Generative AI in education

AQA's response to the Department for Education's call for evidence

Alex Scharaschkin
Director for Assessment Research & Innovation

August 2023

Executive summary

Artificial Intelligence (AI) has the power to transform education for the better – by cutting teacher workload and improving marking reliability.

We have been exploring potential uses of AI to improve assessment and education – we have found that there is potential for AI tools to make working life better for teachers and complement human marking. But AI tools can generate wrong or misleading information, which are known as hallucinations, and therefore need close human supervision to be effective.

AQA researchers trialled AI tools including ChatGPT, GPT4, LLaMA and Alpaca on a range of publicly available science papers. We found that, with appropriate regulation and human involvement, AI could be used to automate simple marking processes. This in turn could be used to help teachers plan their lessons for students.

In the long run, AI could help teachers by:

- helping teachers create tasks and assessments for students
- assisting students with their own research
- giving students automated feedback
- supporting awarding organisations to create digital assessments.

That said, we need to remember that AI tools are a bit like actors in Casualty: they can learn to use the language of doctors and sound like they have medical expertise, but they cannot perform an operation. They will always need close human supervision. More detailed answers to the questions posed by the DfE consultation are provided below.

Have you or your institution used generative AI in an educational setting? If so, could you briefly describe the ways it was used and the specific tools used.

AQA is an exam board and education charity that sets and marks more than half of all GCSEs and A levels taken in the UK every year. We reinvest our income to help young people fulfil their potential, in particular through cutting edge research into assessment.

We have used generative AI tools as part of our programme of research in this area, and have been exploring these issues for some time.

AQA has found that, in the long term, there is real potential for AI tools to make working life better for teachers and complement human marking, and that there are many potential beneficial uses for AI. However, we have also found that, at present, AI tools still need some refinement.

The main potential benefits of AI as we see them are:

- quicker feedback
- quicker content creation
- improving quality assurance

AI tools could be used to automate simple marking processes to provide quicker feedback, as well as help teachers create informal assessments in class, such as quizzes. This, in turn, would save time for both teachers and awarding organisations.

In addition, currently quality assurance processes for marking high-stakes examinations involve periodically giving work to more than one examiner to make sure that they come out with similar marks. But AI tools could be used to check marks hundreds of times to improve marking reliability.

However, the current limitations of AI tools mean that they are unsuitable on their own to mark or create questions for high stakes exams.

We used ChatGPT, LLaMA and Alpaca to test automated marking capabilities on a range of publicly available science papers not connected to AQA. We found that they were insufficiently accurate for high stake exams.

We also used GPT 4 for a proof-of-concept exploration of how it could be used to create exam questions. This was more promising, but it could not currently be used without significant human oversight for high stakes exams. It could be used to support assessment in the classroom, but this requires further investigation.

AQA has also trialled using ChatGPT to create curriculum summary presentations. These have still needed a lot of work, but there is potential here.

What were the main challenges you faced in using generative AI and how did you address these?

We found there were a number of challenges when it came to using AI tools. These were:

- limitations and inconsistency – they cannot cope with unusual situations and sometimes do not work as expected
- inaccuracies – AI will sometimes generate factually incorrect answers, sometimes known as 'hallucinations'. This could be symptomatic of limitations in the training set or the way in which it interprets the corpus' content
- closed systems – some AI tools are black boxes that are closed to analysis of the way in which they are designed to work
- the potential for bias – real-world data informs current AI, which in turn has the potential to reflect, and sometimes amplify, biases and prejudices in the training data. This highlights the need for reliable and verified training sets to train the AI models.

Take reviews of marking that students and teachers can request after exam results day as an example. In an AQA blog published in March 2021, we said that "a reviewer must be able to explain how they reached a decision, but this is something that AI struggles to do". Although generative AI can produce explanations for marks and attempt to link them to the criteria set out in mark schemes, its effectiveness in doing so depends significantly on factors such as levels of implicit knowledge that mark schemes assume, and the precise ways in which criteria are specified. At present all AI-generated explanations would have to be reviewed by a human expert in any case.

