Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular article

Using generative artificial intelligence to enhance the performance of disadvantaged students in secondary education

Ryan J. Brunton[a,e], Soukaina Rhazzafe[b], Raymond Moodley[c], Stefan Kuhn[d,c,*], Fabio Caraffini[e], Sara Wilford[c], Rachel Higginbottom[c], Simon Colreavy-Donnelly[b], Mario Gongora[c]

[a] *Ditton Park Academy, Berkshire, UK*
[b] *CSIS, University of Limerick, Limerick, V94 T9PX, Ireland*
[c] *School of Computer Science and Informatics, De Montfort University, Leicester, UK*
[d] *Institute of Computer Science, Tartu University, 51009, Tartu, Estonia*
[e] *Department of Computer Science, Swansea University, Swansea, SA2 8PP, UK*

ARTICLE INFO

Keywords:
Disadvantaged pupils
Pupil premium
Feedback
Education
Large language models
Artificial intelligence
Generative AI

ABSTRACT

We show that generative AI can support disadvantaged students, improve grades, and help close the attainment gap between pupil premium (PP) and students with special education needs (SEN). It can also alleviate teacher workload, especially for PP and SEN students, by minimising marking and feedback time, enabling better lesson planning and interventions, which can enhance teacher retention and staffing. We focus on disadvantaged students with SEN and low-income families and use AI for personalised feedback and lesson planning in arts and humanities. This enables school leaders and parents to view the qualitative and quantitative student progress. The results of this study demonstrate the potential of using AI-based systems to help close the attainment gap between disadvantaged students and their peers. The intervention given to these pupils would have been an unreasonable demand on the current teacher workload in the UK.

1. Introduction

It is well recognised worldwide that there is an achievement gap between disadvantaged pupils and their ‘non-disadvantaged’ peers Chmielewski (2019). In the United Kingdom (UK), a study conducted by The Education Endowment Foundation (EEF) in 2022 found that the overall achievement of disadvantaged students, that is, those eligible for student premium funding (see Section 2.1), lagged 19 months compared to their peers who were not eligible for it Edovald and Nevill (2021). Likewise, in Key Stage 4 (for pupils aged 14 to 16 years), among those who qualify for free school meals (a criterion for pupil premium funding), only 28.4 % managed to score a grade 5 or higher in English and mathematics, in contrast to 55.4 % of their counterparts not receiving pupil premium Edovald and Nevill (2021). Research by Third Space Learning Third Space Learning (2024) outlines several drivers for the attainment gap, notably parental participation, the digital gap, and the teacher’s role. With regard to the teacher’s role, the study highlights evidence from previous works that suggests that one-to-one tuition and mentoring, tailored to the pupil’s needs is one of the most effective ways of closing the attainment gap. Indeed, part of this process will require the teacher to conduct a lesson with the pupil, evaluate the pupil’s work and provide tailored feedback in a timely manner. However, given resource constraints and class sizes in state-funded schools in the UK, this is not always possible Edovald and Nevill (2021); Third Space Learning (2024).

Feedback in education is deeply embedded in the social construct of schooling, with teacher evaluations of student work being a regular occurrence. Following evaluation, in the UK, student performance is captured using vague quantitative data that can often be confusing to teachers, parents, and students alike. This can create barriers for parents to support their children effectively and for teachers to effectively target areas for improvement; as a result, these parents are sometimes known as ‘hard to reach’ Harris and Goodall (2008). Educational research

[*] Corresponding author at: Institute of Computer Science, Tartu University, 51009, Tartu, Estonia.
Email addresses: r.brunton@dittonparkacademy.co.uk (R.J. Brunton), soukaina.rhazzafe@ul.ie (S. Rhazzafe), raymond.moodley@dmu.ac.uk (R. Moodley), stefan.kuhn@ut.ee (S. Kuhn), fabio.caraffini@swansea.ac.uk (F. Caraffini), sara@dmu.ac.uk (S. Wilford), rachel.higginbottom@dmu.ac.uk (R. Higginbottom), simon.colreavy@ul.ie (S. Colreavy-Donnelly), drmgongora@gmail.com (M. Gongora).

https://doi.org/10.1016/j.ssaho.2025.102110
Received 14 March 2025; Received in revised form 11 October 2025; Accepted 14 October 2025
Available online 29 October 2025

considers feedback to be the most effective strategy to improve student outcomes Hattie and Timperley (2007). However, a 2016 TES survey revealed that marking and subsequent feedback can take up to 11 hours of additional work per week for teachers. This high workload has contributed to burnout and the shortage of teachers in the profession Ward (2016).

Recent studies have shown that artificial intelligence (AI) has the potential to revolutionise education by supporting the role of the teacher in evaluation and feedback Adıgüzel et al. (2023). One of the key advantages of an AI system is its ability to provide real-time feedback that is both quantitative and qualitative. Thus, teachers can receive timely information on their students' progress and areas for improvement. The AI system can also provide personalised feedback to individual students, allowing teachers to provide targeted support and interventions. This real-time feedback can also help parents better understand their child's progress, their specific strengths and areas needing improvement Povey et al. (2016). As a result, an AI system could address some of drivers of the attainment gap highlighted earlier, notably parental involvement and the teacher's role. However, AI systems are not without their pitfalls, with the most common issue being the quality and effectiveness of the AI system in providing reliable and meaningful feedback. For example, in Ma (2023), the researchers found that in higher education history courses, the use of OpenAI's ChatGPT Large Language Model (LLM) had an accuracy of 90 % when translating ancient texts and historical facts.

It is against this backdrop, that this study aims to explore whether AI systems can alleviate the burden of marking and feedback for teachers, and improve outcomes for students, particularly those from disadvantaged backgrounds and those classified as having special educational needs and disabilities (SEND). Specifically, we ask if AI-generated feedback has helped narrow the grade gap for disadvantaged students. We answer this question by examining a sample of year 10 history students. The hypothesis to be tested is that personalised feedback provided by AI will lead to an improvement in student performance, measured by grades achieved. Government legislation since 2011 has been focused on providing extra funding to schools to help them close the attainment gap, with schools receiving ¥1050 per disadvantaged pupil and costing the UK government ¥1.65 billion in 2023 Department for Education (2024). Indeed in 2019, the UK Prime Minister announced his government's flagship 'Levelling-Up' policy which included several targets for education UK Government (2024b). As part of this study, research was conducted in a secondary school in England, with the sample students all being classified as SEND, varying from mild learning difficulties to autistic spectrum disorder (ASD). It has been shown that SEND students require a lot more personalised feedback, which enables them to make equitable progress compared to their neurotypical peers Keen et al. (2016).

The importance of feedback for learning has been widely emphasized and is a central issue in learning research and theory. A broad distinction can be made between behaviourist approaches which observe the reaction of the learner to feedback stimulations and constructivist approaches, which emphasize the role of the learner in producing and absorbing feedback Molloy and Boud (2014). Whilst we present statistical results about the effects of AI-generated feedback here without discussing the students' learning process in detail, we maintain a constructivist view of learning and see the role of the AI system as helping students to build their knowledge.

In this paper, students were given four assignments. ChatGPT was configured to assess the students' submissions, provide a grade, give feedback, and set a task for the students to work on as part of the post-assignment remedial activities. Since the subject under consideration in this study is history and the assessments were largely essay-based, this paper examines situations in similar subjects, mainly in arts and humanities. The students' progress and attainment following the use of AI were compared with their previous progress and attainment. In principle, if an AI-assisted learning management system could improve pupil outcomes, then this should give a statistically significant improvement in grades. Further, from a teacher's perspective, the AI system could help reduce workload, and create a virtuous circle, i.e., improve recruitment and retention of teachers, provide higher quality teaching and learning, and enhance pupil outcomes. Skinner et al. (2021).

The following sections concisely present the rationale, research question and hypothesis as well as the aim and objectives of the study.

1.1. Rationale

As outlined in the introduction this study sought to address two critical problems: firstly, poor performance of SEND students and those from disadvantaged back-grounds which leads to the widening attainment gap between these students and the rest of the cohort. Secondly, teacher workload is increasing due to ongoing enhancements to policy and outcome targets which is impacting teacher retention. The proposed solution is to improve efficiency, quality and personalisation of feedback by using an AI based system. In this way, the application of AI enables a level of feedback that is difficult to achieve otherwise due to the considerable effort required, helps close the attainment gap, and helps reduce the workload of teachers.

