ChatGPT: The Response Must Be Pedagogical

Jean-Philippe Boucher

At the time of writing, GPT-4 has been available for less than 48 hours and already we know that it can pass the majority of standardized tests as well as the exams of several professional orders (OpenAI, 2023). Compared to its predecessor GPT-3.5, released four months earlier in November 2022, the new version does better on the vast majority of tests and scores up to two to three times higher. We can only speculate on what capabilities generative artificial intelligence tools (GAIs) will have come fall 2023. Do we really want to embark on a technological arms race against Silicon Valley? What if our safe-conduct were pedagogical? In this article, I want to approach the phenomenon of GAIs from a pedagogical perspective, particularly with regard to the thorny issue of assessment validity.

Research is dead, long live generation!

Let's first quickly clarify what GPT-4 is before going any further. GPT-4 is the heart of ChatGPT. It is a language processing model that has access to a pile of data and is able to generate an answer to a specific question using this mass of information. What makes GPT-4 impressive lies mainly in two aspects:

1. It is able to analyze and understand complex issues.
2. It is able to provide an answer in a near-perfect textual form, and the format of the answer can be adjusted according to our needs (word count, writing style, etc.).

More concretely, GPT-4 understands questions such as "Can you write me a 600-word text linking Rousseau's work to global warming?" and then provides answers that (approximately) meet the required length and specified topic. The quality of the answers varies, but in many cases, this application is able to produce an answer that requires only a little human editing to produce an acceptable school assignment.

GPT-4 also has memory. So, if the answer is not perfect, it is possible to refine the question before doing the editing. To take the example of Rousseau again, we could ask to adjust the answer so that the text is written in the first-person singular and to add three quotations; GPT-4 would probably be able to satisfy both requests.

The major advance of GAIs like GPT-4 is that the answer provided is generated; it is not simply retrieved from the data. There is no pre-written text of 600 words on each of the possible topics in its data. The ground swell movement that is taking place concerns a probable replacement of tools that search for content, such as a search engine or Wikipedia, by GAIs that generate content on demand. This is a real paradigm shift.

The elusive sentience

As far as we can tell, because it is privately owned by OpenAI, GPT-4 remains a mere computer program that depends on the knowledge produced by humanity. It is no closer to achieving sentience, the ability to reason, than the first artificial intelligence programs of the last century. Thus, it is totally dependent on its pile of data and is limited to generating content from it. Since GPT-4 is not sentient, it has no metacognition and cannot validate or correct itself

without human help. It is also unable to display divergent thinking on the content at its basis.

Another characteristic of GPT-4 is that it doesn't like to admit that it doesn't know something. As a result, sometimes it affirms things that are completely false, either with regard to facts or references. To maintain a mystical side, GAI designers indulge in a little anthropomorphism and call these mistakes "hallucinations." In other words, GPT-4 is not perfect, and it is regularly wrong, but since it knows no doubt, it writes its output with as much confidence as the most obvious truth.

In short, GPT-4 is like someone analyzing everything available on the Internet on a given topic and synthesizing it into a syntactically perfect aggregate. It will be aesthetic, smooth, researched, and unique in its formulation. But if there is anything untrue or unheard of, it will be due to hallucinations.

Evaluation, fraud and GPT-4

GPT-4 has enormous implications for a number of evaluation practices in the college network. If students have a tool in their pocket that can write a 600-word argumentative essay in seconds, it's a big game changer. We hear a lot of talk about solutions, including purchasing detection tools, training on how to use GAIs in the classroom, or adding resources to support faculty. I have to admit that most of these solutions do not convince me. I don't like the fact that most solutions lead us back to technology instead of focusing on our strength: pedagogy.

It is true that the arrival of GPT-4 is disrupting the established order, but I believe that it is possible to use this disruption positively, much as teachers of another era did with other technological advances. In fact, I think GPT-4 could help a (genuine) transition to competency-based assessment. Instead of playing cat and mouse with GPT-4 to try to detect fraud, I believe our salvation lies in evaluations that cannot be completed by a GAI like GPT-4. However, I realize that this idea is of varying difficulty depending on the discipline being taught.

I see two major paths to achieving GAI-proof evaluations. First, in the upstream design stage of our evaluations, and second, through continual process-oriented grading.

Evaluation design

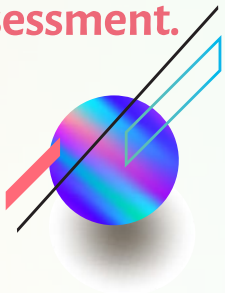
The *Report on the State and Needs of Education: Evaluating so It Truly Counts* (Conseil supérieur de l'éducation (CSE), 2018) states that in the logic of the competency-based approach, it is the evaluation at the end of the course or cycle that should make it possible to certify the extent to which learning has been achieved. However, it is prudent to base the evaluation judgment on several learning traces.

