FirstPass pilot

Final report (December 2023)

An Assessment Innovation Fund project

We're NCFE. We're shaping smarter learning.

We're NCFE: an educational charity and leader in vocational and technical learning. **We combine over 170 years of education experience** with deep insight, working with a network of expert collaborators to shape smarter solutions around the greatest learning needs. In doing this, we're working for a fairer education system for all learners to power inclusivity and choice.

In 1848, we were born from the belief that no learner should be left behind. Today, we're taking up that cause with fresh energy. Our vision and goals will be achieved through:

- Creating education for a fairer world
- Moving towards a smarter education eco-system
- Using our influence to shape real change
- Promoting the idea that potential is personal.

Bolton College is one of the North West's largest providers of vocational training and further education. The college enables individuals, businesses and communities within Bolton and surrounding areas to reach their potential. They provide courses for a wide range of people of all different ages and are proud to be able to offer hundreds of courses to suit everyone.

Table of Contents

Executive Summary

The FirstPass platform is designed to support students and teachers with the formative assessment of open-ended questions. The objectives for the FirstPass pilot were designed to be broad so as to assess the general efficacy of the platform to support practices relating to the formative assessment of open-ended questions:

1. To test the functionality and usability of the platform with students and teachers in multiple educational institutions.
2. To identify any technical or operational issues that need to be addressed.
3. To evaluate the platform's effectiveness in supporting the formative assessment of open-ended questions within the context of the pilot.
4. To gather feedback from participating colleges and students to inform the future development of the platform.

Method

Bridgend College, Coleg Cambria, Glasgow Kelvin College, Sunderland College, Leeds City College, and Bolton College took part in the pilot, which included a total of 65 teachers and 263 learners who used Fist Pass. Learner and teacher survey data and platform usage metrics combined to provide the data set for analysis.

Key Findings

- 82% of teachers stated that FirstPass will be an effective AI service for supporting them as they undertake formative assessment.
- All teachers stated that a shared subject topic classifier was easier to train than doing the training on their own. When training subject topic classifiers on their own, 67% of teachers stated that their classifiers accurately labelled previously unseen text. For classifiers that were trained on a shared basis, this rose to 75%.
- 92% of learners stated that FirstPass correctly detected what they entered on the screen as they composed their responses to open-ended questions.

- 92% of learners said that the feedback from FirstPass helped them to compose better answers. All learners welcomed the immediacy of the feedback from the platform.

The FirstPass pilot study offers a compelling case for the adoption of AI-powered formative assessment tools. The platform's real-time feedback and crowdsourcing model show promise in addressing longstanding challenges in open-ended question assessment. While the study is not without limitations, its findings contribute to the ongoing conversation about the role of technology in education and the potential for AI to support teachers and learners. Overall, the FirstPass pilot study represents a significant step toward more efficient and effective formative assessment practices, with implications for educators and students alike. It paves the way for further exploration and development of AI-driven solutions in the field of education.

Introduction

Bolton College's FirstPass platform is designed to support students and teachers with the formative assessment of open-ended questions. The platform has several notable properties. The first is the platform's ability to deliver real-time feedback to students as they compose their responses to open-ended questions. Secondly, a crowdsourcing model is used to support the development and expansion of the knowledge base that underpins the delivery of all open-ended questions on the platform. That knowledge base grows as teachers offer labelled sentences to each subject topic classifier and as students respond and submit their work to their teachers. And thirdly, the platform's ability to label or annotate previously unseen text improves as additional subject specific text is offered to it.

What was the initial challenge or problem that needed to be addressed?

Bolton College identified the following challenges or problems regarding the assessment of open-ended questions.

The problem set for students

Here is a list of day-to-day problems that were encountered by students at Bolton College.

1. Students stated that there was a delay in receiving feedback from their teachers. Waiting times varied from a single day to over a week. If there was a significant delay it became harder to return to the question and amend the answer.
2. Teacher feedback is not always consistent, and it varies from teacher to teacher; and the type of feedback that is received by one student and another varies too.
3. When students compose responses to open-ended questions they do so with no real-time feedback, so they are not aware if their answers are adequate or appropriate.

The problem set for teachers

Here is a list of day-to-day problems that were encountered by teachers at Bolton College.

1. Teachers valued the use of open-ended questions as a device for conducting formative assessment with their students; but they stated that marking and providing feedback to student responses to open-ended questions was time consuming.
2. Whilst existing online services could be used to annotate student work or be used to deliver written or oral feedback to students these services could not undertake the same task on behalf of the teacher. Even though multiple-choice questions can be scored by optical readers and computers, the large number of students and the measurement based on human factors (bias, fatigue, inconsistency, and validity-reliability) make it difficult to conduct an accurate appraisal of information provided by open-ended questions.
3. With large-scale exams, the human effort that is required to evaluate student responses to open-ended questions increases significantly. Çınar et al (2020, p.3822) argue that the formative assessment of open-ended questions by computers allows teachers to evaluate and analyse open-ended responses in a much more detailed and objective manner, and in a shorter period than they can currently do.

