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

DEFENDING ASSESSMENT SECURITY IN A DIGITAL WORLD

Preventing E-Cheating and Supporting Academic Integrity n Higher Education

cradle

DEAKIN UNIVERSITY

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

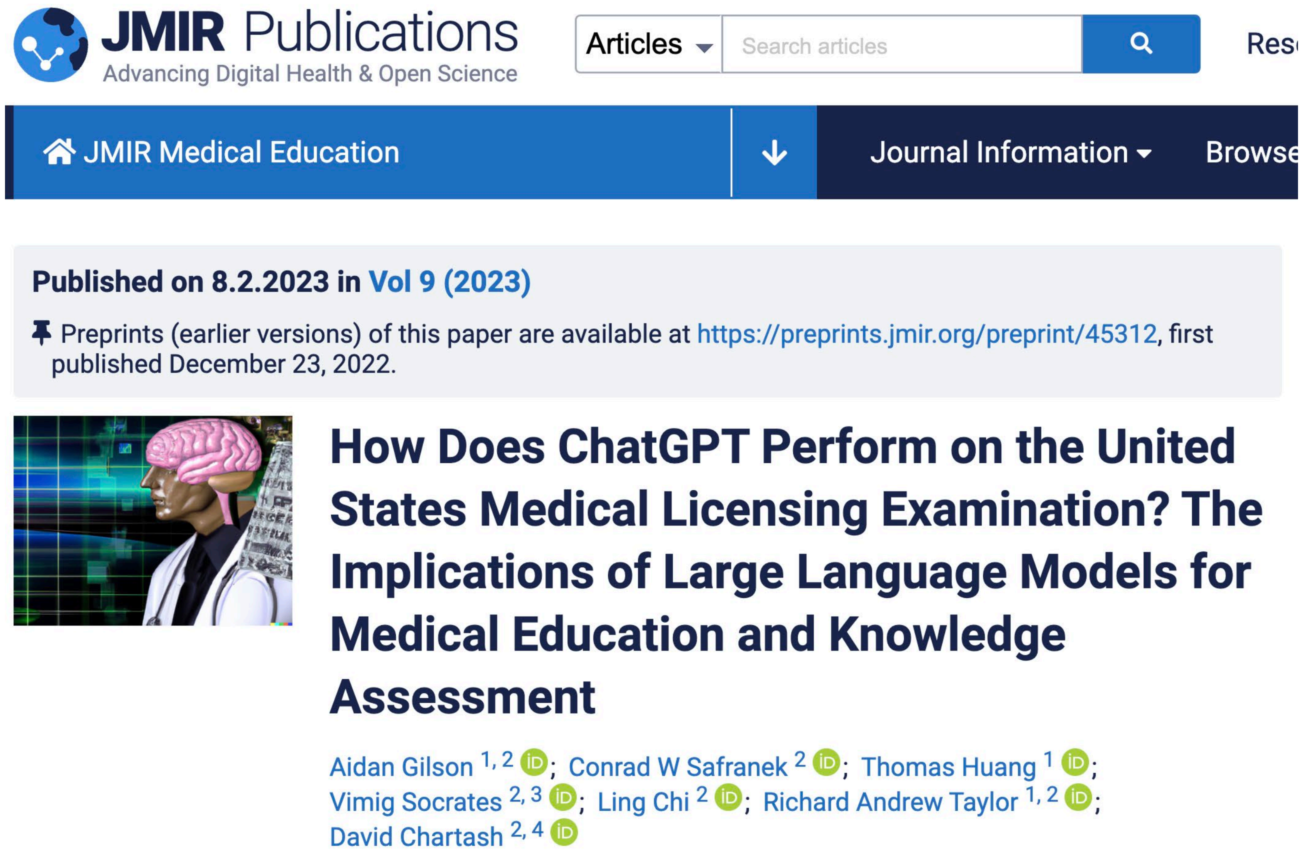
Three things to take from this presentation

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.”