The CSE report warns against the accumulation of multiple evaluations throughout the session, as by accumulating points on a report card, we also keep track of early errors in the final score (the addition of these points), so that this grade does not always adequately reflect the student's actual learning (2018). In fact, those who experience difficulties at the beginning find themselves particularly penalized by the accumulation of points (2018).

The competency-based approach advocates a more holistic evaluation—called *certification*—at the end of the learning sequence, to attest to the students' achievement of the competency, and is not in phase with overly fragmented evaluation practices. The idea behind this vision of evaluation is that we cannot really assess a competency by breaking it down into multiple subcomponent evaluations. Just because I can run, pass a ball between two cones, and kick it into a goal does not mean I can play soccer. Soccer involves teammates, opponents, and most importantly, a ton of micro-decisions that need to be made quickly. So, if

I want to assess the skill of playing soccer, I can't just conduct evaluations outside of the context of the sport itself by targeting the subcomponents of soccer.

In fact, I think GPT-4 could help a (genuine) transition to competency-based assessment.



If we have a fragmented evaluation practice, we are usually forced to conduct small evaluations focused on subcomponents of competencies or evaluations that only touch on knowledge, omitting skills and attitudes. By compartmentalizing in this way, not only do we move away from the assessment advocated by the competency-based approach, but we also facilitate fraud by GAIs. The latter are much more successful with specific, circumscribed problems that focus on knowledge. The more encompassing the assessments are of the entire competency, the less vulnerable we are to the use of GAIs.

Indeed, even before the advent of ChatGPT, the CSE report recommended changing superficial learning evaluation practices that allow students to take shortcuts in their academic engagements (2018). A shift to less fragmented assessment, geared more toward certification, is therefore not only a defense against GAIs, but also a pathway to increased engagement and learning for the student community.

I'll also add in bulk a few other defenses that affect the content of evaluations. First, and this may change eventually, but for now, GPT-4's pile of data is limited to the years before 2021. Therefore, on any current topic, it is unable to produce an adequate answer. The choice of topics covered in our evaluations can thus protect us from the risk of fraud.

Secondly, the results GPT-4 provides suggest that anything with little Internet presence is unknown to it. Thus, if we work with "local" or authentic pedagogy case studies with real clients in close proximity to our colleges, such as small non-profit organizations (NPOs) or entities internal to our colleges, GPT-4 will probably not know what to say about them. Conversely, the more we use global and international case studies, the easier it is for GAIs to cheat.

Process-oriented grading

The duality of *process* and *product* is a much-studied topic, especially where performance is easily measurable and comparable, such as in management and sports. It is generally recognized that focusing goals on process rather than product has a positive effect on motivation and ultimately on performance (Williamson *et al.*, 2022).

That's positive news, because in my opinion, the best defense against fraud by GAI is a process-oriented rather than a product-oriented evaluation practice. This may seem to be in direct contradiction to the previous argument of not fragmenting assessment and having a more certification-based approach if we perceive this practice as the product and especially as the only time we can assess. However, an assessment practice is not limited to summative evaluations. In fact, the CSE report recommends basing evaluative judgment on multiple learning traces (2018).

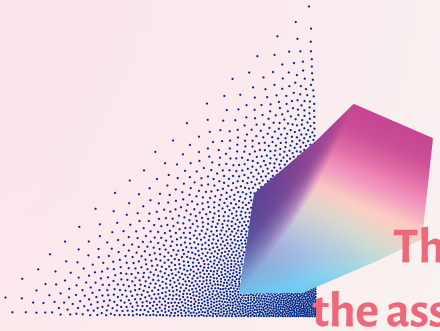
As Scallon (2014) so aptly puts it, the right path is probably somewhere between 15 weekly tests each worth 6.66% and a final exam worth 100%: to assess a skill, one cannot simply add up various pieces of information collected along the way, nor can one rely on a single situation, however complex, placed at the end of the sequence. The certification process must be situated between these two extremes.

So what does a process-oriented assessment practice that is not a collection of small, focused evaluations look like? And what are these learning traces? It simply looks like a sequence of formative assessments that eventually lead to a certification evaluation. Formative assessments not only allow us to support and guide learners, but also to adopt an evaluation posture thanks to the learning traces that we obtain.