There are two major strands to the FirstPass platform. The first is the crowdsourcing element which enables multiple teachers and subject specialists to train subject topic classifiers through a collaborative and participatory practice in different settings across the UK. The second is ability to offer real-time feedback to learners as they compose answers to open-ended questions which have been set by their teachers. The narrative of both strands runs throughout the length of this report.

The pilot study proposed two hypotheses. The first stated that the utilisation of crowdsourcing to train subject topic classifiers on the FirstPass platform will lead to the creation of higher-quality open-ended questions compared to questions generated solely by individual educators. And the second stated that the provision of real-time feedback to students through the FirstPass platform will lead to improved student performance in composing responses to open-ended questions compared to traditional feedback delivery methods with delayed feedback.

This hypothesis posits that the collaborative effort of multiple educators and subject specialists in crowdsourcing subject topic classifiers will result in the development of training data which can support the delivery of open-ended questions to students as part of the formative assessment process. Furthemore, well trained subject topic classifiers can enable the delivery of a comprehensive suite of open-ended questions. It also suggests that the diversity of perspectives and expertise brought by the crowd will support ethical AI practices.

The second hypothesis stated that the provision of real-time feedback to students through the FirstPass platform will lead to improved student performance in composing responses to open-ended questions compared to traditional feedback delivery methods with delayed feedback. This hypothesis suggests that the introduction of real-time feedback in the FirstPass platform will have a positive impact on student performance by allowing them to make immediate adjustments and improvements to their responses. It implies that the timely feedback will result in better answers from students, contributing to their learning and comprehension of the subjects they are studying.

Crowdsourcing

One of the defining traits of Bolton College's FirstPass platform is its use of crowdsourcing. This trait enables multiple teachers across multiple education settings and in different parts of the UK to set up and train the subject topic classifiers that are hosted. The rationale for adopting a crowdsourcing approach is wide ranging. Some of these were explored during the project. All teachers stated that they welcomed being part of a wider group of teachers who worked collaboratively to support formative assessment practices.

The nature of the FirstPass platform will mean that teachers in multiple institutions will be able to train subject topic classifiers that can be used to present open-ended questions to students across the UK. As more teachers and students engage with each classifier and question on the platform, the more accurate FirstPass becomes at supporting teachers and students at they complete their formative assessments, either as part of the learning or as in preparation for an end point assessment.

Literature Review

The following literature review was undertaken to identify and synthesise the current knowledge base for designing Education services built with Artificial Intelligence (AIED) through a collaborative and participatory model. The review also served to highlight any gaps in knowledge in this field of study.

AIED services often take shape through participatory and collaborative processes. Wang, et al (2012, p.10-11) go on to state that annotation projects, each require a different form of motivation to achieve the end goal of annotation. They highlight obvious motivations of the participant such as profit or altruism; but they state 'that the space of possible motivations and dimensions of crowdsourcing have not been fully explored.' In the case of the education sector the primary motives for participants are very likely to be the desire to improve the student experience, to support student wellbeing, to support students as they journey through their studies, to raise attainment levels and so forth. These motivations are well placed to support the participatory model that underpins the FirstPass pproject. Whilst Wang et al (2012, p.15) focus on the potential benefits, Ye and Kankanhalli (2017, p.103) have examined social exchange theory to review some of the more problematic areas of crowdsourcing platforms such as: the difference between an individual's actual participation and planned intention; the costs and benefits of participating; and the effect of trust on participation. Their work has highlighted that when individuals participate in an exchange they do so when they expect the rewards to exceed costs. Individuals tend to maximise benefits and minimise costs from exchanges. The qualitative feedback from the teachers who participated in the pilot gave strong suppprt to using a crowdsourcing model. Teachers stated that their active involment in the training of the subject topic classifiers gave them a say in the design and use of a new and emerging AIED service. They also statated that their voices were represented in the training models on the FirstPass platform.

Whilst Ye and Kankanhalli (2017, p. 102) state that crowdsourcing models can be very productive, they do caution project teams before they embark on asking people to participate. One of the first concerns that they have is sustaining involvement by individuals on a crowdsourcing platform. However, since crowdsourcing platforms are entirely dependent on the voluntary behaviour of their members, they must understand what

motivates them to become and remain active throughout a variety of tasks. An important distinction between platforms that thrive and those that fail is the level of ongoing activity of their members (Boons, Stam and Barkema, 2015, p. 720).

A participatory process is important when crowdsourcing labelled data; however, Zheng et al (2019, p.867) state that project teams need to guard against deteriorating model accuracy as a wider pool of individuals are involved in the annotation process. But they go onto argue that repeated labelling by the crowd leads to an improvement in model accuracy, especially regarding label inference. Label inference takes place when a natural language processing or classification model has acquired a sufficient volume of annotated data that allows the model to infer a label to unseen unlabelled text. The means of production for new and emerging AIED services are dependent on leveraging the strengths that are inherent in large numbers of people who fulfil tasks that were previously carried out by the few. The participatory model lends itself particularly well to the education sector where the larger group are motivated by shared goals (Eskenazi et al 2013 p.316-17).