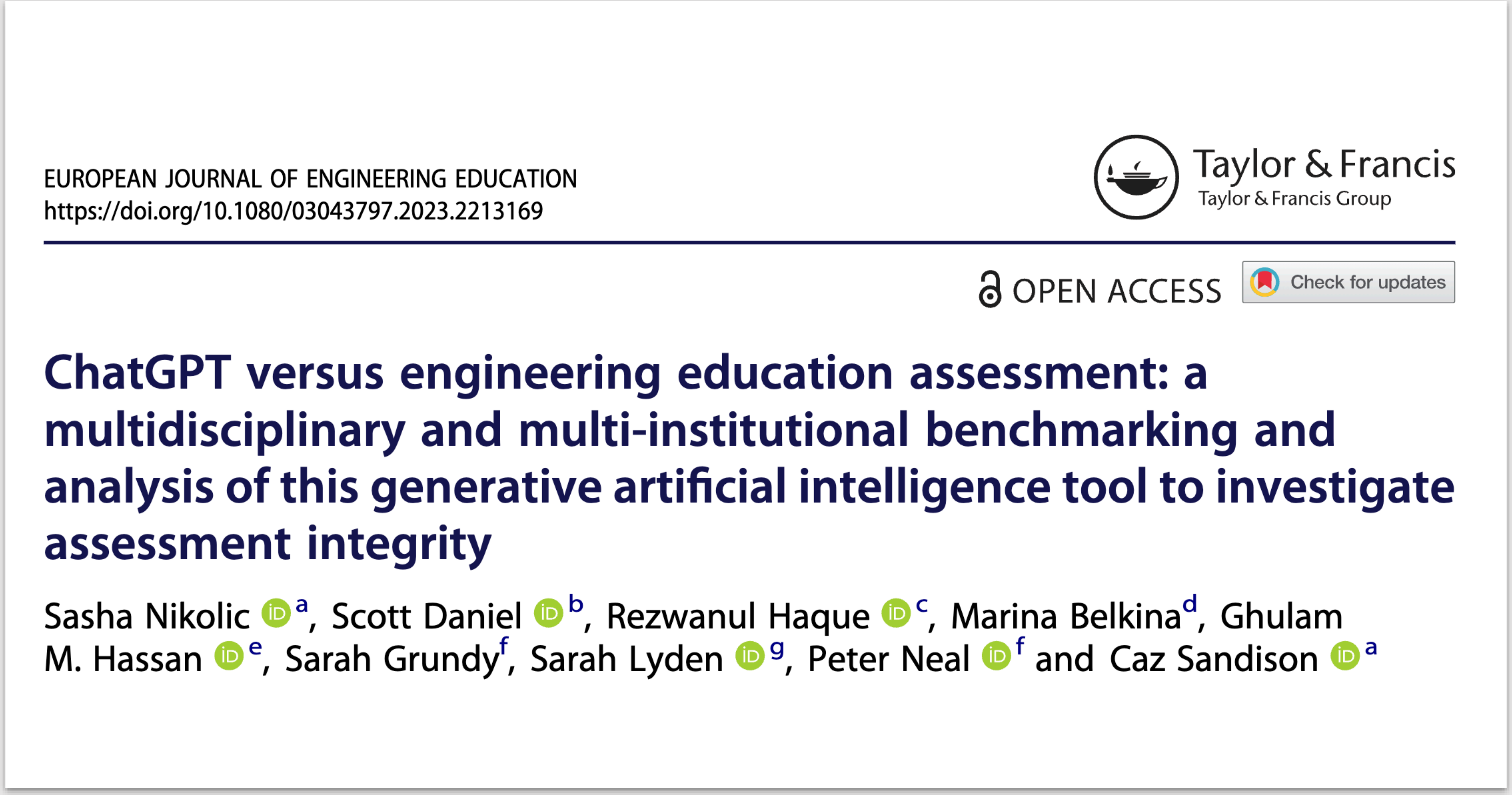


Published on 8.2.2023 in Vol 9 (2023)

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/45312, first published December 23, 2022.

How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment

Aidan Gilson [1, 2]; Conrad W Safranek [2]; Thomas Huang [1]; Vimig Socrates [2, 3]; Ling Chi [2]; Richard Andrew Taylor [1, 2]; David Chartash [2, 4]



EUROPEAN JOURNAL OF ENGINEERING EDUCATION
https://doi.org/10.1080/03043797.2023.2213169

