



Department
for Education

Use Cases for Generative AI in Education

User Research Report

August 2024



Government
Social Research

Contents

List of figures	4
List of tables	5
Executive Summary	6
Education hackathons	6
Hackathon Outcomes	7
PoC Development	7
User Research	8
Introduction	10
Project Aims	10
Project structure and phases	10
Summary of PoC build	13
Outline of the 3 strands of user research	16
Hackathon Findings	18
Summary	18
Summary of key use cases	22
User Research Findings	29
Introduction	29
Methods	30
Results	31
Conclusions	44
User research conclusions	44
Suggested areas of future focus	45
Further work in progress	48
Recommendations for delivery of future hackathons	49
Annex 1: Summary of pre-Hackathon consultation findings	52
Overview	52
Aims	52
Methods	53
Findings	55
Enablers and barriers to AI use	63

Next steps	64
Annex 2: Full Hackathon findings report (November 2023)	65
Introduction	65
Potential barriers to Generative AI adoption	66
Use cases for Teachers	68
Use cases for School Leaders and Administrators	77
Use case for Students	80
Annex 3: AI mini-hackathons project summary	83
Background to the project	83
The aims of the hackathon	83
What did we do?	84
What did we learn?	84
Annex 4: Membership of steering committee	88
Annex 5: Methods for user research	89
Methods	89
Sample and sampling strategy	89
Materials and Procedure	90
Annex 5: Prompts used in the Hackathons	92
Generating age-appropriate lesson plans	92
Generating effective questioning materials	94
Marking a KS2 essay based on a modified mark scheme and providing personalised feedback	96
Providing feedback on students' work in a way meaningful to students	97
Prompt used to generate newsletter articles, using standardised data fields (parent-carer communications)	99
Prompt used to simplify text for students (lesson materials)	100
Hallucination in creating end-of-year reports, with the hallucination highlighted in bold italics	102
Writing end-of-year reports for other teachers based on key information	103
Prompting the LLM to create a vocabulary list with simple prompting (language learning assistant)	105
Prompting the LLM to correct a student's work by providing steps to follow and one-shot prompting	105
Adjusting data literacy policies in line with school values	106

Providing detailed context for adapting school policy	109
Creating policy-based scenarios for testing with staff	110
Generating newsletters based on short user inputs	112
Mandarin language assistant	114
Annex 6: Key Terms	117
Annex 7: References	118

List of figures

Figure 1: Screenshot of the essay upload user interface	14
Figure 2: Screenshot of the teacher feedback user view	14
Figure 3: Screenshot of the student feedback user view	15
Figure 4: Screenshot of the student task user view	16
Figure 5: Opportunities and risks associated with using GenAI tools for assessment and feedback	32
Figure 6: Needs identified by teachers based on their perceived barriers and facilitators to GenAI adoption in schools	39
Figure 7: Breakdown of survey stakeholders by role	53
Figure 8: Venn diagram to show the overlap in teachers' use case choices for time saving and strengthening practice	56

List of tables

Table 1: 8 use cases selected as having potential to develop to PoC	12
Table 2: Participants in each strand of user research	17
Table 3: Use cases explored during the hackathons	19
Table 4: Summary of user feedback for generating age-appropriate lesson plans	22
Table 5: Summary of user feedback for generating effective questioning ideas	23
Table 6: Summary of user feedback for generating lesson materials	23
Table 7: Summary of user feedback for assigning marks to work submitted by students	24
Table 8: Summary of user feedback for generating personalised formative feedback and addressing student misconceptions	26
Table 9: Summary of user feedback for generating drafts of statutory policies.....	27
Table 10: Summary of user feedback for generating pupil or class data analysis and synthesis.....	27

Executive Summary

From September 2023 to March 2024, Faculty AI, the National Institute of Teaching (NlOT) and ImpactEd Group (representing the AI in Schools Initiative) have worked with the Department for Education (DfE) to deliver the Use Cases for Generative Artificial Intelligence in Education project. The project explored potential applications for Generative AI (GenAI) in the education sector as part of a wider effort to ‘transform a teacher’s day to day work’ – reducing workload and improving educational outcomes by automating routine tasks. The project team ran 4 work-streams at different stages of the project: pre-hackathons user engagement; delivery of the hackathons; build and testing of the PoC; and user research. Roles on this project included:

- Faculty AI: Project leadership; tool development and testing (‘the development team’)
- National Institute of Teaching: User research; initial consultation to support selection of use cases (‘the research team’)
- ImpactEd Group, AI in Schools Initiative: Involvement of teachers, leaders and students in hackathons and user testing.

This report sets out the user research findings of the project and has been published alongside the Technical Report which details the experimentation and technical development work conducted as part of the project.

This is an independent report and is not intended to represent the policy position of the Department.

Education hackathons

In October 2023, 60+ participants came together to deliver the first ever hackathons in education to explore GenAI. This event provided a huge amount of insight into the needs and perspectives of teachers, school leaders and administrators as regards the use of GenAI in education. It also enabled data scientists and engineers from across education, government and the private sector to collaborate to make progress on solving some of the most difficult associated challenges. The impact of these events was clear, and the participants made significant progress with some of the use cases, meaning that they were able to use GenAI tools to help users complete tasks such as drafting content for school websites or creating lesson materials. They also generated important learnings applicable to the future development of tools using GenAI.

“[The participants developed...] a very practical [solution] that nobody came here thinking about, but the fact you’ve got teachers, the fact you’ve got computer scientists, you’ve got researchers and experts all together, and they’ve actually created examples of it right here that could be utilised. It’s quite astonishing.” – Education Secretary, Gillian Keegan

“What the AI was able to do was to be fed examples of existing best practice policies to learn from, and then tailor to the school context, and that potentially shortens the time it takes school leadership to create policies, but also reviews, checks quality, checks for adherence to best practice or legislation. So, I think that’s really exciting.” – Tom Nixon, Head of Government Practice at Faculty

“[ChatGPT initially] wasn’t producing the outputs that we were hoping for, so we’re feeding it even more information. But what that really highlighted for us was the fact that we need these models to be trained in subject disciplines for each individual use case scenario for each individual academy. That could have potentially a great amount of impact on teacher planning time, adapting lessons for each individual student.” – Jonathan O’Donnell, Computing Consultant at the Harris Federation.

Hackathon Outcomes

Ahead of the hackathons, the project team co-designed a list of use cases for GenAI in education, engaging over 700 stakeholders from the education sector (involving teachers, leaders, administrators, teacher educators, and students), and aiming for a school-led approach in selecting use cases most useful to schools. These use cases were reviewed and a shortlist of twelve use cases was created. One of the expected outcomes was the classification of the twelve use cases explored during the hackathons into three groups, namely: use cases which were ready to be used in schools; use cases which would first need to be developed into a PoC tool; and use cases which were beyond the models’ capabilities, or too risky to deploy.

Although some of the twelve use cases explored in the hackathons had significant potential for impact, there were no cases in which the participants thought that the solution developed was ready for use in schools. There are ways in which GenAI can support teachers in some limited versions of some of the use cases identified, such as coming up with suggestions for lesson activities. However, even the most successful use cases had remaining challenges to be addressed before they could be widely used by teachers or solve the problem entirely, such as the need for access to additional data sources or integration with other tools, and the team then progressed to development of the PoC tool. It is important to note that the hackathons were restricted to the use of GenAI models (mostly GPT-3.5 and GPT-4) and did not explore wider EdTech tools, which may be able to successfully solve some of the use cases.

PoC Development

To support these objectives, the development team developed a PoC GenAI tool that marks students’ work and provides personalised feedback and a revision activity. This tool was developed as an experiment and there was no intention that it would be rolled out in schools, but instead was designed to provide an opportunity to investigate a specific application of GenAI, the performance of the tool and users’ responses to it. As

such, the learnings from this process, such as the importance of allowing teachers to customise outputs based on their professional judgement, the ability of LLMs to provide feedback, and the potential for GenAI tools to effectively assess model outputs, are expected to be broadly applicable to the development and implementation of any GenAI tool in this educational context.

The development team then spent several months developing a PoC tool focused around the areas of feedback and generating revision activities. This was developed to the point where a user can upload a piece of student work, have the errors automatically detected and linked to the national curriculum, generate feedback for the student and teacher, and generate a series of revision activities tailored to help the student address the errors in their work. Through experimentation with different models, architectures, prompts and reference data, the team drew a number of key insights which are applicable to the development of a broad range of GenAI tools in education. For example, the positive impact of blending deterministic language modelling and GenAI based approaches on performance, and the importance of structuring reference documents such as the national curriculum to enable Large Language Models (LLMs) to interpret them. These insights are outlined in detail in the accompanying Technical Report.

User Research

Alongside this technical development, the research team completed a programme of user research to understand users' perspectives on the PoC tool and their views on the use of GenAI in education more generally. The aim of this research was to put teachers' views at the forefront of the tool development, and to gain their perspectives of the potential of tools in this area. Although teachers' perspectives were varied, the key themes that emerged included general positivity about the potential impact of GenAI, and many teachers were able to see the potential time saving and standardising benefits of using GenAI tools for feedback. However, for many this was tempered with concern about the potential risks. For example, some teachers reported that they would not trust a PoC tool to provide feedback, and others were concerned that teachers may become over-reliant on the tool. Teachers also reported a need for time, training, funding, and expert help to increase their knowledge in the use of GenAI in their practice, as well as a need for guidance on how they should be using AI, and how to do so safely.

It is important to note for both the user research and technical experimentation work, that the insights and conclusions drawn reflect the current state of GenAI technology and users' adoption of it. This is a rapidly advancing area, and increasing interest in – and adoption of – AI tools for educational purposes in recent years has resulted in a fast-growing body of research in this area. However, such has been the speed of development that even recent large-scale reviews of the literature (e.g., Ng et al., 2023; Zhai et al., 2021; Zhang & Tur, 2023) are in danger of becoming out of date within a few years. This highlights the importance of up-to-date research, bringing teachers' and students' perspectives to the fore in this fast-changing landscape. For example, in the

time since the hackathons, new EdTech tools have come to market that address (or partly address) some of the use cases explored, and a key part of the challenge for schools and educators in the adoption of GenAI is navigating this changing landscape.

Introduction

Project Aims

The central aims of this project were to:

- put teachers' voices at the heart of the future of AI in education,
- broaden the evidence base on the existing strengths and limitations of GenAI tools within educational contexts,
- understand teachers' requirements for AI to meet their needs and effectively enhance their role in improving students' learning,
- disseminate these learnings for the sector, and
- investigate the potential to improve the performance of generative AI models using education-specific datasets.

Project structure and phases

To meet the project objectives outlined above, the project team ran 4 work-streams at different stages of the project: pre-hackathon user engagement (all); delivery of the hackathons (all); build and testing of the PoC (Faculty AI and AI in Schools Initiative); and user research (NIoT).

Pre-hackathon user engagement

The first of these work-streams was the initial user consultation and development of a shortlist of use cases. We used a co-design approach with stakeholders from the education sector, where teachers, school leaders, administrators, teacher educators and students took part in a consultation via surveys and stakeholder group meetings. 710 school-based practitioners took part in a 10-minute online consultation survey, and an additional c.20 stakeholders (teachers, leaders, and secondary school and sixth form students) took part in one of 4 online stakeholder group sessions to discuss the potential uses of AI in schools. This research is summarised in more detail in section 3 of this report, and the full report into the consultation work is provided in Annex 1.

The findings of the consultation, alongside the DfE's Call for Evidence on Generative AI in Education, were then used to develop a list of use cases, or applications of generative AI in education, with the further input of a team of teacher educators with digital expertise. To narrow this to a shortlist of use cases for testing in the hackathons, this larger group was assessed and prioritised based on iterative feedback from DfE as well as consideration of whether:

- LLMs could be expected to perform the given task in line with users' needs,

- the use of LLMs for a given purpose reflected users' priorities and views on potential efficacy of workload reduction and outcome improvement,
- there were any safety considerations, cultural or organisational blockers that would make using an LLM tool for a given purpose difficult.

The proposed shortlist was reviewed by the project's Steering Group, and 12 use cases, representing potential GenAI applications for teachers, school leaders, administrators and students were selected for testing in the hackathons. Membership of the project's Steering Group is listed in Annex 3.

Delivery of the hackathons

On the 30 and 31 of October 2023, the team held the hackathons, bringing together potential users, education policy experts, and data scientists. Participants tested different approaches to:

- tasks such as 'plan a French lesson for Year 9 students on the past tense' for a selection of use cases,
- further engineer their prompts to improve the GPT models' outputs, and
- assess whether the final outputs produced would be usable in a school environment.

The 60+ participants demonstrated clear enthusiasm for the application of GenAI in their own contexts, and the events provided significant insight into their needs and perspectives. The Education Secretary and Minister for the School System and Student Finance also hosted a roundtable discussion with teachers, school leaders and education policy experts. In two days of events, participants made significant progress with some of the use cases, as well as finding important learnings applicable to the future development of tools using GenAI. The process and findings of the hackathons are outlined in more detail in section 3 of this report.

Of the 12 use cases explored in the hackathons, 8 were selected as having potential for development to PoC. These 8 use cases were assessed against selected criteria: innovation, learning potential, practicality, feasibility, novelty and strength of evidence.

Table 1: 8 use cases selected as having potential to develop to PoC

Use Case	Description
Lesson plan or activity adaptor	Adapt existing lesson plans to the context required and tailor lesson activities to specific classes.
Feedback and revision activity generator	Review student work and provides them with both feedback and a personalised activity to develop and consolidate their learning.
Question generator	Generate graded, lesson plan aligned questions automatically, based on information from sources such as lesson plans, objectives, etc.
SEND support tool	A tool able to support teachers to adapt lesson content to meet the needs of students with Special Educational Needs and Disabilities (SEND).
Essay marker	A tool able to support the marking of English essays in bulk and provide insights for teachers to better understand how performance varies across a class.
Lesson activity generator	A tool able to generate a variety of lesson materials including differentiated activities, quizzes and scripts for a lesson.
Parent and carer communications tool	A tool able to generate communications with parents and carers, for example school newsletters or emails about upcoming school events.
Policy document generator	A tool able to support the generation of school policy based on submitted characteristics about a school, any existing policy documents and national legislation or guidance.

A candidate for PoC was then selected based on this assessment, combining 2 of the proposed use cases: essay marker and feedback and revision activity generator. The proposed PoC was 'A tool which reviews the student's work and provides them with both feedback and a personalised activity to develop and consolidate their learning'. This was limited to Year 4 literacy work, comparing the students' errors against the national curriculum. Focusing on Year 4 work was due to this year group not being close to any national assessments, and as their writing was expected to generally be accurate enough to be well-suited to processing by an LLM, while being short enough examples to limit the required processing power where possible.

Although there were clear emerging challenges with GenAI marking, during the hackathons it was also clear that support with giving feedback to students is highly valued by teachers, and that a PoC tool that addressed this would be impactful. Other expected benefits to developing this PoC included:

- that there is no existing generalisable tool that performs both the feedback and activity generation functions together based on the English curriculum,
- developing the feedback and activity generation functions have wider applications than these use cases, and
- the PoC produced would demonstrate a broader application of GenAI in education, with both teachers and students as potential users.

Following the selection of a use case for development to PoC, the project team progressed to the build and testing of the PoC, and the delivery of the user research.


Summary of PoC build

The PoC has been developed to test the potential of the technology to support educators and teachers and facilitate student learning and development. Specifically, the PoC was designed to test GenAI's ability to take a piece of Year 4 writing, assess it against the national curriculum, and produce personalised feedback and a revision activity based on the student's writing. The PoC was designed to enable two main user journeys; one that allows the user to upload their own pupil work, while the other uses a bank of pre-processed essays. These only differ at the start, and so this document will describe only the user upload journey, which is more realistic to how an end user would interact with the tool if this were developed to deployment and integrated into their workflow. It is important to reiterate that the objective of this work was to explore the potential of GenAI in this context, rather than to develop a tool to Minimum Viable Product (MVP) or deployment.

The PoC has 4 key features:

1. **Student Essay:** when a user accesses the tool, they are presented with the essay upload page. This allows them to upload details of the specific task a pupil was set, such as a creative writing task where the pupil was asked to practise their skills in developing suspense in an everyday situation, and the pupil's piece of work. The user will copy and paste these into the relevant input fields and click the 'Submit essay' button. The tool analyses the work and, based on guidance from the national curriculum and other materials, identifies areas for improvement. The user can then hover over any of the highlighted areas for more information, including the type of error, the correction and the Year group associated with the error according to the national curriculum.

Essay Task

 Please provide an essay task. The task cannot be empty.

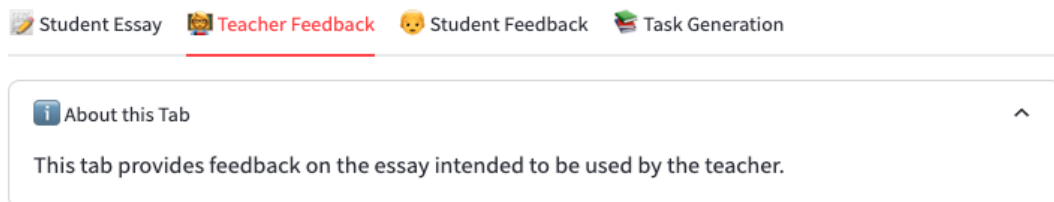
Input the essay task here.

Student Essay

 Please provide an essay text. The essay text cannot be empty.

Figure 1: Screenshot of the essay upload user interface

2. Teacher Feedback: the user then navigates to the 'Teacher Feedback' tab to view an assessment of the pupil's work intended for the teacher. This generated feedback gives a summary of how well the pupil did in relation to the task, as well as specific details on their spelling, punctuation and grammar based on the errors that the tool has detected.



Teacher Feedback

GENERAL FEEDBACK

The student's essay on the history and significance of pointe shoes in ballet is a delightful read. The student has shown a clear understanding of the topic and has managed to convey their passion for ballet and pointe work. The essay is well-structured, with a clear introduction, body, and conclusion. The student has also demonstrated a good use of vocabulary and has included some interesting facts about pointe shoes and ballet.

SPECIFIC FEEDBACK

Spelling: The student has made a few spelling errors, particularly with homophones, which are introduced

Figure 2: Screenshot of the teacher feedback user view

3. Student Feedback: navigating to the 'Student Feedback' tab, the user sees feedback intended for the pupil. The language here is encouraging and focuses on

how the pupil can improve, rather than with a list of errors. The user can share this directly with the pupil or use it as the basis for their own feedback.

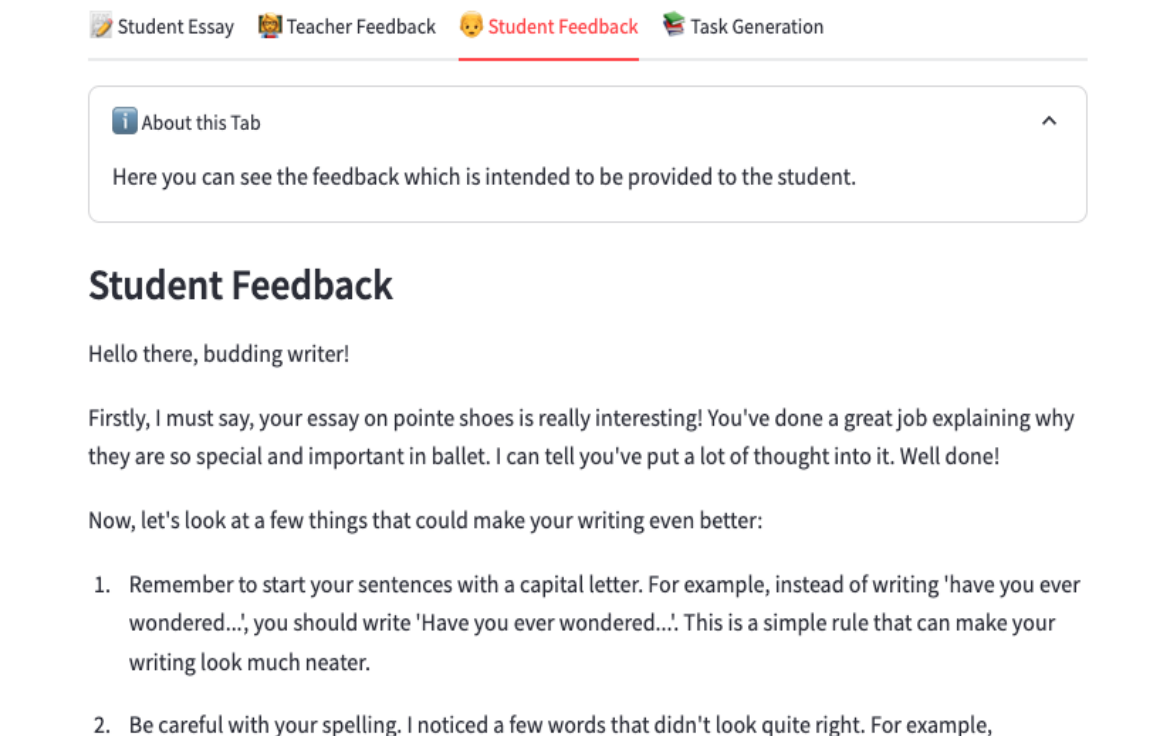


Figure 3: Screenshot of the student feedback user view

4. Task Generation: finally, the user can view a selection of formative worksheets or practice exercises for the student to complete that the tool has generated based on the errors and feedback. Four varieties of worksheet are generated; the first focuses on the most important errors as defined by their order in the national curriculum, and rest specifically on spelling, punctuation and grammar respectively. These could be used directly as they are, or as a first draft for the user to refine.

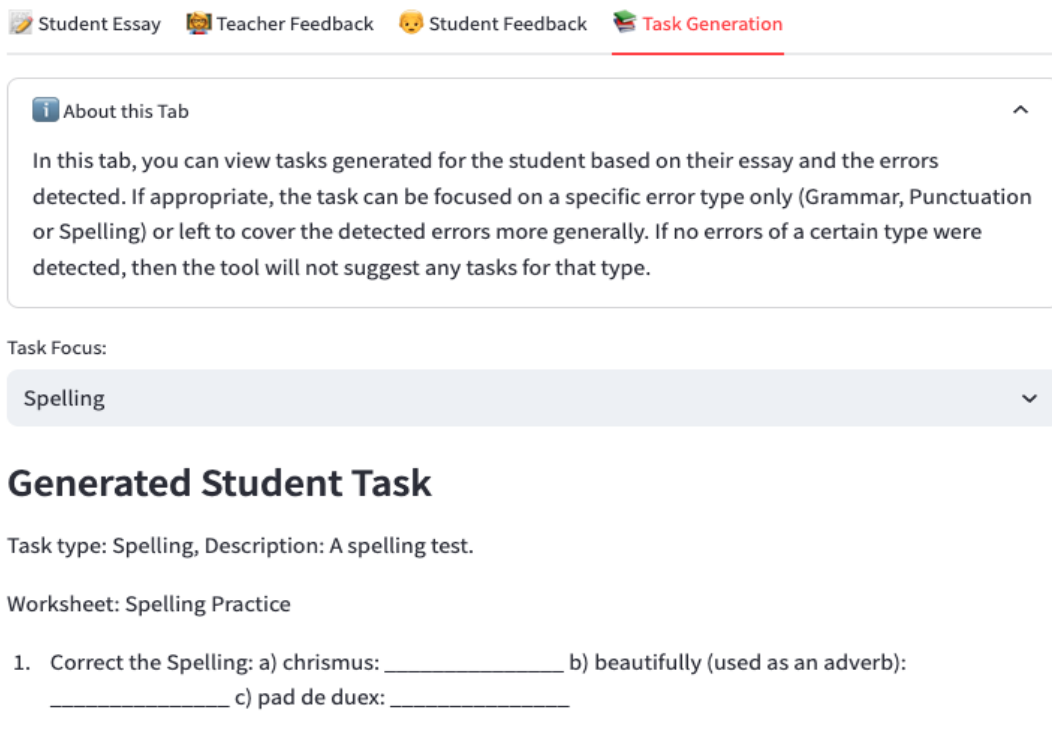


Figure 4: Screenshot of the student task user view

A detailed overview of the PoC tool is provided in section three of the technical report published alongside this report.