1.2. Research question and hypothesis

Based on the study rationale, the research question for this study is: how might the use of AI enhance the performance of disadvantaged students in secondary education, notably in an arts and humanities context? Given this, the hypothesis of this study is: Using an AI-based system to provide personalised feedback to students leads to improved student outcomes as measured by grades achieved.

1.3. Aim

The aim of this study is to investigate whether the use of AI systems in teaching and learning can help improve outcomes for students, particularly those from disadvantaged backgrounds and those classified as having special educational needs and disabilities (SEND). This research is timely since, despite the growth in the field of AI in education, studies demonstrating significant, measurable academic gains for disadvantaged students at the secondary level are scarce.

1.4. Study objectives

The primary objective of this study was to investigate whether the use of an AI-based system to provide personalised feedback to students improves outcomes for students, notably those classified as pupil premium and/ or students with special educational needs.

The secondary objectives of this study were:

- To evaluate students' work and provide customised feedback using an AI based system .
- To assess the impact of AI usage on alleviating teacher workload and improving retention by reducing marking and feedback time.
- To investigate to what extent AI can help close the attainment gap between pupil premium students and students with special educational needs.
- To demonstrate that a richer qualitative and quantitative student feedback enhances the usability and understanding for teachers, students and parents.

1.5. Paper organization

The remainder of the paper is divided as follows.

- Section 2 surveys the literature to provide background information and presents essential definitions for the English education system.
- Section 3 describes the methodological framework used in this study.
- Section 4 presents the results.
- Section 5 discusses the results.
- Section 6 draws the conclusions of this work.

2. Background and definitions

2.1. Terminology

This investigation is situated within the framework of the English educational system, in which several definitions are routinely employed. We adopt these definitions and elucidate them in this section for the sake of clarity and the benefit of readers unfamiliar with the system.

Pupil premium (PP)

PP is a grant given to schools for pupils who fall into at least one of the categories listed below.

- Pupils who are recorded as eligible for Free School Meals (FSM), or have been recorded as eligible in the past 6 years (referred to as Ever 6 FSM).
- Children previously looked after by a local authority or other state care, including children adopted from state care or equivalent from outside England and Wales Department for Education (2024).

PP is often used to define a group of pupils rather than a single individual.

English as an additional language (EAL)

EAL describes a diverse and heterogeneous group of learners who speak English as an additional language. In England, such learners are defined as those who have been 'exposed to a language at home that is known or believed to be other than English' Department for Education (2020).

Special educational needs and disabilities (SEND)

SEND is a term used to describe learning difficulties or disabilities that make it difficult for a child or young person to learn compared to children of the same age.

Technologies

Learning Management Systems (LMS) are platforms designed for the administration and delivery of educational content and activities online. They are used by educational institutions, corporations, and various other entities to facilitate access to online learning resources for students or employees. Canvas and Moodle are prominent examples of open source LMS options widely used worldwide Firat (2023).

Intelligent tutoring systems (ITS), in contrast to traditional classroom teaching and massive open online course (MOOC) platforms, prioritise automating and personalising the learning process for students, which requires several essential components, including automatic grading models and personalised guidance tools Almasri et al. (2019).

2.2. AI in education

AI refers to computational systems that are capable of mimicking human behaviour and intelligence to make decisions McCarthy (2007), showing human-like ability to extract insights from past experiences and outcomes Hooda et al. (2022). The use of AI is emerging as a transformative force in redefining pedagogical approaches and addressing the diverse needs of learners. Educators are increasingly exploring the potential of AI to promote educational transformation, revolutionise classroom dynamics, and optimise student outcomes. Further, with AI applications in student performance prediction and intelligent tutoring systems among many others, there are numerous issues and concerns that need to be addressed through the implementation of AI ethics principles that should be built into the systems from the outset Bahroun et al. (2023); Crompton and Burke (2023); Yu (2024).

The work of Moodley et al. (2020) proposes a method to forecast student performance using an AI approach based on data mining and probability theory to address persistent absenteeism—a significant barrier to a pupil's educational and personal growth. In this study, the authors identified specific days and dates with low attendance rates and proposed recommendations to implement effective interventions.

A recent study by Muthuselvan et al. (2024), introduces a novel method for predicting student academic performance through advanced data analysis and AI techniques. By integrating diverse factors such as demographic attributes (e.g., gender), educational indicators (e.g., SSLC scores), social factors (e.g., number of siblings), and logistic details (e.g., travel time). This method offers a comprehensive view of student academic progress, empowering educators to tailor interventions and support strategies effectively.

The RadarMath system developed by Lu et al. (2021), stands out in the landscape of ITS educational technology for its innovative fusion of AI-driven automation and personalised guidance. By integrating these features, it creates an interactive and adaptive learning environment that enhances mathematical comprehension and improves learning outcomes. The platform employs AI models to optimise assessment processes, offering automatic grading services for both text- and formula-based maths questions. Additionally, its personalised learning guidance, powered by an education-centric knowledge graph, customises learning paths to each learner's specific knowledge state.

Beyond that and with the use of Generative AI, Intelligent Teaching Systems (ITS) can now generate personalised course content and plans, adjusting difficulty and content based on student performance, while also providing real-time feedback and facilitating interaction with parents and teachers for comprehensive educational support Khosravi et al. (2022); Yu and Guo (2023).

2.2.1. Generative AI for education

Generative AI is a set of computational methods that use AI technologies and algorithms to recognise patterns and relationships between data points to create novel and meaningful content, such as text, images, and audio Feuerriegel et al. (2024). LLMs are a type of generative AI that have been extensively trained on large text sets to produce new text data. These models have shown remarkable efficacy in various text processing tasks, including but not limited to language translation, question-answering, and text generation Zhou et al. (2024).

A survey conducted in May and June 2023 by Ghimire et al. (2024), at a medium-sized research university in the western United States, suggested that the idea of employing generative AI models and large language models (LLMs) as a personalised tutor tailored to individual learning needs received enthusiastic support from the respondents. Faculty members from across the university participated in an investigation of educators' perspectives on the integration of generative AI tools and LLMs in educational practices. Using a mixed-methods approach, the survey collected quantitative data using the Likert scale Likert (1932) and qualitative insights through interviews, providing a comprehensive understanding of educators' attitudes towards these technologies.

Although the survey revealed a considerable level of enthusiasm and optimism among faculty members regarding the potential of generative AI tools and LLMs to revolutionise education, it also uncovered some concerns, such as the potential for automation to overshadow critical thinking skills and the challenges in maintaining academic integrity in AI-driven learning environments. However, respondents recognised several benefits, including the potential to automate repetitive tasks, such as grading assignments, offering personalised feedback, crafting test questions, and even identifying flaws and biases in students' arguments, thus enabling educators to devote more time to meaningful aspects of teaching, as well as enhancing student engagement through personalised learning experiences.

These findings have important implications for educational policy, curriculum development, and future research in the field, as educators navigate the opportunities and challenges presented by AI integration in education.

While the survey did not focus on one specific LLM, many prominent models and applications were considered, such as ChatGPT, powered by the GPT model, which has gained widespread attention for its remarkable ability to produce coherent, structured, and informative answers, and has become the fastest-growing user application in history, achieving 100 million active users by January 2023, a mere two months after its introduction Lo (2023). Other popular LLMs include Gemini, Llama, BERT, and Phi-3 John Rush (2024).

2.2.2. ChatGPT for education

A review of ChatGPT's uses in educationJauhiainen and Garagorry (2023) found that its application has predominantly been studied in higher education, focussing on creating teaching materials and assessment tasks.

Research has shown that ChatGPT can help teachers organise exercises, activities, and quizzes and obtain information tailored to students' learning styles Rahman and Watanobe (2023). Its ability to support teachers by reducing their workload is particularly significant. Automating the creation of educational materials and assessments, allows educators to dedicate more time to direct student interaction and personalised instruction Karaman and Göksu (2024); Rahman and Watanobe (2023). It can explain solutions to complex problems, generate new exercises, and provide instant feedback, thus saving valuable time for teachers, as shown by Kasneci et al. (2023); Liang et al. (2023). In instructional design, ChatGPT offers a new dimension by providing instant feedback on the design and planning of educational activities, including course curricula, schedules, lesson plans, and evaluation studies Farrokhnia et al. (2024); Lo (2023); Zhai (2022). Its ability to create innovative materials and organise multimedia presentations, encourages active learning and interaction between teachers and students Mondal et al. (2023).