Taylor & Francis
Taylor & Francis Group

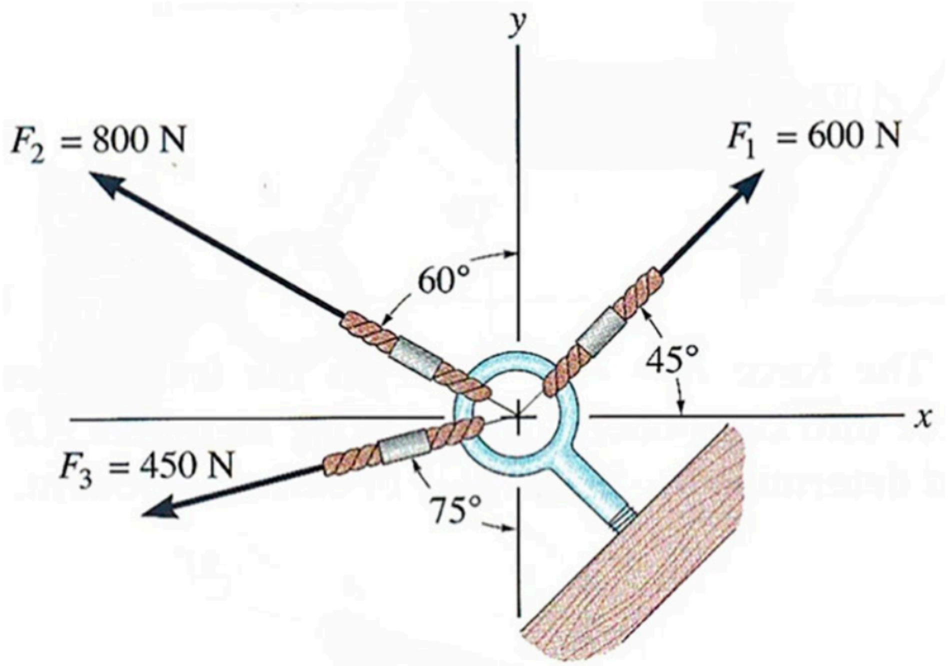
OPEN ACCESS

Check for updates

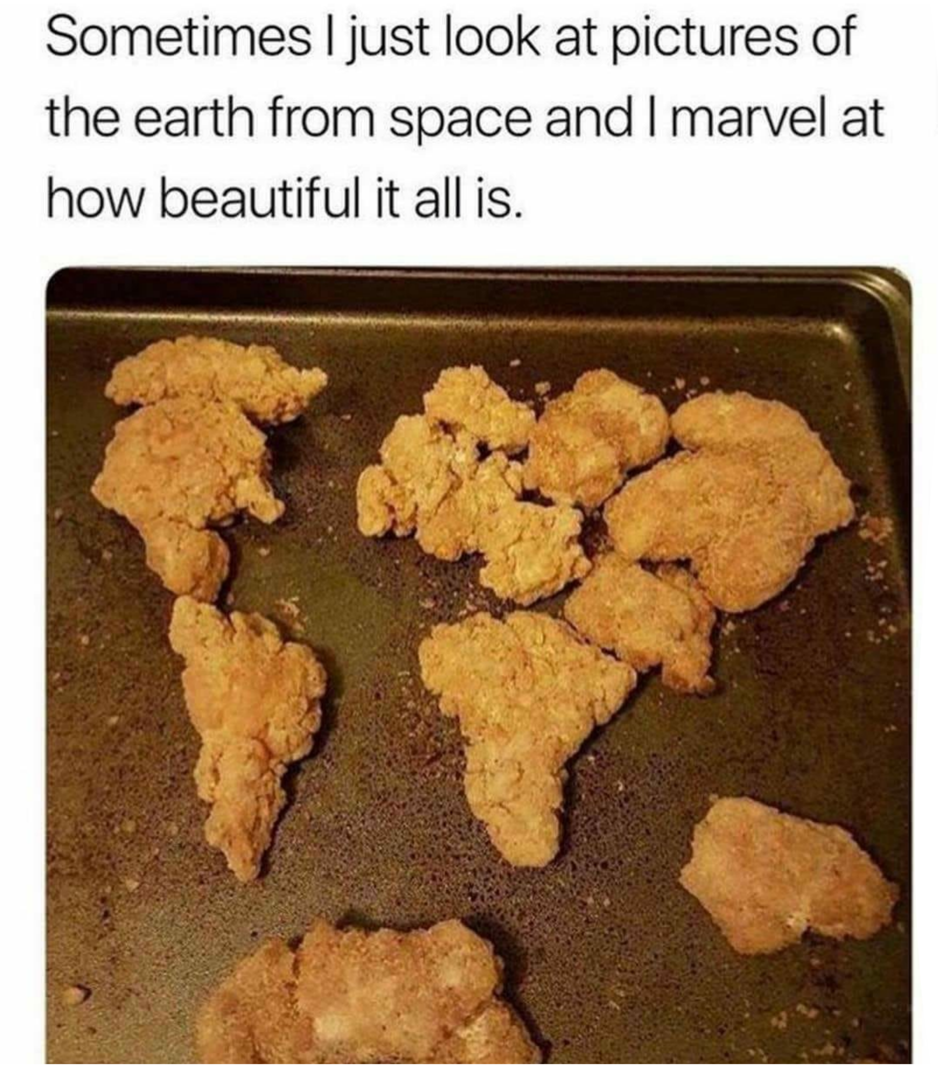
ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity

Sasha Nikolic [a], Scott Daniel [b], Rezwanul Haque [c], Marina Belkina[d], Ghulam M. Hassan [e], Sarah Grundy[f], Sarah Lyden [g], Peter Neal [f] and Caz Sandison [a]

Table 1. Example of translating a figure into ChatGPT (Image from (Merriam and Kraige 1993)).

Original question	Question modification in ChatGPT
The screw eye shown below is subjected to three forces F_1, F_2 and F_3.  y F_2 = 800 N F_1 = 600 N 60° 45° x F_3 = 450 N 75° Calculate the x and y components for each of the three forces F1, F2 and F3 and their resultant R by Transferring to your written paper and completing the table below showing your calculations	The screw eye is subjected to three forces F1 = 600 N East 45 degrees North; F2 = 800 N North 60 degrees west; F3 = 450 N South 75 degrees West. Calculate the x and y components for each of the three forces F1, F2 and F3. Use calculated vector components to find the magnitude and direction of resultant force

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.

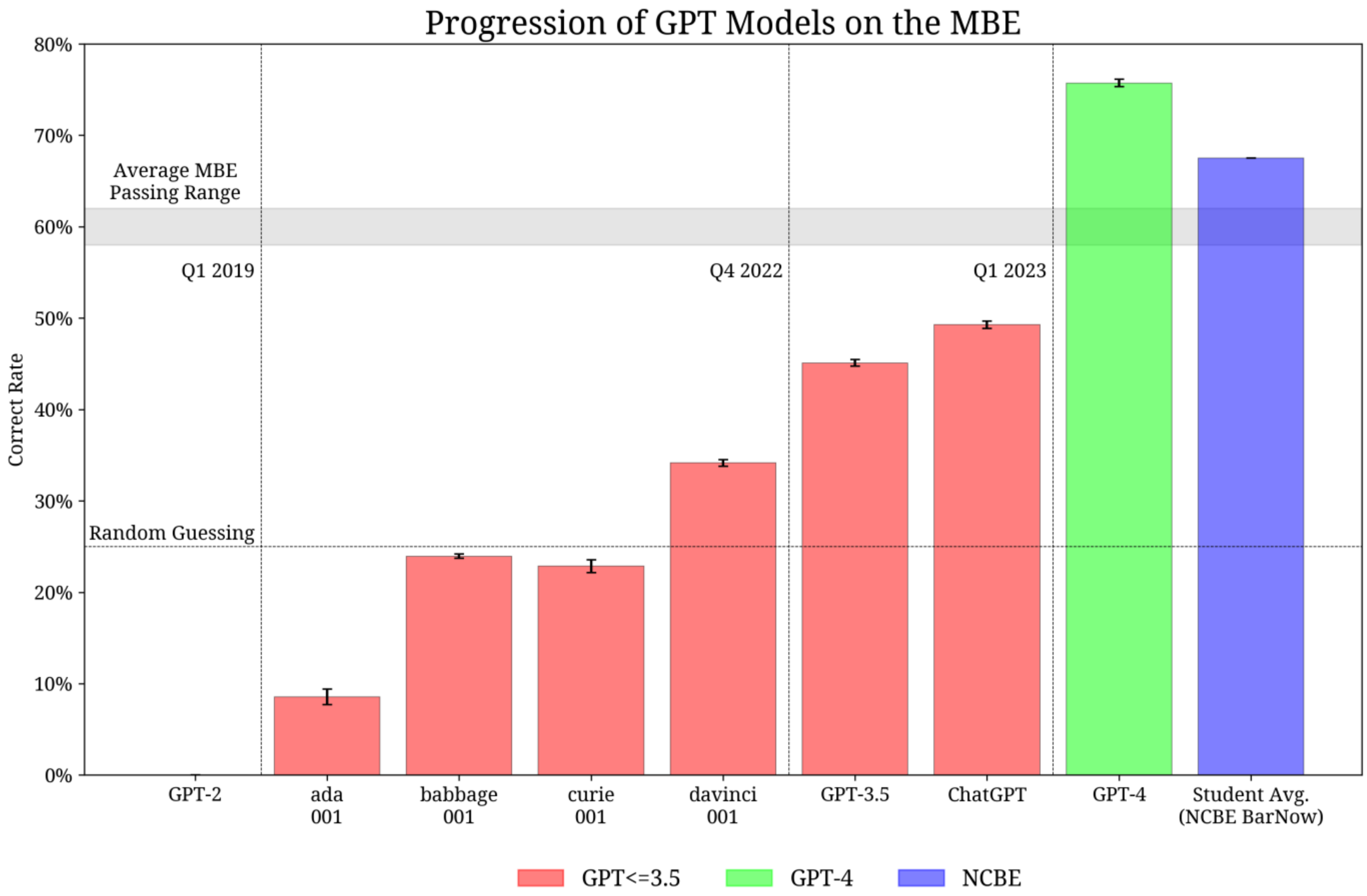
This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the 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