Virtuality lends itself well to the crowdsourcing element of the FirstPass platform. Virtuality has been defined in a variety of ways. implies distance between collaborators, requiring asynchronous communication. To some, computers can display information differently from the way it appears in reality. In recent years a variety of tools have made complex information more easily accessible to a wide range of people. According to this concept, an organisation or community should have a flexible structure that can adapt to changing circumstances to optimise individual and collective performance. For others, the term also encourages knowledge and expertise to be shared. The crowdsourcing element of FirstPass gets a boost from Virtuality. Having a virtual space means that teachers can contribute and collaborate without the limitations of geographical boundaries. Teachers commented favourably on this, saying that it gave them the opportunity to contribute positively to an AIED platform.

Virtual communities are communities of people who actively participate in, or are associated with, activities that are centered around a particular interest. A community support system, whether it be distributed or localised, is an essential element to implementing a continuous meta-improvement strategy. There are two types of virtual communities: professional

societies and virtual educational communities. There are professional associations that enable their members to better understand their fields and improve their performance of duties related to society or community. There is a growing trend towards virtual educational communities which offer an interactive environment through mutual sharing and action learning, going beyond the typical push nature of distance learning. Among professional societies and educational institutions there exists a mutuality of interests and an existing organisation infrastructure which may be used to facilitate the maintenance and development of these environments (Bieber et al, 2002, pp. 11-35).

With regards to the FirstPass platform the cooperative and collaborative traits of crowdsourcing allow subject specialist teachers to set up and train subject topic classifiers that can be shared by many teachers across the UK and further afield. And as students engage with the open-ended questions on the platform, and if their answers are quality assured, their responses are added to the training dataset of each subject topic classifier that underpins each respective open-ended question. This in turn furthers the reliability of the FirstPass platform to support students and teachers with the formative assessment process.

The emergence of new education technology and the services that underpin them have come about through the acquisition of new knowledge and through a participatory model. Sfard (1998, p.5) argues that both are required to stimulate learning at an institutional level within an education setting. Acquisition focuses on stimulating the thinking capacity of individuals to develop new knowledge, and participation identifies new knowledge as a product of cultural practices and shared learning processes (Owusu-Agyeman, 2019, p.695).

AIED services often take shape through participatory and collaborative processes. Wang, et al (2012) state that: It is an accepted tradition in natural language processing (NLP) to use annotated corpora to train machine-learned models for common tasks such as machine translation, parsing and summarisation. Since machine-learned system performance is dependent on the quality of the input annotation, much work in annotation has centred on defining high quality standards that were reliable and reproducible, and finding appropriately trained personnel to carry out such tasks (p.12).

Impact of Real-Time Feedback

One of the key merits of the FirstPass platform is its ability to offer real-time feedback to learners at the point of action. Learners do not have to face a delay in receiving feedback. This enables learners to reflect on, adapt and improve their answers to open-ended questions before they submit them for final review and commentary from their teachers. Previously, we have found that 93.8% of learners stated that they were able to submit a better answer with the real-time feedback that they got from the FirstPass platform.

FirstPass offers learners real-time textual and graphical feedback. This offers learners the opportunity to reflect, edit and improve upon their answers before submitting them to their teachers. On receipt of the student work, teachers can add final commentary and feedback. The ability to offer learners effective real-time feedback from FirstPass reduces the time that learners would typically have to wait before they received feedback from their teachers.

Real-time feedback has been received positively by teachers and learners. Prior to the current AIF pilot, qualitative feedback was received from learners and teachers at Bolton College. Learners stated that the real-time feedback that was offered to them enabled them to identify what they had covered and were further work was required to enhance their response to the open-ended questions that they had received from their teachers. Learners also welcomed the immediacy of the feedback that they received from the platform. Teachers stated the FirstPass platform allowed for improved submissions from their learners and that they spent less time giving remedial feedback to their learners. Learners welcomed real-time feedback because it allowed them to garner a better understanding of their work; particularly in relation to the strengths of their work and in identifying areas where they could make improvements. Immediate reinforcement enabled learners to amend their responses whilst remaining engaged with the task. From the teacher's perspective, learner feedback demonstrated enhanced metacognition.