These capabilities make ChatGPT a viable instrument for modernising educational methodologies and enhancing learning outcomes. This is substantiated by the study conducted by Hashem et al. (2023), which demonstrated a significant increase in student motivation regarding listening skills as well as an increased interest in learning. These findings imply that ChatGPT can play a pivotal role in improving student engagement and optimising educational results.

However, as mentioned in Hashem et al. (2023), the initial responses of ChatGPT may lack depth, so providing detailed and precise instructions can improve the clarity of its responses and reduce the chance of oversimplification. Well-designed prompts demonstrate how collaboration between AI and human expertise can ensure that the context of tasks is maintained and enhance the quality of generated responses. This highlights the importance of balancing the input of ChatGPT with that of teachers, recognising their crucial roles in integrating AI into lesson planning, while also preventing burnout.

2.2.3. Bridging learning gaps with ChatGPT

In recent years, intensive, in-person tutoring has demonstrated significant positive effects on students' learning outcomes at a reasonable cost. During the COVID-19 pandemic, learning deficits were particularly pronounced for disadvantaged children, of low socioeconomic backgrounds, who were more affected by school closures than their more advantaged peers. These children face disadvantages in access to digital learning technology, the quality of their home learning environment, the support they receive from teachers and parents, and their ability to study autonomously Betthäuser et al. (2023). The global negative impact of the pandemic on education highlighted the effectiveness of tutoring programmes in narrowing the educational gaps that widened during this period. Therefore, many governments, including the UK, introduced online tutoring to counter these negative effects Hupkau et al. (2024).

By providing scalable and personalised learning assistance, ChatGPT can help address these educational challenges Sullivan et al. (2023). It can offer real-time support, answer questions, and explain difficult concepts in a variety of subjects, making it a valuable tool for students who may not have access to in-person tutoring. It can also adapt to each student's learning pace and style, providing tailored resources and exercises to reinforce understanding. Furthermore, ChatGPT can be available outside of traditional school hours, offering students flexibility and additional support whenever they need it. This level of personalised and accessible assistance can help bridge the gap for disadvantaged students, ensuring that they receive the support necessary to succeed academically.

2.2.4. Issues

As we have seen, AI holds great promise for improving educational outcomes. It raises questions about data privacy, plagiarism, the role of human agency in the learning process and other considerations related to the ethical and practical implications of its adoption. As educators navigate these complexities, it is imperative to approach AI integration thoughtfully, ensuring that it aligns with pedagogical goals and ethical principles.

Most of today's AI applications revolve around machine learning, where programming skills are not directly needed. Instead, these applications learn and adjust based on user behaviours and trends, making data essential for their functioning Tahiru (2021). However, this also means that issues like privacy, trust, ethics and social impact must be carefully considered when using AI in education Internet Society (2017), as it poses several ethical risks to students' personal information, including privacy infringement and data leakage risks. This can lead to the disclosure of personal information due to the secondary exploitation of students' private data and network fraud caused by data trafficking Huang (2023). Therefore, it is critical to implement strategies to enhance student personal data protection, which include various approaches, such as raising awareness among students about self-protection through targeted classroom sessions, involving them in case studies, and engaging them in activities related to managing their data, providing them with legal remedies against infringement of data privacy and improving IT industry self-regulatory mechanisms to develop a consensus about user privacy protection in line with GDPR and data protection legislation Huang (2023).

As AI develops and becomes more ubiquitous, with students expecting an instant response to a question from a sophisticated chatbot, the role of the human teacher may change to become a facilitator and supporter, rather than teaching students directly. As a cost saving exercise, the use of AI may be attractive to education budget managers. The current furore and concerns regarding the employment of physician associates to more cheaply 'replace' doctors with less qualified practitioners Kmietowicz (2023), may indicate a possible future whereby teaching associates support students rather than qualified teachers, who may be retained for producing updated content for the AI system to deliver, but whose numbers will likely be small relative to the associates who would provide the bulk of the direct student contact.

Another concern is students cheating on their non-exam assessments or coursework. Anecdotally, students using ChatGPT and other AI systems can be identified through the almost perfect grammar produced by these systems, the 'made up' references and the inaccuracies that are both plausible and incorrect. Whilst these issues are being explored, with technology being developed to stop plagiarism, this is still in its infancy, and the more convincingly human the AI gets, the more difficult it will be to identify. One way that the issue of plagiarism can be overcome is by students being educated on the effective use of AI to support and engage their learning, rather than relying on AI to create work without a learning process. By using AI as a tool, students can extend their learning towards greater understanding of the topic. However, it must be acknowledged that some students will cheat, game the system, or act strategically to improve their results and that no amount of education or training can prevent bad actors from attempting to circumvent the rules.

Digging deeper into education, individual subjects are looking at how their area is being impacted by AI. In Cooper (2023), the authors found that the use of LLMs runs the risk of "creating a single truth" where research is assumed to be correct without proper grounding. An additional issue is that there is little room for innovation and new knowledge. This is particularly concerning when it is considered that all LLMs rely on previous knowledge in order to learn, and if relied upon by students and academics, may result in new knowledge being reduced in favour of re-using and re-purposing old knowledge.

Furthermore, while Generative AI tools, like ChatGPT, excel in their efficiency and provide rapid feedback, they exhibit some limitations. In a study conducted at the University of Queensland, Australia Li et al. (2024), ChatGPT was instructed using a set of prompts, to mark the assessment of college students' coding assignments and a self-reflective essay based on a marking rubric. Results showed that ChatGPT can increase written assessment marking efficiency, reduce costs, and potentially decrease bias by offering a moderating perspective. However, it struggled with nuanced or creative responses, often producing variable results even with clear rubrics and prompts. ChatGPT was able to evaluate both coding and reflective assessments and distinguish between different quality assignments, demonstrating high consistency and accuracy for higher quality assessments, similar to human marking. It was also able to generate detailed justifications based on the rubric and assessment task description. For lower-quality submissions, however, the model showed reduced accuracy and inconsistency in aligning feedback with marks, likely due to the overall difficulty of discerning borderline pass/fail assessments while marking. Additionally, its reliance on training data introduces challenges in addressing context-specific or subjective tasks, unlike the more objective coding task for example. These findings highlight the importance of combining AI systems with human oversight to ensure meaningful, reliable and pedagogically sound feedback, making AI a powerful tool for enhancing education.

A broader survey Xia et al. (2024), that reviewed 31 articles on Generative AI's role in transforming higher education assessments, found that 26 % of the reviewed articles emphasized the need to balance AI-assisted methods with human-centred approaches. This balance allows human-centred assessment to focus on areas where Generative AI falls short, such as higher-order thinking skills like creativity and critical analysis, while taking advantage of its strengths in tasks like improving writing skills through proofreading, critique and editing. Additionally, the survey highlighted that Generative AI tools offer significant benefits including perceived unbiased feedback, immediate and diverse responses as well as enhancing students' self-assessment capabilities. However, it also poses a significant challenge to academic integrity by raising ethical concerns such as potential misuse.

Given these ethical challenges, the need for regulation is clear. Accordingly, UNESCO has published guidelines calling for the regulation of Generative AI in education UNESCO (2023), to address the problems that Generative AI might pose to children, as well as to support educators in the best way. Although regulation is challenging, it highlights the critical need to develop strategies to maximise the effectiveness of Generative AI and the use of tools like ChatGPT in the classroom. For this study, we do not aim to evaluate particular strengths and weaknesses of ChatGPT, nor do we aim to compare ChatGPT to other large language models. This is justified since the general results should be transferable because all Generative AI models have similar aims and ChatGPT is generally considered a leading product.

Looking at recently published literature (from 2024 and 2025) not included in the review Xia et al. (2024), we note that their focus is different from our study. Er et al. (2025) compares human feedback with AI-generated feedback and finds human feedback superior. In contrast, we are in a scenario where AI allows one to increase the amount of feedback, given the resource constraints in schools. Banihashem et al. (2024) compares peer feedback with AI-generated feedback in a higher education setting, not in a school, as in our case, with a wide range of students. Peer feedback is the topic of Shin et al. (2025) as well. Here, secondary school students receive peer review and AI-generated feedback, showing that peer feedback is generally preferred by students. Meyer et al. (2024) compares giving no feedback, but still asking students to do exercises, with giving AI-generated feedback. The study compares cognitive and affective-motivational outcomes. In contrast, in our study students do additional exercises and we look at marks in national GCSE exams. Nazaretsky et al. (2024) examines if students value different feedback differently, showing that AI-generated feedback, if revealed as such, is considered less valuable by students than feedback of unknown origin.