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

GPT-4 Passes the Bar Exam

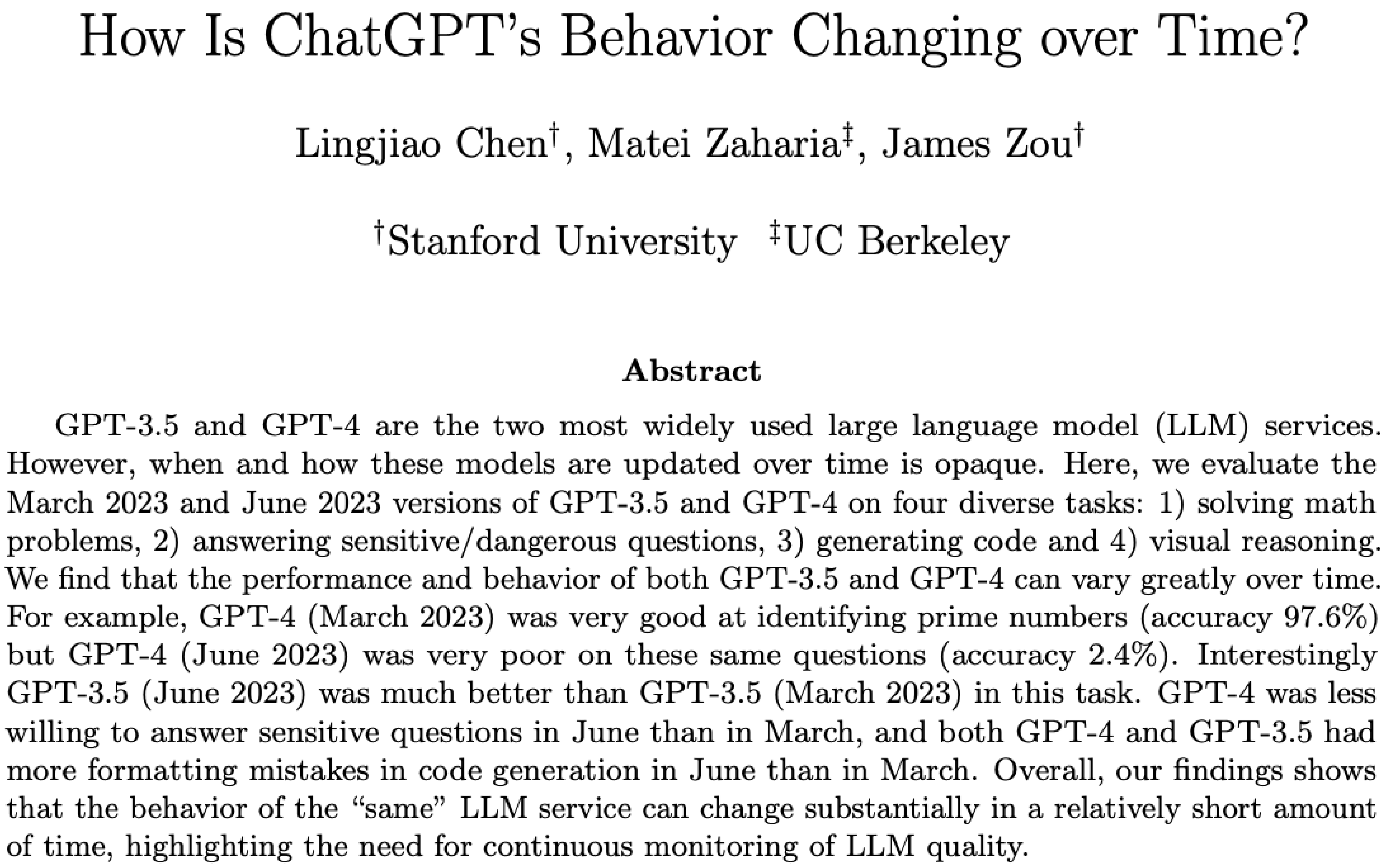
Daniel Martin Katz[1,2,3,4,*], **Michael James Bommarito**[1,2,3,4], **Shang Gao**[5], **and Pablo David Arredondo**[2,5]