Outline of the 3 strands of user research

Alongside the development of the PoC tool, the project team conducted 3 strands of user research: ‘superuser’ engagement (Faculty AI and AI in Schools Initiative); user testing of the PoC tool (Faculty AI and AI in Schools Initiative); and user research (NIoT). Each of these strands of user research were designed to contribute to the project’s overall objectives (2.1, page 6).

‘Superuser’ engagement: Co-design of PoC

The ‘superuser’ engagement involved building a cohort of users who are supportive proponents of using GenAI in educational contexts, with significant amounts of experience as classroom teachers and curriculum leaders. These users were drawn from the AI in Schools Initiative hackathons participants and the stakeholders from the consultation phase of this project. This took part in a series of sessions to help the team to understand users’ requirements and priorities when it came to the marking and feedback tool, and once the initial PoC had been developed, to give iterative feedback on the tool.

User testing: Teacher evaluation of PoC performance

The user testing strand involved asking 8 primary school teachers to rate the quality and accuracy of the outputs of the PoC tool for specific examples of student writing. Teachers were sent examples of feedback which the tool had generated and were asked to review and rate these outputs for a variety of criteria including whether the feedback generated by the tool was accurate and met their general expectations for what feedback to students should look like such as tone or length. Teachers also explored the tool's functionality by directly interacting with the tool. A detailed summary of the findings of the 'superuser' engagement and the user testing is provided in section 4 of the technical report published alongside this report.

User research: Experiences and perspectives study

These two strands focused on teachers' and students' perspectives on the PoC tool and its outputs. The NIoT conducted user research exploring users' experiences and perceptions of AI in broader terms, with a particular focus on how these may vary depending on their school's context and their previous experiences of using AI.

Teachers (N = 12) and secondary school students (N = 9) were recruited from different geographic regions of England and took part in a series of interviews and focus group discussions. This research used the PoC as a platform for discussion to explore: the barriers and facilitators to users' adoption of AI for feedback in schools, and understand teachers' perceptions of the opportunities AI for feedback would offer and the associated risks and challenges. A detailed summary of the findings of this user research is provided in section 4 of this report.

Table 2: Participants in each strand of user research

Group	Number of Participants	Roles
'Superuser' Engagement	5	Secondary school classroom teachers and AI and technology leads
User Testing of the PoC Tool	8	Primary school classroom teachers
User Research	21	Teachers and secondary school students

Hackathon Findings

Summary

Working with NIoT and DfE, Faculty explored potential applications for GenAI in the education sector as part of a wider effort to reduce teacher workloads and improve educational outcomes by automating routine tasks. As part of this exploratory work, Faculty held two days of 'Generative AI in Education Hackathons', inviting participants from across the education sector to test a range of education-related use cases for LLMs.

During the hackathons, over 60 potential users, data scientists and education experts were brought together to test different approaches to education-based tasks. They engineered prompts to improve the GPT models' outputs, and assess whether the final outputs produced would be usable in a school environment. For example, one task involved planning a French lesson for Year 9 students on the past tense. Ultimately, the purpose of the hackathons was to generate a set of findings to inform the identification of 3 groups of use cases:

- use cases for GPT models which are currently ready for schools (potentially with some associated guidance/instructions provided),
- use cases which could be good candidates for a PoC tool with some additional tooling/functionality, and
- use cases which are currently not possible for GPT models or too risky to take forward into schools.

Twelve GPT use cases were tested in the Generative AI in Education Hackathon and were then assessed and prioritised based on iterative feedback from key project stakeholders as well as consideration of the following key elements:

- technical feasibility – whether LLMs were expected to be able to perform the given task in line with users' needs,
- expected impact – whether the use of LLMs for a given purpose reflected users' priorities and their views on potential efficacy in terms of workload reduction and outcome improvement, and
- risk assessment and feasibility – whether there were any safety considerations, cultural or organisational blockers that would make an LLM tool for a given purpose difficult to roll out.

For some of the most successful use cases the participants rated the solutions developed highly in terms of their future potential for impact. However, there were no cases in which the participants thought that the solution developed was currently ready for use in schools. Even for the highest scoring use cases, there were remaining challenges to be addressed, often related to the quality or consistency of outputs, safety

or privacy concerns, and the need for access to additional data sources or integration with other tools, although it is possible that the desired results may have been achieved using different models or existing edtech tools. As a result, the team prioritised developing a comprehensive PoC model that would provide valuable learnings on approaches to optimising a model for education, rather than focussing on creating guidance.

Table 3: Use cases explored during the hackathons

Category	Use Case	List of suggested tasks provided to hackathon participants
Lesson planning	Generating age-appropriate lesson plans	<p>Plan a French lesson for Year 9 students covering the past tense.</p> <p>Plan a History lesson for Year 1 students comparing the lives of Elizabeth I and Queen Victoria.</p> <p>Plan a Maths lesson for Year 6 students covering Ratio and Proportion.</p> <p>Plan an English lesson for Year 13 students covering the structure of A Midsummer Night's Dream.</p>
	Generating effective questioning ideas	<p>Generate a group of questions across a range of difficulty which I could ask Year 10 students in a GCSE History lesson covering World War 2.</p> <p>Generate a set of questions that I can ask a group of Year 6 students to test whether they have understood the concept of fractions.</p>
	Generating lesson materials	<p>Create two short quizzes that I can use as part of a Year 10 Biology lesson on ecosystems. One quiz should be made of short-answer questions and the other should be made of long-answer questions.</p> <p>Generate a range of independent learning activities of varying difficulty for Year 8 students in a lesson about Lord of the Flies.</p>
Assessment	Assigning marks to work submitted by students	Mark a collection of writing exercises completed by Year 4 students for spelling and grammar.

Category	Use Case	List of suggested tasks provided to hackathon participants
	Generating personalised formative feedback and addressing student misconceptions	<p>In the form of short paragraphs, provide personalised feedback on completed Plants worksheets to Year 1 students.</p> <p>Analyse these pieces of Geography work submitted by Year 5 pupils and provide me with a list of common misconceptions.</p> <p>My Year 8 History class often confuse James I and Charles I; suggest strategies to help them remember the differences.</p> <p>Analyse these pieces of Maths work submitted by Year 11 pupils and provide personalised feedback on areas of strength and weakness in the form of short paragraphs.</p> <p>Review these mock A-Level Sociology essays submitted by Year 13 students and suggest specific improvements which would improve the quality of their work.</p>
	GenAI as a teaching aid	Generate a series of GCSE Physics questions for Year 9 students to practise. Ask each question one-by-one and provide feedback on correct and incorrect answers.
Report writing	Writing end-of-year reports	Generate an end-of-year report for Student X. Student X is high-achieving with no concerns about behaviour. They enjoy English and achieve good marks but struggle with Maths and dislike Science.
SEND	Generating tips for SEND intervention	Generate a list of recommendations for a teacher supporting a pupil in Reception with speech, language and communication needs, with reference to the latest SEND guidance and academic literature, to be reviewed by a SENCO.

Category	Use Case	List of suggested tasks provided to hackathon participants
Communication assistant	Generating parent-carer communications	<p>Create a letter to parents reminding them that PE lessons are every Monday and that pupils need to remember to bring their kit to school.</p> <p>Create a reminder for parents that a nursery will be closed for a bank holiday.</p> <p>Create an article for the school newsletter about Year 6's recent trip to London Zoo.</p>
Policy generation	Generating drafts of statutory policies	<p>Create a first draft of an explanation of my school's uniform policy for publication on the school website, according to DfE guidance.</p> <p>Create a first draft of an explanation of my school's remote education policy for publication on the school website, according to DfE guidance.</p>
Data analysis	Pupil or class data analysis / synthesis	<p>Generate a report to be sent to Student X's form tutor about their behaviour over the past two terms. In the report, highlight any recurring patterns.</p> <p>Generate a report for a headteacher about Year 7 Maths formative assessment results over a school year.</p>
Language learning	Language learning assistant	<p>Can you help me revise for my GCSE German exam? I need you to test me on vocabulary to do with hobbies.</p> <p>Can we have a conversation in Spanish about travel and holidays in the style of an A-Level speaking exam?</p> <p>I can't remember how to conjugate the French verb avoir, can you explain to me how the conjugation works?</p> <p>Is this sentence right? Le weekend dernier, je suis aller au cafe et j'ai manger les frites. J'aime frites.</p>

Summary of key use cases

Use cases for Teachers

Generating age-appropriate lesson plans

Teachers report spending significant time on the generation of lesson plans, as good planning and structuring are key to ensuring that students make progress and that lessons themselves are effectively delivered. During the hackathons, teachers explored using base GPT models to both plan age-appropriate lessons – in Key Stage 3 (KS3) French or A level computing – and adapt an existing lesson plan for Year 7 History to cater for different ability levels.

Results: The outputs from using base GPT models to generate lesson plans were not particularly positively rated by teacher participants of the hackathons. They scored the outputs of their testing at an average of 3.7 out of 5 for time saving. However, meeting national standards, improving outcomes and likelihood of use were scored between 2 and 2.3 out of 5. In addition, usability was scored at 1.7 out of 3, emphasising the need for improvement in functionality or additional tooling in future. A PoC with additional functionality, allowing an LLM to refer to any necessary contextual data (which would be synthetic during any development work), could surmount the above drawbacks.

Table 4: Summary of user feedback for generating age-appropriate lesson plans

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Generating age-appropriate lesson plans	1.7	3.7	2.3	2.0	2.3
Average across all use cases	2.1	3.7	3.2	3.2	3.6

Generating effective questioning ideas

During the hackathons, teachers explored: using base GPT models as a diagnosis agent to understand misconceptions among students; a tool to develop effective questions for teachers to evaluate understanding; and a chatbot to gauge student understanding. Successful approaches included employing an ‘inner monologue’ technique (requiring base GPT models to outline their reasoning for returned outputs) and prompting base GPT models with a combination of a topic specific lesson and overarching learning objectives to ensure questions generated were relevant.

Results: On average, this use case was rated joint highest by users at 4 out of 5 for whether the final output met established national standards such as the national

curriculum. Users also rated output for this use case highly in terms of time saving (4.7 out of 5) but moderately in terms of usability (2.3 out of 3), improving outcomes (3.7 out of 5) and likelihood of use (3.3 out of 5).

Table 5: Summary of user feedback for generating effective questioning ideas

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Generating effective questioning ideas	2.3	4.7	4.0	3.7	3.3
Average across all use cases	2.1	3.7	3.2	3.2	3.6

Generating lesson materials

During the hackathons, teachers explored ways for GPT models to assist in the generation of lesson materials. Specifically, teachers tested whether the models could generate homework for students based on a lesson PowerPoint and transcript, adapt text extracts to a lower reading age and develop progressive worksheets to boost exam confidence in English as an Additional Language (EAL) students.

Results: Overall, user scoring was positive with participants giving the model average scores of 3 out of 3 for usability, 4 out of 5 for following national standards, 4.5 out of 5 for time saved, and 5 out of 5 for improved outcomes. They also rated their likelihood to use such a model for generating lesson materials as 4.5 out of 5. Considerations for future development include providing guidance for prompting and delivering an efficient and effective user experience for users.

Table 6: Summary of user feedback for generating lesson materials

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Generating lesson materials	3.0	4.5	4.0	5.0	4.5
Average across all use cases	2.1	3.7	3.2	3.2	3.6

Assigning marks to work submitted by students

In the hackathons, teachers tested a range of prompting approaches to improve GPT models' effectiveness and accuracy when presented with Year 4 essays and asked to mark them, assign grades and provide feedback based on the national curriculum.

The group applied a variety of approaches such as using more detailed prompts to specify what an output must contain, and applying a modular approach to prompting – requesting that the LLM generate an exemplar essay based on a real-world mark scheme, and use this to benchmark the grading of a pupil essay. In addition, the group was able to improve the model’s output by reducing its ‘temperature’, described by OpenAI as “a parameter that controls the “creativity” or randomness of the text generated”, as well as experimenting with the more recently released GPT-4 model.

Results: It was evident from the hackathons user feedback, as well as our previous user engagement, that the development of an accurate AI marker could save time for teachers and schools. However, teachers scored this use case low on meeting national standards (2.5 out of 5) and moderately on usability (1.8 out of 3), time saving (3.3 out of 5) and outcome improvement (3 out of 5). Likelihood of use however was scored slightly higher at 3.8 out of 5 suggesting that, should the associated challenges be resolved in future, a tool such as this may be useful but the bar for accuracy is high.

Table 7: Summary of user feedback for assigning marks to work submitted by students

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Assigning marks to work submitted by students	1.8	3.3	2.5	3.0	3.8
Average across all use cases	2.1	3.7	3.2	3.2	3.6

Generating personalised formative feedback and addressing student misconceptions

Teachers participating in the hackathons investigated whether GPT models could be useful for generating specific feedback for individual Year 4 students on pieces of English work. The aim was to investigate a scenario in which the models were asked to provide positive feedback, suggesting ways for students to improve their work. The team began by using different prompt structures with GPT-3.5 to elicit specific feedback on each essay (asking for both teacher facing and student facing feedback) and extracting this into more structured formats (specifically, JSON files) before testing the viability of using these JSON files to request different kinds of outputs.

Results: The group felt that the inaccuracies observed could be mitigated in future with: better quality mark schemes for the model to interpret; more expert input; and more training data for the model to see a range of performance and greater context. Despite these potential areas for improvement, the group expressed doubts that it would be

possible to improve accuracy to the level desired by teachers. LLMs therefore may be better used to aggregate or swiftly structure human feedback to aid lesson planning, rather than to produce feedback unaided.

Table 8: Summary of user feedback for generating personalised formative feedback and addressing student misconceptions

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Generating personalised formative feedback and addressing student misconceptions	1.8	2.3	2.0	2.0	2.3
Average across all use cases	2.1	3.7	3.2	3.2	3.6

For the full breakdown of use cases for teachers – including GenAI as a teaching aid, supporting students with SEND, and writing end of year reports – readers should consult Section 3 of the Summary of Hackathons Findings Report (November 2023), included here as Annex 2.

Use cases for School Leaders and Administrators

Generating drafts of statutory school policies

School leaders are often tasked with creating or updating school policies. This is a time-consuming process with the necessary preparation time impacting staff leave periods and involving multiple meetings across the Senior Leadership Team. School leaders and administrators examined whether GPT models could act as a support tool to generate drafts of statutory school policies for further review before implementation.

Results: School leaders and administrators scored this use case positively across almost all criteria – rating the generation of draft statutory policies between 4 and 5 out of 5 for time saving, meeting established standards, improving outcomes and likelihood of use. Usability was rated as an average 2 out of 3, indicating users felt that outputs were somewhat usable but would require further adjustment by a human before being trusted.

Table 9: Summary of user feedback for generating drafts of statutory policies

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Generating drafts of statutory policies	2.0	4.0	4.0	4.0	5.0
Average across all use cases	2.1	3.7	3.2	3.2	3.6

Pupil or class data analysis and synthesis

During the hackathons, school leaders and administrators explored whether GPT models could be used to ingest a synthetic pupil data set (such as dummy IDs, target grades, and reading ages) and analyse the specific assessment data to identify personal and group-level capability gaps. The group tested whether GPT models were able to rank students overall but also perform granular, question-level analysis of mock exam scores. This could enable class teachers to identify patterns and variation in understanding and develop next steps much faster.

Results: Despite the concerns and difficulties encountered during the exploration of this use case, school leaders and administrators scored this use case moderately. School leaders and administrators scored this use case at or above 4 out of 5 for likelihood of use and whether outputs met established national standards. They rated this use case at an average of 3 out of 5 for both time saving and outcome improvement. Usability was also scored moderately at 2 out of 3. This indicates that there is still an appetite among this user group to test this use case further with alternative approaches to those considered in the hackathons.

Table 10: Summary of user feedback for generating pupil or class data analysis and synthesis

Use case	Usability	Time saved	Meeting national standards	Improving outcomes	Likely to use
Pupil or class data analysis and synthesis	2.0	3.7	4.0	3.7	4.3
Average across all use cases	2.1	3.7	3.2	3.2	3.6

For the full breakdown of use cases for school leaders and administrators, including generating parent-carer communications, readers should consult Section 4 of the Summary of Hackathon Findings Report (November 2023) included here as Annex 2.

Use case for students

Language learning assistant

During the hackathons, Faculty tested the viability of using GPT models as an assistant for students learning modern foreign languages such as French, Spanish and German. The group of students in attendance explored how GPT models could help them understand grammar concepts, correct their work and practice conversations in their target language.

Insights: Due to time constraints on the day and to ensure robust safeguarding by avoiding direct contact with students, the student user group did not complete the same post-use case survey as the teachers and school leaders and administrators groups. However, there were still key insights drawn from the student hackathons session.

One clear finding was that a useful language learning assistant would need to be able to accomplish several tasks concurrently as what appears to be a simple request can spread into various areas of competence. For example, in the case where a student asked GPT-3.5 to have a practice conversation in their target language using a certain vocabulary list (Spanish GCSE vocabulary), the conversation evolved into asking the LLM which mistakes they had made, before requesting exercises to strengthen their understanding of the grammar points they struggled with.

The group also found that hallucinations can undermine users' trust in model outputs. Occasionally, the LLM made mistakes when correcting answers to a list of multiple-choice questions it had created or hallucinated and corrected non-existent mistakes in student responses to short answer questions. The more advanced GPT-4 model was able to correct some of these errors and was substantially better at correction than GPT-3.5. The students agreed that GPT models could be usable for practising conversations, but trust in the model remained a major concern, and they indicated that they would be more likely to use ChatGPT as a supplemental tool to the other methods they use for language learning.

For the full breakdown of use cases for students, readers should consult Section 5 of the Summary of Hackathon Findings Report (November 2023) included here as Annex 2.

User Research Findings

Introduction

Teachers' uses for GenAI in education

As noted in section 3 of this report on the hackathons and the interim report on the stakeholder consultation, there is a broad scope for the potential use of GenAI in schools. These range from the facilitation of administrative tasks such as data management and report writing to more creative tasks such as material generation and the production of exemplar essays. While much of the research in this field predicts positive outcomes resulting from GenAI adoption, such as saved time and teacher empowerment (Kim et al., 2020; Wang et al., 2024; Zhang & Tur, 2023), there are also a wide range of concerns about its implementation in the classroom, including issues relating to ethical considerations (Akgun & Greenhow, 2021; Nazaretsky et al., 2022). Many teachers still have limited knowledge about GenAI (Chounta et al., 2022), and may be put off using it in their classrooms if it is not perceived to be useful, useable, and trustworthy (Choi et al., 2023; Department for Education, 2023, 2024).

GenAI for feedback and the current study

Feedback is one of the most powerful stages of the learning process, where teachers support their students to close the gap between where they are and where they should be in their learning (Hattie, 2012). However, marking and providing feedback often take up a significant amount of teachers' time (OECD, 2018). Workload is one of the main reasons why teachers consider leaving the profession (Räsänen et al., 2020), and a national survey in England found that 46% of teachers felt they spent too much time on marking (Department for Education, 2024). Research has also shown that the more time teachers spend on these tasks, the lower their wellbeing tends to be (Jerrim & Sims, 2021), highlighting the importance of finding ways to support teachers with this process. One potential method that has recently garnered interest is to use GenAI tools to generate feedback on students' work. In recent years, a range of GenAI tools have been released for this purpose and one estimate suggested this could save teachers around three hours of marking time a week if implemented successfully (Bryant et al., 2020). However, much of the existing research on the efficacy of these tools has been focused on higher education (e.g., Lee, 2023), produced mixed results (Aloisi, 2023; Cavalcanti et al., 2021), and highlighted concerns relating to their reliability (e.g., Li et al., 2023). We still have limited knowledge about the potential accuracy or reliability of these tools, or their potential impact on learning and teaching in schools.

Although there is some evidence relating to teachers' broad perspectives on the use of GenAI in education (Department for Education, 2023) there is less comprehensive

research that covers the practical, pedagogical, psychological, and social considerations raised by teachers and students in relation to GenAI's use for feedback and across education more broadly. This qualitative study therefore used semi-structured interviews (teachers) and focus group discussions (students) to examine teachers' and students' perspectives about GenAI following a trial of a PoC tool designed to support teachers in providing personalised feedback to students. The research focused on the opportunities and risks presented by the use of GenAI tools for feedback, and the facilitators and barriers that could help or hinder teachers' adoption of GenAI tools in schools.

Methods

This study employed a qualitative methodology, involving interviews with teachers and focus group discussions with students, across several schools in England. Full details of the methods used are in Annex 4.

In January and February 2024, teachers (N = 12) with varying levels of experience with GenAI trialled a new PoC GenAI feedback tool and took part in a 45 minute one to one interview about their views on GenAI. The PoC tool was developed by Faculty and was designed to provide feedback on the spelling, punctuation, grammar, and vocabulary of Year 4 students' written work, and produce a personalised revision activity for students that aimed to address a key area for improvement.

In addition, secondary school students (N = 9) took part in online focus group discussions about their perspectives on GenAI use in school, focusing on teachers' use of GenAI for feedback. We chose to speak to secondary over primary students as this was deemed more appropriate to elicit nuanced discussions on the subject of GenAI, through the medium of remote focus group discussions.

Ethical considerations

Ethical approval for the study was granted by the NIoT's ethical review board. All participants shared written agreement prior to the study. Teachers received an information sheet and had opportunities to ask questions, and written agreement was provided by participating teachers and a senior leader at their schools. For students, headteachers provided written agreement for their participation, and parents and carers were informed of the study's aims and methods and were given one week in which to opt out on behalf of their child. At the start of the discussions, students were presented with information about the study by an NIoT researcher and were offered the chance to either withdraw or provide verbal assent to take part. For safeguarding purposes, students were joined in the (physical) room by a member of school staff for the online focus group discussion.

Results

These results are supplemented by the additional user research designed to inform the development of the PoC tool (see section 4). The PoC was not designed for development to deployment, but rather to support learning about how GenAI can be applied to an education context. As a result, the findings from this research were not used to inform subsequent phases of development of the tool as they would usually be in the development of a product or service.

The results are presented in three main sections: teachers' prior experiences with GenAI; opportunities and risks associated with using GenAI for feedback; barriers and facilitators to GenAI use in education.

Teachers' prior experience with AI

To contextualise their views on these subjects, we begin with a brief overview of the teachers' attitudes towards AI in general. Despite very few self-reported knowledgeable or confident users of AI in the sample, most teachers had experimented with AI in one or more aspects of their professional life. This is broadly in line with survey data suggesting that in November 2023, 42% of teachers had used AI to support their roles (Fletcher-Wood, 2023). In our sample, these activities ranged from AI for report writing (Teacher 8) and idea generation (Teacher 7) to AI for adapting learning tasks for individual needs (Teacher 6) and generating materials (Teacher 3):

"It might sound a bit weird, but I've got lots of PowerPoints with bullet points on them and I don't really want the kids just copying that down. So I've actually asked ChatGPT to basically turn all my bullet points into paragraphs. So the kids can then re-read them back and highlight the key points."

