Generative AI and Classical Studies

by Neville Morley

How is the advent of so-called 'Generative AI' (GenAI) – tools like ChatGPT that can almost instantaneously produce complex, relevant, accurate and human-like text in response to detailed 'prompts' – affecting the study of ancient history and classical studies at university level, and how should we as teachers respond to this? Over the last two years it's been difficult to ignore the almost apocalyptical rhetoric around this technological development; not just the suggestion that what is currently a glorified auto-complete tool might develop into an existential threat to humanity, but shorter-term claims that it will render assessment by coursework obsolete, undermine the credibility and worth of humanities degrees, and transform every aspect of learning and society. This last point often comes hand in hand with the argument that universities *must* incorporate GenAI into their teaching forthwith, to help prepare students for the future, and should pay large sums to the consultancy firms that are pushing this argument…

Over the last year, I've been running a small project (funded by the University of Exeter's Education Incubator scheme) to explore the impact of GenAI on the assessment of historical skills. I had encountered a number of essays that seemed to show clear evidence that the student had used ChatGPT to generate much if not all of the text: passages alleged to come from ancient sources that didn't show up in any searches and didn't seem at all familiar (and at times were entirely implausible), either without proper citations or with citations that didn't lead to anything remotely similar, and plausible-looking but entirely fake bibliography. I couldn't prove that these were the product of GenAI – the students concerned either claimed that they'd got their notes confused, or didn't engage with the academic misconduct process at all – and it didn't really matter, as the essays were extremely poor by the standards of normal assessment criteria without any need for additional penalties. But it raised the question of whether the practice of using GenAI was becoming more widespread, including among students who might be more competent in tidying up the output before submitting it. At the same time, I was adamantly opposed to the proposed response that we should simply abandon coursework and revert to in-person unseen exams as the only reliable way of 'AI-proofing' our assessment process.

What Students Are Actually Doing

The first phase of the project was to survey students, and follow up the results in small focus groups, to get a sense of how (and how far) they were actually making use of GenAI. I engaged with Exeter undergraduates studying ancient history or history, purely because it was easier to get funding for a clearly targeted project, but I feel reasonably confident that the results are likely to be applicable to students in classical studies and other essay-based humanities subjects. A recent national survey for HEPI did not differentiate between disciplines in its results – and some of the students I talked to did suggest that some of their housemates, studying in social science departments, were far more willing to use GenAI for coursework than they were…

I've published a longer summary of the results of this research in a blog post for the Royal Historical Society, so I won't repeat that in too much detail. The key headlines are that our students are very aware of the advent of GenAI and the hype about its impact, from their friends, housemates and social media; the majority expressed anxiety about its potential impact on their job prospects, and feel in need of guidance and perhaps training – though only a minority argued that it should be integrated into their studies; most have played around with some of the tools for fun. Only a few had any confidence that they understood how GenAI works, though most felt (perhaps over-optimistically) that they could distinguish between its outputs and those written by humans.

When it comes to assessment, we might take heart from the fact that only around 10% of students suggested that they might use GenAI to produce coursework, and most had strong views both about the quality of the output (not analytical, not critical, not written like a student) and the ethical question (half felt that *any* use of GenAI in assessment was cheating). The bad news is that their idea of what 'using GenAI to produce coursework' involves is focused on getting the tool to do the actual writing; a quarter of them would happily ask ChatGPT and its ilk to produce an essay plan for them and to 'read' their work and offer feedback, and nearly half would use it to support their research by generating summaries of topics, debates and individual publications. Indeed, the latter point was put forward as a key advantage of GenAI, that it could 'level the playing field' for students with dyslexia by allowing them to engage with a much wider range of scholarship than would otherwise be possible.

GenAI and Assessment

The second phase of the project was focused on developing and piloting some new approaches to assessment, to explore whether GenAI could be usefully integrated or successfully excluded. The survey data suggested that the use of such tools to generate essays and exam answers was less of a problem than I had expected – or at any rate that other problems were more pressing. It's also the case that the more I've looked at GenAI outputs and engineered my own, the more confident I feel that they can be dealt with simply by setting sufficiently taxing tasks (focus on analysis and interpretation and on engagement with recent scholarship), by requiring proper referencing even in 'take-home' papers, and by applying existing assessment criteria firmly. We can't prove that a given essay was produced using ChatGPT, unless the student has been very careless – there is little reason to believe that AI-powered AI-detectors are reliable – but it is unlikely to be worth more than a bare pass in any case, unless we are excessively lenient markers. It is regularly claimed that the *next* generation of GenAI will overcome all its current problems, especially once it can draw on live data from the internet, but the way in which it works, generating text on the basis of statistical probabilities without any reference to (or concept of) truth, makes this implausible. Its outputs are simply not very good, a bland compilation of conventional ideas (a colleague of mine helpfully refers to it as an 'averaging machine'), and in all likelihood the majority of students to rely on it will be those who are struggling and desperate, rather than cunning cheats.

