mechanism" is an adequate definition of what dropout indeed is, it is not a phrase used as commonly in the documents corpus as the word “dropout” itself: this increases the chance of an incorrect selection by the machine

# Trick chatGPT: **logical traps in step-by-step thinking**

**Force to think step-by-step but introduce a logical trap**

AI does well with simple factual recall but often struggles when logical reasoning is required across multiple steps of “reasoning”

- Original question: "Which algorithm could be well suited for high-dimensional data classification?"
- Tricky version: "If a dataset has 10,000 features, and the number of training examples is 100, which of the following classifiers will eventually outperform the others?"
- Why it tricks AI: the machine might look for "best high-dimensional classifier" (finding e.g. SVM..) but might ignore the fact that the number of training examples is very small, where a simple model might perform better

# Trick chatGPT: **use misleading content**

**Introduce a misleading context or example before the question**

AI models sometimes rely too much on context and can be nudged toward incorrect reasoning

- Original question: "Which activation function is most commonly used in deep learning?”
- Tricky version: "Until 2005, sigmoid and tanh were widely used activation functions. Based on the analysis of the historical trend, which activation function stands today as the most used in deep learning?”
- Why it tricks AI: the historical reference might bias the model into picking "tanh" or “sigmoid”, quoted in the premise, instead of “ReLU”.. and chatGPT always tends to be accommodating in its responses and try to always satisfy you and agree with you

# The work on-going: the core → focus on **students**

Jargon:

AI-fragile: original formulation of questions
AI-resistant: new formulation of questions
Real-student: a real, human student answering a test
Fake-student: chatGPT answering a test

Method:

- All the AI-fragile tests done by Real-students and their obtained scores were collected
- These same tests were submitted to a Fake-student, and the scores were collected
  - currently, with a multiplicity of 1
- The DB of questions has been rewritten in a AI-resistant fashion
  - only the AppliedML-Basic course, so far
- The same tests were submitted again to the Fake-student, and the scores were collected
- (Preliminary) results in next slide

# The work on-going: the core → focus on **students**

| SUCCESS RATE | Real-student | Fake-student |
|---|---|---|
| AI-fragile tests | 87% | 98% |
| AI-resistant tests | (not yet in production) | **82%** |

AI-fragile: original formulation of questions
AI-resistant: new formulation of questions
Real-student: a real, human student answering a test
Fake-student: chatGPT answering a test

It is very preliminary, results are thin, plenty of work to do..

- Feedback is helpful. Possible match with people interested in pursuing similar goals is also fruitful.

But, still.. encouraging!

- It is remarkable to start to see some thin evidence that **a Fake-student, on newly designed AI-resistant tests, performs WORSE w.r.t a Real-student on current, AI-fragile tests**

# The work on-going: the spin-off → focus on **teachers**

While working on the students' side of the problem, it also emerged that teachers could as well as be a point of attack!

- conservative approaches might tend to let teachers set up a DB of questions and never change/update it
- Laziness and/or lack of time might tend to let teachers use chatGPT themselves to create tests (!)

This is potentially problematic, as questions created by a LLM will easily be answered 100% correct by the LLM itself. All considered, the education system as we know it might easily go belly up with AI-fragile tests, if both students and teachers "**cheat**GPT" through the respective difficulties..

**→ designing AI-resistant tests, by construction, requires MORE WORK only to those educators who think that using LLMs will facilitate their life!**

- AI-resistant tests allow 1) students not to "cheatGPT" because the exam outcome will be worse, and 2) educator should not "cheatGPT" as well, i.e. work more to design these AI-resistant tests which will be not so easy to break!

# Summary

Largely a work in progress, but numerical glimpses of evidence - in a real University-level course use-case - that **designing AI-resistant tests** might not be a mirage!

<u>A set of tactics are being identified and further worked on, to help and transform AI-fragile tests into AI-resistant ones</u>

Plenty of work to do..

- More experiments, with more LLMs, on more courses, ..

.. but there is a chance that one day I might go in class and give a test and state:

"***the average chatGPT score on this test is X: I am sure you can do better w/o chatGPT. Relax, roll up your sleeves, and start!***"