Looking at these papers, our study is novel in so far as it focuses on disadvantaged students and measures the influence of AI-generated feedback on marks in a national examination system.

3. Methodology

3.1. Research methodology

The study followed the well-recognised Action Research methodology, as outlined in Coghlan (2019). This approach was selected because the lead researcher was conducting research in his own organisation and focused not only on being able to describe, understand, and explain the issue, but also on using the data to resolve the underlying issues, thus improving the outcomes for the participants, the researcher, their organisation, and society at large.

This section details the experimental approach used in this study, with the data generated presented and discussed in Section 4.

3.2. Proof of concept

3.2.1. Sample and approach

An initial proof-of-concept experiment was conducted to understand the viability of the larger study. As this is a pilot study, variability among students and teachers for example, in the nature of the interventions and support, is revealed in the data. Using purposeful sampling, a sample of nine underperforming, disadvantaged students in Year 10 (15 years old) were given four assignments to complete in their History course, a month before their year-end exams. The purposeful sampling technique was used as studies have shown it to be effective when resources are limited and when selecting individuals or groups of individuals that are especially knowledgeable or have experience in the phenomenon of interest, in this case, experience of being disadvantaged and/or having special education needs, and underperforming at school Palinkas et al. (2015). In addition to this sample being selected as students met the criteria of being disadvantaged pupils who were underperforming, they also met the criteria set out by the UNESCO guidance of students being over the age of thirteen when interacting with generative AI systems UNESCO (2023). The UNESCO guidance for generative AI in education and research emphasizes the importance of a 'human-centred approach'; therefore during this research, students were regularly checked on and supported with the use of the AI-generated content Miao and Holmes (2023). To ensure the 'human centred appoach' all feedback was checked by the teacher for quality assurance but not edited in any way, and the students had time to speak to the teacher to clarify interpretation of feedback. ChatGPT was optimised to grade assignments by developing appropriate prompts, provide relevant, customised feedback, and provide a remedial or extension task for students to complete to improve their writing, knowledge, and overall grade. The control group was the rest of the year group studying GCSE History. These students received the same lessons as the experimental group; the experimental group was given an added task as an intervention which was then marked by AI. The control group from other classes had the same teachers and resources (apart from the AI intervention). There were also PP and SEN students in these classes. Some general information about the demographics of the school in question can be found in UK Government (2024a).

3.2.2. Outcome and lessons learnt

Unfortunately, these nine students did not complete a quarter of the assignments, and most of them did the assignments the day before the assignment was due. Consequently, the students performed poorly in their exams (Summer year 10 mock exams). This was not unexpected, as the student sample was chosen because they had known barriers that prevented them from progressing as expected.

As a result, the researchers realised that more needed to be done for them to improve their performance in their GCSE exit exams, which were to be held the following year. Students were spoken to about their exam results and were told about the opportunity they were part of, including how AI could help them if they completed the assignments on time. Parents were also contacted for support and given a detailed presentation of the study, including the aims and objectives, and how this could improve their child's learning outcomes. Parents and students were also reminded that AI is just a tool, and ultimately it is the teachers and other stakeholders who play a key role in developing the student's attitudes and skills. The students were more engaged and wanted to gain feedback from AI based on the poor exam results they had achieved a few months before. This could be a common pattern in maturity and developing study skills prior to a set of exams.

3.3. Main experiment

In light of the issues encountered with the proof-of-concept study, the main experiment was conducted as follows:

1. At the start of the academic year (September), the same nine students who were part of the initial study were asked to reflect on their previous year's exam grades, where they had failed to submit assignments to be marked and respond to feedback given by the AI system.
2. For the main study, consent was obtained again from both the students and their parents.
3. Over a six week period, students were given four extra assignments in GCSE History. These were:
   - Life in 1920s Germany
   - Stresemann and recovery
   - Explain why there were challenges to the Weimar Republic
   - Nazi control and the road to dictatorship 1933–1939
4. Assignments were spaced two weeks apart to allow students to write an essay and to have time to react to the feedback generated by the AI system
5. The AI system leveraged OpenAI completions (Playground) turbo 3.5 which was trained using mark schemes and other relevant assignment criteria
6. To ensure student anonymity, their names were removed from their assignments upon submission and changed to student letter codes, e.g,. 'Student A' this meant their names were not shared with 'OpenAI' and from this point all data stored for the experiment was anonymised.
7. Students submitted their assignments on Microsoft Teams, which were then copied into the OpenAI Playground to review the students' answers and provide feedback
8. Students then used the feedback to make improvements or to expand their knowledge where it was found to be lacking
9. Students took their winter mock exams (year 11 mock exams) in December, which were marked by the school's history department in the usual manner for all students
10. The data from these mock exams were then compared to exams that were taken six months earlier (Summer year 10 mock exams)

Fig. 1 gives a schematic view of the research workflow.

It should be noted that the main experiment was completed using largely a positivism paradigm in that the lead researcher did not teach or have direct lesson contact with the students, but processed their work using the AI system and submitted the system-generated feedback to the students without providing any additional thoughts or inputs to both the students and teachers. The results of the main experiment, which was conducted in July and December 2023, are discussed in Section 4.

3.4. Ethical considerations

Whilst the use of AI may appear to be the panacea for the future of learning, it should be noted that there are significant practical and ethical issues that should be considered. Firstly, there are huge resource and environmental implications for the sustainable use of AI Van Wynsberghe (2021). In the context of this study, and beyond the wider implications, this also includes the cost of training the system, teacher time allocation for oversight and the provision of technical support. This begs the question of whether the 'human centred' approach may be resource and time costly and could add significant pressures on teachers. This may particularly be the case where individualised learning applies to large class sizes. However, it is important to also note that the use of

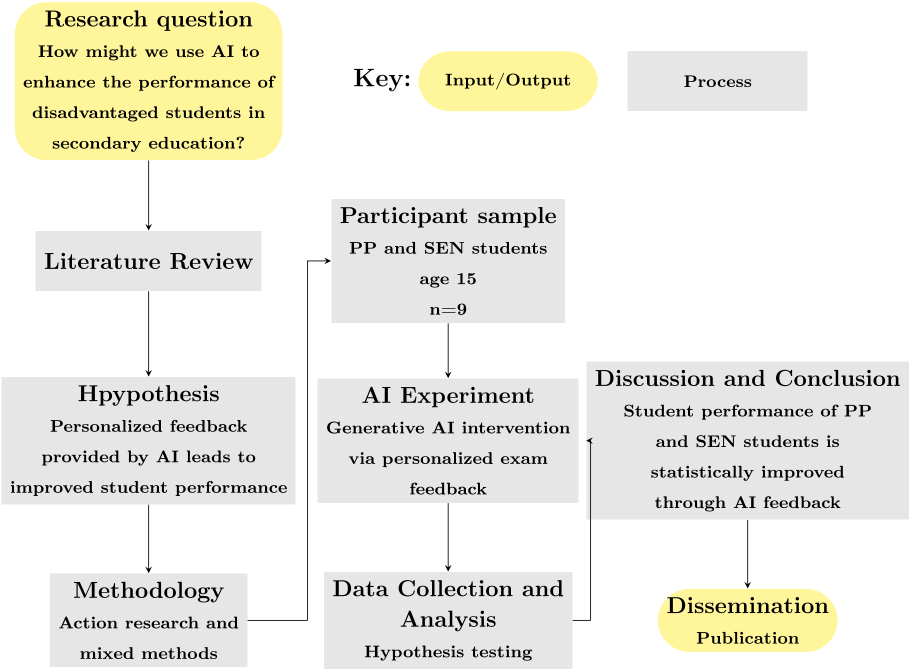
Research question
How might we use AI to enhance the performance of disadvantaged students in secondary education?
Key:
Input/Output
Process
Literature Review
Participant sample
PP and SEN students
age 15
n=9
Hpypothesis
Personalized feedback provided by AI leads to improved student performance
AI Experiment
Generative AI intervention via personalized exam feedback
Discussion and Conclusion
Student performance of PP and SEN students is statistically improved through AI feedback
Methodology
Action research and mixed methods
Data Collection and Analysis
Hypothesis testing
Dissemination
Publication

Fig. 1. Flow chart describing the research process and AI intervention for this study.