However, engaging with GenAI does have benefits in its own right. In one of my modules this spring, on Greek Historiography, I offered students the option, alongside a conventional essay (emphasising detailed interpretation and engagement with the sources…), of developing a short GenAI answer and offering a critical analysis of the process and the results. This resulted in some genuinely thoughtful engagement with the topic, and the colleagues whom I asked to evaluate a selection of these essays anonymously (to keep the process completely separate from the marking) both felt the exercise was worthwhile. Perhaps predictably, take-up of this option was low amongst second years, loth as ever to risk their marks by doing anything new, but a substantial number of first years did try this out. I am satisfied that the exercise did test students' knowledge and understanding of Greek Historiography and relevant scholarship, as the basis for the critical evaluation of the GenAI text, and at the same time broadened their experience and analytical skills.

My feeling, supported by some student feedback, is that this would have been improved if I had spent a bit more time in class talking about GenAI and demonstrating how to evaluate its outputs; as I suggest below, this would not necessarily be a distraction from 'proper history'. This was reinforced by the second exercise, this time in a 24-hour take-home paper, where students had to write a commentary on either a tricky passage of modern scholarship or a ChatGPT-generated passage (both selected by me) – the idea being that neither task lent itself to GenAI assistance. This might indeed be the case, but the answers were relatively disappointing; most students who attempted the second exercise disparaged the GenAI output in very vague terms, almost by rote, rather than *showing* how it was deficient as an account of ancient historiography. In retrospect, they needed to have been given more information about the workings of GenAI, and seen more worked examples of critical analysis of its outputs, so they could develop their own.

GenAI and Student Learning

Overall, GenAI seems like much less of a threat to the integrity of our assessment processes than feared – and certainly not enough to abandon decades of research into effective assessment practices. We can limit its usefulness to any students inclined to cheat or cut corners by emphasising analysis and interpretation and requiring the proper support of arguments with evidence and references. GenAI outputs are general, descriptive and bland, when they are not actually fictional, and its assertions are not properly developed or supported.

But this is precisely why we do need to be very worried about the fact that many students are increasingly relying on it for their 'research'. Relying on a GenAI summary of a topic as a starting-point may be no worse than relying on a Wikipedia article or a general piece on the internet; a vague distillation of what people have tended to say about the topic over the last century or so, with no idea of recent research. It is more of a problem if they don't go beyond this, aiming simply to flesh out the points they've been given. I did encounter a significant number of essays this year which presented the same sequence of ideas in relation to a given topic, often with the same examples; this work was not plagiarised in the sense of unacknowledged verbatim copying of existing text, since the wording was often quite different, but clearly it followed the same basic structure without question (or acknowledgement).

Still more alarming is the idea that many students are relying on AI summaries of publications as a key part of their research process; indeed, some colleagues are being encouraged to direct students, especially those with dyslexia or other learning issues, towards tools like Scholarcy. More research is needed on this issue, but there are already indications that such summaries are not reliable; 'abstractive' summarisers based on GenAI can introduce ideas and arguments that are not present in the original, while 'extractive' summarisers like Scholarcy use only sentences and phrases from the original but present these as decontextualised pieces of information rather than offering a clear account of an article's overall argument.

Further, this approach indicates that students (and those promoting such tools) may think of reading as just a matter of extracting content from a text, something that can be delegated to a bit of software to save time – that their task is to offer points from as many pieces of scholarship as possible, without necessarily having a sophisticated understanding of any of them. In other words, AI and GenAI may be undermining our students' learning, and limiting their acquisition of key skills of comprehension, analysis and interpretation, rather than undermining our assessment of these.

What Can We Do?

We can't stop students using GenAI if they choose; currently, at least, it is too pervasive and easily accessible to most of them. What we can do is make it clear why we think they shouldn't; not just that it's against the rules, and ethically questionable, but simply because it's rubbish. This means that we need to be sufficiently familiar with it ourselves, to be able to offer a plausible or authoritative account of its limitations and shortcomings, to counteract the breathless hype students are receiving from friends and social media and the apparently self-confident, assertive rhetoric of GenAI outputs themselves. We need to show rather than tell; general assertions about its deficiencies will be less effective than detailed analysis of concrete examples – which doesn't have to be a digression from the main subject matter of whatever we're teaching, but can be used as an exercise in using material already learnt as the basis for critique, as well as modelling how to go about such a critique.

We also need to understand why students might be turning to GenAI, and to think about whether there are things we can do differently here. Are we failing to teach them proper research practices or to explain why they need to develop reading skills? Are we giving them the wrong idea about how they should research topics – creating an impression that they need lots of references regardless of how well they understand them? Are we failing to give them the feedback they need, so they turn to a bullshitting text generator instead? By addressing such questions, we have the opportunity to develop our pedagogical practices in ways that can preserve the integrity of teaching and assessment and properly prepare our students for their future careers.

Neville Morley, University of Exeter
N.D.G.Morley@exeter.ac.uk