As we said in a blog published earlier this year, "there is a chance that the AI systems simply perpetuate popular myths, as they have no real-world context to draw upon beyond 'what's talked a lot about on the internet'. As a result, their explanations might appear convincing while not being based on facts."

AI tools can produce the appearance of being intelligent but are not themselves intelligent. In this respect, they are somewhat like actors in 'Casualty', as we said in a presentation at the Wellington College Festival of Education on Thursday 6 July: they can learn to use the language of doctors and sound like a doctor, but they are in no way replacements for doctors (let alone able to perform medical procedures or treatments).

More positively, in a research paper published earlier this year, we said that if we can address the issues of AI unreliability, low explainability and bias, then we can, in the long term, put AI-

assisted exams on a par with more traditional exams. This would, of course, require a high degree of public confidence in the underpinning approach.

AQA, at our exam paper processing facility in Milton Keynes, scans millions of exam papers and sends them to examiners every summer. In the long term, AI could have a role in this. High stakes exams are not about to be marked by robots. But we can imagine a future in which professional examiners and teachers use AI tools to help them do their jobs more easily, and focus on what is really important.

The combination of ubiquitous digital devices and rapidly evolving capabilities of AI will transform what teachers teach, how they teach it, and how awarding organisations assess young people.

What was the result of your use of these tools, including any impacts?

AI has a clear potential to help reduce teacher workload, which is crucial when it comes to improving teacher recruitment and retention.

As our Head of External Affairs Reza Schwitzer noted, quoting an Education Endowment Foundation report in a Schools Week column on 12 May (the Secretary of State is right to back AI to reduce teacher workload), marking is the single biggest contributor to unsustainable workload, and "AI could be used to automate simple marking processes and provide better or quicker feedback and information". In this scenario, AI would be aiding teachers rather than replacing what they do; and marking assessments as well as formulating them.

We have found AI tools to be currently limited in the generation of robust exam questions or the marking of responses to questions that require students to provide more comprehensive and detailed answers. But AI has the potential to be of benefit in producing items for assessments that operate in a different model, namely drawing (relatively templated) items from a large item bank to administer to students in a form which doesn't require human marking (multiple choice items or factual responses); many assessments for professional qualifications currently work like this.

We have found that there is definite scope to apply AI to item generation (with appropriate human-in-the-loop quality control and oversight) for formative assessment products, that can interact in a personalised manner with learners to help improve their learning outcomes, provided they are educationally well-designed. They can also be used to support teachers' classroom practice.

This suggests AI will need to improve further in order to provide the full breadth of real-world positive impact in the classroom and broader education.

How do you think generative AI could be used to improve education?

Workload is a big factor in teacher recruitment and retention. The Education Endowment Foundation found, in its 2016 report, that "marking was identified as the single biggest contributor to unsustainable workload in the Department for Education's 2014 Workload Challenge – a consultation which gathered more than 44,000 responses from teachers, support staff and others".

We believe that AI could be used to automate simple marking and provide better and quicker feedback and information. It could help teachers with shorter pieces of feedback and support them to analyse and distribute information quickly to senior leaders and parents.

Indeed, given that marking is only a means to an end (such as certifying a broad level of competence, or providing specific feedback to students), it is possible to envisage applying large language models directly to the (largely linguistic) information generated during teachers' (formal or informal) assessment interactions with students (which may be observational, oral, written, or digital), to achieve these ends more directly. To do this in an educationally valid, robust, and trustworthy way would require a programme of significant cross-disciplinary research and development. At AQA we wish to explore this further, as the benefits for supporting teaching and learning, and re-imagining the fundamentals of educational assessment, could be substantial.

And, as well as marking assessments, AI could also be used to create them. For example, it could help by making a bespoke on-screen quiz on a topic, with the AI tool then marking those responses and offering feedback. This would ease a teacher's workload and enhance a student's learning experience.

Teachers could use AI tools to create tasks. An interactive chatbot could follow a teacher's instructions to create a lesson with audio, video and text resources around a given topic.