[1] Illinois Tech - Chicago Kent College of Law (USA)
[2] CodeX - The Stanford Center for Legal Informatics (USA)
[3] Bucerius Law School (Germany)
[4] 273 Ventures (USA)
[5] Casetext, Inc. (USA)
[*] Corresponding Author: dkatz3@kentlaw.iit.edu


Progression of GPT Models on the MBE
Correct Rate
80%
70%
60%
50%
40%
30%
20%
10%
0%
Average MBE Passing Range
Q1 2019
Q4 2022
Q1 2023
Random Guessing
GPT-2
ada 001
babbage 001
curie 001
davinci 001
GPT-3.5
ChatGPT
GPT-4
Student Avg. (NCBE BarNow)
GPT<=3.5
GPT-4
NCBE

GenAI 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 shows 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

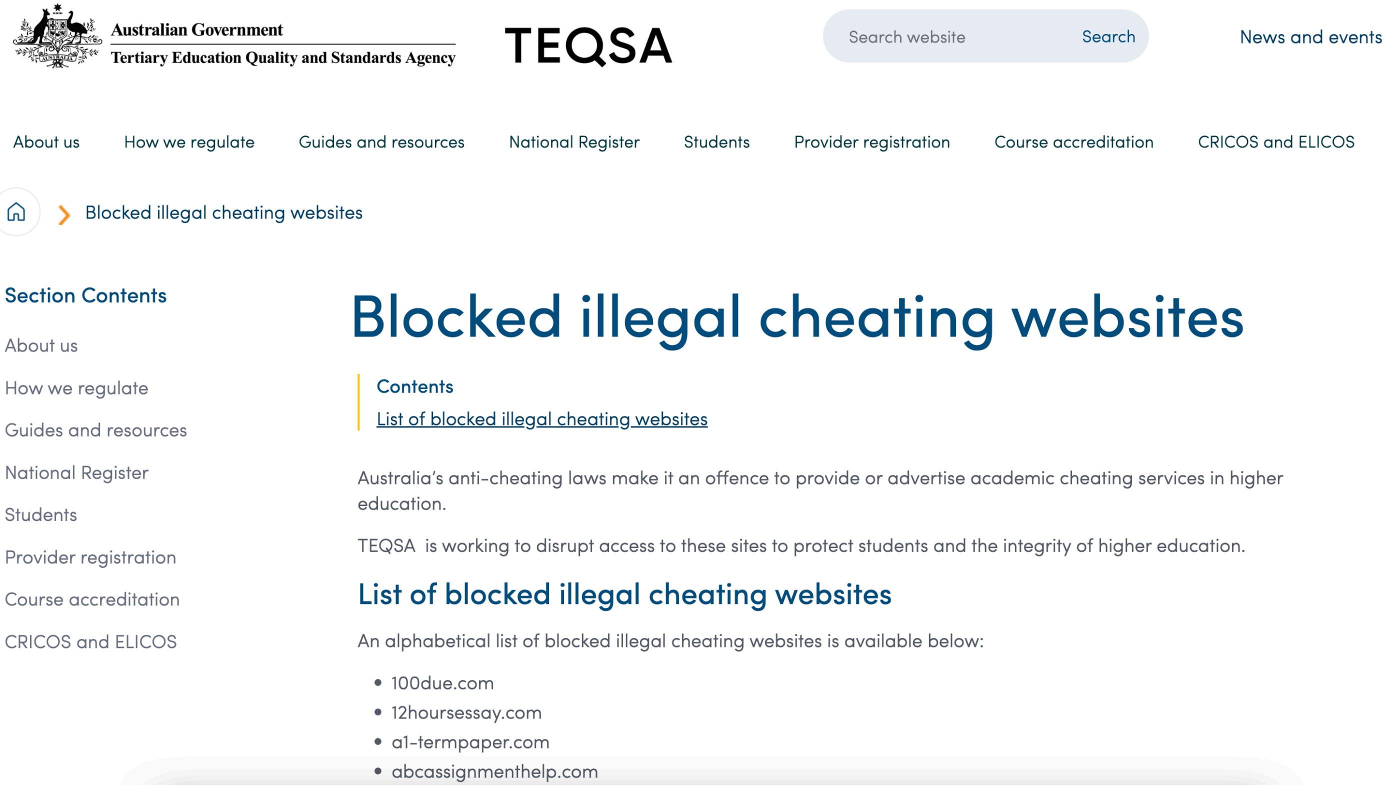
Banning GenAI in educational settings

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

PHILLIP DAWSON
DEFENDING ASSESSMENT SECURITY IN A DIGITAL WORLD
Preventing E-Cheating and Supporting Academic Integrity in Higher Education
ROUTLEDGE

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?



Blocked illegal cheating websites

Contents
List of blocked illegal cheating websites

Australia's anti-cheating laws make it an offence to provide or advertise academic cheating services in higher education.

TEQSA is working to disrupt access to these sites to protect students and the integrity of higher education.