Teacher 3, Secondary

"I said 'Oh, let's hold on a second' [...] 'what if we ask Chat GPT maybe to sort of give some ideas?' So we put the prompts in again, similar to my talk for writing cycle that I've done previously."

Teacher 7, Primary

Although these teachers had at least considered how they might use AI, some initially reported negative feelings towards AI:

"It's probably something I've kind of avoided [...] Yeah, I am someone that mistrusts technology in that way."

Teacher 5, Secondary

In certain cases, teachers also reported that the negative states of uncertainty and intimidation surrounding AI were precisely what had motivated them to learn more about it:

"One of the reasons why I'm quite interested in AI is because I'm also quite scared of AI and quite intimidated by it. So, I kind of want to understand a little bit, kind of get to grips with it."

Teacher 4, Secondary

"I went on an AI webinar and I spoke with the headteacher and I said, 'We don't really know anything about this', I find it fascinating."

Teacher 6, All-through Special School

Despite having limited experience with AI, teachers were generally open to exploring its potential in education. Many also felt some trepidation and lack of certainty about what it would mean going forward.

GenAI for feedback

Although it will not be deployed, teachers were generally very interested in discussing the new PoC tool and what it could mean for teaching.

"It's got my head spinning now"

Teacher 7, Primary

Teachers identified several positives about the tool such as being able to choose a specific focus for the feedback (Teacher 4), but also several areas for improvement. For example, the feedback was deemed "a bit wordy" (Teacher 8), and several teachers questioned how it would cope with handwritten work.

Conversations about the PoC tool acted as starting points for exploring teachers' perceptions of the opportunities and risks associated with using GenAI tools for assessment and feedback. It also sparked wider debates about the heart of what it means to be a teacher. These are summarised visually in figure 5.

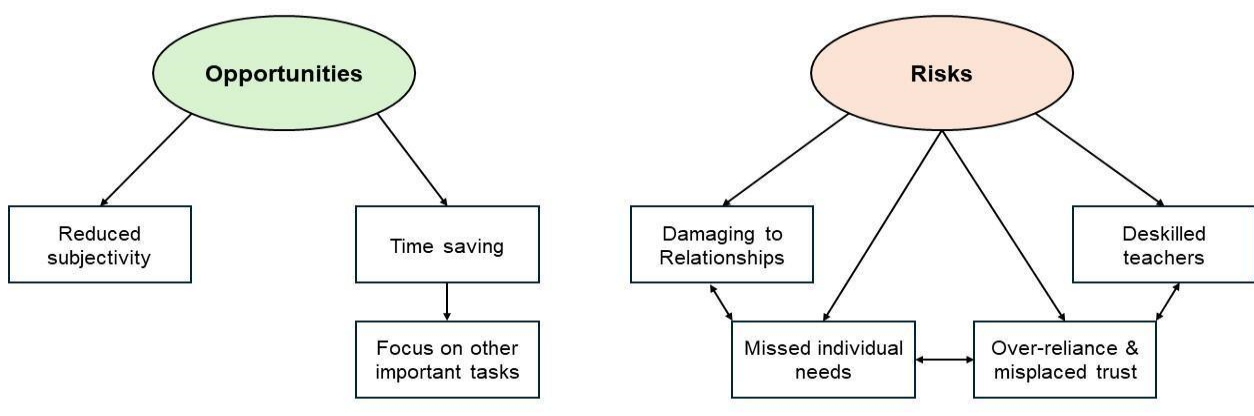


Figure 5: Opportunities and risks associated with using GenAI tools for assessment and feedback

Opportunities

Reducing subjectivity

Teachers described two main opportunities that could arise from using GenAI feedback tools. The first was the potential for standardising outputs, thereby reducing subjectivity and inter-teacher discrepancies in judgements, and increasing the consistency of feedback. This was mentioned by a small number of teachers in relation to moderation of work:

“Last year I went into a room with all the top teachers in [city]. They're all there, all Year 6 teachers, and discussing [...] every single text. No one was quite sure. Is it this level? Is that level? [...] If you want to create, you know, the AI, the formula for that, then it takes out all the subjectivity.”

Teacher 8, Primary

“Moderation is hard because you're personally involved. So that's why, if you're a teacher, moderating your own work is completely different [...] You want to fight for that child, so that's why I think if you remove that [...] that's why I think the AI system could be quite strong.”

Teacher 12, Digital Learning Lead across a MAT

This suggests that using GenAI to support feedback could help to remove the subjectivity and bias from teachers' own assessments. Similarly, one teacher suggested that GenAI could be used to support inexperienced teachers with the uncertainties of working with a new year group:

“If there's a teacher moving to a new year group, having a tool that could then help them with where the mistakes are – what is expected for that particular year group [...] and you kind of lose some of those discrepancies between the teachers as well.”

Teacher 1, Primary

Time saving

The second key opportunity was that GenAI feedback tools would potentially save teachers time on marking and that this time could be used for other tasks, which would support the learning of their students:

“It's the case of it saving you the time. So, then you can spend that time analysing how to therefore take [their] journey to the next step. So the teachers' role becomes less of the assessor and more of the next step, the teaching again, it puts the onus back on track, doesn't it? OK, so the computer program's telling me that it's assessing you as this, so therefore I am a teacher and actually now I can teach you.”

Teacher 8, Primary

This suggests that the time saved on a task such as feedback could be reinvested into the learning process by supporting learners based on the feedback generated and teaching accordingly. Some teachers therefore felt that there were opportunities to save time and strengthen practice using GenAI for feedback, however, implicit in these hypothetical opportunities was the assumption that the GenAI tools would be accurate and reliable.

Risks

Teachers also identified a range of risks that they felt should be considered before implementing GenAI for feedback. Teachers generally discussed these risks in greater depth than they did opportunities. Firstly, and in contrast to the positive potential for time saving, some teachers expressed concerns about whether such tools would save time due to their lack of trust in the tool:

“I think I would end up going back over and reading their essays anyway to see if the AI was correct.”

Teacher 3, Secondary

This suggests that in order for the time to be saved, teachers would need to trust the accuracy of the tool, although it remains important to note that this time saving does not remove the need for a ‘human in the loop’, and the importance and centrality of the teacher’s professional judgement.

Changing role of the teacher

Teachers also expressed risks that may exist regardless of the tool’s capacity to give precise, appropriate, and accurate feedback. One such concern was that using a GenAI tool for feedback would mean changing the role of the teacher and the learning process in a significant way. This theme reoccurred across the interviews, with teachers reflecting on the centrality of teachers’ involvement in students’ work for the learning process:

“I don’t think I could let it go in that way because these are my students and I should be the one giving them that feedback.”

Teacher 3, Secondary

“It’s our job to know their barriers to learning. It’s our job to know how to deliver that feedback in a way that will actually ensure that the progress happens.”

Teacher 11, Primary Special School

“The next step is that discussions around marking through AI and then you just think, well, if the students are writing it through AI and then we’re marking it through AI, then the whole thing is pointless. Like what? None of us need to be engaging in this activity at all, It’s an empty hollow exercise. If the students aren’t doing the work and we’re not doing the work, then what was the point of writing or setting that assignment?”

Teacher 2, Secondary

In this quote, Teacher 2's worries around her own changing role are further compounded by her fears around broader issues arising from students' GenAI use, thus making the entire feedback process redundant in her eyes.

Overreliance and deskilling teachers

Another risk was that using GenAI for feedback may deskill teachers, especially those new to the profession:

“What are they [teachers] in the classroom for? [...] It doesn't require any skills whatsoever from that teacher other than some basic ICT skills.”

Teacher 10, Secondary

“I think my concern is you would have a lot of teachers that would just rely on that and they would actually lose their professional judgement [...] You could have, you know, ECTs that come into school and only use that and that's not gonna develop them professionally.”

Teacher 11, Primary Special School

“You need to be able to identify these errors yourself. You can't just be relying on technology all the time [...] My fear with that would be that you raise a generation of lazy teachers.”

Teacher 9, Primary

Concerns about teachers becoming over-reliant on such tools were also echoed by others. While some teachers reported that they would be likely to trust such feedback tools, they also reported concerns that both they and their colleagues could become lazy and excessively trusting of the technology, in part due to their unsustainable workloads:

“I think I could get quite too reliant on it if I had really positive experiences at the beginning, you know, because I have four Year 7 classes – that's over 120 Year 7 students. I don't have the time to go through every single test, right? What did they put for this question? What did they put for this? So, I think, you know, anything I can do to save my time – I'm all for it.”

Teacher 4, Secondary

“I think there could be a danger, as I said, like teachers getting a bit, you know, lazy [...] I think that I would get quite reliant on it.”

Teacher 4, Secondary

“There's the temptation, I guess, for somebody [...] to not actually really have a proper look at it and see what they've done well and what they've not done well. So [...] that personal interpretation of what they've done right and what they've

done wrong, they might miss that because they might not take the time to actually read a summary of what they've done.”

Teacher 10, Secondary

The concern about losing track of where students are in their learning was shared by many of the teachers:

“I know that obviously the whole point of AI is to take that job away from me, but as a teacher, I think that's quite an important part, and I'd worry that somehow AI would lose the sort of nuance of what the kids should be doing”

Teacher 3, Secondary

Teacher-student relationships and individual needs

Many teachers expressed a concern that handing feedback over to GenAI would remove an intrinsic aspect of the teacher-student relationship. Some teachers suggested that the acts of work submission, feedback, and response to feedback are parts of an important cycle between student and teacher where the student can open up to their teacher, helping their teacher to better understand them:

“What I find really problematic about using AI to mark a students' work is that there is no relationship in that and so many students want, they want you to read their work because this isn't just about them producing a piece of quality work [...] If they thought that you were just going to run that through an AI marker, I think their investment in that is gone. They want you to read their work. They want you to know and understand who they are as an individual. They want to impress you often. They want to interest you in who they are.”

Teacher 2, Secondary

“I think, to not even read it as a human would be really detrimental to the relationship”

Teacher 2, Secondary

“I would lose that kind of rapport, I suppose, to some extent with the kids and that kind of ongoing conversation that is there.”

Teacher 3, Secondary

“I think the downside for me personally is I don't think I'd get to know my students and their quirks as effectively”

Teacher 5, Secondary

This highlights teachers' perceptions of feedback as being more than an academic exercise and how using GenAI to provide feedback could result in negative socio-emotional consequences. Additionally, relating to teacher-student relationships, some teachers expressed the importance of knowing one's students both academically and behaviourally when providing feedback.

“You still need to read it all through yourself to see what the AI has put, and whether or not it has focused on what you want it to focus on for that student as an individual learner.”

Teacher 2, Secondary

Teacher 2’s concerns indicate a need for teachers’ involvement in the process to ensure that outputs are appropriate for the individual learner. However, the perceptions of teachers in special educational settings suggested that even such mitigations may not make written feedback via AI a viable option for many students with special educational needs:

“There are a lot of other issues that we need to take into consideration for our pupils whenever we are doing any learning, and you could have someone who gets really thrown by any negative marks on their paper or rather than a verbal [feedback] or there are people who don’t like verbal feedback. So, it’s all individualised that way.”

Teacher 6, All-through Special School

“They never read it. They never respond to it. At the moment the feedback needs to be immediate. You know, in the lesson, not after, and especially in a special school because they don’t really have the capacity to go back and reflect. And also, if there’s something that they found particularly tricky and you’re asking them to reflect upon it, you could actually trigger behaviour.”

Teacher 11, Primary Special School

While both describe the importance of individualised and carefully considered feedback, Teacher 11 also suggests that the written feedback produced by GenAI would likely be inaccessible for many of her students. This further demonstrates the perceived importance of having teachers at the heart of the feedback process who understand their students on a social, academic, and behavioural level.

Students’ views on using GenAI for feedback

Similar themes came up in the student focus group discussions. Regarding opportunities, some students noted that an GenAI feedback tool could save teachers time that could then be used to cover more content when teaching:

“I think it’d be easier for the teachers just to put it in, so then it’s not so time consuming as they don’t have to hand mark it [...] I think more content could be covered instead of just focusing on marking a particular piece of homework.”

Student 3, Focus Group 2

However, there were also concerns among students, with many suggesting that an important part of the feedback process is the teacher’s understanding of their students’ learning needs:

“AI wouldn't know [...] the way you learn like a teacher would [...] unless there was a way where you could tell it beforehand.”

Student 2, Focus Group 2

“Artificial intelligence won't ever really be able to match how a teacher has known you through the class.”

Student 4, Focus Group 2

This echoes the teacher's sentiments and shows that students value the personal level of understanding that teachers bring to the feedback process. One student also described the sense of demotivation that may arise from knowing that their work would not be checked.

“And demotivation from that – that no one's ever going to check it. And also, it doesn't show the teacher [...] [It] doesn't matter if you're improving or not when you're like learning good stuff like that.”

Student 1, Focus Group 1

While it is possible that the academic subject in question will influence the extent to which GenAI feedback would damage the teacher-student relationship (for example, maths tasks are less likely to facilitate the sharing of personal information than literacy or humanities), a student's motivation to complete their work to a high standard may nonetheless be diminished if they are aware that it won't be read.

Teachers' needs for GenAI uptake in schools

The interviews also included opportunities for teachers to explain potential barriers and facilitators to GenAI adoption in their schools. In many cases, these were two sides of the same coin, so are analysed together below as 'Needs' rather than discreetly. The most commonly cited facilitators that would support the use of GenAI in schools included time, training, expertise, and funding (see figure 6).

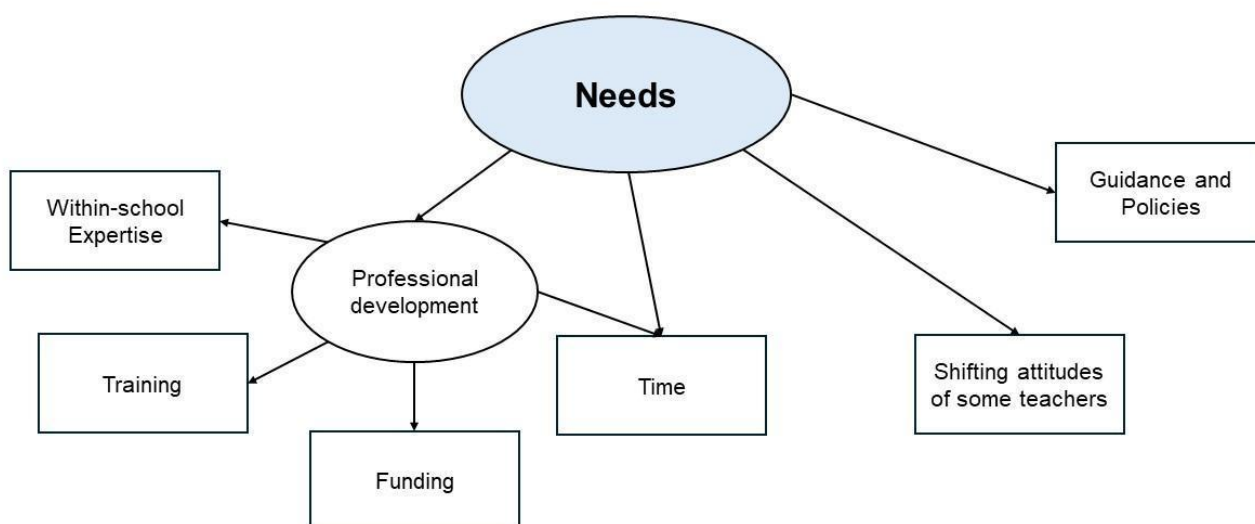


Figure 6: Needs identified by teachers based on their perceived barriers and facilitators to GenAI adoption in schools

Shifting attitudes of teachers

Perception of teachers was most commonly cited in relation to the implementation of GenAI, and several teachers suggested there would be a degree of scepticism unless it was clear and evidenced how the tools would support their practice:

“It’s about convincing us that it’s worth putting the time and effort into learning how to use it.”

Teacher 3, Secondary

“I think all staff to some extent, you know, feel a bit of trepidation because it’s technology and the children stereotypically understand the technology better than most teachers.”

Teacher 8, Primary

“I think it would have to be made very clear that it is to benefit workload and not add to it.”

Teacher 9, Primary

If teachers’ concerns about the time investment required and the efficacy of the tools can be allayed, they may be much more likely to implement them. However, until those needs are met, this culture of hesitance within some schools may delay the adoption of new technology.

Some teachers suggested that individual differences between teachers may influence their eagerness to adopt GenAI in their classrooms. One teacher felt that their age played a factor in confidence around uptake of GenAI and other technology, and expressed their preference for ‘old school’ ways of working:

"I think my age probably says it all really. I'm not. I'm not really super tech savvy. The youngsters in my faculty are far more tech savvy than I am [...] I'm very precious about marking and things [...] and although I don't like it, I feel like I need to do it and I do it and I'm very old school."

Teacher 3, Secondary

Teachers of a range of ages also picked up the point about teachers' years of experience and age, sharing the view that older teachers may be less interested in GenAI:

"I think also a lot of teachers are quite hesitant for change, to change, so I feel... don't mean to be rude, but maybe that older generation of teachers would be quite thrown by [...] you know, sort of worried about it. You know, there's a lot of teachers that do struggle with their IT and technology."

Teacher 4, Secondary

While our data does not offer extensive insights into how teachers with more years of experience feel about GenAI, these quotes suggest that discussions around age-related differences in willingness to adopt new technologies occur within schools. Whether based on truth or stereotypical judgements, these show that some teachers perceive differences in colleagues' readiness to adopt GenAI based on age. One teacher suggested that teachers with more years of experience may be less likely to stay in the profession if forced to implement GenAI technologies:

"I am seeing teachers hitting around 60 and thinking 'I can't do this anymore' [...] thinking 'it's not worth me now having to engage with the whole new system of working. I can't do that again.'"

Teacher 2, Secondary

Time

There were tangible facilitators to GenAI use identified in the interviews. Teachers widely reported that to effectively implement GenAI tools in school, they and their colleagues would require time set aside to learn about the tools and become confident in their use. Without this time, it was suggested that teachers would likely disengage from the GenAI learning process:

"I think time would be something that teachers are saying. I don't [have] time for this and kind of throw it off because of that."

Teacher 4, Secondary

"The most key element, is if you want us to engage with this, you have to carve out time into our timetables in order to be able to play around with it. Create it. Try it out. And so much of our timetable, I mean, this is the inbuilt catch 22, isn't it? The kind of irony of situation is so much of our time is so busy already. There is so little time to take on board new things. So, every inset day we are bombarded with more initiatives, more ideas."

Teacher 2, Secondary

"If you're asking someone to go off and do it themselves and saying there's information – go and look at that yourself it's not gonna happen [...] there will have to be a dedicated time set aside because people are just very busy and they're not gonna go off unless they can see [...] the potential in an allotted time."

Teacher 6, All-through Special School

This demonstrates the somewhat circular situation that teachers find themselves in, whereby they do not feel that they have enough time to invest in learning how to use tools that are designed to help save them time in the long run.

Training and expertise

Teachers also expressed a need for training and someone with expert knowledge to support the transition towards GenAI adoption:

"I feel as long as the right training was in place in terms of how to use it, with it being technology, [...] I think any teacher that's worth their salt would welcome it with open arms because it's seen as a tool to support them."

Teacher 8, Primary

"[We would need] someone with a lot more knowledge on which are the best [AI tools] out there. Which ones lend themselves better to our primary and our primary curriculum and yeah, certain areas and I know there is training out there, it's just tapping into it and having somebody, I guess, championing it."

Teacher 1, Primary

Guidance

Some teachers seemed to be at a crossroads with regards to their own sentiments towards adopting GenAI, in part due to uncertainty about what others might think if they knew that teachers were using it:

"So actually, the first time I used it to do my references I, I kept it quite quiet because I didn't know if morally that was the right thing to be doing – whether the people would see that as me obviously not taking the time to obviously write those references individually and bespoke"

Teacher 10, Secondary

"I don't think parents would like it [...] I think society is very much like, you know, teachers need to be working really hard. And I think that when you, when you hear AI, your first impression is oh, you know, making your life easier, you know, cheating, you know, that sort of thing because it does have those negative connotations [...] so I think you'd have to be really careful how you sold it to parents if you were gonna use it in school."

Teacher 4, Secondary

This suggests that some teachers may find themselves conflicted about whether they should be using GenAI. Moreover, if GenAI tools are to be recommended for use by teachers, such guidance will need to clarify exactly how teachers should be using them:

“We haven’t had anything on how teachers might use AI.”

Teacher 2, Secondary

“I think the only thing that you would necessarily need from someone like Ofsted is the OK that things are done like that. Kind of like their approval, in a way [...] because there is always that fear that, you know, the teachers ultimately are the ones that are accountable.”

Teacher 9, Primary

“Schools need a policy of what are the boundaries that we are working in, what’s safe practice so that we know that we’re putting our children and children’s data, in particular in a safe environment that we can play and explore. And I think if we don’t make sure that is at the forefront of what we do, the danger is that we will have lots of these apps and new things going out and places where we’re sharing data.”

Teacher 12, Digital Learning Lead across a MAT

These quotes suggest the need for greater clarity and guidance about the boundaries of GenAI in education and teachers’ roles in a GenAI-assisted school system. It is possible that with these clarifications, some of the risks identified above (such as, risk of over-reliance etc.) may be partially mitigated.

Limitations

This qualitative study was conducted with a sample of 12 teachers and 9 students. While we aimed to be comprehensive during these interviews and focus groups, the findings from this small sample should not be generalised across the education sector. Although we attempted to recruit participants from a diverse range of geographic locations within England, we note that all participating schools were rated either Good or Outstanding by Ofsted, and two-thirds had below the national average levels of students eligible for free school meals. Moreover, 8 of the 9 schools were part of Multi Academy Trusts (MATs). As the needs and priorities of schools in different contexts may vary significantly, we recommend that future research explores GenAI adoption in schools deemed to be inadequate or requiring improvements, those in areas of greater deprivation, and those not part of MATs.

Taking these limitations into consideration, the data revealed issues that were important and concerning to teachers and will likely be transferable to other schools across the sector. Some of these, such as time saving, mistrust in GenAI, and a need for professional development activities were broadly in line with existing research, but our

research also includes perspectives on lesser explored issues, such as the potentially detrimental impacts to teacher-student relationship that may occur if GenAI tools are used for feedback.

Conclusions

User research conclusions

Teachers' and school leaders' insights reveal that there is not one uniformly adopted stance on the idea of using AI in schools. Most interviewees were positive about some aspects of AI but cautious about others. The teachers and students in our study expressed high levels of interest in AI for feedback and saw a number of important opportunities, but also discussed their concerns and risks in depth.

Although many users were able to see the potential time saving and standardising benefits of using AI tools for feedback, they also highlighted considerations that would have to be addressed before any widespread rollout. Some do not have universal solutions, for example, with some teachers reporting that they would not trust the tool at all, and others suggesting that teachers may become overly trusting and therefore not take the time to check its judgements. Interviewees were concerned that these issues could result in educators losing the ability to make accurate professional judgements, and potentially lose track of where their students are in their learning journeys. A common theme in the interviews was that feedback is more than an academic exercise – it is also social, and to take that task out of the hands of teachers may risk damaging the relationship between teacher and student, which is a key predictor of students' academic development.