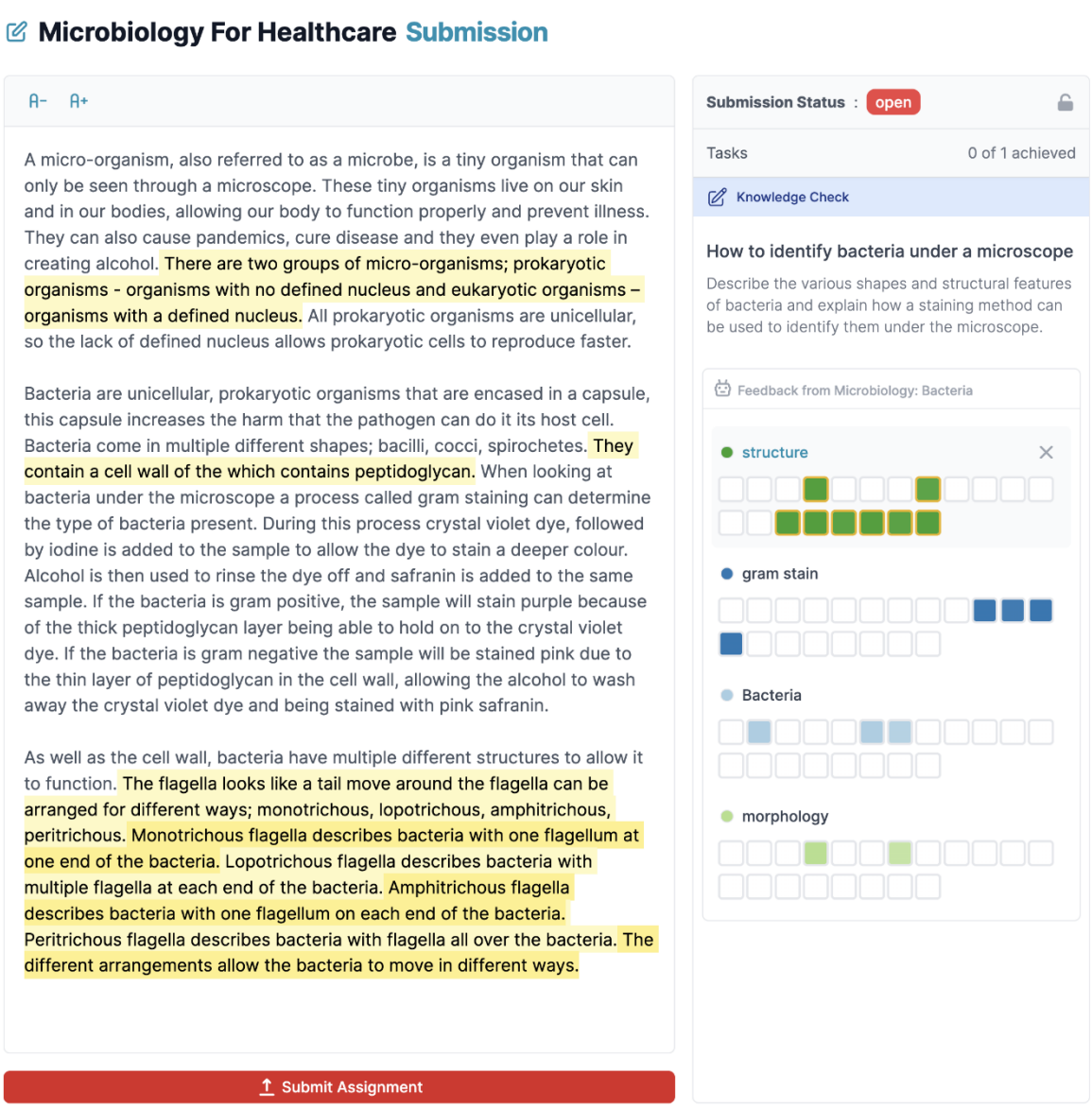


Figure 1: learner real-time feedback screenshot from FirstPass.

A small-scale pilot with 110 students who participated in the initial trial of the FirstPass service was completed at Bolton College. All students welcomed the real-time feedback that they received from the platform. Level 1 Business students played particular attention to the feedback element of the service. The teacher for this group stated that they were new to the topic and that could explain why they played closer attention to the feedback panel. Students in the trial stated that the feedback panel was useful because they knew what was done and what needed to be completed in their answer. Level 3 students referred to the feedback panel when they had completed typing their free form text answer. They then adjusted or added to their answer when they knew what topics they needed to write more about in their answer.

The Head of Learner Support at the College who was invited to review FirstPass as part of the intial trial group stated that the platform would be particularly useful for Autistic learners who find it difficult to absorb feedback from their teachers and support workers. The ability for FirstPass to provide real-time feedback as well as hints was welcomed for this user group. Further tests will need to be conducted to assess the validity of this statement.

Teachers and learners have welcomed the feedback that they receive from the FirstPass platform, but they have also stated that the teacher should offer final commentary on submitted work.