AI tools could also be used to generate prompt feedback for peer-assessment, or for automatic feedback in self-assessment. We can envisage a role by which teachers use AI tools with information extraction and summarisation, such as: "write a personalised report for the parent or guardian of student X, based on the following information …" The teacher would use this information as a first draft of a report for the student's parents or guardians.

And, at a macro level, AI tools could be used to help school leaders, as well as policymakers, by providing analysis and insights drawn from learner attainment and performance data more quickly and effectively than is currently possible. This has the potential to provide personalised insights and suggested interventions at a whole school, group or individual level.

What subjects or areas of education do you believe could benefit most from generative AI tools?

Perhaps most obviously, we believe that generative AI could become part of the taught Computer Science curriculum.

That said, we need to consider the extent to which knowledge of AI should be considered as a basic literacy, a part of digital literacy, alongside literacy itself and numeracy. It doesn't seem far-fetched to see AI literacy as another skill that everybody should have. Given how little time in schools, particularly in Key Stage 4, is set aside for ICT and digital fluency, featuring AI tools in the classroom in other subject areas should be encouraged as part of greater uptake of EdTech and digital tools in general.

We are already researching whether new on-screen and on-demand literacy, numeracy, digital fluency and oracy assessments could support young people to be confident and competent in these skills, which they need to ensure they are not excluded from opportunities when they leave full-time education.

Moreover, AI could help students do research. AQA's Student Advisory Group has told AQA that students are using generative and other AI tools to help with study and revision. Such tools can be useful to, for example, summarise large amounts of information.

Some schools abroad are using AI answers as part of critical thinking studies to compare human answers. But confident and independent use of AI by students requires experience and presents problems such as factual errors that harm learning.

Other schools have been using them as part of English literature classes: for example, where pupils ask an AI tool to generate a piece in the style of a particular poet – and then discuss how it does and doesn't match that style, using those discussions to hone their prompts and the poem the tool generates: improving both the students' AI literacy and their literary criticism skills.

All this being said, as we argued in a Times Red Box column, If exams stay, don't fear ChatGPT, in March 2023, we must retain the ability to test what students actually know. Externally marked exams, and the use of exam conditions, are therefore more important than ever in retaining fairness and reliability in how we assess our young people.

AQA would be keen to work with the government, other assessment organisations and schools, to examine these issues further, including how to sensibly encourage a culture of using generative AI for good, whilst also retaining the ability to test whether students themselves understand the subject content they have been taught.

What are your main concerns about using generative AI in educational settings?

Our concerns can be grouped into three categories.

1 – Intrinsic limitations of AI tools

AI tools are currently unreliable and they can make factual errors. If they are unregulated, they have the potential to generate discriminatory, biased or harmful content.

AI tools are closed to analysis of the way in which they work.

Assessors, educators and the government will have to make sure there is no significant impact, legal or otherwise, on the rights of data subjects.

2 – The data used to train AI tools

We need to consider what data is used, who it belongs to, and how it is used. Data protection matters and there are risks such as using training data that may include personal data, and what impacts this might have on a person's ability to control the use of their personal information, object to it being processed, and request that it be deleted.

The legal basis for processing huge amount of personal data in training AI is problematic, and case law is largely yet to be developed.

3 – The use of AI tools

As we have already said, AI tools such as ChatGPT strengthen the case for externally marked exams and the use of controlled conditions in other assessments.

At present, there is a lack of an accountability framework when it comes to the application of AI in education (for example use of data sets to train AI models, arriving at judgements about student and teacher performance, commercial services etc.). We broadly support the government's White Paper about AI regulation; it is important to protect against the potential harms of AI while we explore harnessing its tremendous potential. Moreover, one potential approach would be for AI development to follow broadly the same process as drug patents. This would mean trialling in controlled environments and rolling it out only when approved. With the right will, this could be done at speed, as the pandemic showed.

There is a subtle risk of the tail wagging the dog – reducing education to what generative AI makes possible, rather than using it to do education differently.

Conversely, there is a risk that subjects that require creativity, assimilation and evaluative skills may be devalued if students misuse AI. This in turn will reduce the potential for thinking and wider skill development as part of the learning process, which would deskill future citizens.