When discussing potential barriers and facilitators to AI adoption, interviewees reported a need for time, training, funding, and expert help to increase teachers' own expertise in AI use. Further, teachers reported needing support and guidance to feel that AI use is acceptable, important, and safe.

However, regardless of what guidance and professional development opportunities are in place, it is possible that certain teachers may not be swayed towards AI implementation. Further, discussions about willingness to adopt AI often veered towards the subject of age and experience, with some teachers perceiving older teachers as being more reluctant to adopt AI tools. This suggests that first, we need more research on approaches to develop teachers' confidence in using AI at all stages of their career, and second, it would be helpful to understand further how discussions around AI may play into perceptions of difference and possibly stereotyping, related to age and other factors among staff.

The methodology we employed facilitated debates about the use of AI and the role of the teacher that may not otherwise occurred, had there not been practical examples to discuss, such as inviting teachers to trial a PoC AI tool before taking part in an in-depth interview. We therefore recommend that future research in this area consider employing similar methods to capture the developing picture.

The findings both highlight the importance of this critical area, and express the need for caution when moving forwards, emphasising both a need for tools that are accurate and fit for purpose, and a greater understanding of the consequences arising from their implementation by teachers in practice. Future research should focus on the hypothetical perceptions of stakeholders and should also implement short- and long-term trials with AI in the field of education, measuring academic, behavioural, and psychological outcomes, offering recommendations for the mitigation of any negative consequences.

Suggested areas of future focus

Some of the most significant findings from the work conducted as part of this project related to the degree to which GenAI has the potential to benefit the education sector over the long-term and the enthusiasm of many in education to take advantage of this powerful technology. Further detailed exploration of the potential benefits of GenAI, in its current state as well as following its inevitable further development, will enable government to continue to shape the policy landscape and how the technology is adopted by the sector. However, it is also clear that there are a range of cultural, logistical and technical constraints that prevent schools and individuals from fully exploiting the potential of GenAI.

Following the hackathons, Faculty, NIoT and ImpactEd Group conducted a review of the project's findings up to that point, to identify a number of these constraints. The list of potential barriers to the adoption of GenAI in schools has been updated in light of the further findings from the second phase of the project, including the PoC build, the user testing and the user research programme. Linked to these barriers, we have suggested a range of areas for focus for future work that may support the broader take up of GenAI, as well as mitigating risks and increasing overall impact.

Perceptions of the impact of AI on the role of teachers, and a lack of clear expectations regarding its use

Some teachers raised concerns that using an AI tool for feedback would change the role of the teacher and this would affect the learning process in a significant way. Teachers also raised the importance of close human interaction to students' personal development as well as their learning. This concern was not limited to giving feedback but was echoed by teachers and school leaders in discussions of other use cases, including lesson planning and writing student reports.

Teachers also raised concerns about whether using AI to support or replace elements of their role is the right thing to do in terms of best educational practice, as well as whether it is morally right. They were also uncertain about what other teachers, and parents or students would think if they were found to be using AI, as well as authorities such as their school leaders or Ofsted.

Suggested areas of focus:

- exploration of the role of GenAI in teachers' and school leaders' work, with a special focus on their interactions with students and how these may be affected by the introduction of GenAI;
- development of further guidance for teachers and school leaders on best practice in the use of GenAI including an assessment of the existing evidence and its limitations; and
- commissioning of related research where gaps are identified in the evidence base.

Challenges of integration with existing systems, and disparities in access to technology

Most of the GenAI use cases in this project would require a regular feed of existing, context relevant data into an LLM based service. This would require LLM integration via Application Programming Interface (APIs) with existing systems that capture this data at source, such as information management systems and school databases. This may be more straightforward for some applications than others: for example, enabling access to a bank of lesson plans is likely to be more achievable than to student personal data, given challenges with privacy, frequent updates, and the systems used to store this type of data. In some cases, commercial data management solutions or APIs developed by the edtech provider can close this gap, but this may not always be possible, or may become prohibitively expensive.

Schools' varying digital maturity and student access to technology were also raised as potential barriers to take up of GenAI. Varying levels of digital maturity across schools would need to be considered and possibly provisioned for if such tools were to be adopted on a national scale. For example, for schools in some areas internet bandwidth could come under strain with increased use of LLMs. And for the student-facing use cases, it is important to note that not all students have access to the same devices or internet availability at home, creating a divide between those who are able to spend extra time practising use of GenAI, and those who are not. There are examples of edtech tools which students can access through a mobile app or browser, but even these cases may preclude some students. Schools are better equipped post Covid-19, but access remains varied.

Suggested area of focus:

- exploration of the digital and data infrastructure challenges that schools face in accessing GenAI tools, and how these may differ across different schools or different types of school.

Lack of reliable information on GenAI tools' effectiveness, accuracy, and safety

Confidence in the accuracy of AI tools

During the hackathons the teachers and leaders expressed reluctance to adopt tools that have not been tried and tested. Specifically in relation to the PoC feedback tool, teachers expressed concerns about whether the tool was accurate. A lack of accuracy would risk negating the potential benefits of the application of GenAI, for example with the PoC tool, if the feedback given is inaccurate the tool will be unable to deliver the potential benefits around improving marking consistency and saving teachers' time. Even if the tool is accurate, a perceived lack of accuracy on the part of users could result in reduced take up and equally negate the potential benefits.

Budget constraints and lack of information on value for money

Effective adoption of GenAI may require additional financial resourcing for schools, for example to enable schools to use commercial off the shelf GenAI tools which have an associated licensing fee. Depending on the tool, the costs can be prohibitive, and this challenge may be more acute for schools outside of large MATs if they are less able to negotiate due to their smaller size. In addition, teachers and schools' technology leads can find it difficult to access objective assessments of edtech tools, relating to their overall efficacy, impact on outcomes, and safety or data privacy questions.

Data protection and intellectual property challenges

Use cases that require LLMs to be trained on or use student owned or personally identifiable data may require agreement from parents or students, or may not be permissible depending on school policies. These challenges cover both data protection challenges such as ensuring privacy and gaining agreement to the use of personal data, as well as challenges related to intellectual property, including the training of AI models using student-owned data and the required agreement. Even for models deployed securely, where student data is kept within the school or MATs environment, ensuring sufficient protection of students' data can be challenging and complicated to navigate. Appropriate and sufficient guidance would need to be provided from DfE to assist schools and the edtech sector to navigate the UK General Data Protection Regulation (UK GDPR) and the use of AI, and would need to be updated on a regular basis given the rapid pace of evolution of technology.

Suggested area of focus:

- exploring options for producing guidance or kitemarking GenAI tools to enable schools to make informed judgements about their safety, data privacy, effectiveness and impact on outcomes and accuracy, including ensuring that schools are clearly informed where the evidence base for a tool is limited or in development.

Strengthening teachers and school leaders' confidence in using GenAI tools

If teachers and school leaders do not have access to high quality advice and training which helps them to be confident in using GenAI effectively and safely there is a risk that teachers disengage from incorporating GenAI tools into their practice and lose access to the potential benefits. Teachers may not be aware of the tools available or their benefits, or may perceive them as prohibitively complicated or time-consuming to learn about. Individual differences between teachers, such as their prior experience with other technology, may also affect their confidence and readiness to adopt GenAI tools.

Suggested areas of focus:

- development and testing of training for teachers and school leaders in the use of GenAI, including the potential benefits, how it can be applied in schools, and practical guidance on how to support their school or teachers on the journey to implementing GenAI; and
- building cohorts of technology leads who can support their colleagues and school leadership in implementing GenAI tools in their specific contexts.

Further work in progress

Since the beginning of the Use Cases for Generative AI in Education project, two additional projects have been conducted in response to the interim project findings.

AI Readiness in Schools – Data Systems Project

In light of project findings relating to the challenges schools face with integration of GenAI tools with existing systems, and how this is preventing schools and teachers from fully exploiting the potential of GenAI, the AI Readiness in Schools – Data Systems Project was launched in January 2024. This project is engaging with several schools and MATs to explore how they could be best readied for GenAI adoption, producing analysis and case studies that can guide further workstreams and guidance for schools and Trusts.

Mini hackathons for schools

Similarly, interim project findings identified teachers' and school leaders' views that a lack of guidance and training around how to use GenAI in their context prevents them from identifying potential applications of this technology and being confident that they can use it appropriately and safely. As a result, a programme of 'mini hackathons' for schools was established, delivering a small number of Proof of Concept (PoC) training sessions to trial approaches to:

- providing an introduction to AI and key concepts required to apply it in an education context,
- explaining strategies, ideas and tools for teachers to use AI,
- giving an overview of the risks and how they can be mitigated, and
- delivering a 'mini hackathon', where teachers are provided with a list of tools and invited to explore ways they might use this in their own practice.

Teachers are then invited to share their learnings with the group, and with their peers across their school. Details of this are in annex 3.

Recommendations for delivery of future hackathons

In addition to the specific learnings related to the application of GenAI in education via each use case, there were a number of lessons learnt relevant to the delivery of hackathons, which would be valuable in the design and planning of any similar future events:

- **User consultation ahead of the hackathons:** Engagement with a range of users ahead of the hackathons provided the team with a strong initial understanding of users' needs, their perspectives on the use of GenAI, and the key use cases that were most important to them. This had a clear impact on the success of the hackathons, as the use cases selected for experimentation were closely aligned to users' expectations and their priorities, helping to build buy-in from participants and to ensure that the findings of the events were relevant. The combination of surveys and focus groups was particularly effective, reaching a high number of respondents and enabling ranking of use cases, while also providing detailed insight into users' perspectives.
- **Allowing flexibility for the use cases selected:** While participants were presented with a list of use cases, they were also encouraged by the group facilitators to interpret them in a way which was most relevant to their own specific contexts. This had a number of benefits, including very high levels of user engagement with the process and the solutions developed, ensuring that users were able to give very specific feedback as to the potential effectiveness of a solution, and enabling users to provide real test examples drawn from their own experience.
- **Engineering work ahead of the hackathons:** Before the events, a team of engineers and data scientists worked with the education policy experts to prepare for the hackathons. The key elements of their preparation (aside from ensuring that all participating data scientists and engineers could access the platform used) were the preparation of the environment, including access to the GPT API, provision of example datasets relevant to each use case, and provision of Jupyter notebooks with all essential code to start prompting GPT. This ensured that the teams could

begin their prompt engineering work immediately, with no need to search for or process data, or set up a coding environment.

- **Ability to upload data:** In some of the hackathon sessions, the users were keen to make use of their own datasets, for example in the lesson planning use case, where the teachers were keen to use their school's curriculum or lesson planning guidance as reference data for GPT. In some cases, this was still possible, such as when the data was available for download from the internet. However, provision for the upload of data – of course subject to considerations like data privacy, and protection of Intellectual Property (IP) – would be beneficial in planning for future hackathons.
- **Composition of hackathons teams:** The teams in the hackathons included a balance of users (teachers, school administrators and school leaders), data scientists, engineers and policy experts. Each of the hackathon groups included one facilitator, one policy expert, 2-3 users and 3-5 data scientists and engineers. This composition was adjusted based on feedback from stakeholders with experience in running government hackathons and proved to be very productive.
- **Attempting to solve multiple use cases:** Each of the teams outlined above attempted to solve three to four use cases over the two days of the hackathons. This was challenging, as often the teams felt that they did not have sufficient time to fully explore all the different approaches that they had designed for each use case. It may have been more productive to have smaller groups, but each focusing on just one or two use cases throughout the two days of the hackathons, potentially giving a higher likelihood of participants being able to solve a use case.
- **Involvement of maintained schools and smaller MATs:** The user participants were drawn from large MATs such as Outwood Grange, Harris and Star, and brought huge value to the hackathon groups given their extensive knowledge of teaching practice and school processes, their judgement as to whether particular solutions would work in their school context, and their enthusiasm for finding solutions to difficult problems. However, had the participants been drawn from other types of school including maintained schools and smaller MATs (as the participants for user research were), we could have ensured that their feedback was representative of a broader range of school contexts.
- **Student participation:** Bringing in students as participants was challenging in terms of the logistics needed ahead of the hackathons, with a range of insurance, safeguarding, travel and accompaniment implications. However, the insight gained from the feedback provided by students was very valuable, and although only one use-case for students was explored, there were some novel findings that were also relevant to other applications of GenAI (such as the ways in which students use existing AI tools to practise concepts they have learnt in lesson). Some of the practical difficulties of student participation could be mitigated (while still benefiting from the clear value add of students' perspectives) by holding a separate hackathon that was focused on students held in a school.

- **Size of the hackathons event:** The hackathons were all held in the same location over two days, and although this had some benefits (e.g. reducing travel requirements), it meant that the event itself was large and complex, with over 60 participants and attendees, including members of the press. At certain points it was challenging for the attendees and organisers to maintain focus on the problem-solving required, and the high levels of event management involved took up a significant amount of the project's available resource in the lead up to the event. Although there were additional benefits to this size of event, such as the press coverage increasing awareness of the potential of GenAI in education, a series of smaller events may better deliver the key aims of the hackathons.
- **Opportunities for networking and building a cohort of users:** Throughout the hackathons, a number of opportunities for networking over coffee, breaks and lunch were built into the timeline. This helped maintain focus during the intense days, but also provided time for the participants and attendees to jointly reflect on the interim findings of the hackathons and to 'cross-pollinate' and spread effective practice between hackathons groups. It also enabled participants to share their experiences and ongoing initiatives related to the use of AI in schools that they were involved in, leading to the development of a group of 'super-users' who were key in the testing of the eventual PoC once developed.

Annex 1: Summary of pre-Hackathon consultation findings

Overview

The Generative AI in Education project is a collaboration between the National Institute of Teaching, Faculty AI, the Department for Education, and the AI in Schools Initiative. The project is exploring how Generative AI could be used to reduce school staff workload, increase practitioners' effectiveness, and improve learning in schools. Stage 1 of this project was a school led, co-design phase with stakeholders from the education sector. Teachers, school leaders, administrators, and students took part in a consultation via surveys and stakeholder group meetings.

We found that teachers were most keen on the idea of using tools to support them with lesson planning, marking and assessment, and making better use of class and pupil data. Likewise, school leaders were keen on implementing tools that could support them with data analysis, whilst also reporting that they would like support with writing and updating policies, documents, risk assessments, and parental communication. Administrators were keen to use tools to streamline and improve their understanding of data, develop timetables, and draft risk assessments. Finally, in discussion groups, students showed interest and support for several AI use cases, but expressed concerns about AI replacing teachers and other school staff.

Aims

Stage 1 was a school led, co-design phase with stakeholders from the education sector. Teachers, school leaders, administrators, and students were consulted via surveys and stakeholder group meetings. The first aim of this phase was to understand how stakeholders felt that AI could support school staff to save time and strengthen practice. The second aim was to gauge students' perceptions on how AI could support their learning. The findings from this phase were used to inform the shortlisting of use cases for the hackathons in stage 2.

This consultation phase of the project engaged stakeholders in a dialogue about the potential uses of AI in schools, seeking their input as key stakeholders and collaborators, rather than research participants. We do not report direct quotes from stakeholders. This initial stage will feed into more focused research into users' perspectives about AI during the user testing phase in stage 3 of the project.

Methods

Initial Generation of Use Cases

In collaboration with education experts, teacher trainers, and teachers, we generated an initial long list of use cases. Use cases were categorised depending on the target user (for example, teacher, leader, administrator or student), and whether it would help with saving time, strengthening practice, or both.

Consultation Sample

In total, 710 school-based practitioners (394 teachers, 133 school leaders, 97 administrators, 48 trainee teachers, and 31 teacher educators – see figure 7) took part in a 10-minute online consultation survey. An additional 20 stakeholders (teachers, leaders, and secondary school and sixth form students) took part in one of 4 online stakeholder group sessions to discuss the potential uses of AI in schools. Most stakeholders who participated in the group meetings reported that they had an interest in AI or had already been experimenting with AI solutions to school-based tasks.

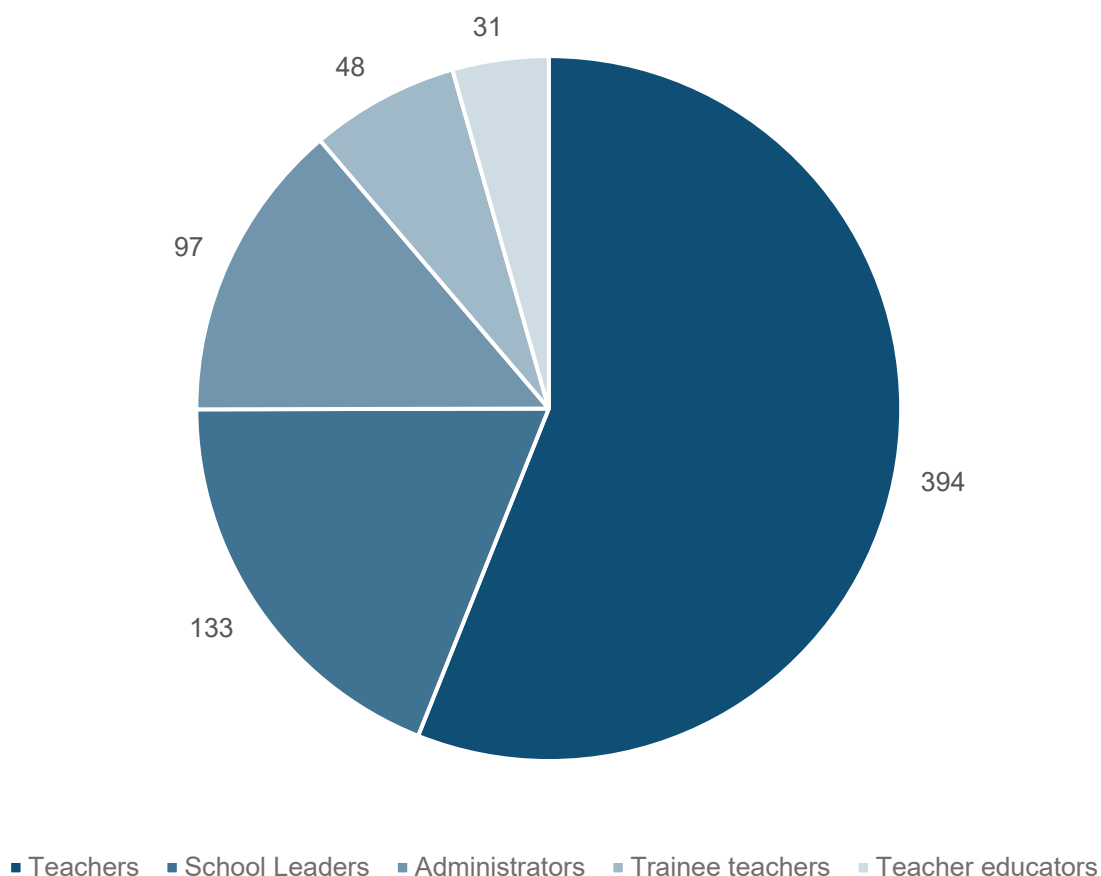


Figure 7: Breakdown of survey stakeholders by role

Survey

The consultation survey ran from 5 to 13 October 2023. Respondents were recruited via emails to the NIoT's network of 250+ schools and their teacher training and development programmes. After agreeing, all respondents identified the job title that best described their role. Teachers, trainee teachers, and teacher educators were automatically directed to respond to questions about AI for teachers. School leaders and administrators were directed to respond to separate sets of questions about AI for their respective roles.

Reducing workload

Respondents were asked: "If there was a magic tool that could cut some of the time that [teachers, school leaders or administrators] spend on the following tasks, what do you think people would choose to use it for?" before being asked to choose three use cases from a list. We chose the framing of a "magic tool" for the opening questions to gauge where school staff felt support was needed, without being biased by preconceived notions about AI. Later questions then explicitly referred to AI and asked whether there were any barriers and facilitators to its implementation in schools.

Professional Development

After being asked to reflect on things they would like to get better at in their practice, respondents were asked "If there was a magic tool that could strengthen teachers', school leaders' or administrators' practice in one of the following areas, what do you think most people would choose to use it for?" before again identifying three use cases from a list.

Teachers and leaders were shown different lists of use cases for time saving and professional development, although several use cases appeared in both lists (see Tables 1-2). Administrators were shown the same list of use cases for both time saving and professional development (see Table 3).

Stakeholder group meetings

Teacher, leader, and student stakeholders were invited to take part in consultation group discussions. Stakeholders were recruited through the NIoT's regional teams, who directly reached out to schools. The group discussions (1 with teachers, 1 with school leaders, and 2 with students) were carried out online, lasted approximately 45 minutes, and were facilitated by education experts from the NIoT. The aims were to engage in further dialogue about the potential impact and feasibility of the existing use cases and to generate any additional use cases for consideration. For the teacher and leader groups, we shared preliminary findings from the surveys and asked stakeholders for their views on the most popular use cases, before having a broader discussion about barriers and facilitators to using AI in schools. For the two student groups, we discussed several use

cases where AI could potentially help students in their learning. These discussions were not recorded.

Interpretation of results

For teachers, leaders, and administrators, we generated use case ranking lists based on how many times each use case was selected during the survey (see Tables 1-3). Stakeholder discussions were summarised immediately after the online sessions and are used to complement the main findings below.

Findings

Teachers

Table 11 shows that teachers who responded to the survey were most keen on the idea of using time saving tools for marking. Although teachers in the consultation group meetings agreed that using AI for marking could save time, they also expressed concerns about its reliability and shared stories about erroneous AI marking at their schools. Despite these concerns, school leaders in the group discussions were particularly keen to use AI to support teachers with their marking and assessment. They were interested in using AI to provide individual feedback to pupils, as well as providing collective, group-level feedback by feeding in all pupils' work and looking for common misconceptions or mistakes.

The second most popular use case for saving time was data entry and analysis of pupil progress or attainment. Using AI to help with various data entry and analysis requirements was also a popular choice in the consultation group discussion. One teacher in the group described data related tasks as the most disliked, uncomfortable, and laborious thing teachers have to do, thereby offering a rationale for using AI in this sphere. In general, using AI for data tasks was considered feasible and one teacher also felt that improving teachers' understanding of their class's data could aid all other aspects of their practice. The next 3 most popular use cases for time saving were lesson planning, differentiating, and writing reports.

Teachers who responded to the survey were also keen on using tools that could strengthen their practice in differentiating, designing lesson resources, framing or reframing concepts for students who had not understood, lesson planning, and making sense of data. This highlights a substantial overlap in the top survey choices for time saving and professional development categories (see Figure 8).

