Don’t fear the robot: future-authentic assessment and generative artificial intelligence

Phillip (Phill) Dawson
Centre for Research in Assessment and Digital Learning (CRADLE)
Deakin University, Melbourne, Australia

@phillipdawson
Disclaimer

• I’m a standards-based assessment person
• I think getting this right matters because evidencing learning matters
• Broader questions about genAI also matter, but they aren’t the focus here.
• I don’t have an easy fix
• I’m interested in long-term approaches not short-term hacks
• My mum helped me cheat in grade 4
Three things to take from this presentation

- GenAI can do a lot of what we currently assess
- We probably can’t and shouldn’t ban it
- Assessment needs to prepare students for their future, not our past
“By performing at a greater than 60% threshold on the NBME-Free-Step-1 data set, we show that the model achieves the equivalent of a passing score for a third-year medical student.”
ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity

Sasha Nikolic da, Scott Daniel db, Rezwanul Haque dc, Marina Belkina d, Ghulam M. Hassan de, Sarah Grundy ef, Sarah Lyden eg, Peter Neal ef and Caz Sandison a
The screw eye shown below is subjected to three forces $F_1$, $F_2$ and $F_3$. Calculate the $x$ and $y$ components for each of the three forces $F_1$, $F_2$ and $F_3$ and their resultant $R$ by transferring to your written paper and completing the table below showing your calculations.
GPT-4, why is this meme funny?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
GPT-4 Passes the Bar Exam

Daniel Martin Katz\textsuperscript{1,2,3,4,*}, Michael James Bommarito\textsuperscript{1,2,3,4}, Shang Gao\textsuperscript{5}, and Pablo David Arredondo\textsuperscript{2,5}

\textsuperscript{1} Illinois Tech - Chicago Kent College of Law (USA)
\textsuperscript{2} CodeX - The Stanford Center for Legal Informatics (USA)
\textsuperscript{3} Bucerius Law School (Germany)
\textsuperscript{4} 273 Ventures (USA)
\textsuperscript{5} Casetext, Inc. (USA)
*Corresponding Author: dkatz3@kentlaw.iit.edu
GenAl is currently the least capable it will ever be

https://giphy.com/channel/gifodyssey
How Is ChatGPT’s Behavior Changing over Time?

Lingjiao Chen†, Matei Zaharia†, James Zou†

†Stanford University ‡UC Berkeley

Abstract

GPT-3.5 and GPT-4 are the two most widely used large language model (LLM) services. However, when and how these models are updated over time is opaque. Here, we evaluate the March 2023 and June 2023 versions of GPT-3.5 and GPT-4 on four diverse tasks: 1) solving math problems, 2) answering sensitive/dangerous questions, 3) generating code and 4) visual reasoning. We find that the performance and behavior of both GPT-3.5 and GPT-4 can vary greatly over time. For example, GPT-4 (March 2023) was very good at identifying prime numbers (accuracy 97.6%) but GPT-4 (June 2023) was very poor on these same questions (accuracy 2.4%). Interestingly GPT-3.5 (June 2023) was much better than GPT-3.5 (March 2023) in this task. GPT-4 was less willing to answer sensitive questions in June than in March, and both GPT-4 and GPT-3.5 had more formatting mistakes in code generation in June than in March. Overall, our findings show that the behavior of the “same” LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLM quality.
Three things to take from this presentation

GenAI can do a lot of what we currently assess

We probably can’t and shouldn’t ban it

Assessment needs to prepare students for their future, not our past
Banning GenAI in educational settings

- A ban is a type of restriction
- Restrictions that don’t work are theatre
- How do we enforce GenAI restrictions?
Restricting access to GenAI

- In Australia, federal legislation allows site blocking for contract cheating sites. 200 are blocked at a national level.
- Many institutions block contract cheating sites.
- How effective are these blocks?
Detecting GenAI

• Unclear how effective GenAI detectors are
• Claims of very high detection rates haven’t been verified by independent researchers
• Claims that GenAI detection won’t be possible in long term
• If you take one thing from this talk: don’t upload student work to random GitHub/HuggingFace ChatGPT detectors or ChatGPT itself
What is easy in the short term probably won’t work in the long term.

<table>
<thead>
<tr>
<th></th>
<th>Short-term</th>
<th>Medium-term</th>
<th>Long-term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Ignore</td>
<td>Might get away with it momentarily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Ban</td>
<td>Problematic</td>
<td>Becomes risky</td>
<td></td>
</tr>
<tr>
<td>3. Invigilate</td>
<td>Where appropriate</td>
<td>Where appropriate</td>
<td>Where appropriate</td>
</tr>
<tr>
<td>4. Embrace</td>
<td>Being mindful of equity issues</td>
<td>Where appropriate</td>
<td></td>
</tr>
<tr>
<td>5. Design around</td>
<td>Risky</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Rethink</td>
<td>Requires time and effort</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Jason Lodge, Sarah Howard & Jaclyn Broadbent
Three things to take from this presentation

GenAI can do a lot of what we currently assess

We probably can’t and shouldn’t ban it

Assessment needs to prepare students for their future, not our past
Through a validity lens

• Validity is why we should address genAI
• GenAI is a secondary concern to validity
• Restrictions that can’t be enforced hurt validity
• Validity includes what we assess

Image: https://learningspy.co.uk/assessment/when-assessment-fails/#post/0
Authentic assessment

• Authentic assessment: let’s represent the world outside education in assessment.

• Aims to improve validity

• Not an anti-cheating panacea (Ellis et al 2019)
Future-authentic assessment

“assessment that faithfully represents not just the current realities of the discipline in practice, but the likely future realities of that discipline”

Do students need to do it without genAI...

Every time?
Sometimes?
Never?
Cognitive offloading
Future evaluative judgement

- Making judgements about quality of work
- Constituting standards

CRADLE suggests...

This three-page CRADLE Suggests covers...

- Enacting principles of good assessment design, in a world of genAI
- Adapting current assessment practices to account for genAI
- Research from five books from the CRADLE team
- Under a CC-BY licence

Enacting principles of good assessment design, in a world of genAI

Deploy resources to assure assessment when it matters most. Assessment at the most basic level requires us to assure that all students have achieved all learning outcomes at least at a threshold level. This function, as distinct from other purposes of assessment, needs to be protected to a high level of integrity. Assessment for assurance may therefore involve a focus on meeting program learning outcomes and therefore require fewer, but more carefully planned and protected, occasions of assessment.

In early years of a program, for students to learn from undertaking such tasks, they should also receive improvement-focused comments. They then have opportunities to act on this information and demonstrate improved understanding in subsequent work. Such tasks could authentically incorporate genAI. Specific, forward-looking, contextualised feedback comments from teachers, peers or others (including genAI) can prompt student learning as well as promote effective and appropriate engagement with genAI.

Focus on evidencing that outcome are met

Rather than adopting a cheating or academic integrity framing of genAI, we suggest an assessment framing, focused on evidencing outcomes. For most outcomes this will require a rang

https://tinyurl.com/CRADLEgenAI
Three things to take from this presentation

- GenAI can do a lot of what we currently assess
- We probably can’t and shouldn’t ban it
- Assessment needs to prepare students for their future, not our past
1. What types of task are now infeasible as assessment of learning? Would we still keep them as assessment for learning?
2. What learning outcomes are easy to assess now? What outcomes are hard to assess?
3. What new learning outcomes do we need, and what outcomes can we let go of?
4. Which outcomes suit a scaffolding approach? Which suit reverse scaffolding?
5. What extraneous cognitive load can we allow students to hand over to GenAI?