If at all, have these concerns impacted your use of generative AI? Please explain how.

Yes – we have been proceeding carefully. AQA has been researching and implementing ethical AI principles.

We do not think that careful and responsible development is a barrier to swift innovation.

Are there specific subjects or areas of education where you believe generative AI should not be used? Why?

AI could be used as a tool when students have to assimilate or evaluate research, or when they have to think creatively or develop problem-solving skills. These are areas where higher-order skills need to be developed for future use. However, teachers and students should be aware of AI's limitations, in particular bias and unreliability, and the need for constant quality assurance. AI should only be used when quality assurance is possible.

Generative AI should not be used when the focus of learning is what generative AI is good at, such as summarising a text. Just as it is right to teach pupils arithmetic even if we can all use calculators, we believe that children and young people need to achieve certain levels of literacy and numeracy to make a success of their lives.

So, using AI to sidestep essential learning like learning to read, interpret and summarise information should be avoided at all costs.

If any, what are your views regarding ethics, data privacy and security when using generative AI in education?

We have been talking about the need for ethical and regulated AI use in assessment since 2020, and we published a research paper, The future of standardised assessment: Validity and trust in algorithms for assessment and scoring, earlier this year. We believe in responsible AI use and follow ethical principles, frameworks and practices for internal AI development when we work with partners.

These are still experimental technologies and there is no actual 'need' for generative AI in education. It follows then that AI ethics, data privacy and security considerations should be addressed before any high-stakes use.

One potential approach would be for AI development and use to mirror the way we trial drugs: discovery and testing in laboratories, trialling under controlled conditions, and rolling out after independent approval – all under regulation.

How do you see the role of generative AI in education evolving in the future?

Generative AI is only going to keep developing. We believe that AI offers strong benefits to education in providing more support for students and reducing teachers' workload. Singapore is already developing AI systems to help with marking and providing learning paths for students.

> ‘The objective is simply to develop increasingly student-tailored curricula that reflect different abilities and learning styles… a technology-driven future where adaptive learning, powered by artificial intelligence, has a major role to play.’
>
> [Extract from AQI’s article - Singapore: Global education pace setter looks for another gear to stay relevant in a challenging world | AQi powered by AQA]

We cannot take for granted that positive AI developments will spread evenly across the country. Without centralised planning or at least a central fund, schools that have the money will benefit the most as they will be able to afford the most advanced systems, with schools with less money left behind. So we need to have consistent quality standards and equality of opportunities.

What support do education staff, pupils, parents or other stakeholders need to be able to benefit from this technology?

A good starting point would be for schools to have equal access to safe and secure services that utilise AI rather than simply what is currently available on the open market. School leaders, teachers and students need to have enough time for training to get familiar with the technology. It is reasonable to assume that educators will increasingly use generative AI tools, so equal access and guidance will be critical for productivity and fairness.

It is also important that teachers, students and parents have the opportunity to learn how to use AI tools safely and productively. That is why we need guidance and training. Parents and teachers will need to understand the systems that students are accessing and guard against malicious software or false information.

What activities would you like to see the Department for Education undertaking to support generative AI tools being used safely and effectively in education?

There is an urgent need to make sure that students have equal access to reliable and verified services that utilise AI through training, exemplification and guidance. Commercial products currently on sale have not been developed for our education system, and they do not necessarily have the best interests of students and teachers in mind. That is why we would call

on the Department for Education to sponsor the development of good AI that is in line with the government's white paper.

Is there anything else you would like to add on the topic of generative AI in education?

AI will continue to develop and so advice, guidance and resources to support its effective educational use will have to be reviewed and updated. The current Joint Council for Qualifications guidance, AI Use in Assessments: Protecting the Integrity of Qualifications, which was only published earlier this year, is already being updated. AI developments continue to progress rapidly, so any educational and assessment systems will have to be future-proofed through principles for effective AI use, as opposed to specific task guidance, related to current products and capabilities.

AQA is happy to be re-contacted about further research on this topic.