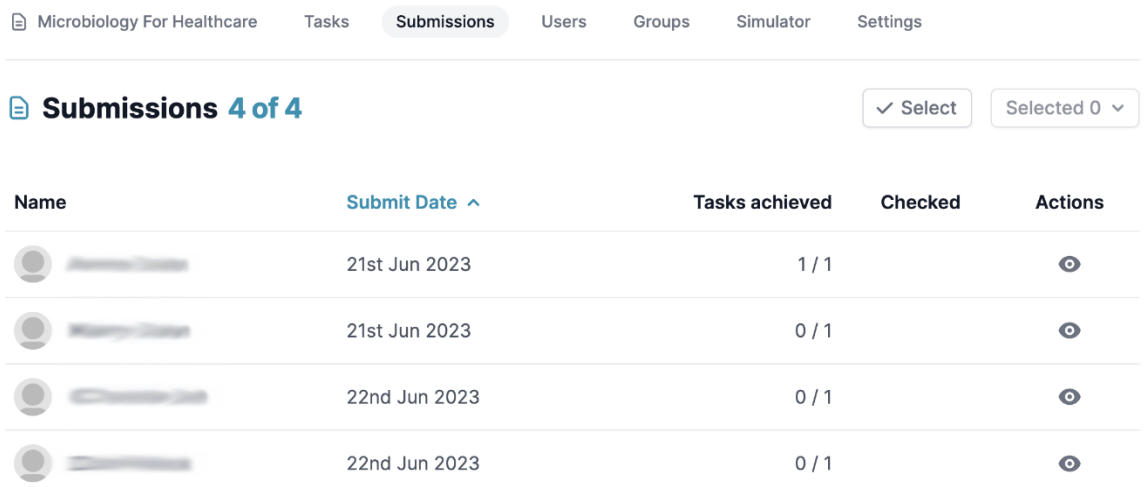


Figure 2: Assignment submissions from learners on FirstPass

For teachers, it removes the burden of marking a high volume of scripts over a short period of time; and for students they garner real-time feedback at a place, time and pace that suits their needs. In the case of FirstPass one simply needs to type a response to an open-ended question to showcase how FirstPass offers real-time feedback to a student. This in turn leads teachers to discuss how the platform could be used to support the delivery of effective feedback to the student.

The FirstPass platform is used as a tool to inform student writing. Teachers cite FirstPass' ability to provide real-time effective feedback to their students as one of the most compelling reasons for using the service to support the formative assessment process; especially if it enables students to submit work that is more comprehensive in answering open-ended questions. The ability for FirstPass to structure, organise, label and process free-form text is welcomed by teachers who have a high workload around formative assessment (Martins et al, 2019, p.183).

A benefit of open-ended questions is that they do not impose boundaries on respondents' responses. Open-ended questions may also help the teacher gain a deeper understanding of the learner (Schonlau et al, 2019, p.563). Readers can interact with texts more effectively by asking open-ended questions. They may allow comprehension to be assessed and when coupled with real-time feedback, open-ended questions may be more effective than using closed questions. The use of open-ended questions, coupled with real-time feedback, may provide an enjoyable method of externalising comprehension monitoring. However, until now, students have not been able to receive real-time feedback on their responses to open-ended questions (Smith et al, 2020, p. 1228).

Methodology

Within the framework of the FirstPass pilot, two primary components play a pivotal role in the context of this study. These components represent the fundamental aspects of our research methodology, as they are central to the examination of the platform's functionality and impact. This section describes the interplay of these components within our research framework.

Crowdsourcing

The first component of the FirstPass platform under investigation pertains to its crowdsourcing element. This facet of the platform operates as a collaborative and

participatory mechanism involving multiple teachers and subject matter experts. These individuals collectively engage in the process of training subject topic classifiers across diverse educational settings throughout the United Kingdom. The involvement of this diverse pool of teachers and subject specialists is integral to our research methodology, and their contributions are central to the insights derived in this study.