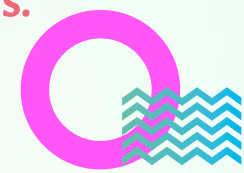
For example, a three-step assignment could consist of two formative submissions followed by a third submission that leads to certification. The first two submissions would be less and less approximate versions of the final work. Thus, as educators, we have two learning traces and a final version of the assignment. This allows us to use a qualitative grading approach and to apply the principles outlined in the CSE report: when assigning a final grade, all of the information one has about a student's developed abilities (including formative assessments) should be considered in the exercise of professional judgment, including observing progress and ensuring that the expected performance is sustained over time. (2018)

It is important to note here that a three-step submission should not triple the time spent on grading; rather, a process-oriented grading approach should be adopted. The two formative submissions allow for student support and guidance, but also provide an opportunity to start grading before the certification evaluation. Thus, throughout the process, one should observe progress made and ensure that the expected performance is sustained over time (CSE, 2018) by recording learning traces and beginning to fill out the grading tools (descriptive evaluation grid or other).

Broadly speaking, instead of doing a correction blitz every month, the aim would be to do a little bit every week. With process-oriented assessment, there should be no surprises at the certification evaluation since the act of correcting is started before it. Thus, the correction of the certification evaluation becomes a validation of our judgment of the last few weeks rather than a completely new comprehensive correction. The evaluation is therefore no longer based solely on the precise moment of the certification evaluation, the product, but also makes room for everything that came before, i.e. the process.



The more encompassing the assessments are of the entire competency, the less vulnerable we are to the use of GAIs.



Beyond this example, we can also use several other strategies to obtain learning traces, such as having students keep logs, using progress reports, and holding short weekly meetings for teamwork.

A GAI can produce aesthetic finished products, but it can hardly produce learning traces like the two incomplete formative submissions given as examples. Nor can it keep a log or speak at a team meeting.

Pedagogy as a means of defense

Overall, I would say that in order to counteract GAI fraud, we need to "humanize" our assessments. What I mean by *humanizing assessments* is to think of them as an act of learning, not an act of grading or ranking. The first argument for this transition is relentless: GPT-4 is already in the 90th percentile for ranking-style assessments (standardized tests, order exams, etc.), so we've already lost the battle on that front.

More importantly, I believe this transition is beneficial to the student community. By using more comprehensive assessments, formative submissions, learning traces, feedback on these traces and a holistic focus on the learning process, we are aligned with the competency-based approach and contribute to students' mastery of these competencies. The learning and assessment phases become more aligned and less distinct. To reuse the soccer analogy, the sport itself should be taught and assessed by playing it as often as possible; it should not be deconstructed into three five-week blocks where you first teach and assess how to run, then pass, and finally shoot a goal. At some point, if possible as regularly as possible, learners need to be in full execution of the skill so that they can learn it and we can assess them.

The prophetic revolution in education

While there is no doubt that GAIs will have an undeniable impact on our society, and while I must admit that their impact on educational practices is sudden and severe in some cases, particularly for standardized tests, I am in opposition to those who see in these GAIs a revolution in education and in the role of the teacher.

To date, no technological advance can claim the title of "the one that revolutionized education," and not for lack of prophets who have repeatedly made that prediction for centuries. The reality is that education has not changed much in the last 2000 years. It remains a group of students learning with a teacher in a recognized and appropriate physical location, and that is probably because it is still the best way to do it. GAIs will probably join the manuscript, the printing press, the radio, the television, the photocopier, the calculator, the computer, the Internet, massive open online courses, and all the others in the group of "ex-pretenders" to the education revolution. And it will probably happen sooner rather than later.

I would even make a riskier prediction and say that, contrary to what some may believe, the actual integration of GAIs into the classroom will be minimal and that they will become just another tool in the array of what students use alongside the calculator, the computer and the Internet. ▬

References

Conseil supérieur de l'éducation (2018). *Report on the State and Needs of Education: Evaluating so It Truly Counts.*

OpenAI (2023). GPT-4 [openai.com/research/gpt-4].

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences,* Saint-Laurent, Éditions du renouveau pédagogique.

Williamson, O. *et al.* (2022). "The performance and psychological effects of goal setting in sport: A systematic review and meta-analysis," *International Review of Sport and Exercise Psychology* [Online].

Jean-Philippe Boucher has been teaching computer science at Cégep Garneau since 2008. He has a master's degree in College Teaching from the Faculty of Education at the Université de Sherbrooke and has been a member of the *Pédagogie collégiale* editorial committee since 2019. He has been caught twice already in the excitement of "AI that will revolutionize the world," once in 2005 during his university career and then ten years later related to the prowess of AlphaGo. So once bitten, twice shy, he has tempered expectations of the real changes AI will cause, especially in education.

jpboucher@cegepgarneau.ca