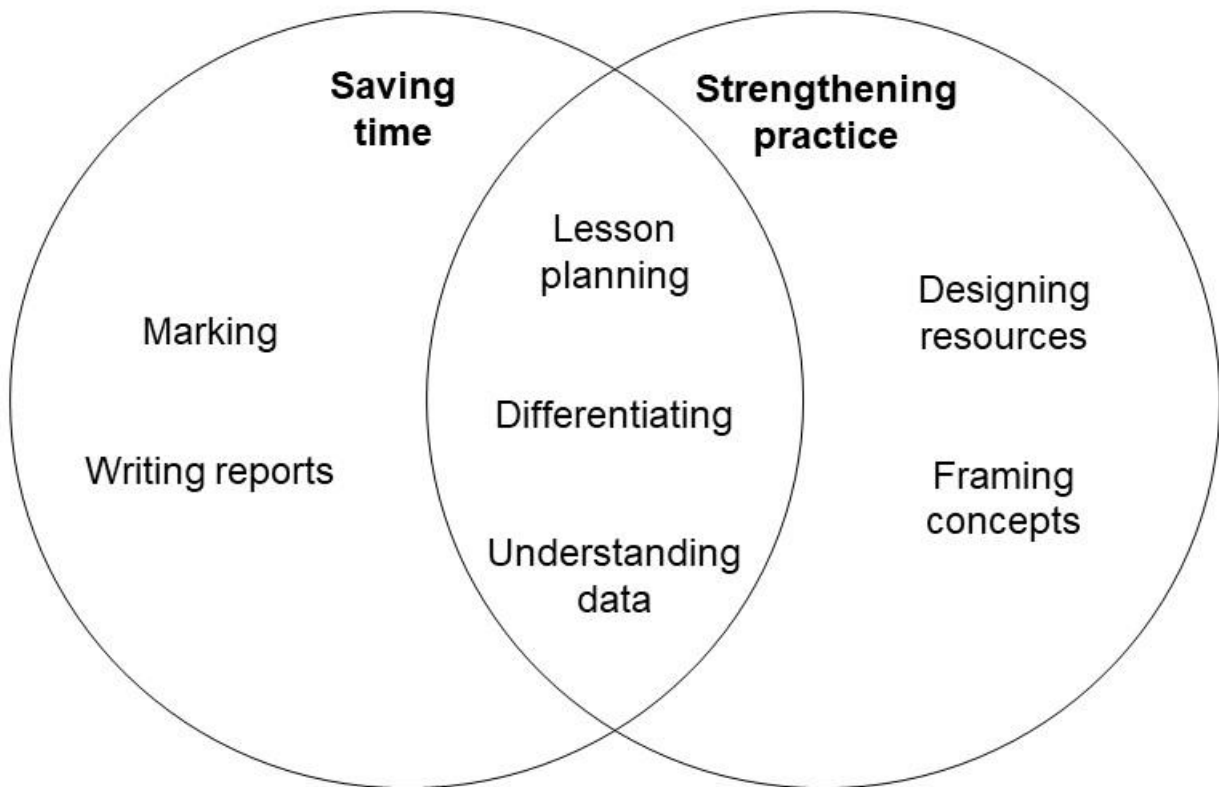


Figure 8: Venn diagram to show the overlap in teachers' use case choices for time saving and strengthening practice

Note: Teachers, trainee teachers, and teacher trainers were asked to choose 3 areas that they and their colleagues would most like tools to help them save time on and strengthen their practice. The diagram includes the five highest ranked use cases for each category from the consultation survey.

Teachers in the consultation groups felt that this overlap in the survey findings was not surprising. One suggested that it was almost inevitable that teachers would choose to strengthen their practice in areas that would ultimately save them time, reduce their workload and “give them their lives back”. This strong emphasis on using AI for reducing workload and improving wellbeing was a theme that ran throughout the discussion.

Teachers, trainee teachers, and teacher trainers were asked to choose 3 areas that they and their colleagues would most like tools to help them save time on and strengthen their practice. Use cases are grouped into 'time saving' and 'professional development' below, and are ranked in order of how often stakeholders selected them.

Time saving:

1. Marking
2. Data entry and analysis
3. Lesson planning
4. Differentiating
5. Writing reports
6. Developing schemes of work
7. Producing materials
8. Producing formative assessments
9. Communicating with pupils' families
10. Creating display materials
11. Updating parents and carers on progress

Professional development:

1. Differentiating learning materials and approaches to meet different students' needs
2. Designing lesson materials
3. Framing (or reframing) concepts
4. Lesson planning
5. Making sense of data
6. Generating formative assessments
7. Responding to misconceptions and errors
8. Measuring pupil participation in lessons
9. Designing homework tasks
10. Expanding subject knowledge
11. Generating effective behaviour management
12. Creating briefing materials for support staff
13. Creating seating plans
14. Preparing for challenging parent conversations
15. Using effective questioning

16. Providing career guidance to pupils

School leaders

Table 12 reveals that leaders who took part in the survey were most interested in using tools to help them save time on making sense of class or cohort data, creating policy documents, updating and disseminating risk assessments, communicating with families, and organising school trips. Of these, the leaders in the consultation group discussions were most enthusiastic about how AI could save them time in gaining insights from data. Leaders commented that such a tool may not only save time when producing insights but also reduce the time taken to train newer staff in data management systems and reporting. Although potentially beyond the capabilities of large language models, leaders were interested in the possibility that AI could help with timetabling, which they agreed consumed an enormous amount of time.

In terms of improving professional practice, school leaders were generally most keen on tools that could support them with making better use of their school data. The 5 most common choices were updating core documents automatically from data, making sense of class or cohort data, generating suggested actions from school data, attendance tracking and analysis, and intervention tracking and management. Leaders in the group discussions expressed support for each of these potential uses of AI, but felt that updating core documents from data streams would be less valuable in terms of professional development than the others.

Finally, leaders in the group discussion were in agreement that focusing on the development of AI tools for reducing class teachers' workload would be a greater priority than for reducing leaders' own workloads. On this point, leaders reported concerns around teacher retention and therefore welcomed the potential for AI solutions to support teachers in their roles.

School leaders were asked to choose three areas that they and their colleagues would most like tools to a) help them save time on, and b) strengthen their practice. Use cases are grouped into 'time saving' and 'professional development' below, and are ranked in order of how often stakeholders selected them.

Time saving:

1. Making sense of class or cohort data
2. Creating policy documents
3. Updating and disseminating risk assessments
4. Communicating with families
5. Organising school trips
6. Developing timetables
7. Curriculum mapping
8. Keeping up to date with the latest research
9. Recommending interventions
10. Summarising Educational, Health and Care Plans (EHCPs)
11. Staff recruitment
12. Budget planning

Professional development:

1. Updating core documents automatically from data
2. Making sense of class or cohort data
3. Generating suggested actions from school data
4. Attendance tracking and analysis
5. Intervention tracking and managements
6. Identifying opportunities for staff wellbeing
7. Tracking professional development
8. Developing departmental subject knowledge
9. Pupil Premium spend tracking
10. Updating parents/carers on pupils' progress
11. Developing timetables

12. Matching teachers to classes
13. Recommending interventions
14. Identifying pupil risk of becoming youth not in employment, education or training (NEET)
15. Summarising EHCPs

Administrators

Table 13 shows that school administrators who completed the survey were most interested in using tools to help them save time and improve their professional practice in whole pupil data analysis, attendance tracking, drafting risk assessments, and data storage and file structure. They also reported that they would like a tool that could save them time on timetabling, and a tool that could strengthen their practice in making sense of management information system data.

Administrators were asked to choose three areas that they and their colleagues would most like tools to a) help them save time on, and b) strengthen their practice. Use cases are grouped into 'time saving' and 'professional development' below, and are ranked in order of how often stakeholders selected them.

Time saving:

1. Whole pupil data analysis
2. Attendance tracking
3. Developing timetables
4. Drafting risk assessments
5. Data storage and file structure
6. Summarising meeting notes
7. Stock ordering
8. Drafting and formatting newsletters
9. Making sense of Management Information System (MIS) data
10. Staff recruitment
11. Behaviour tracking
12. Preparing social media content
13. Diary management
14. UK GDPR request collation
15. Transition administration
16. Library systems
17. Examination entry process
18. Writing code for school website

Professional development:

1. Whole pupil data analysis
2. Attendance tracking
3. Drafting risk assessments
4. Data storage and file structure
5. Making sense of MIS data
6. Developing timetables
7. Summarising meeting notes
8. Stock ordering
9. Behaviour tracking
10. Drafting and formatting newsletters
11. Staff recruitment
12. Transition administration
13. Preparing social media content
14. Writing code for school website
15. Diary management
16. UK GDPR request collation
17. Library systems
18. Examination entry process

Table 11: Summary of top five time saving use cases across teachers, leaders and administrators

Teachers	Leaders	Administrators
1. Marking	1. Making sense of class or cohort data	1. Whole pupil data analysis
2. Data entry and analysis	2. Creating policy documents	2. Attendance tracking
3. Lesson planning	3. Updating and disseminating risk assessments	3. Developing timetables
4. Differentiating	4. Communicating with families	4. Drafting risk assessments
5. Writing reports	5. Organising school trips	5. Data storage and file structure

Table 12: Summary of top five professional development use cases across teachers, leaders and administrators

Teachers	Leaders	Administrators
1. Differentiating learning materials and approaches to meet different students' needs	1. Updating core documents automatically from data	1. Whole pupil data analysis
2. Designing lesson materials	2. Making sense of class or cohort data	2. Attendance tracking
3. Framing (or reframing) concepts	3. Generating suggested actions from school data	3. Drafting risk assessments
4. Lesson planning	4. Attendance tracking and analysis	4. Data storage and file structure
5. Making sense of data	5. Intervention tracking and managements	5. Making sense of MIS data

Students

Students shared their views on AI in the stakeholder meetings, in which they discussed several use cases at length. Students were very positive about the idea of using AI to act as a historical figure to explain historical events. They felt it could be helpful and engaging for students who learn in different ways and make things more memorable. Students were also positive about having AI explain difficult concepts in a friendly and alternative way and felt that this would act as a complement to teachers who might not know what examples to give to students. The possibility of using AI to provide personalised feedback on work before submitting it to a teacher received a mixed response from students. One group struggled to imagine how it would work whereas a student in the second group was more positive, suggesting that this could enable students to improve their work before being assessed. This student was quick to note that he would not be comfortable with AI being used to mark work instead of the teacher, particularly for more subjective and creative subjects.

Students were less keen on the idea of having an AI tutor, deeming it impersonal and potentially error prone. While it was acknowledged that an AI tutor could be helpful when students feel stuck on something at home, it was also felt that it would not be as helpful as a real teacher. Students also had mixed opinions about using AI to gain careers advice. While some felt that it could be useful as a first step that could potentially suggest a greater range of options than a teacher, students generally felt that the tool would not

know enough about the individual student to give sound advice. Although the students stressed that they would not want AI to replace the role of student counsellor, they also acknowledged that some nervous or introverted students may find it easier to interact with AI than a counsellor, thereby suggesting that it could be more useful for some students than others.

In summary, the student consultation revealed that while some use cases for AI may be beneficial, overall, students felt that AI tools would be most helpful when used in conjunction with teacher support (such as using AI as a first stop for career advice before following up with a teacher or counsellor).

Enablers and barriers to AI use

Stakeholders were also asked to share their thoughts on what might facilitate or prevent the use of AI in schools.

Enablers

Stakeholders felt that training is required to increase teachers' understanding and ability with AI tools. Some teachers felt that AI may be more effective if schools were to create their own bespoke LLMs. Teachers also felt that these LLMs will be better if schools are able to feed all data about pupils (such as personal data, all assignments, grades) into them so that the tool is better informed. Teachers shared the perspective that they need guidance from Ofsted and exam boards on the appropriate use of AI and what good practice looks like, including how outstanding leadership approaches decisions around AI. Teachers noted that many students are already using AI, with engagement perceived to be high, so this should encourage teachers to follow suit in order to "speak the language" of those they teach and help students identify AI hallucinations, deep fakes, and other risks.

Barriers

Teachers noted that there is a level of fear and anxiety surrounding AI. One teacher suggested that some of this fear may have arisen due to a perception that using AI for teaching-related tasks might not be supported by important bodies such as Ofsted and exam boards. School leaders noted that high costs could prevent AI use, and that any tool would need to be demonstrably effective and impactful relative to the cost investment to the school. Leaders also identified ethical and data protection issues surrounding the use of AI, including ensuring users' understanding of these factors. Leaders also expressed a need to have trust in the AI organisations themselves, and their liability to manage and use the data responsibly and ethically. A further consideration was that users will need to review and sense-check all outputs carefully, including being alert to potential biases that might have been introduced by the tools.

Next steps

The findings from this school-led consultation were used to inform the selection of use cases for the Generative AI in Education Hackathon project. Subsequently, the results of the hackathons will be used in the development of a PoC model to be user tested in the final stage of the project. Findings from this consultation will also feed into qualitative research in the next stage, where we will return to explore the key findings about AI use in schools in more depth.

Annex 2: Full Hackathon findings report (November 2023)

Introduction

Working with the National Institute of Teaching (or NIoT) and the Department for Education (or DfE), we are exploring potential applications for GenAI in the education sector as part of a wider effort to ‘transform a teacher’s day to day work’ – reducing workload and improving educational outcomes by automating routine tasks. As part of this exploratory work, we held Generative AI in Education Hackathons, inviting participants from across the education sector to test a range of education-related use cases for LLMs.

During the hackathons, potential users and data scientists were brought together to test different approaches to tasks such as ‘Plan a French lesson for Year 9 students on the past tense’ for a selection of use cases, to further engineer their prompts to improve the GPT models’ outputs, and assess whether the final outputs produced would be usable in a school environment. Ultimately, the purpose of the hackathons was to generate a set of findings to inform the identification of three groups of use cases:

1. Use cases for GPT models which are currently ready for schools (potentially with some associated guidance/instructions provided).
2. Use cases which could be good candidates for a PoC tool with some additional tooling/functionality.
3. Use cases which are currently not possible for GPT models or too risky to take forward into schools.

The GPT use cases tested in the Generative AI in Education Hackathons were selected from a larger group originally generated through a combination of consultation with DfE, project stakeholders, and user research by NIoT. This larger group of use cases was assessed and prioritised based on iterative feedback from DfE as well as consideration of three key elements:

1. Technical feasibility – i.e. whether LLMs were expected to be able to perform the given task in line with users’ needs.
2. Expected impact – i.e. whether the use of LLMs for a given purpose reflected users’ priorities and their views on potential efficacy in terms of workload reduction and outcome improvement.
3. Risk assessment and feasibility – i.e. whether there were any safety considerations and/or cultural or organisational blockers that would make an LLM tool for a given purpose difficult to roll out.

The resulting selection of twelve use cases tested in the hackathons represents potential Generative AI applications for teachers, school leaders, administrators and students.

This report details the findings made during the event including descriptions of the tasks tested by users, commentary on the approaches used for prompt engineering and summaries of each use case's benefits and drawbacks.

Technical approach and model terminology

The two LLMs used during the hackathons were GPT-3.5 and GPT-4 (referred together as 'GPT models'). These two models power the popular ChatGPT chatbot. Using the two underlying LLMs directly enables using extended capabilities not offered in ChatGPT, such as deeper customisation of the output and the ability to use them in conjunction with other tools. GPT-3.5 was released before GPT-4, and while less powerful than GPT-4, its higher output speed and cheaper cost makes it relevant for this hackathon.

Potential barriers to Generative AI adoption

Alongside the user engagement with teachers, school leaders and students ahead of the hackathons, it was clear from the use case testing sessions in the hackathons that there are a range of cultural and technical constraints that could prevent schools and individuals from fully exploiting the potential of GenAI. This section of the report sets out the broader dependencies and constraints identified for the implementation of GenAI in educational settings.

Data protection

Appropriate and sufficient guidance should be provided from DfE to assist schools and the edtech sector to navigate UK GDPR and the use of AI.

Varying user needs and contexts

Differences in school and MAT curricula, lesson plan formats, behaviour processes and other resources will have to be understood and designed for if proposing a scaled Generative AI solution. There may be too much variance to provide one central solution without it becoming too complex both for a technical team to maintain and for users to navigate intuitively.

Confidence in using AI

AI upskilling should be included as part of teacher training programmes. This will enable teachers to confidently use AI to get the best out of the technology, and to equip them with the knowledge they need to flag any inappropriate, unsafe or inaccurate outputs.

Budget constraints

Effective adoption of GenAI may require additional financial resourcing for schools. However, the extent will depend on the service model that is rolled out. Cloud services and subscription services come at a cost that may present an additional burden which some schools may not be able to prioritise or afford.

Acclimatising school policies and cultures

Many schools may not yet have policies in place for safe and effective use of AI. Ethical concerns were expressed, particularly around data sharing, and should be worked through and addressed through policy, guidance and regulations or legislation as appropriate. Schools can then adapt these for their own context.

Trust in AI

Teachers and leaders express reluctance to adopt tools that have not been tried and tested, however we cannot learn about AI until there is some degree of adoption. Research on AI in education (for example, the quality of a lesson based on an AI generated lesson plan vs human generated lesson plan, students' subsequent engagement and attainment) has progressed more slowly than the speed at which AI has become more common in the past few years. While edtech technology providers will conduct research on their products prior to release, teachers should be supported with training and guidance to experiment with and evaluate tools in a low risk way and with a degree of caution. They should also be equipped to share findings and contribute to the collective understanding on a school and sector level.

Technical considerations

Integration with existing systems

During the hackathons, a number of GenAI use cases were explored that would require a (regular) feed of existing context-relevant data into an LLM based service, such as adapting existing lesson plans for students with SEND, or providing personalised feedback to students. This would require LLM integration via APIs with existing systems that capture this data at source, such as information management systems and school databases. Varying levels of digital maturity and technology interoperability across schools would need to be considered and potentially provisioned for, if such tools were to be adopted on a national scale.

Access to technology

Not all students have access to the same devices or internet availability at home, creating a divide between those who are able to spend extra time practising use of GenAI, and those who are not. Schools are significantly better equipped post Covid-19, however access remains varied.

Network and infrastructure

Internet bandwidth could come under strain with increased use of LLMs, this would need to be provisioned for. In addition, for some use cases, schools would ideally be able to host the 'enterprise' version of ChatGPT within their own IT infrastructure for safe and flexible use, unless a service was to be provided from the centre (such as DfE provides a user interface to interact with LLMs and therefore are also responsible for hosting).

Accessibility

Any user interface that is built as part of a GenAI service would need to adhere to accessibility guidelines to ensure all users can use the service.

Use cases for Teachers

Generating age-appropriate lesson plans

What we explored

Teachers report spending significant time on the generation of lesson plans, as good planning and structuring are key to ensuring that students make progress and that lessons are delivered effectively. During the hackathons, teachers explored using base GPT models to both plan age-appropriate lessons (in KS3 French or A level computing) and adapt an existing lesson plan for Year 7 history to cater for different ability levels.

Results

Labour intensive inputs

Reasonably successful strategies for pushing GPT models towards providing a usable output included providing detailed context as part of the prompt, breaking down requests made to the model into step-by-step instructions and including images such as screenshots from source documents as inputs. However, the group found that overall, base GPT models could only support lesson planning which met teachers' needs after a significant amount of input and prompt engineering. This was reflected in users' average scoring of GPT for this use case as low.

In addition, lesson plans need to be appropriate for the context – meeting any established formatting requirements, matching the school's scheme of work, aligning to the national curriculum and reflecting the context of students (including their previous learning). The need to input this data to generate high quality lesson plans creates a process which may be too labour intensive for base GPT models to be of significant use by teachers for this purpose.

Access to data required

To support teachers' lesson adaptation needs, an LLM would need to be able to refer to a school's existing data (including lesson plans, student data or potentially transcripts or recordings of lessons). Users commented that a tool geared towards lesson adaptation – as opposed to lesson planning – would potentially hold more benefit for teachers in well-resourced Multi Academy Trusts which already have a significant amount of high quality, requirement aligned resources. A significant remaining drawback with base GPT models is that a user may still have to input too much of this data for the tool to be time saving on a daily basis.

Conclusion

Candidate for PoC

A PoC with additional functionality allowing an LLM to refer to any necessary contextual data (this data would be synthetic during any development work) could surmount the above drawbacks. However, it should be noted that the outputs from using base GPT models to generate lesson plans were not particularly positively rated by teacher participants of the hackathons. While they scored the outputs of their testing at an average of 3.7 out of 5 for time saving, meeting national standards, improving outcomes and likelihood of use were scored between 2 and 2.3 out of 5. In addition, usability was scored at 1.7 out of 3 emphasising the need for improvement in functionality or additional tooling in the future.

Generating effective questioning ideas

What we explored

Effective questioning can be challenging for teachers to master. The users in the hackathons were clear that there is a significant difference between how well new teachers perform at this compared to their experienced colleagues and this was supported by our pre-hackathons user engagement. A tool that automatically generates a range of questions, suitable for students of all abilities, could be useful for all teachers, especially new teachers or those teaching outside of their specialties.

During the hackathons, teachers explored using base GPT models as:

1. a diagnosis agent to understand misconceptions among students,
2. a tool to develop effective questions for teachers to evaluate understanding, and
3. a chatbot to gauge student understanding.

Successful approaches included employing an 'inner monologue' technique (requiring base GPT models to explain their thought process with returned outputs) and prompting base GPT models with a combination of a topic specific lesson and overarching learning objectives to ensure questions generated were relevant.

Results

Further training and a user interface required

Base GPT models performed notably well at generating ‘application’ style questions, where students are asked to apply their knowledge. The questions integrated real world examples upon request and were aligned to examples of exam requirements. As with the similar use cases on lesson planning and adaptation, generic GPT-4 was able to generate questions well, but a significant amount of input was still required from the teacher which precludes this use case from being deliverable without a supporting user interface.

For example, some of the best performance in terms of output from GPT-4 for this use case came from providing the model with an example lesson outline and transcript to generate questions suitable for students with a range of abilities. One user highlighted this in our survey stating: “If this is trained and given enough data this could be an extremely useful tool for students to use, direct questioning and teaching. Based upon current understanding, this would require a lot of teacher time to generate individually.”

Conclusion

Candidate for PoC

On average, this use case was rated joint highest by users at 4 out of 5 for final output generated during the hackathons and met established national standards such as the national curriculum. Users also rated the output for this use case highly in terms of time saving at 4.7 out of 5, but moderately in terms of usability at 2.3 out of 3, improving outcomes at 3.7 out of 5, and likelihood of use at 3.3 out 5.

Should this use be taken forward to PoC stage, extra tooling could include:

- a supporting user interface for teachers to provide their own lesson transcripts or integration with existing systems to enable ingestion of the required contextual data; and
- the development of a single, validated and customisable prompt for generating questions per subject or lesson objective or the standardisation of the single quotes technique explored during the hackathons.

Generating lesson materials

What we explored

Teachers can spend significant amounts of time generating lesson materials and activities as part of their lesson planning process, and activities need to be varied by topic and age group to meet the needs of classes and individual students.

During the hackathons, teachers explored ways for GPT models to assist in the generation of lesson materials. Specifically, teachers tested whether the models could generate homework for students based on a PowerPoint and transcript, adapt text extracts to a lower reading age and develop progressive worksheets to boost exam confidence in EAL students.

Results

Success on adaptations and differentiation

The group found that the use of the model for these tasks was mostly successful. The team was able to create a brief, functional demonstration of a basic GPT-based tool for generating differentiated homework using lesson transcripts, which received a positive reception. In addition, GPT-3.5 was successfully able to provide guidance on adapting text extracts to a lower reading age and could implement some of the suggested adaptations. The team applied this effectively by including the desired adaptations (user or GPT generated) within prompts for the model. The final task tested (developing progressive worksheets) was less successful and base GPT models performed poorly at creating questions of different difficulties which matched real-world examinations.

The main issue the group faced while testing this use case during the hackathons was the lack of ability to ingest and lift text from PowerPoint for model training. A workaround was devised in which PowerPoint could be converted to images (recognisable by GPT-4) and ingested into the model. Since the hackathons however, this issue has been negated by the fact that OpenAI has updated its models with additional file format ingestion capabilities, including for PowerPoint.