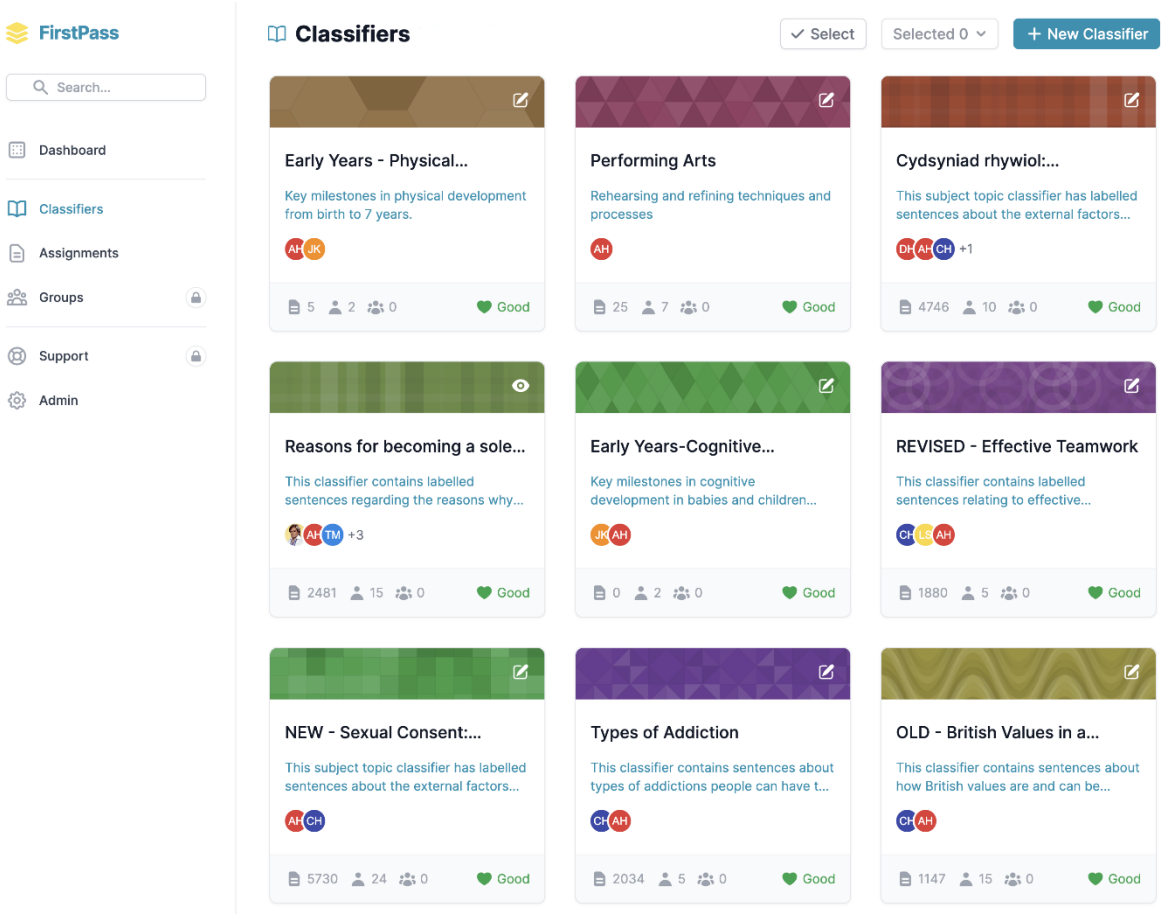


Figure 3: Subject Topic Classifiers in First Pass

Impact of Real-Time Feedback

The second aspect of the FirstPass platform that forms the focal point of our research methodology is its capacity to provide real-time feedback to learners. Effective feedback is designed to support learners as they craft responses to open-ended questions posed by their respective educators. Our study delves into the intricacies of how this real-time feedback feature operates within the educational context, shedding light on its effectiveness and implications.

Throughout the entirety of this research report, both the crowdsourcing and real-time feedback strands, remain inextricably intertwined. Their narrative serves as a foundational element that guides the exploration and analysis presented herein, offering a comprehensive understanding of the FirstPass platform and its impact on the formative assessment landscape. This dual-pronged approach within the FirstPass platform forms the backbone of our research methodology, and our subsequent analysis and findings are informed by the multifaceted interactions and outcomes stemming from these two key components.

Objectives

The objectives for the FirstPass pilot were designed to be broad to assess the general efficacy of the platform to support practices relating to the formative assessment of open-ended questions:

2. To test the functionality and usability of the platform with students and teachers in multiple educational institutions.
3. To identify any technical or operational issues that need to be addressed.
4. To evaluate the platform's effectiveness in supporting the formative assessment of open-ended questions within the context of the pilot.
5. To gather feedback from participating colleges and students to inform the future development of the platform.

These objectives were assessed through several tests that were undertaken during the pilot. These are listed as follows:

- **Test 1**: Individual teachers set up and train their own subject topic classifier.
- **Test 2:** Course teams share the training of one of their subject topic classifiers with a like-for-like course in another college or colleges in the pilot.
- **Test 3:** Course teams set up one open-ended question for their students which is based on a classifier that has been trained during tests 1 and 2. Students in each course answer the open-ended question.

- **Test 4:** Course teams set up one open-ended question which uses an alternative feedback mechanism. Students in each course answer the open-ended question.

Qualitative and quantitative feedback was received from students and teachers at the end of each test or activity.

Sample

Bolton College issued a call for Expressions of Interest (EOI) to further education (FE) colleges in the UK. Following on from the EOI, the following colleges signed up to take part in the FirstPass pilot: Bridgend College, Coleg Cambria, Glasgow Kelvin College, Sunderland College, Leeds City College, and Bolton College as the host and lead college. A total of 65 teachers and 263 learners participated in the pilot study.

Data collection

The following activities were planned, and quantitative and qualitative data collection was undertaken at the end of each activity or test. Upon completing each test, all participants were asked to complete an online survey.

TEST	DESCRIPTION
1	Teachers individually set up and train subject topic classifiers.
2	Teachers train shared subject topic classifiers.
3	Learners respond to an open-ended question.
4	Teachers offer and learners receive different modes of feedback.

Views and opinions were also sought from all students and teachers in the form of two online open-ended conversations which were held prior to the first test, after the second test and following on from the final activity of the pilot. Additional data such as the number of participating students and teachers, the number of subject topic classifiers, and the size of the training dataset for each classifier was garnered from the FirstPass platform. The hours logged by data labellers was gathered through timesheet data. In addition, all participating

colleges agreed and signed a data privacy agreement which detailed the collection, processing and deletion of student and staff data.

Data analysis

As the sample size was small, descriptive statistics was used to summarise the data that was collected at the end of each test which enabled the project team to identify patterns, and to make interpretations. These descriptive statistics provided a quick snapshot which allowed for easy comparisons and insights to be made.

In. order to assess the general viability of the FirstPass platform to support formative assessment practices across different vocational areas, teachers across all colleges were given the scope to setup and train their own choice of subject topic classifiers for their chosen specialist subject areas, and they also had the freedom to create their own formative assessment task for their learners.