Table 1
Results of students from the main experiment in year 10 and year 11 mock exams.

Student		PP	EAL	SEN	Year 10 Mock (July)	Year 11 Mock (December)
Code	Gender					
A	F	Y	Y	Y	U	4
B	F	Y	Y	N	U	6
C	F	N	N	Y	U	5
D	M	Y	N	Y	2	4
E	M	Y	Y	N	4	5
F	M	Y	N	Y	2	3
G	F	Y	N	N	U	3
H	M	N	Y	Y	2	7
I	F	N	Y	Y	2	7
mean±std					1.3 ± 1.33	4.8 ± 1.45
median					2	5
interquartile range					2	4

AI for providing feedback, streamlining administration and supporting continuous learning may reduce a teacher's workload Uddin (2024).

The key here is how well AI is integrated into existing systems, the support and training offered to teachers, and the avoidance of reliance on AI to provide answers for everything, which may impact the ability of students in critical thinking and analysis Cooper (2023). Further, there may be issues around deep learning (or lack thereof) through the quick and observably comprehensive (although not always accurate) responses to relatively simple prompts. The use of AI tools, particularly LLMs provides tempting opportunities for plagiarism. Students find that outputs from AI are perfectly written, with excellent punctuation, grammar and structure, even if the detail may be lacking in accuracy. Students may not have sufficient prior learning to discern the accuracy of the AI outputs and may utilise them uncritically. This means that not only are students in danger of not learning much, but they are also subject to the perception that AI can provide answers better than they can (and therefore gain higher grades). This dishonesty in passing off AI produced work as their own, not only undermines the ability of the education system to assess students according to ability, but it also means that there is a possibility of a significant number of students passing assessments in subjects they have little or no understanding of.

In addition, the lack of privacy and concerns about GDPR and fundamental rights issues when using AI should be considered. The ability of AI systems to monitor both students and teachers, including student identification, progress, grades and so on, means that any assurances that student work, identification etc., will not be used in further training of AI systems cannot be verified. The key issue here is that there is a lack of agreement about who should regulate AI in education or how it should be regulated Berendt et al. (2020). There needs to be a balance of risks and benefits of using AI technologies, including the right to opt out of data collection, ensuring that data is accurate and that it is not re-purposed without permission or used for profiling and surveillance.

The use of AI as part of teaching might also encourage students to use AI more broadly, including in unethical manners. In particular, students may be tempted to use AI to produce work for submission. Clear communication about the use of AI, why it is used, and why other uses are considered unethical is therefore imperative to ensure students understand the situation and act ethically.

4. Results

Results of the students' winter year 11 mock exam in the main experiment and summer year 10 mock exams are presented in Table 1. From the data provided in Table 1, it can be seen that in the year 10 mock exams, the average mark was 1.3, with four students receiving a 'U' (ungraded mark). In contrast, in the year 11 mock, the same group received an average mark of 4.8, with no students receiving an ungraded mark.

Table 2
Average grades for mock tests.

Group	Year 10 Mock (July)	Year 11 Mock (December)	Grade Points (±)
Whole year group (101)	2.9	4.1	1.2
Other PP & SEN (25)	2.6	2.7	0.1
Main Experiment students (9)	1.3	4.8	3.5
Non PP & SEN (67)	3.4	4.4	1

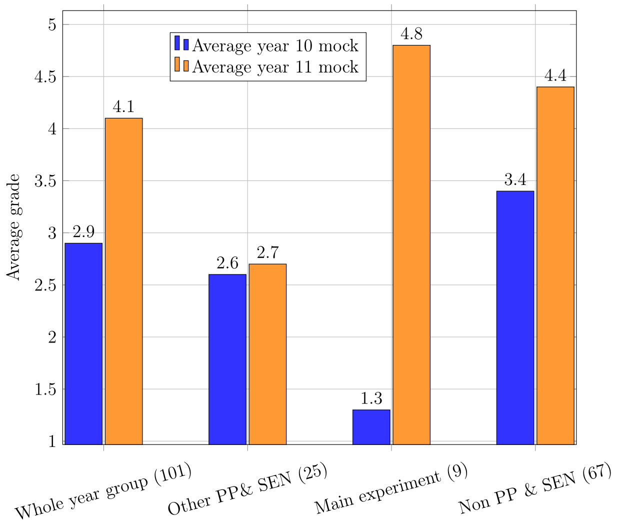

Fig. 2. Development of marks for various groups of students. A much higher improvement is noted for students participating in the experiment, showing the impact of generative AI on student attainment.

It can be argued that other factors could have influenced this outcome beyond the AI system. Studies show that students typically improve their exam results from year 10 to year 11, driven in part by student maturity, more learning, and greater emphasis on studying and exams in year 11 versus year 10 Kemp and Berry (2023) 'using data in the classroom' that on average over a space of four months students with no intervention would only make less than half a grade in progress. Although this may well be the case, it should be noted that all students will be exposed to these factors at school, with studies showing that typically non-pupil premium students derive greater benefit from these drivers and their year 11 exam results improve to a greater extent compared with their pupil premium peers Edovald and Nevill (2021).

To verify the impacts of these other factors, we compare the results of the other students in the year group, serving as a control group, with the students who participated in the main experiment in Table 2 and Fig. 2. As noted in Table 1 students from the main experiment improved their exam marks from 1.3 to 4.8. In contrast, other PP & SEN students, those who were not part of the main experiment, improved slightly from 2.55 to 2.7. The non-PP & SEN students improved from 3.4 to 4.4, which is more than the other PP & SEN students, but significantly less than the students in the main experiment. Data for the whole year group, including the experimental group, showed that the exam results improved from 2.9 to 4.1. Overall, the results of students from the main experiment group show a clear outperformance of all other student groups for GCSE History. In Fig. 3, we compare the improvements to expected improvements of typical students of different attainment levels.

To reinforce our findings, we present historical data gathered in preceding years within the School (Fig. 4). It is evident that, in prior years, students with PP and SEN exhibited lower performance, both generally and comparatively, in relation to their non-PP and SEN counterparts.

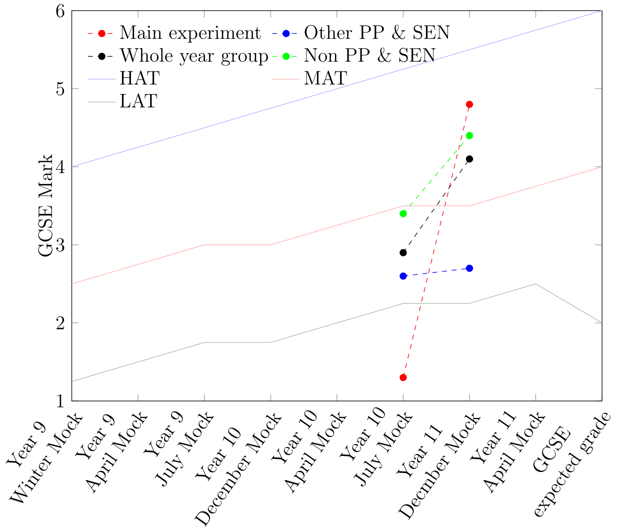

Fig. 3. Development of marks for the students in the experiment, compared to a typical expected development for students with low (LAT), medium (MAT), and high (HAT) attainment. Source DPA: GCSE History flight path.

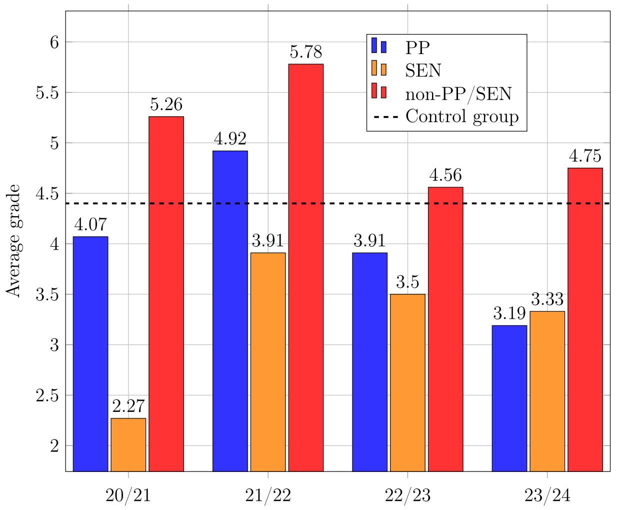

Fig. 4. Historical grades collected at Ditton Park Academy from the academic year 2020/2021 to 2023/2024.