More work required on progressive worksheets

Other challenges remain in terms of improving output quality. For the task of developing progressive worksheets, the model will likely need to be fed additional, existing exam questions to improve the creation of differentiated questions and further fine-tuning (or extra tooling in the form of a RAG-based approach where the model retrieves additional knowledge from a data source), could be beneficial.

Conclusion

Overall, user scoring was positive with participants giving the model average scores of 3 out of 3 for usability, 4 out of 5 for following national standards, 4.5 out of 5 for time saved, and 5 out of 5 for improved outcomes. They also rated their likelihood to use such a model for generating lesson materials as 4.5 out of 5 which is encouraging.

Considerations for future development include the need to provide guidance for prompting and to deliver an efficient and effective user experience for users.

Writing end of year reports

What we explored

Writing end of year reports is time consuming, with teachers reporting in our hackathons participant survey that in some cases, schools have reduced the amount of comments required in reports in an effort to reduce workload. The time and effort required to ensure that end of year reports contribute to improving educational outcomes is significant, creating potential for an AI tool which supports this process to be beneficial.

During the hackathons, teachers tested the ability of GPT-3.5 to create end of year reports based on two types of inputs. In one scenario, the model was asked to create end of year reports solely based on assessment data (such as grades and marks) and in the other scenario, the model was asked to work based on feedback provided by a teacher.

Results

Grade data is insufficient

While the group found that GPT was able to summarise text data about the student's behaviour reasonably well, the end of year report generated based solely on assessment data in particular was not considered to be of usable quality. The model was not able to incorporate enough of the teachers' professional judgement that users felt was critical to an effective end of year report.

In some instances, the model exhibited hallucination, and, in others, outputs were considered to be too generic. Other remaining challenges highlighted by users included the need to not only develop some level of integration (such as via an API) with internal school databases or information management systems but also to ensure that reports are created with the appropriate tone, formatting and notices to parents informing them of the use of AI.

Extensive training required

While the group agreed that using GPT models in this fashion could save time, much more additional work would need to be done to reach a usable output including further training on examples and improvement of summarisation, potentially via the inclusion of a 200-word limit. Further exploration would also require more quantitative training data, and a robust way to measure the quality of the outputs.

Conclusion

This use case was scored negatively by participants in the hackathons with scores of between 1 and 2 out of 5 on all criteria and we would classify it as one which does not represent a good candidate for development of a PoC.

Assigning marks to work submitted by students

What we explored

Giving accurate feedback to students is key for their learning and progression but can also be time consuming. Teachers sometimes need to mark large volumes of work on the same topic, resulting in fatigue or affecting time assigned to other tasks. As such, even assuming that AI is not used as the singular marker for examinations, there could be significant benefits to developing a model which is able to mark accurately, including saving teachers' time, improving efficiency and improving the quality of marking through a reduction in the overall burden on teachers. An AI marker could also have a significant impact on learning outcomes for students through more qualitative time with teachers, and potentially more personalised feedback from AI. In the hackathons, teachers tested a range of prompting approaches to improve GPT models' effectiveness and accuracy when presented with Year 4 essays and asked to mark them, assign grades and provide feedback based on the national curriculum.

The group applied a variety of approaches such as using more detailed prompts to specify what an output must contain and applying a modular approach to prompting – requesting that the LLM generate an exemplar essay based on a real-world mark scheme and use this to benchmark marking for a pupil essay. In addition, the group was able to improve the model's output by reducing its 'temperature', described by OpenAI as "a parameter that controls the "creativity" or randomness of the text generated", as well as experimenting with the more recently released GPT-4 model.

Results

Success on ranking, feedback and benchmarking

Users found that in most cases GPT-4 was able to rank students' work, identifying the highest and lowest scoring students. The model was also able to give feedback on students' work using the correct tone and in a way that would be meaningful to the student. The team were able to push GPT-4 to produce brief assessments of whether the quality of a given student's work was below, at or above expectations for a Year 4 pupil.

More work needed on accuracy

The main challenge flagged by users was that the model as used in the hackathons was not able to consistently mark students' work with a sufficient degree of accuracy. While the model may have been able to provide appropriate feedback and a matching assessment for the level at which it interpreted a piece of work to be – often, this level was incorrect when compared to real-world, established standards.

Considering how this challenge could be resolved in future, the group working on this use case were able to identify 3 key resources which could be provided to the model to improve performance:

1. **Machine-readable mark schemes:** the group were able to identify that the model struggled to interpret and apply the mark scheme provided to it during the hackathons. A solution to this could be the provision of mark schemes that can be more easily 'read' by the LLM, which could be reasonably expected to improve marking accuracy.
2. **Few-shot prompting:** providing the model a few examples of the desired input (correct and incorrect answers to a question) and output (human-awarded mark) can be used as conditioning for other unseen question/answer pairs by the model, making its output more in line with human marking.
3. **Fine-tuning based on marks from experienced markers:** fully marked scripts that were known to be accurate (i.e. marked by an experienced human marker) could be used as a 'ground-truth' for GPT-3.5 and GPT-4 models to train with so they awards marks more closely aligned with those given by human markers.

Conclusion

It was evident from the hackathons user feedback (as well as our previous user engagement) that the development of an accurate AI marker could result in time-savings for teachers and schools. Overall, teachers scored this use case low on and meeting national standards (2.5 out of 5) and moderately on usability (1.8 out of 3), time-saving (3.3 out of 5) and outcome improvement (3 out of 5). Likelihood of use however was scored slightly higher at 3.8 out of 5 suggesting that, should the associated challenges be resolved in future – a tool such as this may be useful but the bar for accuracy is high.

Generating personalised formative feedback and addressing student misconceptions

What we explored

The provision of personalised feedback is critical for students to understand where they need to improve; however, providing feedback for entire classes can be immensely time-consuming for teachers. Teachers participating in the hackathons investigated whether GPT models could be useful for generating specific feedback for individual Year 4 students on pieces of English work. The aim was to investigate a scenario in which the models were asked to provide positive feedback, suggesting ways for students to improve their work.

The team began by using different prompt structures with GPT-3.5 to elicit specific feedback on each essay (asking for both teacher-facing and student-facing versions) and extracting this feedback into more structured formats (specifically, JSON files).

The team then divided into smaller teams testing the viability of using these JSON files to request different kinds of outputs. For example, one team presented the model with a JSON file and aimed to produce an improved version of the student's essay as an exemplar version, another team attempted to use JSON files to present a teacher-friendly

grid of feedback and a third team focused on matching JSON files to specific examples in student work – offering the opportunity to play these back to students and demonstrate where improvements could be made.

Results

Significant customisation required for accuracy

While much of the feedback generated looked convincing at first glance, it was almost entirely inaccurate. Therefore, even though some of the approaches created outputs that were helpfully structured for students and/or teachers, the investigation of this use case resulted in the conclusion that generic GPT models would not yet be a viable tool for generating personalised feedback.

Indeed, similarly to writing end of year reports (which was another use case requiring extreme tailoring to students), this use case was scored negatively by users for potential time saved, likelihood of use and adherence to national standards (2-2.3 out of 5 on all four criteria). Usability however was scored more favourably at 1.8 out of 3.

Conclusion

Ultimately, the group felt that the inaccuracies observed could be mitigated in future with better quality mark schemes for the model to interpret, more expert input and more training data for the model to see a range of performance and greater context. Despite these potential areas for improvement though, the group expressed doubts in survey responses that it would be possible to improve accuracy to the level desired by teachers. LLMs therefore may be better used to aggregate or structure human feedback swiftly in order to aid lesson planning, rather than to produce feedback unaided.

GenAI as a teaching aid

What we explored

Generative AI tools have the potential to enhance the student experience and improve learning outcomes. During the hackathons, teachers tested whether GPT models could assist teachers in the classroom by testing student understanding and providing key stage-appropriate information on specific subjects. As both teachers and students could be considered the users of GPT models in this way, teachers role-played as students during testing, investigating how the model would respond to a student trying to develop their understanding of key Geography concepts.

Results

Success on generating a dialogue with students

While generic GPT models were adequate at providing appropriate responses in this context, the group found that a more successful approach was to provide the models with

an extended prompt structure before any student engagement. The model was instructed not to provide answers directly but instead to test students' knowledge and build on elements of their understanding iteratively. These instructions were able to push the model to keep responses brief and maintain an active dialogue. Overall, the group felt that the outputs of the task were student-appropriate, could be used as part of wider lesson aims and overall, close to 'ready for classroom' status.

Additional considerations raised by participants included the need to examine how any activities involving GPT models as a teaching aid could be effectively presented/set up for a class, how to normalise its use as part of classwork and avoid it becoming a 'toy' to play with, and how to ensure the accuracy of any responses delivered to users.

Conclusion

Overall, the group scored the outputs of this use case highly on time saving potential, improvement of educational outcomes for students and likelihood of use – giving an average of 4.3 out of 5 for each of these criteria. For adherence to national standards, the outputs scored 3.5 on average however, it was noted that usability was scored poorly, with an average of 2.3 (potentially reflecting the concerns and considerations previously mentioned). As a next step, the group recommended potentially targeting model training to focus on a specific corpus of information or textbook to further improve outputs to users and usability; however, overall the use case was perceived positively.

Supporting teachers to adapt content to meet the needs of students with Special Educational Needs and Disabilities (SEND)

What we explored

A key aim of our exploration of potential uses for Generative AI in education is the improvement of educational outcomes for pupils. Tailoring strategies and resources to students with special educational needs and disabilities (SEND) using tools such as GPT may represent a way to not only bolster their learning, but also improve inclusivity within the classroom, widening access to learning materials and content. As part of the hackathons, teachers explored a range of ways in which Generative AI could support themselves and special educational needs co-ordinators (SENCOs) by suggesting potential teaching approaches or interventions to better meet these students' needs.

Results

Further engineering required for guidance and interventions

The use of GPT as an advisor, suggesting potential teaching approaches, interventions and adaptations to better meet the needs of students with SEND did not test so well within the hackathons. In response to prompting, GPT was only able to provide very generic advice on interventions – even when prompts were further engineered to request

more specificity in responses and specific guidance documents were provided as context.

Hackathons participants agreed that further development to improve the model's performance could include applying a more in-depth, RAG-based approach to support information retrieval from a document library of SEND guidance. There are however questions about the ultimate efficacy of this approach considering the provision of specific guidance documents in the hackathons appeared to have little effect on the model's ability to differentiate its recommendations.

Conclusion

Given recent advances with GPT and the potential for impact on an important student group, further exploration could be carried out on this use case. Responses to our survey indicated that using GPT models to support students with SEND may be useful however the outputs of the hackathons were received with middling sentiment. Usability was scored on average at 2 out of 3, while time-saving (3.3 out of 5), meeting national standards (3 out of 5), outcome improvement (2.8 out of 5) and likelihood of use (3.3 out of 5) were all rated at around average.

Use cases for School Leaders and Administrators

Generating drafts of statutory school policies

What we explored

School leaders are often tasked with the generation of new policies or, more frequently, policy updates. This is currently a time-consuming process with the necessary preparation time impacting staff leave periods and involving multiple meetings across the Senior Leadership Team (SLT).

School leaders and administrators therefore examined whether GPT models could act as a support tool for this process, generating drafts of statutory school policies for further review before implementation.

Results

LLM can be used to compare policies and inform key elements

An initial finding from the group discussion was that expectations for the contents of school policy documents varied across schools and the group began by investigating GPT-3.5's ability to compare and summarise policies in use in different schools. The LLM was effective, helping to provide a wide view of the necessary items to include. The group concluded that the output would be useful as preparation for wider SLT sessions around policy; however, moving forward additional capability (e.g. using models with

larger input capabilities) could improve the ability to compare multiple, large policy documents even more efficiently.

Policy only as good as training materials

Participants also spent time testing GPT-3.5's ability to adjust generated drafts of behavioural and data literacy policies to better align with specific school values to better meet the needs of different audiences. Here, users found that the LLM was helpful for generating tailored policies but noted that the quality of outputs was heavily dependent upon the quality of exemplar materials used for training.

Longer prompts with detailed context required

In both the cross-school policy analysis task and the policy adaptation tasks, providing detailed context as part of the prompt given to the model was effective. Providing longer prompts, describing must-haves and must-nots, pushed the model to provide more tailored outputs.

Finally, participants tested GPT-3.5's ability to support professional development by creating different policy-based scenarios for testing with staff. In this case, users felt that GPT-3.5 needed a lot of further development to refine the output appropriately, ensure that scenarios were consistently updated for policies updated frequently (such as examination-related policies) and to integrate the model's use as part of the current policy generation workflow. Using GPT-4 did not yield any significant improvements.

Conclusion

School leaders and administrators scored this use case positively across almost all criteria – rating the generation of draft statutory policies between 4 and 5 out of 5 for time saving, meeting established standards, improving outcomes and likelihood of use. Usability was rated as an average 2 out of 3, indicating users felt that outputs were somewhat usable but would require further adjustment by a human before being trusted.

Pupil or class data analysis/synthesis

What we explored

The important, yet time-consuming nature of data entry and analysis in education meant that these tasks were among the most popular potential uses of AI for both reducing workload and strengthening practice in the pre-hackathons consultation stage.

During the hackathons, the group of school leaders and administrators explored whether GPT models could be used to not only ingest a synthetic pupil data set (which included dummy IDs, target grades, reading ages etc.) but also analyse the specific assessment data to identify personal and group-level capability gaps. Specifically, the group tested whether GPT models were able to rank students overall but also perform granular, question-level analysis of mock exam scores (which could enable class teachers to

identify patterns and variation in understanding and develop next steps much faster than is currently possible).

Results

Code generation a potentially better goal for data analysis

The results were mixed. Base GPT models are not currently conducive to direct analysis of student data, due to accuracy and ethical concerns. However, participants pivoted to considering whether GPT would be more helpful at generating code for the R programming language, which could in turn be used to analyse student data offline, or using other systems.

GPT-generated R code was somewhat successful at ranking the synthetic student data according to some predefined criteria, however, was not usable in terms of the granular, question-level analysis which would support identification of group-level capability gaps. For this reason, users felt that this approach would not add much value to existing systems. Concerns were that a process involving the combination of GPT models and R would inherently require teachers to then train in the use of R – adding extra complications. In an attempt to address this point, code for Microsoft Excel was also generated using GPT. However, this did not produce accurate or reliable results.

Conclusion

Overall, despite the concerns and difficulties encountered during the exploration of this use case during the hackathons, school leaders and administrators scored this use case moderately on average when surveyed. Likelihood of use and whether outputs met established national standards were both scored at or above 4 out of 5 while average scores above 3 out of 5 were returned for both time-saving and outcome improvement. Usability was also scored moderately at 2 out of 3. This indicates that there is still an appetite among this user base to test this use case further with alternative approaches to those considered in the hackathons.

Generating parent-carer communications

What we explored

During the hackathons, a group of school leaders and administrators also explored the potential for GPT models to support the generation of communications to parents and carers. Specifically, participants tested prompts designed for creating a school newsletter for distribution to parents and carers based on limited information.

Results

Highly usable outputs with minimal prompting

Overall, this use case was one which LLMs was mostly able to address successfully with users finding that GPT-3.5 was able to generate relevant and usable outputs based on the small number of user inputs provided. First drafts of newsletters were also generated in a consistent tone which was aligned to expectations for content delivered to parents.

From a technical perspective, particularly effective approaches included the inclusion of mock newsletter articles in the prompt given to the model, developing standardised data fields that users would need to complete (e.g. date, location or year group) and providing explicit templates to standardise article formatting.

In our associated survey of participants, the consensus from the group was that application of GPT-3.5 for generating parent-carer communication could save a considerable amount of time for users (rated at 4.5 out of 5, on average) and this use case also scored the joint highest rating for the likelihood of using Generative AI (5 out of 5, on average). The usability of hackathon outputs scored highly at 2.5 out of 3 on average. Scoring less highly were the outputs' ability to meet national standards (rated at 3.5 out of 5, on average) and improve outcomes (rated at 3 out of 5, on average). However, it should be noted that this use case is one which neither depends on the National curriculum/exam board specifications etc. nor one for which outputs are student-facing.

Conclusion

There still remain a number of challenges before AI can be used in this way. For example, users wanted the model to be able to deliver outputs that were pre-formatted, for any tool to be integrated within their existing workflows (e.g. enabling interaction between GPT and Word) and to be able to bring in relevant images automatically. Participants also provided feedback that further collaboration with headteachers and school administrators would be key to refine user requirements but, should extra tooling and functionality become available, there could be potential to expand this application of Generative AI to cover all parent and carer communications (e.g. generating letters, reminders or permission slips).

Use case for Students

Language learning assistant

What we explored

Student-facing use cases for Generative AI in education are on the rise, with commercial solutions currently being used in schools for a variety of purposes. During the hackathons we tested the viability of using GPT models as an assistant for students learning modern foreign languages such as French, Spanish and German. The group of students in

attendance explored how GPT models could help them understand grammar concepts, correct their work and practice conversations in their target language.

Results

A complex product with several requirements

An immediate finding was that a useful language learning assistant would need to be able to accomplish all of these tasks concurrently as what appeared to be a simple request could spiral into various areas. For example, in the case where a student asked GPT-3.5 to have a practice conversation in their target language using a certain vocabulary list (Spanish GCSE vocabulary), the conversation evolved into asking the LLM which mistakes they had made, before requesting exercises to strengthen their understanding of the grammar points they struggled with.

A variety of approaches were explored, ranging from simple prompt engineering to RAG. Users found that, in some instances, no prompt engineering or additional data was necessary as outputs were high-quality and met user needs. For other tasks, like asking the LLM to act as a language tutor and correct a sentence, more advanced prompt engineering was necessary). Outputs were improved by increasing the specificity and length of single prompts and providing multiple steps for the LLM to follow (i.e. Tree of Thought prompting, a technique where an LLM explores multiple intermediate thoughts simultaneously to then choose the best one).

Hallucinations lead to low trust in model outputs

Occasionally, the LLM made mistakes when correcting student answers to a list of multiple-choice questions it had created or hallucinated and corrected non-existent mistakes in student answers to short answer questions. It was noted that the use of the more advanced GPT-4 model was able to correct some of these errors, improving performance at correction to be substantially better than GPT-3.5.

Overall, users reported that they felt base GPT models could be usable for practising conversations, creating practice exercises and finding mistakes in their work; however, trust in the model was still a major concern. Given the LLM had made mistakes during testing, students doubted the reliability of the output and as such, indicated that they would be more likely to use ChatGPT as a supplemental tool to the other methods they use for language learning.

Conclusion

If an LLM-based language learning assistant tool were to be built, RAG could be applied to improve output quality. A model could make reference to GCSE and A-Level specifications and it may be possible to provide better differentiated outputs as a result. Further user testing is necessary to meet different learning styles (e.g. for students who would prefer to use visual inputs or to rehearse new vocabulary aloud). Additionally, GPT

may struggle with other languages also offered at GCSE level such as Mandarin, Hebrew and Urdu. When testing similar prompts in Mandarin for example, the model ignored certain mistakes and provided explanations of lesser quality. This suggests that, depending on the target language, additional fine tuning of the model may be necessary.

Annex 3: AI mini-hackathons project summary

Background to the project

In November 2023, the National Institute of Teaching co-hosted a series of generative AI hackathons with teachers and leaders. This collaborative venture, in partnership with the AI organization Faculty and the Department for Education, aimed to engineer creative solutions to pressing challenges. As we're all aware, our educators are often stretched thin, with workload pressures contributing significantly to the sector's retention and recruitment crisis. The hackathons aimed to address this critical issue by focusing on the scarcity of time available for educators to fulfil their extensive responsibilities.

During these sessions, participants used generative AI (via ChatGPT) to try and streamline common tasks, discovering both time-saving applications and limitations of the technology.

In March 2024, 3 additional hackathons were organised across three trusts within the NIoT's network—Oasis Community Learning (at Oasis Hobmoor Primary Academy), Windsor Academy Trust (at Windsor High School), and Learn AT (at Meadowdale Primary School).

The aims of the hackathon

The aim of each hackathon was to:

- i. To gain more understanding of the value of a 'hackathon' approach as a tool for professional development.
- ii. To assess how, when and why generative AI may be able to support teachers and leaders with some of the challenges they face.
- iii. To build awareness of the benefits and limitations of AI within the education sector.

These sessions welcomed a diverse group of participants, from AI novices to seasoned user, across a range of different school roles from classroom teachers, SENCOs, Heads of Department, Senior Leadership, Trust Leadership and Professional Services roles. In total 17 individuals from across the 3 trusts took part.

All 3 hackathons resulted in positive feedback for their impact on educational practice. Participants were often visibly moved by the possibilities of generative AI to reclaim time and enhance support for students. The hackathon process sparked rich and detailed conversations that bridged new insights with deep-rooted expertise in leadership and teaching. For a detailed look at the outcomes, please refer to the evaluation data in the appendix.

However, amidst the enthusiasm, there was at times a noticeable absence of critical examination of potential unintended consequences, as the wave of positive sentiment seemed to overshadow caution.

What did we do?

- Each hackathon was led by a facilitator from the NIoT and a data scientist from Faculty and lasted two hours, scheduled at a time to suit the school and held on their site. A researcher from the NIoT was also present to record observations of what was happening during the hackathons.
- Before the hackathon, all participants completed a short survey to identify their current usage of generative AI and their attitudes towards it. The survey also asked participants to identify 3 priority areas for development and 3 frustrations within their current role.
- The data was analysed and shared with participants during the session to help identify the lines of enquiry for the hackathon.
- The hackathon began with a short introduction to generative AI before moving on to identifying a problem/challenge to work through using ChatGPT.
- For two of the three hackathons, the data scientist would act as scribe, taking input from the group to write a prompt into ChatGPT. The output of each prompt was then discussed and analysed by the group.
- In the third hackathon, which had a larger group of participants, the data scientist and NIoT facilitator supported participants as they worked on finding solutions to problems in pairs or threes.
- All hackathons ended with a reflection on what had been gained; where generative AI could be of use and what were its limitations. A short evaluation was completed by all participants, with a follow up survey planned for July.

What did we learn?

The hackathons emerged as a profound learning experience, revealing the potential of generative AI not only as a practical tool but also as a catalyst for educational innovation.

- What worked well?
 - Gathering staff at different levels of the hierarchy (and in some cases from different schools) in one room and sharing experiences.
 - Participants seemed to value the opportunity to learn about the background of AI and prompting techniques from an expert data-scientist. This supported their understanding and made them keen to learn more in their own time. This is evident in one evaluation survey response which stated that 'It made me more aware of how I already use AI in my work. It made me aware of the benefits of AI and how it could reduce workload in schools for leaders and teachers.'
 - Participants enjoyed being involved in the process rather than simply being told what to do with AI (this may be particularly true for those who are not hugely tech-savvy).
 - Participants demonstrated positive attitudes towards AI and were keen to develop their ability/confidence in using it in their work.