Findings

This section details the key findings of the research pilot for the two key strands of the project, namely crowdsourcing and the impact of real-time feedback on learning within the formative assessment process.

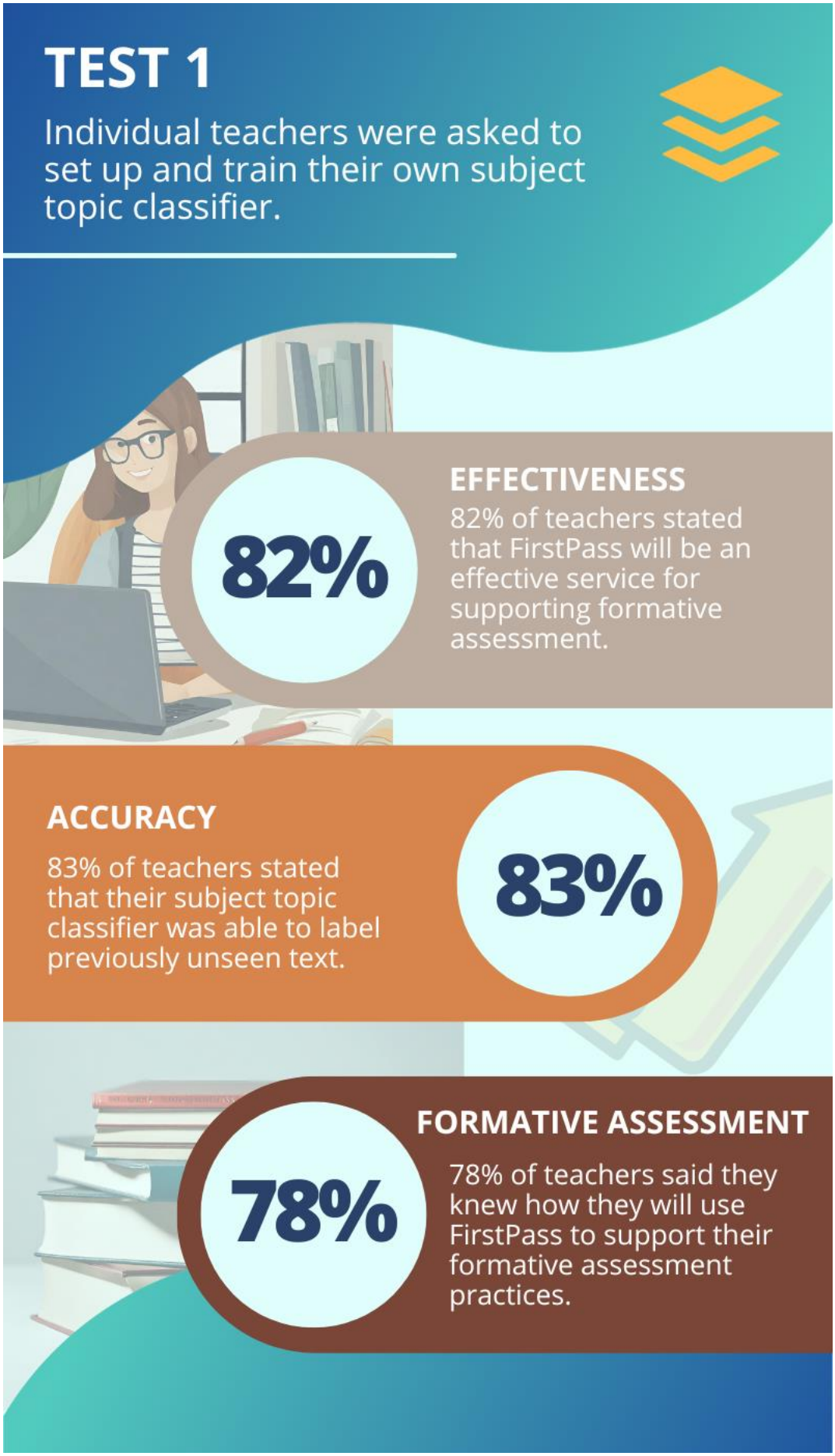
Crowdsourcing

This section details the key findings for Tests 1 and 2 which focused on the training of subject topic classifiers, and in particular, the platform's ability to support the crowdsourcing of subject topic classification.

Key Findings for Test 1

As part of Test 1 individual teachers were asked to set up and train their own subject topic classifier.

The FirstPass platform was found to be easy to use and the user interface was intuitive to use


TEST 1
Individual teachers were asked to set up and train their own subject topic classifier.
82%
EFFECTIVENESS
82% of teachers stated that FirstPass will be an effective service for supporting formative assessment.
ACCURACY
83% of teachers stated that their subject topic classifier was able to label previously unseen text.
83%
78%
FORMATIVE ASSESSMENT
78% of teachers said they knew how they will use FirstPass to support their formative assessment practices.