The control group, depicted as a dashed black line, comprises nine students with SEN/PP in the academic year 2023/2024. It was observed that, during this period, the control group successfully narrowed the performance gap relative to their peers without SEN/PP requirements and achieved notably better results compared to earlier years. Additionally, it should be acknowledged that grade inflation occurred during the COVID-19 pandemic as a consequence of teacher-assessed grading Jack (2024).

4.1. Results validation

Paired hypothesis testing was employed to analyse the history grades from the Year 10 Mock (July) and Year 11 Mock (December) for the nine students involved in the experiment, trying to substantiate the observable quantitative difference in grades with a quantitative approach. We formulated a null hypothesis, H_0, to represent the case where there was no statistical difference between the results for Year 10 Mock (July) and Year 11 Mock (December), i.e., the differential grade is null, and chose a confidence level of 95 % (i.e. a significance level $\alpha = 0.05$) for all statistical analyses.

Table 3
Results of hypothesis testing (with $\alpha = 0.05$).

Test	r	p-value	null hypothesis
Paired t	1.89	0.00046549	rejected
WSR	0.87	0.0088488	rejected

Following an initial Shapiro-Wilk (SW) test Shapiro and Wilk (1965) to test for normality of the distribution of the differences in the grades in years 11 and 10 we could not reject the assumption of normality (p-value is 0.2309) and performed a paired *t*-test Hsu and Lachenbruch (2014) which confirmed the observed difference in the grades (Table 3). The results of the Wilcoxon Rank Sum tests are shown in Table 3, and they confirm our observations.

We are aware of the limitations that our small sample size may impose on the SW test; we calculated its a priori statistical power to be 0.2627. This value falls below the generally accepted threshold of 0.8, indicating that the normality assumption required by the *t*-test is poorly met. Because of this uncertainty, we also performed the Wilcoxon Signed-Rank (WSR) test to eliminate the assumption of normally distributed data, as it is uncertain. Again, WSR supports the observation that the difference between the sample change and the expected change is large enough to be statistically significant (Table 3).

From the data presented in Table 3, it is evident that the p-values are considerably lower than α, suggesting a clear rejection of H_0 in favour of the alternative hypothesis. Furthermore, the observed effect sizes ‘r’ are large. Thus, it is concluded that the history grades of the Year 10 Mock (July) are statistically different from the Year 11 Mock (December) for the nine students within the experiment, validating our conclusion that the grades obtained in the Year 11 Mock are better than the grades obtained in the Year 10 Mock.

Despite the modest sample size, as reasons for this detailed in Section 3.2.1, the observed effect size and statistical significance are remarkably robust, indicating a substantial impact that merits further investigation on a larger scale (as further commented on in Section 6.1).

We also note that the improvement in grades from the Year 10 Mock to the Year 11 Mock falls within a range of 0.1 to 1.0 points for students who were not part of experiment. Indeed trends within this range are consistent with the literature, where PP pupils typically make slower progress and fall in the lower end of the range, while their non-PP peers make better progress, and fall in the upper end of the range Edovald and Nevill (2021). Given that all students in the experiment were PP pupils, we formulated a second statistical test with a null hypothesis H_0 to represent the case where there is no statistical difference between their progress and that of their peers who were not part of the study. As a result, if this null hypothesis is accepted, then it can be concluded that the use of AI to improve student performance has no impact.

Students within the experiment saw their progress from the Year 10 Mocks to the Year 11 Mocks improve by 3.5 points on average. Analysing the data using the Inter-quartile range (IQR) we note that the lower bounds and upper bounds for the 0.1 to 1.0 range are -0.35 and 1.45 respectively. Given that 3.5 is significantly above the upper bounds, it can be concluded that the results were impacted by drivers beyond the normal progression from Year 10 to Year 11. As a result, we reject the null hypothesis, and conclude that the use of AI to improve student performance has made a significant, positive impact.

4.2. Qualitative feedback

Apart from quantitative data presented before, we also collected qualitative feedback from students. Table 4 shows the selected

Table 4
Qualitative feedback from students.

'*I have found the feedback useful and being in a routine*'
'*and doing work in a timely fashion has really helped prepare me for the exam.*'
'*I missed one of the lessons and I was struggling to do the assignment,*
but the feedback from the AI was able to fill in gaps and misunderstandings.'
'*I felt so much more confident in my exam doing that little extra work.*'
'*I was getting feedback for things I never realised when making mistakes.*'
'*I found the feedback easy to read and personalised to me, the bullet points helped.*'

statements from the students. They confirm that the ideas behind our study were recognised by the students.

5. Discussion

These students were selected as an intervention group due to a range of struggles based on SEN and PP disadvantaged backgrounds. We can see that in Year 10 their grade average was 1.3 among all nine of them. Even though we tried to give them support with the use of AI they struggled to use it effectively (see Section 3.2.2) and compared to other SEN and PP students in the year group they achieved the average grade of 2.7.

In the year 11 mock exam, with more time and supportive encouragement, they completed each task weeks before their assessment and were able to use the feedback to improve their exam technique and knowledge check. We can see their progress from an average grade of 1.3 to 4.8 as a significant improvement. When compared to the rest of the PP and SEN students in the cohort (18 PP only, 4 SEN only, 3 both) these students started on a higher average grade of 2.6 and made a slight improvement to 2.7 grade average.

Therefore, we can see that with effective support from the teacher and engagement from students in the use of AI, students can make significant progress. If this technology were implemented into a workable learning management system, students would have access to their own personal AI assistant supporting them in their learning, allowing all to reach their full potential, and closing the attainment gaps for vulnerable students.

As with all studies examining a sample to confirm a hypothesis, there are limitations to our findings. We note at least the following points:

- We have limited sample sizes (nine students doing the AI exercises). We show in Section 4.1 that our results are statistically significant, but this does not exclude a slim chance of them being a random result. More studies to confirm the results are welcome.
- External influences may affect the results. The study was conducted over one year, which means there could be other influences (e. g. maturity of students or family situations) changing over this time. It could also be that more specific learning influences (e. g. external tutoring) were taking place. Whilst we did check for any such changes and did not notice any significant issues, the study is clearly taking place in a real-world setting and is not fully controlled.
- The AI system used is not fully reliable and can produce incorrect results or hallucinations Lappin (2024). This can influence the results, but we believe that minor glitches will not render the exercise useless, since students still spent time on learning and reflecting on their results. In particular for the essay exercises this is part of the learning process.
- The study was conducted with history students. It may not apply the same way to other subjects, e. g. mathematics, where single incorrect results from the AI could have a much larger impact.
- Technology changes and may become outdated. We believe that AI-based tools will improve and the results should still be valid, but substantial changes in what AI can offer may create a completely new scenario where the study is not applicable.
- The experiment was conducted in an arts subject, where submissions are often essays, and feedback is on those essays. LLMs are very suitable for dealing with this style of assessment. Whilst they are not necessarily restricted to them (for example, coding exercises can be handled by LLMs Ma et al. (2024)), the use of AI in science and technology subjects requires a separate methodology and study.

6. Conclusion

We have shown that increasing student feedback by using AI technology for marking can help students improve their results. We have demonstrated this using a group of PP and SEN students. The academic performance of these students has shown a statistically significant improvement compared to that of the control group.

The difficulties facing educators and the future of AI in education include students' reading abilities. An AI tool, when giving feedback, will potentially use challenging language or subject-specific language. Therefore, the next steps in our research will include using AI to support language and vocabulary development. This will be part of a model of support to help students and teachers use AI effectively.

In any case, the development of AI in education will not replace teachers in their role, but can be used by teachers to support individual and whole-class interventions based on AI feedback.

6.1. Future work

This research would benefit from a larger sample to test more accurate significance and to look at varying differences of PP and SEN pupils across the UK. A randomized controlled trial would be part of this. Different subjects and age groups should also be included here. Finally, the assessment of qualitative aspects of student engagement and critical thinking is a future challenge.