- Less confident and novice AI-users valued the opportunity to see ChatGPT in action. One participant reported that the hackathon format was '...particularly with engaging less confident colleagues'.
 - The hackathon format was seen as useful for both primary and secondary colleagues, with respondents to the evaluation survey reporting 'working collaboratively', 'prompt [generation]', and '[using] AI' to be particularly useful.
 - In the evaluation survey, participants strongly agreed that the Generative AI hackathon met the aim of assessing how, when and why Generative AI may be able to support teachers and leaders with some of the challenges they face, and found the solutions developed using ChatGPT to be particularly useful.
- What worked less well?
 - Situation where there were too many participants and the data-scientist's time and support was spread more thinly.
 - Making sure that all participants' priorities were explored in the session (e.g., in one session one participant's priority was to explore how AI could support with KS2 literacy, but the main focus for the prompts ended up being about a Year 3 Science lesson plan).
 - 22% of respondents to the evaluation survey raised the need to a longer hackathon session to allow deeper exploration of using GenAI and a comparison of different GenAI models.
- What are the potential benefits and uses for sessions like these (for PD)?
 - Participants are able to share ideas in a group format and learn from a more knowledgeable other. One participant stated that they 'loved this approach and being with colleagues discussing how to use [GenAI] and what others have been doing with it'.
 - Using AI is an effective way generating ideas/starting points to develop in a range of areas (e.g., lesson planning, resource creation etc.). Using AI in this way could be useful in curriculum group meetings etc.
 - The topic of AI lends itself perfectly to this form of PD. Evaluation feedback shows that participants liked the collaborative approach of the hackathon format, and it could be used to support staff collaboration and innovation in other areas for development - 'a great CPL opportunity. Great to have discussion with colleagues'.
 - Hackathon format would be beneficial for Professional Services staff and Support staff (eg. TAs) - provides time to experiment with GenAI in a supportive environment, building confidence/skill in prompting
 - Beneficial for staff working in subject/phase groups to find a creative solution to a common problem. One participant reported that 'As an organisation and an academy, we use solution focused approaches already so it is a very useful idea'.
 - In the evaluation survey 89% of respondents reported they are highly likely to use or implement the generative AI solutions developed during the hackathon in their work.

- What are other potential benefits and uses for sessions like these?
 - The hackathon format can be used to develop collaboration across different subjects/phases/roles in school. One such response in the evaluation survey reported finding ‘the collaboration with others across different subjects [useful], and how they develop their prompts and the discussion around its implementation’.
 - The hackathon also provided dedicated time for colleagues to meet and discuss ideas: ‘great way to dedicate time to sit down, discuss and explore ideas collaboratively, which we often don’t get chance to do’.

- What are key considerations for taking these forwards?
 - The sustainability of the hackathon format, using data scientists is questionable. However, it could be that the hackathon is the starting point for ‘train the trainer’ training whereby a number of ‘super-users’ are trained how to run a GenAI hackathon, before leading similar sessions/disseminating their new knowledge to others in their schools/trusts/networks. There could be the opportunity to engage with teaching and learning leads within settings, induction leads, heads of department/phase etc as well as the different subject associations and professional bodies.
 - Potential risk of GenAI being used without consideration of the possible unintended consequences – e.g. some staff lacking criticality and not taking properly quality assurance measures to evaluate the AI generated content/data.
 - Importance of keeping the ‘human in the loop’ - need to check the outputs for accuracy/relevance and be aware of risk of hallucinations (this would probably need to be made explicit by the facilitator in initial PD sessions)
 - Definite need for a confident/expert user of GenAI to lead the sessions, as highlighted in the evaluation survey: ‘exploring a problem whilst having access to the expert’.
 - Awareness of participants’ experience/attitudes towards GenAI would be useful when planning PD so sessions can be tailored accordingly.
 - Maybe useful to have a particular use-case in mind when planning the session. Staff could bring specific use-cases/problems and exemplar work/exam answers to sessions to increase relevance of the output to school/curriculum context. As an improvement, one respondent to the evaluation survey suggested being provided with prompts and/or topics to explore, then seeing ‘who can generate a usable (or the most effective/appropriate) response fastest’.
 - Respondents to the evaluation survey suggested that the hackathon could be given more time, or it could become a series of progressive sessions, developing knowledge and skills over time.

- What are key questions for future research?
 - Does it require a data scientist to be present or could it just be someone within the school who has some knowledge in the subject area?

- When is the best time in a teacher's career for GenAI to be introduced as a tool? (Debate over whether ITTs/ECTs should be encouraged to develop own 'mental models' of teaching practice before having 'help' from AI)
- Could GenAI be used as a (coaching/'buddy?') for staff to quickly ask for guidance if the person they need to talk to is unavailable? Eg. A situation where a TA/ECT needs to ask how to deal with a particular situation in real-time.
- A need for GenAI training was identified ('I would like more training in using AI and the features we discussed during the session.'). This leads to the question of what is the best way to support schools in using GenAI? This is a pressing question the NIoT is thinking deeply about as it undertakes continuous improvement of the ITE, ECF and NPQ programmes.

Annex 4: Membership of steering committee

- Jisc
- Education Endowment Foundation
- Ofsted
- Ofqual
- Information Commissioner's Office (ICO)
- Centre for Data Ethics and Innovation (CDEI)
- Department for Science, Innovation and Technology (DSIT)
- Office for AI
- Office for Students (OfS)

Annex 5: Methods for user research

Methods

This study employed a qualitative methodology, involving interviews with teachers and focus group discussions with students, across a number of schools in England.

Sample and sampling strategy

Teachers (N = 12) from nine schools, and with an average of 16 years' experience, were recruited via senior school leaders within the National Institute of Teaching's (NfT) network between January and February 2024. To gain a clear picture of how so-called "later adopters" of AI tools would feel about using them in the classroom, we used convenience sampling to recruit mainly teachers with little to no previous experience with AI. Teachers worked in schools in a range of settings (see Table X.1) and were spread geographically across England (North West x 2; South West x 4; North East x 1; South East x 1; West Midlands x 1). The limitations of this sample are discussed in the Conclusion section below.

Table 14: Teacher sample and school contexts

Teacher number	Gender	Experience (years)	School type	Size	Ofsted rating	Urban / Rural	%FSM
1	Female	13	Primary	420	Outstanding	Urban with Significant Rural	8.1
2	Female	11	Secondary	1712	Outstanding	Urban: City and Town	8.8
3	Female	33	Secondary	1712	Outstanding	Urban: City and Town	8.8
4	Female	8	Secondary	1210	Good	Urban with Significant Rural	23.8
5	Female	20	Secondary Grammar	1038	Outstanding	Rural: Village	6.0

Teacher number	Gender	Experience (years)	School type	Size	Ofsted rating	Urban / Rural	%FSM
6	Female	30	All-through Special	330	Outstanding	Urban: City and Town	67.3
7	Male	8	Primary	798	Good	Urban Minor Conurbation	48.0
8	Male	15	Primary	618	Good	Urban: City and Town	17.5
9	Female	10	Primary	618	Good	Urban: City and Town	17.5
10	Male	16	Secondary	200	Outstanding	Urban with Significant Rural	14.4
11	Female	16	Primary Special	59	Good	Urban with Significant Rural	74.6
12	Male	-	Primary	-	-	-	-

Note: Teacher 12 had been working for the last four years as a digital learning lead across a number of schools within a Multi-Academy Trust.

Students (N = 9) aged 13-17 years, from two secondary schools in the South of England were recruited via correspondence with their headteachers to take part in one of two online focus group discussions about their perspectives on AI use in school, with a particular focus on teachers' use of AI for feedback.

Materials and Procedure

Teachers

Teachers were invited to trial a new AI PoC tool designed to provide feedback on Year 4 students' literacy work and then take part in a 45 minute online one-to-one interview. After agreeing to participate, teachers were sent a link to access a PoC tool developed by Faculty and a video tutorial for how to use it. Interviews were subsequently held online and focused on teachers' perceptions of barriers and facilitators to AI use (be they

physical, psychological, structural etc.), risks and opportunities associated with AI use, and the role of the school context in the adoption of AI technologies.

Each semi-structured interview was carried out by one of three members of the research team and focused on teachers' experiences with the PoC tool, their views on using AI for feedback and more broadly within education, and the barriers and facilitators to AI use in their school environments. At the start of each interview, the interviewer made clear to the teacher that despite working on a project about AI in education, they had no agenda beyond giving teachers a voice on the subject. Teachers reported varied levels of engagement with the tool prior to the interview, and in a small number of cases where they had not viewed it at all, the interviewer shared a brief overview of the tool remotely during the interview.

Students

Students were presented with a description of the PoC tool and how teachers may use it. Following this, the NIoT researcher led a discussion about how students would feel about receiving AI-generated feedback, as well as their thoughts on how useful and appropriate this would be for different teaching and learning styles, and different subjects. Students also had the chance to discuss the risks and opportunities associated with implementing tools like this in school.

All interviews and focus groups were recorded and transcribed in Microsoft Teams before being checked by the research team to remove identifying information and to correct inaccuracies in the automated transcription.

Data analysis

Transcripts were thematically analysed (Braun & Clarke, 2006) by the lead researchers and two additional Research Assistants using the NVivo software. The interviews were initially coded by the lead researchers to identify the overarching themes in the data. The four coders then met to discuss and refine the codes in two feedback sessions, before each time returning to the data and recoding in NVivo. Emerging findings were also shared with teacher education experts at the NIoT for further feedback. This iterative process of coding and discussing resulted in several key themes and a number of sub-themes (see below).

Annex 5: Prompts used in the Hackathons

This section includes some of the prompts and outputs developed in the hackathons. Each has three parts:

1. **System prompt:** Sets the behaviour of the assistant and provides instructions on how its response should be. This can include things like tone, answer formatting and data to reference. In a conversation, system prompts are remembered throughout the entire conversation.
2. **User prompt:** A request given to the assistant. Any conversation with the LLM will consist of a series of user prompts intercalated with assistant responses.
3. **Response/Conversation:** The output of the LLM. Where the interaction was a conversation with the LLM, the entire conversation will be presented.

During the hackathons, teams created Jupyter notebooks showing their method of working and the data they used. The prompts presented here are snippets of those notebooks, meaning that not all of the prompts developed in the hackathons are included in this Annex, and some data has been omitted for brevity. The full version of the notebooks will be used to inform the development of a PoC for the next stage.

Generating age-appropriate lesson plans

System prompt

Given a lesson title and 3 learning objectives, create a 1-hour lesson plan following the 5 pillars of learning for a Year 7 class following the national curriculum.

Learning objectives must be clarified. Information from previous lessons must be recapped.

New information must be introduced twice and for a maximum of 5 minutes each time.

Information must be practised through exercises. Regular feedback must be given.

User prompt

Lesson Name: Binary Logic Circuits

Learning Objectives:

- 1) Understand the fundamentals of binary logic circuits, including the concepts of Boolean algebra and truth tables.
- 2) Apply the knowledge of binary logic circuits to analyse and design simple circuits, such as AND, OR, and NOT gates.

- 3) Develop problem-solving and critical thinking skills by exploring different combinations of input signals and predicting the output of a given circuit.

Response

Lesson Plan:

Duration: 1 hour

Introduction (5 minutes):

- Recap the previous lesson on computer systems and binary numbers. Ask students to share their understanding of binary numbers and how they are used to represent information in digital systems.
- Explain that in this lesson, we will be exploring binary logic circuits, which use binary numbers and basic logic operations to process and manipulate information in computer systems.

Body of Lesson:

1) Understanding Binary Logic Circuits (15 minutes):

- Introduce the concept of Boolean algebra, which is used to express logic operations and manipulate logical statements.
- Explain the three basic logic gates: AND, OR, and NOT gates. Briefly describe each gate's function and draw truth tables to show the relationship between input and output signals.
- Show real-world examples of binary logic circuits, such as traffic light systems or vending machines, to help students understand how these circuits are used to control different scenarios.

2) Applying Binary Logic Circuits (20 minutes):

- Provide examples of simple binary logic circuits, such as circuits with multiple AND, OR, and NOT gates. Explain how these circuits can be analyzed and designed using Boolean algebra and truth tables.
- Present students with a problem-solving exercise where they need to analyze a given circuit and determine the output based on different input combinations. Encourage students to work in pairs or small groups to discuss and solve the exercise.
- Circulate around the classroom, providing support and guidance to students as they work on the exercise.
- After the exercise, ask students to share their answers and explanations, providing feedback and clarifying any misconceptions.

3) Exploring and Predicting (15 minutes):

- Present students with a set of logic circuits with incomplete truth tables or missing in-puts. Ask students to predict the output of the circuit for the missing parts.
- This activity will encourage students to think critically and apply their knowledge of binary logic circuits to solve problems that require them to infer missing information.
- Discuss the predictions and provide feedback, emphasizing the importance of logical reasoning and deductive thinking.

4) Conclusion (5 minutes):

- Recap the main points covered in the lesson: the fundamentals of binary logic circuits, including Boolean algebra, truth tables, and the three basic logic gates (AND, OR, NOT).
- Remind students of the significance of binary logic circuits in computer systems and how they are used to process and manipulate information.
- Assign a homework task that allows students to further practice their understanding of binary logic circuits, such as designing a simple circuit or solving additional problems.

End of Lesson.

Generating effective questioning materials

System prompt

You are a UK school teacher.

You need to prepare questions for a Year 9 class using the OCR syllabus and OCR command words.

In order to prepare a lesson on computer science, you need to prepare questions for the students.

There are three types of questions:

- knowledge: Knowledge is defined as the remembering of previously learned material. This may involve the recall of a wide range of material, from specific facts to complete theories, but all that is required is the bringing to mind of the appropriate information.
- understanding: Understanding is defined as the ability to grasp the meaning of material. This may be shown by translating material from one form to another (words or numbers), by interpreting material (explaining or summarizing, and by estimating future trends (predicting consequences or effects).

- application: Application refers to the ability to use learned material in new and concrete situations. This may include the application of such things as rules, methods, concepts, principles, laws, and theories. Learning outcomes in this area require a higher level of understanding than those of comprehension.

User prompt

Provide a understanding question worth 6 using the command words [Define, Describe, Explain, Discuss, Compare, Contrast, Evaluate, Analyze, Examine] that covers one aspect of the contents of a lessons which objectives are:

- Be able to explain the difference between wireless and wired network devices.
- Provide the respective mark scheme outlining number of marks for each point made.

Response

Question:

Discuss in detail the difference between wireless and wired network devices, their respective advantages and disadvantages, and how they operate. (6 marks)

Mark Scheme:

- 1) Definition of a wireless network device (1 mark)
- 2) Definition of a wired network device (1 mark)
- 3) Explanation of their operation: Data transmission through the air for wireless, and through physical cables for wired. (1 mark for explaining the operation of each)
- 4) Comparison of their advantages and disadvantages: Wired networks often provide more stable connections, and are usually faster, but cables can restrict mobility and be difficult to install. Wireless networks provide more freedom of movement, easy to set up, but could experience more interference and security problems. (1 mark for advantages and disadvantages of each)

Each part is worth 1 mark and the final score is the cumulative total of each part. A total of 6 marks can be achieved.

Marking a KS2 essay based on a modified mark scheme and providing personalised feedback

System prompt

You are a Year 4 English teacher.

You need to mark your students work following a mark scheme we provide.

The mark scheme will give a category with four levels where level 4 is the best. Each level has examples of the elements that describe the level but the student does not need to use all of them to qualify.

Compare the level descriptions to the student's work and pick the level that most closely matches the work. Give a list of where they student met the criteria of the level, a list of spelling and grammar mistakes the student made, and a list of what is missing from the criteria of the level above.

Give your answers in a valid json format with the form {"Level": X, "Met Criteria": "Y", "Mistakes": "W", "Missing": "Z"}.

User prompt

The category is Text. The mark scheme is as follows:

Level 1: Sequencing sentences to form short narratives.

Level 2: Correct choice and consistent use of present tense and past tense throughout writing. Use of the progressive form of verbs in the present and past tense to mark actions in progress [for example, she is drumming, he was shouting].

Level 3: Introduction to paragraphs as a way to group related material. Headings and sub-headings to aid presentation. Use of the present perfect form of verbs instead of the simple past [for example, He has gone out to play contrasted with He went out to play].

Level 4: Use of paragraphs to organise ideas around a theme. Appropriate choice of pronoun or noun within and across sentences to aid cohesion and avoid repetition.

Use the mark scheme to mark this writing exercise: {exercise removed for brevity}

Response (with Temperature set to 0)

'Text': {'Level': 3,

'Met Criteria': ['Correct choice and consistent use of present tense and past tense throughout writing',

'Use of the present perfect form of verbs instead of the simple past'],

'Mistakes': ["The word 'present' should be in the present perfect form 'have got' instead of 'got'",

"The word 'picked' should be in the present perfect form 'have picked' instead of 'picked'",

"The word 'urging' should be in the present perfect form 'has urged' instead of 'urged'"],

'Missing': ['Introduction to paragraphs as a way to group related material',

'Headings and sub-headings to aid presentation']}]}

Providing feedback on students' work in a way meaningful to students

System prompt

You are a Year 4 English teacher.

You have written feedback on a student's essay in JSON format. You have broken the feedback down into 5 categories.

Your task is to rewrite your feedback as though you are talking directly to an 8-year-old.

Use encouraging and appropriate language for your student. Use British English throughout. Finish the feedback with a positive note.

The student feedback should have the following Sections and subsections.

Sections:

Positives (Subsections: Word Sentence Text Punctuation Terminology)

Even better if (Subsections: Word Sentence Text Punctuation Terminology)

User prompt

Within both the positive and negative feedback sections, break down your feedback into 5 categories as shown in your teacher feedback (Word, Sentence, Text, Punctuation, Terminology)

Within the 'Even better if' section, if you have examples of how the essay could be improved, then provide the examples.

Teacher feedback:

Student feedback:

Response

Hello there, superstar! I've just finished reading your essay and I must say, you've done a fantastic job! Let's talk about all the great things you've done and some areas where we can make your writing even better.

Positives:

- 1) Word: You've done a brilliant job with your words! You've used a variety of prefixes and you've shown that you understand word families based on common words. Well done!
- 2) Sentence: You've done a great job joining words and clauses using 'and'. Your expanded noun phrases for description and specification are also very good. You've also done a good job expressing time, place and cause using conjunctions, adverbs, or prepositions. Keep it up!
- 3) Text: You've done a good job with your tenses. You've used the present tense and past tense correctly and consistently throughout your writing. You've also used the present perfect form of verbs instead of the simple past. Good job!
- 4) Punctuation: You've used capital letters, full stops, question marks and exclamation marks correctly to mark the end of sentences. You've also started using inverted commas to punctuate direct speech. Well done!
- 5) Terminology: You've shown that you understand a lot of important words like 'noun', 'verb', 'conjunction', 'prefix', 'clause', 'direct speech', 'consonant', 'vowel', and 'inverted commas'. That's fantastic!

Even better if:

- 1) Word: It would be even better if you could show the difference between plural and possessive –s and use Standard English forms for verb inflections instead of local spoken forms.
- 2) Sentence: Try to expand your noun phrases by adding modifying adjectives, nouns and preposition phrases. Also, try to use fronted adverbials.
- 3) Text: Try to introduce paragraphs to group related material and use headings and sub-headings to make your writing clearer.

- 4) Punctuation: Try to use inverted commas and other punctuation to indicate direct speech, apostrophes to mark plural possession, and commas after fronted adverbials.
- 5) Terminology: It would be great if you could use words like 'determiner', 'pronoun', 'possessive pronoun', and 'adverbial' in your writing.

Remember, practice makes perfect! You're doing a great job and I'm really proud of you. Keep up the good work!

Prompt used to generate newsletter articles, using standardised data fields (parent-carer communications)

System prompt

You are a UK school administrator, in charge of preparing newsletters. Teachers send you event reports which you then expand on and include in the newsletter. Write in British English. The report must be written in a consistent and engaging style and tone of voice. You need to follow these steps:

- 1) For each teacher article, read it and understand the core concepts.
- 2) Expand on them using a consistent tone of voice

Limit your writing to 500 words.

User prompt

Expand this report:

1.)

"

What: Charity fundraising (Samosa sale for 1 pound)

When: Every Friday

Who: Year 8 and Year 9 students

Why: Raise money for Water Aid and African Relief Funder"

Response

- 1) Savoury Samosa Sale for Charity

Our industrious Year 8 and Year 9 students have taken it upon themselves to organise a delightful samosa sale, with the noble aim of raising funds for two commendable charities - Water Aid and African Relief Funder. Every Friday, our school corridors will be filled

with the tantalising aroma of these delicious treats, available for just £1 each. This is a wonderful opportunity for our school community to come together, enjoy a tasty snack, and contribute to a worthy cause. We encourage everyone to support our students in this charitable endeavour and help make a difference to those in need.

Prompt used to simplify text for students (lesson materials)

System prompt

You are a UK history teacher teaching to pupils in key stage 3. You adapt your lessons to the reading age of each student. To adapt the texts you use techniques like:

Simplify the language and use synonyms to articulate key arguments from sources.

Use physical timelines and diagrams to contextualise the sources and historical analysis

Use features of a period to create mental images

Add additional definitions or relevant examples that are relatable.

User prompt

The following text is aimed at students with reading age of at least 12. Adapt the passage for a student who has a reading age of 10 with a limited knowledge of the vocabulary and nuances.

How far was Britain changed by the Viking Age?

Following the first raid on Lindisfarne Monastery in AD 793, the Vikings raided Britain for over 200 years. Steadily these raids became larger with more ships and men, some even began to set up camps to over-winter. A turning point came in AD 865, when the Vikings changed tactics. Instead of small hit-and-run raids the Vikings launched a full-scale invasion of England. One by one, the Anglo-Saxon kingdoms fell until only Wessex remained.

Thankfully for Wessex, its King Alfred proved to be an impressive military leader and good ruler. He defeated the Viking army in AD 878 at the Battle of Edington and forced its leader, Guthrum, to convert to Christianity.

Alfred spent the next 21 years of his life creating burhs across Wessex, strengthening the fyrd and creating a navy. In AD 866 he even captured London back from the Vikings.

In AD c.890, Alfred and Guthrum divided England between themselves at the Treaty of Wedmore. Using the old Roman road of Watling Street Alfred took southern England and western Mercia whilst Guthrum took the rest. This area came to be called 'Danelaw' due

to the inhabitants living under Danish laws and customs. The Vikings and Anglo-Saxons who lived mixed and it became very hard to tell who was 'Viking' and who was English.

In AD 927, Alfred's grandson Aethelstan captured the last bit of the Danelaw unifying all of the English for the first time. Despite it not surviving his death, the Danelaw became a permanent part of England in AD 954.

However, the North and East remain greatly impacted by Viking settlement.

By simply looking at a map you can see the evidence of Viking settlement, north of the line from the Treaty of Wedmore exists the majority of the place names from the Viking languages such as suffixes '-by' and '-thorpe'. Even the language has been impacted with simple words like 'egg', 'give' and 'happy'. Trial by Jury also comes from Viking culture and is still used today.

Outside of England the Vikings had perhaps an equal or even larger impact. In the Scottish Islands of Shetland and the Orkneys the Vikings completely took over and more than 50% of modern DNA is Viking compared to under 10% in England. On the Isle of Mann, the Tynwald (Parliament) is descended from the Viking gatherings called things'.

Response

Did the Vikings change Britain a lot?