List of blocked illegal cheating websites

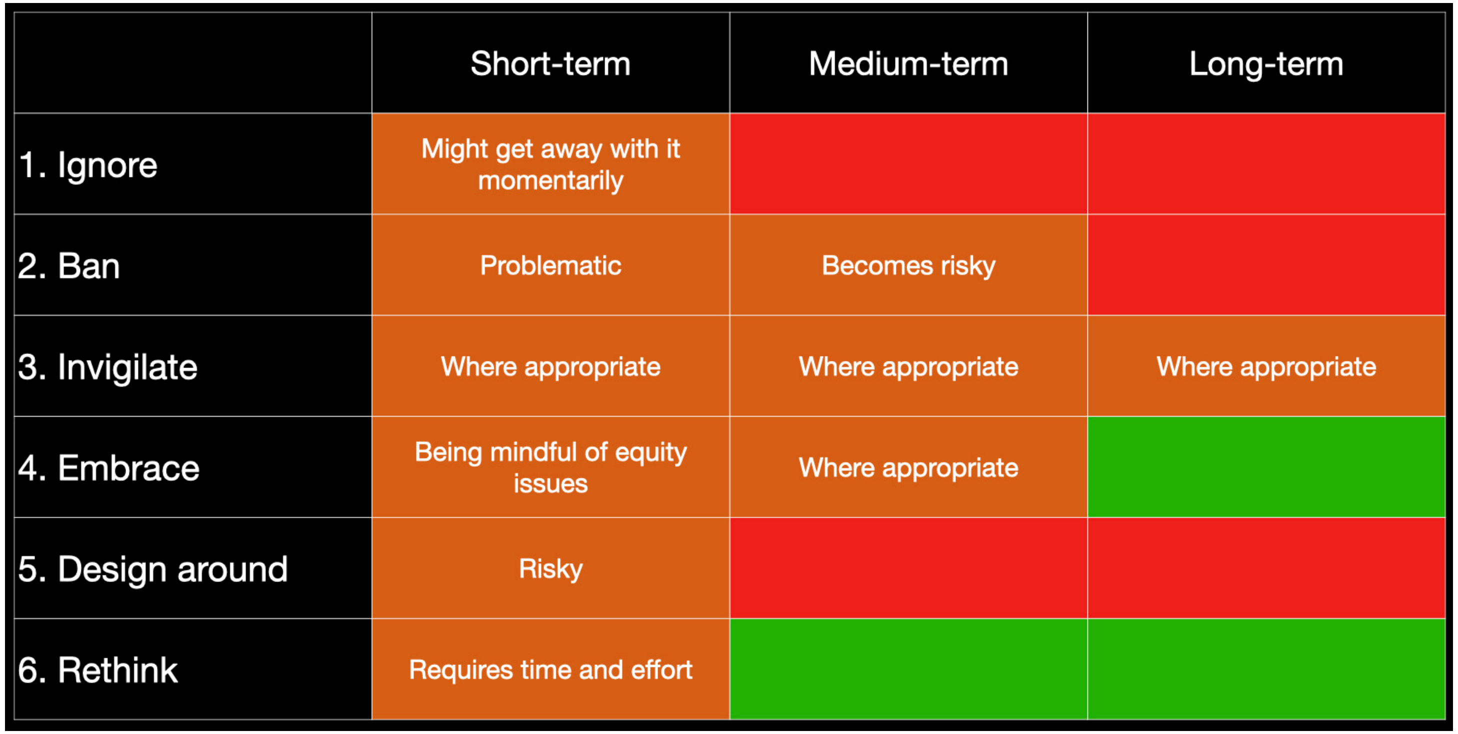
An alphabetical list of blocked illegal cheating websites is available below:

- 100due.com
- 12hoursessay.com
- a1-termpaper.com
- abcassignmenthelp.com

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



	Short-term	Medium-term	Long-term
1. Ignore	Might get away with it momentarily	🟥	🟥
2. Ban	Problematic	Becomes risky	🟥
3. Invigilate	Where appropriate	Where appropriate	Where appropriate
4. Embrace	Being mindful of equity issues	Where appropriate	🟩
5. Design around	Risky	🟥	🟥
6. Rethink	Requires time and effort	🟩	🟩

Jason Lodge, Sarah Howard & Jaclyn Broadbent
https://www.linkedin.com/pulse/assessment-redesign-generative-ai-taxonomy-options-viability-lodge

Three things to take from this presentation

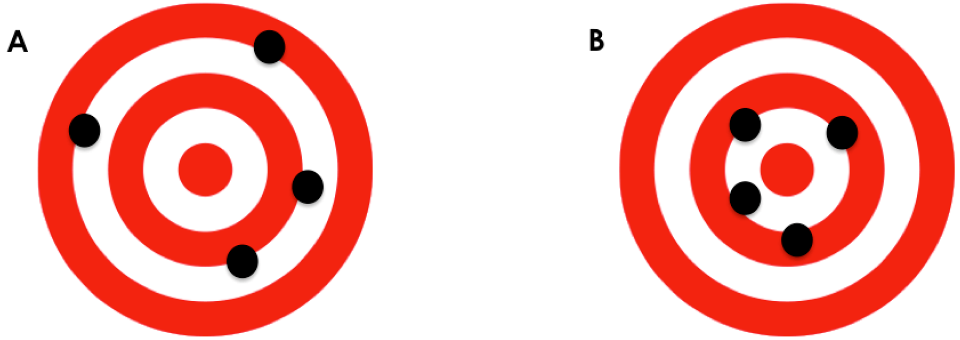
GenAI can do a lot of what we currently assess

We probably can’t and shouldn’t ban it

Assessment needs to prepare students for their future, not our past

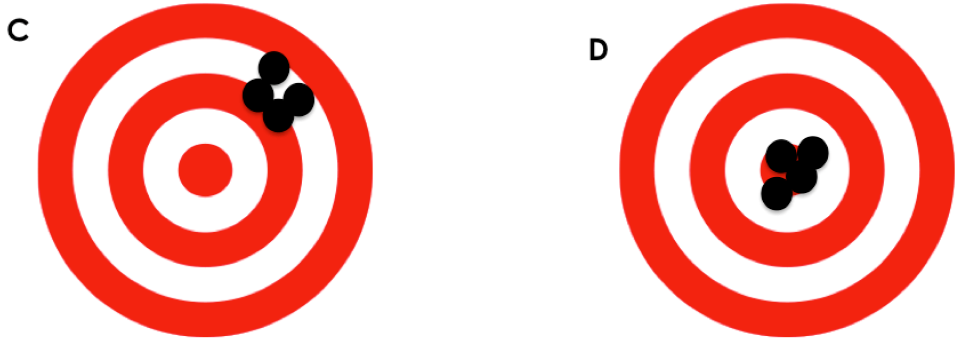
Through a validity lens

- 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



Unreliable & invalid

Unreliable but valid?