Future work should also keep up with the constant changes in AI with regard to improvements and any private companies creating platforms to help pupils like in our sample. Finally workshops and basic training in AI to upskill staff to help them meet objectives which can be difficult with current workloads should become more common practice in schools. As part of this, the long-term impact on student learning and teacher workload can be observed.

CRediT authorship contribution statement

Ryan J. Brunton: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Soukaina Rhazzafe:** Writing – review & editing, Validation, Methodology. **Raymond Moodley:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Stefan Kuhn:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis. **Fabio Caraffini:** Writing – review & editing, Validation, Methodology, Formal analysis. **Sara Wilford:** Writing – review & editing, Methodology. **Rachel Higginbottom:** Writing – review & editing, Methodology. **Simon Colreavy-Donnelly:** Writing – review & editing, Methodology. **Mario Gongora:** Writing – review & editing, Methodology.

Declaration of Generative AI

As explained in the paper, AI was used during the experiment for generating feedback. AI was not used for data processing and during manuscript preparation.

Funding statement

SR acknowledges financial support by Taighde Éireann – Research Ireland, under Grant Number 18/CRT/6223.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank the students of ‘Ditton Park Academy’ for their diligent participation in the investigation and their commendable efforts to complete additional assignments beyond their regular academic commitments.

Appendix A. Prompts & settings

We use the settings in Table A.5 to use the LLM.

Table A.5
ChatGPT settings for this study.

Model	Temperature	Prompt	Supporting files
Turbo 3.5	0.3	Listing 1	Listings 2 and 3

During this research, the Turbo 3.5 model was used for cost efficiency, as Turbo 4.0 was newly released with higher costs and fewer training and research studies on its accuracy.

Temperature was set to 0.3 to ensure consistent grading across trials, while allowing flexibility for essay subjects where broader content might be relevant. High temperatures cause inconsistent feedback and accuracy with excessive randomness, whereas low temperatures result in generic feedback that overlooks unique points or explanations that differ slightly from the mark scheme.

Listing 1 shows the prompt we used.

```
You are a GCSE Edexcel History tutor designed to grade and give
feedback to students’ answers. Make sure you are positive and
professional in your responses. If the student receives less than 6
marks, give feedback in bullet point format to make it easier to
understand. Give advice on next steps for them to move up a grade
with essay structure and any missing or incorrect content.

<upload student’s script>
<upload mark scheme>
<upload marking grid>
```

Listing 1: Prompt template.

A marking scheme and a level-based marking grid were incorporated in addition to the student’s script to improve the model’s marking accuracy and feedback by instructing the AI on mark distribution according to the depth. We report the content of an example of the embedded supporting files below.

```
Indicative content guidance

Answers must be credited according to candidates’ deployment
of material in relation to the qualities outlined in the mark
scheme.
While specific references are made in the indicative content below,
this does not imply that these must be
included; other relevant material must also be credited.

• Stresemann solved the problem of hyperinflation with the
  introduction of a new currency called the Rentenmark.
• The Dawes Plan (1924) temporarily reduced Germany’s reparations
  payments to more manageable annual levels, so enabling a more
  stable environment for economic recovery.
• Under the Dawes Plan US loans boosted the German economy by
  providing investment to stimulate industry.
• Factories using newly-developed mechanisation and assembly
  techniques were constructed, so transforming production.
• Workers spent more money because they generally received better
  wages during this period.
• A lower number of strikes taking place and the removal of French
  soldiers from the Ruhr meant that industrial production increased.
```

Listing 2: A marking scheme example.

```
Grading criteria

Target: Analysis of second-order concepts: causation [AO2];
Knowledge and understanding of features and characteristics [AO1].

AO2: 6 marks. AO1: 6 marks.

Level Mark Descriptor 0 marks No rewardable material.

1-3 marks
• A simple or generalised answer is given, lacking development
  and organisation. [AO2]
• Limited knowledge and understanding of the topic is shown. [AO1]

4-6 marks
• An explanation is given, showing limited analysis and with
  implicit or unsustained links to the conceptual focus of the
  question. It shows some development and organisation of material,
  but a line of reasoning is not sustained. [AO2]
 • Accurate and relevant information is included, showing some
   knowledge and understanding of the period. [AO1] Maximum 5 marks
   for Level 2 answers that do not go beyond aspects prompted by the
   stimulus points.

7-9 marks
• An explanation is given, showing some analysis, which is mainly
  directed at the conceptual focus of the question. It shows a line
  of reasoning that is generally sustained, although some passages
  may lack coherence and organisation. [AO2]
• Accurate and relevant information is included, showing good
  knowledge and understanding of the required features or
  characteristics of the period studied. [AO1] Maximum 8 marks for
  Level 3 answers that do not go beyond aspects prompted by the
  stimulus points.

10-12 marks
• An analytical explanation is given which is directed consistently
  at the conceptual focus of the question, showing a line of
  reasoning that is coherent, sustained and logically
  structured. [AO2]
• Accurate and relevant information is precisely selected to
  address the question directly, showing wide-ranging knowledge and
  understanding of the required features or characteristics of the
  period studied. [AO1] No access to Level 4 for answers which do
  not go beyond aspects prompted by the stimulus points.
```

Listing 3: A marking grid example.

Data availability

No data were generated for this work.

References

Adıgüzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: exploring the transformative potential of ChatGPT. Contemporary Educational Technology, *15*, ep429.

Almasri, A., Ahmed, A., Almasri, N., Abu Sultan, Y. S., Mahmoud, A. Y.Zaqout, I. S. …Abu-Naser, S. S. (2019). Intelligent tutoring systems survey for the period 2000-2018. International Journal of Academic Engineering Research, *3*, 21–37.

Bahroun, Z., Anane, C., Ahmed, V., & Zacca, A. (2023). Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability, *15*, 12983.

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? International Journal of Educational Technology in Higher Education, *21*, 23. https://doi.org/10.1186/s41239-024-00455-4

Berendt, B., Littlejohn, A., & Blakemore, M. (2020). AI in education: learner choice and fundamental rights. Learning, Media and Technology, *45*, 312–324.

Betthäuser, B., Bach-Mortensen, A., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. Nature Human Behaviour, *7*, 1–11. https://doi.org/10.1038/s41562-022-01506-4

Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. American Sociological Review, *84*, 517–544.

Coghlan, D. (2019). Doing action research in your own organization. Sage Publications Ltd.

Cooper, G. (2023). Examining science education in ChatGPT: an exploratory study of generative artificial intelligence. Journal of Science Education and Technology, *32*, 444–452.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. International Journal of Educational Technology in Higher Education, *20*, 22.

Department for Education. (2020). English proficiency of pupils with English as an additional language. Ad-hoc notice. https://assets.publishing.service.gov.uk/media/5e55205d86650c10e8754e54/English_proficiency_of_EAL_pupils.pdf (accessed: 2024-Jun-21).

Department for Education. (2024). Pupil premium: overview. https://www.gov.uk/government/publications/pupil-premium/pupil-premium (accessed: 2024-Jun-21).

Edovald, T., & Nevill, C. (2021). Working out what works: the case of the Education Endowment Foundation in England. ECNU Review of Education, *4*, 46–64.

Er, E., Akcapinar, G., Bayazit, A., Noroozi, O., Banihashem, S.K., 2025. Assessing student perceptions and use of instructor versus ai-generated feedback. British Journal of Educational Technology 56, 1074–1091. https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.13558. https://doi.org/10.1111/bjet.13558.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: implications for educational practice and research. Innovations in Education and Teaching International, *61*, 460–474.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. Business & Information Systems Engineering, *66*, 111–126.

Firat, M. (2023). Integrating AI applications into learning management systems to enhance e-Learning. Instructional Technology and Lifelong Learning, *4*, 1–14.

Ghimire, A., Prather, J., & Edwards, J. (2024). Generative ai in education: a study of educators' awareness, sentiments, and influencing factors. arXiv preprint arXiv:2403.15586.

Harris, A., & Goodall, J. (2008). Do parents know they matter? Engaging all parents in learning. Educational Research, *50*, 277–289.

Hashem, R., Ali, N., Zein, F., Fidalgo, P., & Abu Khurma, O. (2023). AI to the rescue: exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. Research and Practice in Technology Enhanced Learning, *19*, 023. https://doi.org/10.58459/rptel.2024.19023

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, *77*, 81–112.