In the year 793, the Vikings first attacked a place called Lindisfarne Monastery in Britain. For over 200 years, they kept coming back to raid. Their attacks got bigger and bigger, with more ships and men. Some Vikings even started to stay over the winter. In 865, the Vikings changed their plans. Instead of small attacks, they decided to invade England fully. They took over the kingdoms one by one, until only Wessex was left. Luckily for Wessex, their King Alfred was a great leader and ruler. He beat the Viking army in 878 at a battle called Edington and made their leader, Guthrum, become a Christian. For the next 21 years, Alfred worked hard to make Wessex stronger. He built forts, trained soldiers and created a navy. In 866, he even took back London from the Vikings. Around 890, Alfred and Guthrum made a deal. They split England between them. Alfred got southern England and western Mercia, and Guthrum got the rest. The area Guthrum got was called 'Danelaw' because the people there lived under Danish laws. The Vikings and the English people who lived there mixed together, and it was hard to tell who was who. In 927, Alfred's grandson Aethelstan took over the last bit of the Danelaw. This was the first time all of England was under one rule. Even though it didn't last long, the Danelaw became a permanent part of England in 954. However, the North and East of England were still greatly changed by the Vikings. You can see this on a map. North of the line from the deal Alfred and Guthrum made, most of the place names come from the Viking languages, like names ending in '-by' and '-thorpe'. Even some of our words like 'egg', 'give', and 'happy' come from the Vikings. The idea of a trial by jury also comes from the Vikings and we still use it today. The Vikings also had a big impact outside of England. In

the Scottish Islands of Shetland and the Orkneys, the Vikings took over completely. Even today, more than half of the people there have Viking DNA. On the Isle of Mann, their Parliament, called the Tynwald, comes from the Viking meetings called 'things'.

Hallucination in creating end-of-year reports, with the hallucination highlighted in bold italics

System prompt

You are now tasked to write a 200-word paragraph to highlight a student's current behaviour, grade, target grade, in a single cohesive paragraph.

The user input will tell you the behaviour, grade, and target grade of the student.

The behaviour is measured on a scale from 1 to 4 where 1 is the best and 4 is the worst.

The grades are measured on a scale from 1 to 9 where 9 is the best and 1 is the worst.

The enjoyment of a subject is measured on a scale from 1 to 10 where 10 is them really enjoying the class and 1 is not.

In your paragraph, include the student behaviour, grade, target grade or new target grade, enjoyment of subject, homework completion rate, what the student has done well, what the student has found challenging, recommendations for improvement.

With the behaviours, grades and target grades, be factual and non-opinionated.

User prompt

Behaviour: 1 out of 4

Grade: 6 out of 9

Target grade: 5 out of 9

Enjoyment: student does not seem interested in certain topics but generally enjoys the class and frequently partakes

Homework completion: student always submits their homework on time and does their corrections.

Response

The student has exhibited excellent behavior in the classroom, consistently demonstrating attentiveness and engagement. Their current grade stands at a commendable 6 out of 9, which indicates a solid understanding of the subject matter.

However, their target grade has recently been adjusted to a 5 out of 9, suggesting a need for further improvement. The student generally enjoys the class, although they have shown a lack of interest in certain topics. Despite this, they frequently participate and actively contribute to class discussions. Furthermore, the student consistently completes their homework on time and diligently corrects any mistakes. They have shown exceptional commitment to their studies and have consistently demonstrated responsibility in completing their assignments. However, the student has found certain topics challenging and has displayed a lack of interest in them. To improve, it is recommended that the student actively seeks clarification on these challenging topics and engages in additional practice outside of the classroom. With continued effort and a proactive approach, the student has the potential to achieve their target grade and further enhance their understanding and enjoyment of the subject.

Writing end-of-year reports for other teachers based on key information

System prompt

You are a bot which is tasked to compare performances of a student across 3 terms.

The user will provide you with relevant information.

You will need to identify key insights on it.

You are trying to provide key points to consider for a teacher who has not taught this student before.

User prompt

Term 1:

Biology:

Target Grade = 4

Prediction Grade = 6

(Grades span a scale from 1 to 9, where grade 1 is bad and grade 9 is great)

Effort = E1

Homework on time = True

Attendance = 94.2%

Term 2:

Biology:

Target Grade = 4

Prediction Grade = 5

Effort = E2

Homework on time = True

Attendance = 91%

Term 3:

Biology:

Target Grade = 4

Prediction Grade = 5

Effort = E3

Homework on time = False

Attendance = 92.3%

Response

- 1) The student's performance in Biology has been consistently above the target grade. In Term 1, the student exceeded the target grade by 2 points, and maintained a grade above the target in Term 2 and Term 3.
- 2) There is a slight decline in the predicted grade from Term 1 to Term 2 and Term 3. This could be due to a decrease in effort, as indicated by the effort grades moving from E1 to E3.
- 3) The student is generally good at submitting homework on time, with the exception of Term 3. This could be a contributing factor to the decline in the predicted grade.
- 4) The student's attendance is relatively high, with a slight dip in Term 2. However, it improved in Term 3. Regular attendance is likely contributing positively to the student's grades.
- 5) Overall, the student seems to be a high achiever in Biology, exceeding the target grade consistently. However, the decreasing trend in effort and the missed homework in Term 3 could be areas of concern. It would be beneficial to encourage the student to maintain a high level of effort and ensure homework is consistently submitted on time to maintain or potentially improve their grades.

Prompting the LLM to create a vocabulary list with simple prompting (language learning assistant)

System prompt

“You are an expert GCSE Spanish Teacher. The student is learning about travel and holidays”

User prompt

“Please can you suggest a list of vocab I should learn, and present the results in a table”

Response

Of course! Here is a list of vocabulary related to travel and holidays, presented in a table format:

English	Spanish
Airport	Aeropuerto
Airplane	Avión
Train	Tren
Bus	Autobús

Prompting the LLM to correct a student’s work by providing steps to follow and one-shot prompting

User prompt

Is this sentence right? Le weekend dernier, je suis aller au cafe et j'ai manger les frites. J'aime frites.

Simple system prompt

You're a language tutor. You help students improve their language skills.

Simple system prompt output

The sentence should be: Le weekend dernier, je suis allé(e) au café et j'ai mangé des frites. J'aime les frites.

Complex system prompt

You're a language tutor. You help students improve their language skills.

When a student asks you about their work, you take the following steps:

- 1) Read the work carefully
- 2) Write a concise sentence with the meaning the student tried to convey in a higher language level
- 3) Point out the mistakes to the student

Here is an example conversation:

Student: "Is this sentence right? Mañana fui a la biblioteca. Me gustan las libros y quiero leer mucho."

You: "The sentence has a few errors. Here is a better version: Mañana voy a ir a la biblioteca, ya que me gustan los libros y me apetece leer mucho."

In Spanish, libro is a masculine noun and must be used with the 'el' article. Also, 'fui' is a past tense verb"

Complex system prompt output

The sentence still contains a few errors. Here's the corrected version: "Le weekend dernier, je suis allé(e) au café et j'ai mangé des frites. J'aime les frites."

In French, the past participle of "aller" should be "allé(e)" depending on your gender. The verb "manger" should be conjugated as "mangé" in the past tense. Additionally, use the partitive article "des" instead of the definite article "les" before "frites" to indicate an unspecified quantity of fries. Finally, don't forget to include the word "les" before "frites" to indicate that you like them specifically.

Adjusting data literacy policies in line with school values

System prompt

You are a senior leader in a school that specialises in writing statutory policies.

You have been tasked with helping write the first draft of guidances that will be published in the school's website.

This is your school's vision: {school vision redacted}

User prompt

Write the school's policy on digital literacy.

Response

Policy: Digital Literacy

Introduction:

At [School Name], we recognize that digital literacy is an essential skill for our students' success in today's digital world. As part of our commitment to providing an outstanding education that prepares our students for the future, we strive to ensure that all students develop the necessary skills and knowledge to effectively and responsibly use digital technologies. This policy outlines our approach to digital literacy and sets out guidelines and expectations for the safe, responsible, and effective use of technology within our school community.

Policy Statement:

All students at [School Name] will have access to a digitally-rich learning environment that promotes digital literacy as a key component of their education. We believe that digital literacy is more than just technical skills; it encompasses the ability to critically evaluate and use digital resources, communicate and collaborate online, and practice responsible digital citizenship.

Aims and Objectives:

- 1) To equip students with the necessary skills and knowledge to use digital technologies effectively and responsibly.
- 2) To promote critical thinking and digital literacy skills that enable students to evaluate online information and resources.
- 3) To foster creativity and innovation through the use of digital tools and technologies.
- 4) To provide opportunities for students to collaborate, communicate, and connect with others online in a safe and responsible manner.
- 5) To ensure that students understand and adhere to ethical, legal, and safe online practices.
- 6) To support and empower teachers in integrating digital technologies into their teaching practices.
- 7) To establish partnerships and collaborations with parents, guardians, and the wider community to support students' digital literacy development.

Implementation:

- 1) The integration of digital literacy will be embedded across the curriculum and taught explicitly in dedicated lessons to ensure that all students have the necessary skills and knowledge.

- 2) Teachers will receive professional development and support to effectively integrate digital technologies into their teaching practices and encourage the development of digital literacy skills in their students.
- 3) The use of age-appropriate technology and software will be encouraged to cater to the diverse needs and abilities of our students.
- 4) A safe and secure online environment will be maintained, and internet access will be monitored and filtered to ensure the safety of students.
- 5) Clear guidelines for responsible and safe use of digital technologies will be communicated to students, parents, and staff, and regularly reviewed to reflect changes in technology and best practices.
- 6) Regular communication and collaboration with parents and guardians will be encouraged through workshops, resources, and online platforms to support their understanding of digital literacy and their role in promoting its development.
- 7) Opportunities for students to engage in digital citizenship initiatives, such as online safety campaigns and responsible social media use, will be provided to develop their understanding of their rights and responsibilities online.

Monitoring and Review:

The school's leadership team will monitor the implementation of this policy to ensure that digital literacy is effectively integrated into teaching and learning practices. Regular reviews of the policy will be conducted to reflect changes in technology, legal requirements, and educational best practices. Feedback from students, parents, staff, and the wider community will be considered to continuously improve our approach to digital literacy.

Conclusion:

At [School Name], we view digital literacy as an essential skill that empowers our students to thrive in an increasingly digital world. By providing a comprehensive approach to digital literacy education, we are committed to equipping our students with the skills, knowledge, and attitudes they need to be responsible, critical, and creative digital citizens. Through the effective integration of digital technologies, we aim to nurture today's young people and inspire tomorrow's leaders.

Providing detailed context for adapting school policy

System prompt

You are a senior leader in a school that specialises in writing statutory policies.

You have been tasked with helping write the first draft of guidances that will be published in the school's website.

User prompt

Summarise the following policy so 5 year-old students can understand it.

Summarise it under 100 words, and only include content relevant for 5 year-olds.

This is the policy:

Response

Digital Literacy:

At [School Name], we help students learn and use digital technology in a safe and responsible way. We believe that digital literacy means more than just knowing how to use computers and tablets. It also means knowing how to think critically, use the internet safely, and be a good citizen online. Our teachers will teach students these skills in special lessons and throughout their regular classes. We will use tools and software that are right for each student. We will also ask students' families to get involved and learn about digital literacy too.

Creating policy-based scenarios for testing with staff

System prompt

You are a senior leader in a school that specialises in writing statutory policies.

You have been tasked with testing people's understanding of school policy, by asking them multiple choice questions.

The multiple-choice questions have 3 obviously incorrect answers, and 1 correct answer.

Do not use "All of the above" as a possible answer.

This is the policy: {policy redacted for brevity}

For each question, follow these steps:

- 1) Pick 1 point from the policy
- 2) Generate 3 very incorrect answers
- 3) Generate 1 correct answer
- 4) Present them in a random order

Conversation (GPT-4)

User:

Test whether I understand the policy

Assistant:

- 1) What does the Digital Literacy policy at [School Name] aim to promote?
 - a. Only the technical skills required for using digital technologies
 - b. The use of digital technologies for entertainment purposes
 - c. The use of digital resources without critical evaluation
 - d. The responsible and effective use of digital technologies, promoting critical thinking, creativity, and collaboration online

- 2) How does [School Name] aim to support teachers with the Digital Literacy policy?
 - a. By providing extra working hours to teachers for technology integration
 - b. By barring teachers from using traditional teaching methods

- c. By providing professional development and support for integrating digital technologies into their teaching practices
 - d. By making them solely responsible for students' digital literacy development.
- 3) What measure has [School Name] put in place to ensure the safety of students when using digital technologies?
- a. Providing unrestricted internet access for learning freedom
 - b. Discouraging the use of technology unless necessarily required
 - c. Frequent changes of passwords to all school computers
 - d. Maintaining a safe and secure online environment with monitored and filtered Internet access
- 4) How does [School Name] aim to involve parents in supporting the Digital Literacy policy?
- a. By insisting parents supervise all online activities of students at home
 - b. By providing workshops, resources, and online platforms for communication and collaboration
 - c. By making parents responsible for the digital literacy education at home
 - d. By asking parents to limit the use of technology at home
- 5) According to the policy, how often is the Digital Literacy policy reviewed?
- a. Only when there is a significant upheaval in the school management
 - b. Whenever there is a technological upgrade available
 - c. Regularly, reflecting changes in technology, legal requirements, and educational best practices
 - d. When students feel they are having difficulties with the implemented policies.

User:

My answers are:

- 1) a
- 2) b
- 3) c
- 4) d
- 5) b

Did I get it right? Why?

Assistant:

Unfortunately, none of your answers is correct. Here's why:

- 1) The correct answer is d. The policy aims to promote not just the technical skills but also critical thinking, creativity, and collaboration in the use of digital technologies.
- 2) The correct answer is c. The school provides professional development and support to help teachers incorporate digital technology into their teaching methods. It doesn't state anything about barring traditional teaching methods.
- 3) The correct answer is d. The school safeguards its students by maintaining a safe and secure online environment, with monitored and filtered internet access.
- 4) The correct answer is b. The policy aims to involve parents by providing them with resources, workshops, and online platforms for communication and collaboration, rather than solely making them responsible for their child's digital literacy or limiting their technology use at home.
- 5) The correct answer is c. The policy is reviewed regularly, keeping in tune with the changes in technology, legal requirements, and educational best practices, not only when there's a technological upgrade.

Try to study the policy more thoroughly and attempt the quiz again.

Generating newsletters based on short user inputs

System prompt

You are a UK school administrator, in charge of preparing newsletters.

Teachers send you event reports which you then expand on and include in the newsletter.

Write in British English.

The report must be written in a consistent and engaging style and tone of voice.

You need to follow these steps:

- 1) For each teacher article, read it and understand the core concepts.
- 2) Expand on them using a consistent tone of voice

Limit your writing to 500 words.

User prompt

Summarise these teacher reports:

1.)

"

What: Charity fundraising (Samosa sale for 1 pound)

When: Every Friday

Who: Year 8 and Year 9 students

Why: Raise money for Water Aid and African Relief Funder"

2.)

what: Trip to see Romeo and Juliet

where: The Globe, London

when: Last week

who: 65 Year 9 students

why: To experience the magic of Shakespeare's language and build better understanding of the plot in preparation for GCSE study

Impact: Pure enjoyment, build better plot understanding, more success at GCSE

Response

- 1) Charity Fundraising: Every Friday, Year 8 and Year 9 students have the opportunity to make a difference by participating in our samosa sale. For just £1, students can enjoy a delicious snack while supporting two important causes: Water Aid and African Relief Fund. By raising money for these organizations, our students are actively contributing to improving access to clean water and providing aid to those in need. This initiative not only teaches our students the value of giving back, but also instills a sense of empathy and compassion within them. Together, we can make a significant impact and create positive change in the world.
- 2) Trip to The Globe: Last week, 65 Year 9 students embarked on an unforgettable journey to The Globe in London to witness the timeless tale of Romeo and Juliet.

This immersive experience allowed our students to fully immerse themselves in the magic of Shakespeare's language and gain a deeper understanding of the play's plot. By witnessing the live performance, our students were able to connect with the characters and emotions in a way that cannot be replicated in the classroom. This trip not only provided pure enjoyment for our students, but also equipped them with the necessary knowledge and appreciation for Shakespeare's works, which will greatly benefit them in their upcoming GCSE studies. By building a better understanding of the plot, our students are now better prepared to analyze and interpret the play, leading to increased success in their exams. We are proud to offer such enriching experiences that go beyond the confines of textbooks and classrooms, allowing our students to truly engage with the subjects they are studying.

Mandarin language assistant

System prompt

You're a language tutor. You help students improve their language skills.

When a student asks you about their work, you take the following steps:

- 1) Read the work carefully
- 2) Write a concise sentence with the meaning the student tried to convey in a higher language level
- 3) Point out the mistakes to the student

Conversation

User:

'Is this sentence right? 星期后我去兵官吃饺子。我真喜欢饺子'

Assistant:

'The sentence has a few errors. Here is a better version: 星期后我会去饺子店吃饺子。我非常喜欢吃饺子。'

In Chinese, it is more appropriate to use the phrase "饺子店" instead of "兵官" to refer to a place where you can eat dumplings. Additionally, the word "真" can be replaced with "非常" to express a stronger liking for dumplings.'

User:

'Why is using 真 wrong here though?'

Assistant:

'Using 真 in the sentence is not necessarily wrong, but it may not convey the intended meaning accurately. 真 is often used to mean "really" or "truly" in Chinese, but in this context, using "非常" would express a stronger liking for dumplings. It would be more natural and common to say "我非常喜欢吃饺子" to indicate a high level of enjoyment or preference for dumplings.'

User:

'I see. Can you give me an example of where using 真 is appropriate?'

Assistant:

"Certainly! Here's an example where using 真 is appropriate:

这本小说真有趣！我一下午都在看它。

Translation: This novel is really interesting! I've been reading it all afternoon.

In this case, 真 is used to emphasize the speaker's opinion about the novel being truly interesting. It adds emphasis to the statement and conveys a strong feeling of enjoyment or fascination."

User:

'Ok, what about this sentence, is it right? 我学了汉语八月了，因为我的室友也说汉语'

Assistant:

'The sentence is almost correct. Here is a slightly modified version: 我学了汉语已经有八个月了，因为我的室友也会说汉语。

In Chinese, when indicating the duration of learning something, it is more common to use the structure "已经 + duration + 了" instead of just stating the number. Also, we use the verb "会" to indicate that someone can speak a language rather than "说".'

Observations

There are several flaws in this response:

- 1) **Explaining nuances:** The nuances between 真 and 非常 are not explained fully.
- 2) **Omission of mistakes:** It also didn't point out all mistakes (e.g., lack of counter word 个 for in 八月 to express 8 months, 八个月, as 八月 means August instead)
- 3) **Ambiguous explanations:** 'we use the verb "会" to indicate that someone can speak a language rather than "说"' should have specified that 会 just indicates possibility of doing something, and that it should still be accompanied by 说 by writing 会说, maybe comparing it to "speaks" vs. "can speak" in English.
- 4) **Incorrect explanation:** The explanation of its inclusion of 已经 ("already") is incorrect, as the structure it proposes to use with 已经 (duration+了) should be 了+duration+了, which the student already used. Strangely, it uses the correct 了+duration+了 structure in its modified version of the student's answer.

Annex 6: Key Terms

AI: Artificial Intelligence

API: Application Programming Interface

GenAI: Generative Artificial Intelligence

GPT: Generative Pre-trained Transformer, large language models developed by Open AI

LLM: Large Language Model

MVP: Minimum Viable Product

PoC: Proof of Concept

RAG: Retrieval Augmented Generation

MAT: Multi-Academy Trust

IP: Intellectual Property

PII: Personally Identifiable Information

Annex 7: References

Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3), 431–440.

<https://doi.org/10.1007/s43681-021-00096-7>

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithms for assessment and scoring. *European Journal of Education*, 58(1), 98-110.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>

Bryant et al. (2020). McKinsey. How artificial intelligence will impact K–12 teachers. Retrieved 29 Feb 2024 from <https://www.mckinsey.com/industries/education/our-insights/how-artificial-intelligence-will-impact-k-12-teachers>

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y. S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2, 100027.

Choi, S., Jang, Y., & Kim, H. (2023). Influence of Pedagogical Beliefs and Perceived Trust on Teachers' Acceptance of Educational Artificial Intelligence Tools. *International Journal of Human–Computer Interaction*, 39(4), 910–922.

<https://doi.org/10.1080/10447318.2022.2049145>

Chounta, I.-A., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support their Practice in Estonian K-12 Education. *International Journal of Artificial Intelligence in Education*, 32(3), 725–755.

<https://doi.org/10.1007/s40593-021-00243-5>

Department for Education. (2023). Generative AI in education. Call for Evidence: summary of responses. Retrieved 4th March 2024 from

https://assets.publishing.service.gov.uk/media/65609be50c7ec8000d95bddd/Generative_AI_call_for_evidence_summary_of_responses.pdf

Department for Education. (2024). Working lives of teachers and leaders: wave 2 summary report. Retrieved 11th March 2024 from

<https://www.gov.uk/government/publications/working-lives-of-teachers-and-leaders-wave-2/working-lives-of-teachers-and-leaders-wave-2-summary-report>

Fletcher-Wood, H. (2023). How to improve behaviour and wellbeing, and how you're using AI in schools. Teacher Tapp.

Hattie, J. (2012). Visible learning for teachers: Maximizing impact on learning. Routledge.

- Jerrim, J., & Sims, S. (2021). When is high workload bad for teacher wellbeing? Accounting for the non-linear contribution of specific teaching tasks. *Teaching and Teacher Education*, 105, 103395. <https://doi.org/10.1016/j.tate.2021.103395>
- Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My Teacher Is a Machine: Understanding Students' Perceptions of AI Teaching Assistants in Online Education. *International Journal of Human-Computer Interaction*, 36(20), 1902–1911. <https://doi.org/10.1080/10447318.2020.1801227>
- Lee, A. V. Y. (2023). Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. *Studies in Educational Evaluation*, 77, 101250.
- Li, T. W., Hsu, S., Fowler, M., Zhang, Z., Zilles, C., & Karahalios, K. (2023, August). Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research – Volume 1* (pp. 159-176).
- Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British journal of educational technology*, 53(4), 914-931.
- Ng, D. T. K., Lee, M., Tan, R. J. Y., Hu, X., Downie, J. S., & Chu, S. K. W. (2023). A review of AI teaching and learning from 2000 to 2020. *Education and Information Technologies*, 28(7), 8445–8501. <https://doi.org/10.1007/s10639-022-11491-w>
- OECD. (2018). *TALIS 2018 Results (Volume I) Teachers and School Leaders as Life-long Learners*. Retrieved 4th March 2024 from: <https://www.oecd-ilibrary.org/sites/d2a4bf35-en/index.html?itemId=/content/component/d2a4bf35-en>
- Räsänen, K., Pietarinen, J., Pyhältö, K., Soini, T., & Väisänen, P. (2020). Why leave the teaching profession? A longitudinal approach to the prevalence and persistence of teacher turnover intentions. *Social Psychology of Education*, 23(4), 837–859. <https://doi.org/10.1007/s11218-020-09567-x>
- Wang, N., Wang, X., & Su, Y.-S. (2024). Critical analysis of the technological affordances, challenges and future directions of Generative AI in education: A systematic review. *Asia Pacific Journal of Education*, 1–17. <https://doi.org/10.1080/02188791.2024.2305156>
- Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, 2021, 1–18. <https://doi.org/10.1155/2021/8812542>
- Zhang, P., & Tur, G. (2023). A systematic review of ChatGPT use in K-12 education. *European Journal of Education*.



Department
for Education

© Department for Education copyright 2024

This publication is licensed under the terms of the Open Government Licence v3.0, except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3.

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned.

Reference: RR1423

ISBN: 978-1-83870-564-0

For any enquiries regarding this publication, contact www.education.gov.uk/contactus.

This document is available for download at www.gov.uk/government/publications.