Reliable but invalid

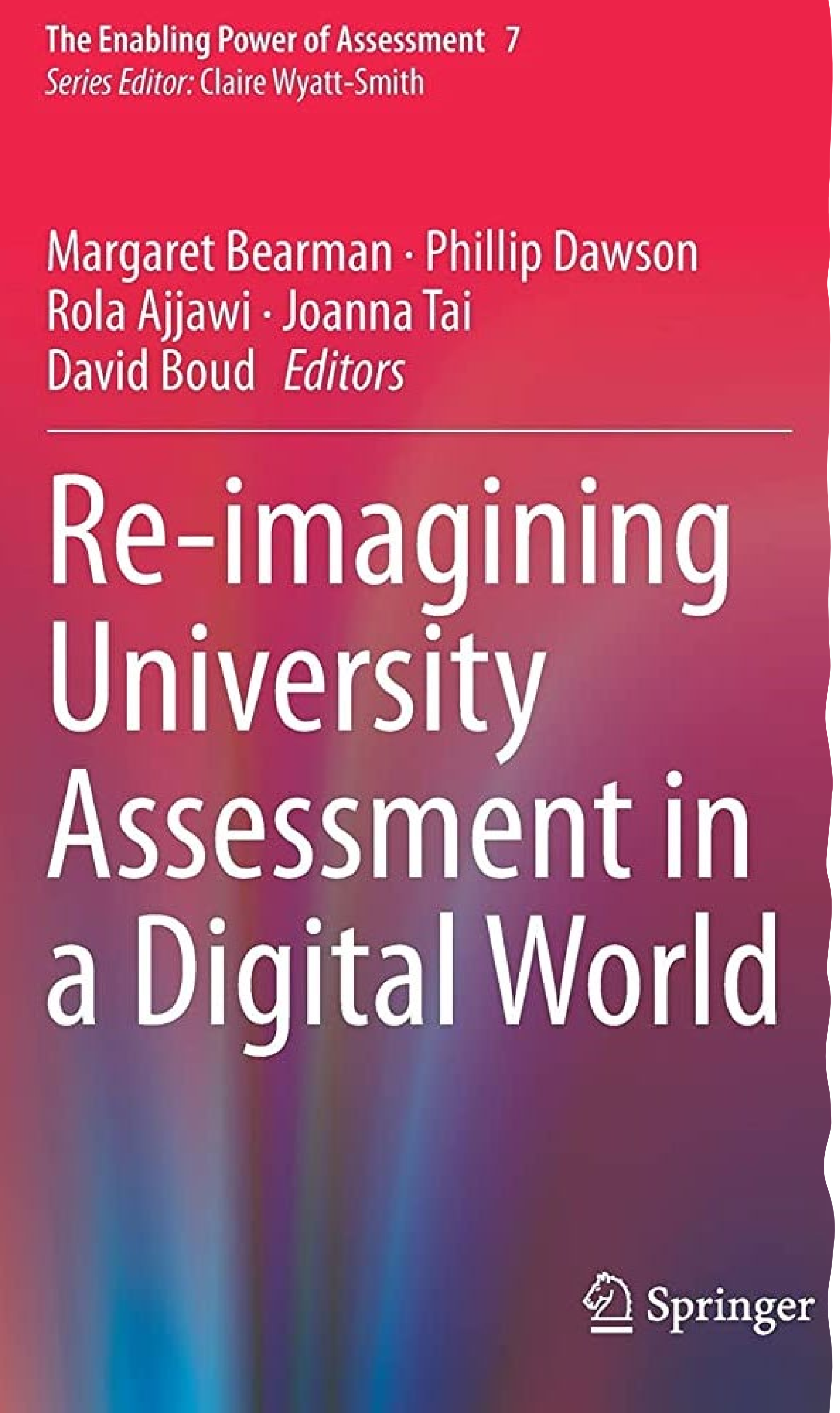
Reliable & valid

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)

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840-854.
Ellis, C., van Haeringen, K., Harper, R., Bretag, T., Zucker, I., McBride, S., Rozenberg, P., Newton, P., & Saddiqui, S. (2019). Does authentic assessment assure academic integrity? Evidence from contract cheating data. *Higher Education Research & Development*, *39*(3), 454-469.