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. Mathematical Problems in Engineering, *2022*, 1–19.

Hsu, H., & Lachenbruch, P. A. (2014). Paired t test. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat05929

Huang, L. (2023). Ethics of artificial intelligence in education: Student privacy and data protection. Science Insights Education Frontiers, *16*, 2577–2587.

Hupkau, C., Gortazar, L., & Roldán-Mones, A. (2024). Online tutoring can transform education for disadvantaged students. https://blogs.lse.ac.uk/europpblog/2024/03/19/online-tutoring-can-transform-education-for-disadvantaged-students.

Internet Society. (2017). Artificial intelligence and machine learning: policy paper. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/ (retrieved May 10, 2024).

Jack, P. (2024). Grade deflation: First-class degrees back to pre-pandemic levels. Times Higher Education, https://www.timeshighereducation.com/news/grade-deflation-first-class-degrees-back-pre-pandemic-levels (accessed: 2024-Dec-5).

Jauhiainen, J., & Garagorry, A. (2023). Generative AI and ChatGPT in school children's education: evidence from a school lesson. Sustainability, *15*, 14025. https://doi.org/10.3390/su151814025

John Rush. (2024). All large language models. https://llmmodels.org/ (Accessed: 2024-Dec-5).

Karaman, M. R., & Göksu, I. (2024). Are lesson plans created by ChatGPT more effective? an experimental study. International Journal of Technology in Education, *7*, 107–127.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D.Fischer, F. …Kasneci, G. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. Learning and Individual Differences, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Keen, D., Webster, A., & Ridley, G. (2016). How well are children with autism spectrum disorder doing academically at school? an overview of the literature. Autism, *20*, 276–294.

Kemp, P., & Berry, M. (2023). Using data in the classroom. Becoming a Teacher: issues in Secondary Education 6e, 169.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S.Kay, J. …Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: artificial Intelligence, *3*, 100074.

Kmietowicz, Z. (2023). Physician associates: BMA survey finds 'shocking scale of concern'. BMJ, *383*, 2931. https://doi.org/10.1136/bmj.p2931

Lappin, S. (2024). Assessing the strengths and weaknesses of large language models. Journal of Logic, Language and Information, *33*, 9–20. https://doi.org/10.1007/s10849-023-09409-x

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: functionality and limitations of ChatGPT in written assessment evaluation. Australasian Journal of Educational Technology, *40*, 56–72. https://doi.org/10.14742/ajet.9463. https://ajet.org.au/index.php/AJET/article/view/9463.

Liang, Y., Zou, D., Xie, H., & Wang, F. L. (2023). Exploring the potential of using ChatGPT in Physics education. Smart Learning Environments, *10*, https://doi.org/10.1186/s40561-023-00273-7

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology.

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, *13*, 410.

Lu, Y., Pian, Y., Chen, P., Meng, Q., & Cao, Y. (2021). Radarmath: an intelligent tutoring system for math education. In Proceedings of the AAAI conference on artificial intelligence (pp. 16087–16090).

Ma, G. (2023). Chance or challenge: the role of ChatGPT in history teaching and historical research in higher education. In 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023) (pp. 869–874). Atlantis Press.

Ma, Q., Shen, H., Koedinger, K., & Wu, S. T. (2024). How to teach programming in the AI era? Using LLMs as a teachable agent for debugging. In A. M. Olney; I. A. Chounta; Z. Liu; O. C. Santos, & I. I. Bittencourt (Eds.), Artificial intelligence in education (pp. 265–279). Cham: Springer Nature Switzerland.

McCarthy, J. (2007). From here to human-level AI. Artificial Intelligence, *171*, 1174–1182.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMS to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: artificial Intelligence, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199. https://www.sciencedirect.com/science/article/pii/S2666920X23000784.

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization.

Molloy, E. K., & Boud, D. (2014). Feedback models for learning, teaching and performance (pp. 413–424). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-3185-5_33

Mondal, H., Marndi, G., Kumar, J., & Mondal, S. (2023). ChatGPT for teachers: practical examples for utilizing artificial intelligence for educational purposes. Indian Journal of Vascular and Endovascular Surgery, *10*, 200–205. https://doi.org/10.4103/ijves.ijves_37_23

Moodley, R., Chiclana, F., Carter, J., & Caraffini, F. (2020). Using data mining in educational administration: a case study on improving school attendance. Applied Sciences, *10*, 3116.

Muthuselvan, S., Rajaprakash, S., Jaichandran, R., Antony, J., Amal, P. U., & Ijas, V. A. (2024). Student academic performance prediction enhancement using t-SIDSBO and triple voter network. Multimedia Tools and Applications, 1–24.

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. In R. Ferreira Mello; N. Rummel; I. Jivet; G. Pishtari, & J. A. Ruipérez Valiente (Eds.), Technology enhanced learning for inclusive and equitable quality education (pp. 284–298). Cham: Springer Nature Switzerland.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. Administration and Policy in Mental Health and Mental Health Services Research, *42*, 533–544. https://doi.org/10.1007/s10488-013-0528-y

Povey, J., Campbell, A. K., Willis, L. D., Haynes, M., Western, M.Bennett, S. …Pedde, C. (2016). Engaging parents in schools and building parent-school partnerships: the role of school and parent organisation leadership. International Journal of Educational Research, *79*, 128–141.

Rahman, M., & Watanobe, Y. (2023). ChatGPT for education and research: opportunities, threats, and strategies. Applied Sciences, *13*, 5783. https://doi.org/10.3390/app13095783

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, *52*, 591–611.

Shin, I., Hwang, S. B., Yoo, Y. J., Bae, S., & Kim, R. Y. (2025). Comparing student preferences for AI-generated and peer-generated feedback in AI-driven formative peer assessment. In Proceedings of the 15th International Learning Analytics and

Knowledge Conference (pp. 159–169). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3706468.3706488

Skinner, B., Leavey, G., & Rothi, D. (2021). Managerialism and teacher professional identity: impact on well-being among teachers in the UK. Educational Review, *73*, 1–16.

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: considerations for academic integrity and student learning. Journal of Applied Learning and Teaching, *6*.

Tahiru, F. (2021). AI in education: a systematic literature review. Journal of Cases on Information Technology (JCIT), *23*, 1–20.

Third Space Learning. (2024). Attainment gap: what it is and 7 strategies to close it in schools. https://thirdspacelearning.com/blog/attainment-gap/ (accessed: 2024-May-14).

Uddin, M. M. (2024). Rejection or integration of AI in academia: determining the best choice through the opportunity cost theoretical formula. Discover Education, *3*, 249.

UK Government (2024a). Ditton Park Academy - compare school and college performance data in England - GOV.UK. https://www.compare-school-performance.service.gov.uk/school/141009/ditton-park-academy (accessed: 2025-Jul-14).

UK Government (2024b). Government unveils levelling up plan that will transform UK. https://www.gov.uk/government/news/government-unveils-levelling-up-plan-that-will-transform-uk (accessed: 2024-May-14).

UNESCO. (2023). Unesco: Governments must quickly regulate generative AI in schools. https://www.unesco.org/en/articles/unesco-governments-must-quickly-regulate-generative-ai-schools?hub = 83250. (Accessed on 27/06/2024).

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, *1*, 213–218.

Ward, H. (2016). Workload: tens of thousands of teachers spend more than 11 hours marking every week. https://www.tes.com/magazine/archive/workload-tens-thousands-teachers-spend-more-11-hours-marking-every-week (retrieved May 10, 2024).

Xia, Q., Weng, X., Ouyang, F., Lin, T., & Chiu, T. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. International Journal of Educational Technology in Higher Education, *21*, https://doi.org/10.1186/s41239-024-00468-z

Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: new demands for teachers' roles. Heliyon, 10 (2), e24289.

Yu, H., & Guo, Y. (2023). Generative artificial intelligence empowers educational reform: current status, issues, and prospects. In Frontiers in education. Frontiers Media SA. p. 1183162.

Zhai, X. (2022). ChatGPT user experience: implications for education. SSRN Electronic Journal, https://doi.org/10.2139/ssrn.4312418

Zhou, P., Wang, L., Liu, Z., Hao, Y., Hui, P., Tarkoma, S., & Kangasharju, J. (2024). A survey on generative ai and llm for video generation, understanding, and streaming. arXiv preprint arXiv:2404.16038.