The Enabling Power of Assessment 7
Series Editor: Claire Wyatt-Smith

Margaret Bearman · Phillip Dawson
Rola Ajjawi · Joanna Tai
David Boud *Editors*

Re-imagining University Assessment in a Digital World

Springer

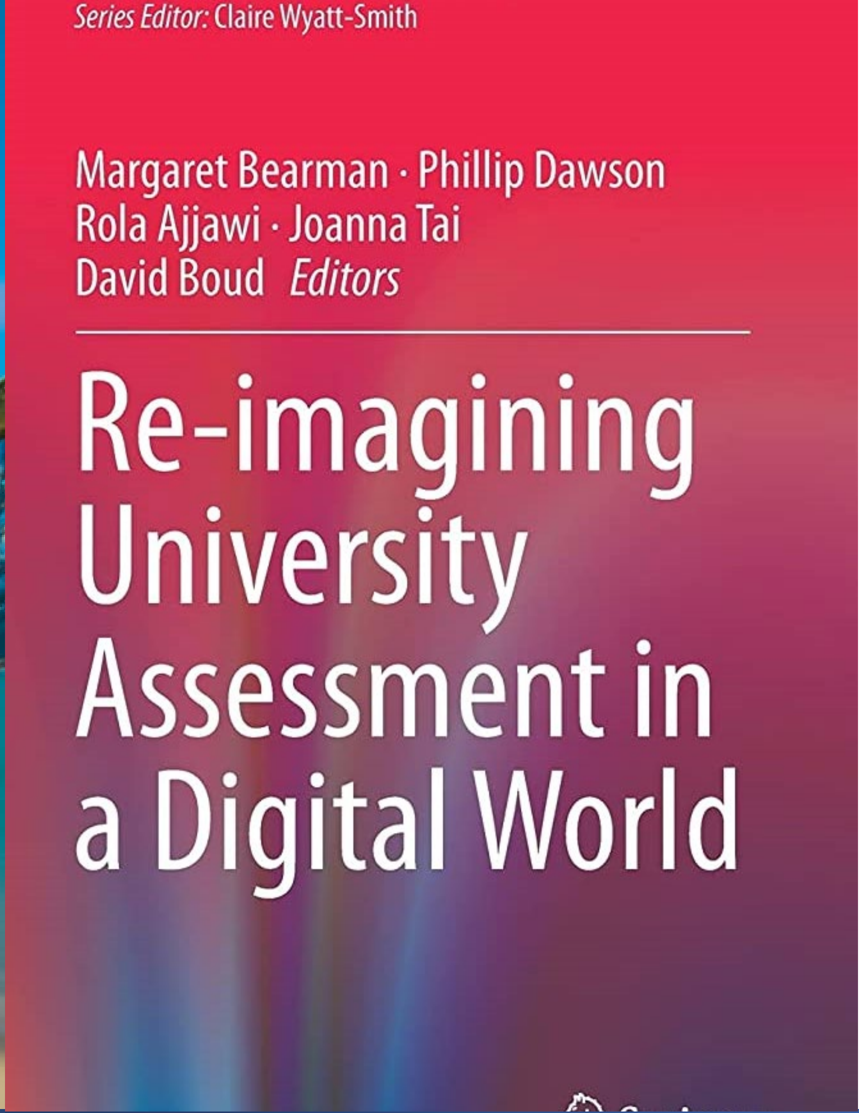
Future-authentic assessment

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

Dawson, P., & Bearman, M. (2020). Concluding Comments: Reimagining University Assessment in a Digital World. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining University Assessment in a Digital World* (pp. 291-296). Springer International Publishing. https://doi.org/10.1007/978-3-030-41956-1_20

Do students need to do it without genAI…

Every time?
Sometimes?
Never?



Series Editor: Claire Wyatt-Smith

Margaret Bearman · Phillip Dawson
Rola Ajjawi · Joanna Tai
David Boud *Editors*

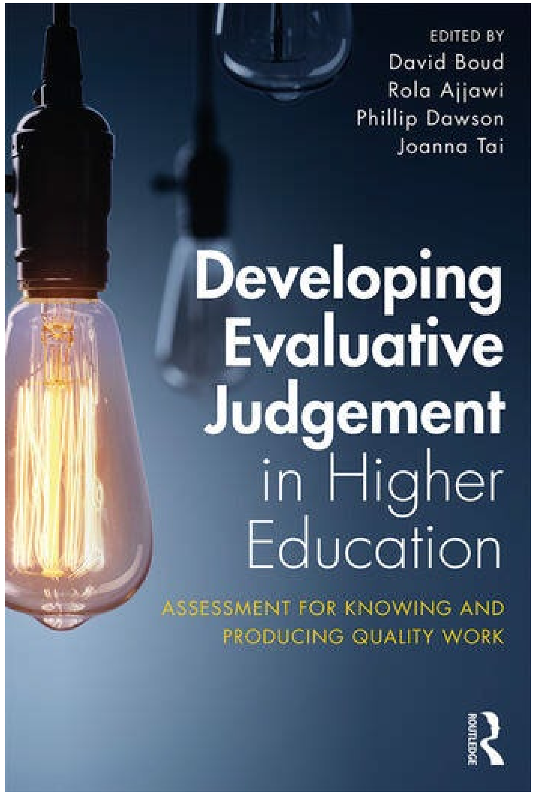
Re-imagining University Assessment in a Digital World

Cognitive offloading

Future evaluative judgement

Making judgements about quality of work

Constituting standards



EDITED BY
David Boud
Rola Ajjawi
Phillip Dawson
Joanna Tai

Developing Evaluative Judgement in Higher Education

ASSESSMENT FOR KNOWING AND PRODUCING QUALITY WORK

ROUTLEDGE

Bearman, M., & Luckin, R. (2020). Preparing university assessment for a world with AI: Tasks for human intelligence. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining University Assessment in a Digital World*. Springer.

CRADLE suggests...

Issue 09 |

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.

Develop student capability to identify 'what good looks like' through assessment

With the rise of genAI, students will need

standards required to be met, such as the task format. Including a variety of submission formats for a task, and/or across a course or program may there improve equitability of assessment. This can balance out whether the use of genAI itself becomes a barrier (for example due to cost) or an enabler (fo example, mitigating the impact of pers circumstances).

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

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

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

Provocations

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?