(62% of respondents, 29% neutral). The overall message conveyed by teachers in the surveys for the initial two tests in the pilot about FirstPass is that they would like to see additional tools and features that would make the platform more useful and user-friendly for training subject topic classifiers. They are looking for ways to make the process of training more efficient, effective, and accessible. They would like to see more customisation options, a better way to organise training on the subject topic classifiers, and improved accessibility. These insights suggest that teachers are looking for a platform that is flexible, easy to use, and able to adapt to their specific needs and requirements for training subject topic classifiers.

At the time of writing the final report, 51 subject topic classifiers have been set up on the FirstPass platform. In FirstPass, a subject topic classifier is a store of labelled sentences about a given subject topic. At the heart of a subject topic classifier is an algorithm that is used to assign labels to previously unseen sentences. The following tables highlights the subject areas and the topics for which classifiers have been set up.

Subjects	Classifiers	Subjects	Classifiers
Pastoral Tutorials	Types of Addiction. Signs and Indicators of Abuse. Sexual Consent: External Influences. British Values in a Further Education College. 5 ways to Mental Wellbeing at a Further Education College. Types of Gambling Activities.	Teaching	Safeguarding Legislation. Contributing to Effective Teamwork. Characteristics of Effective Teamwork in Schools. Characteristics of 'school readiness'.
Performing Arts	Roles and Responsibilities of a Theatrical Agent (Actors). Responsibilities of an actor in the rehearsal of a text-based play.	Early Years	Role of the early years practitioner. Reasons why some children are not immunised. Supporting children's socialisation within play environments. The role of the Key Person in promoting emotional well-being. Signs and Indicators of Abuse.

Subjects	Classifiers	Subjects	Classifiers
			Importance of a healthy balanced diet for babies and young children. Impact a of poor diet on babies & children's health, development and well-being.
Animal Studies	Respiratory system. Describe how mammalian lungs are ventilated. Zoochosis and Stereotypical Behaviour.	Business	Reasons for becoming a sole trader. Fixed and Variable Costs. Disadvantages of Being a Sole Trader. Advantages and Disadvantages of setting up as a Franchisee. Disadvantages of being a Private Limited Company. Industrial Sectors. Types of Taxes.
Science	Homeostasis. Enzymes. Energy Changes. DNA. Cell Structure. Ionic Bonding. Microbiology. Microbiology: Bacteria. Plant Cells.	Computing	Central Processing Units.
Welsh	Cydsyniad Rhywiol: Dylanwadau Allanol (Sexual Consent: External Influences).	Work Experience	Work Experience.

Most teachers stated that their subject topic classifiers did not take too long to train. However, teachers stated that they required more time to train a subject classifier so that it could classify text with a greater degree of accuracy. Future projects may benefit from the support of subject specialist data labellers if the number of subject topic classifiers is going to scale to support teachers across a wide range of contexts for their formative assessment practices.

It was encouraging to see teachers expressing how accurate subject topic classifiers were for accurately labelling previously unseen text (66% of respondents, 29% neutral). However, it is

always possible to refine the training data within a classifier to improve its performance at classifying text.

Summary of Test 1 findings: Individual teachers set up and train their own subject topic classifier.

82% of teachers stated that FirstPass will be an effective AI service for supporting them as they undertake formative assessment, and 74% of teachers stated that they could articulate how the platform the platform could be used for that purpose. 83% of teachers stated that their subject topic classifier was able to label previously unseen text, with the remaining teachers stating that their classifiers needed additional training data before text classification could be undertaken with a high degree of accuracy. When making general comments about the training of subject topic classifiers, a noticeable critique was that classification training was a time-consuming process.

Key Findings for Test 2

For Test 2 course teams shared the training of one of their subject topic classifiers with a like-for-like course in one or more colleges taking part in the pilot.

Whilst most participants of the pilot felt that their choice data labels was manageable, 22% of participants indicated that it was a difficult task to manage. Their insights included the need to edit label names, the ability to delete them or to merge them. This feedback will be incorporated into subsequent versions of the FirstPass platform.

Teachers and data labellers have stated that they do get a writer's block in imagining new ways to phrase text. This can be addressed by using external services such as text generation services such as Bard, Claude, ChatGPT, WordTune. If these services can be used in an appropriate manner, the project team envisages that these tools will be developed and incorporated into the FirstPass platform.

TEST 2

Course teams were asked to share the training of one of their subject topic classifiers with teachers in another college.

100%

AI MODEL TRAINING

100% of teachers stated that shared subject topic classifiers were easier to train.

PRE-TRAINED CLASSIFIERS

50% of teachers would like a library of pre-trained subject topic classifiers.

50%

100%

MODERATOR ROLE

100% of teachers said that a moderator role is required to quality assure the training and performance of subject topic classifiers.

Summary of Test 2 findings: Course teams share the training of one of their subject topic classifiers with a like-for-like course in another college or colleges in the pilot.

All teachers stated that a shared subject topic classifier was easier to train than doing the training on their own. When training subject topic classifiers on their own, 67% of teachers stated that their classifiers accurately labelled previously unseen text. For classifiers that were trained on a shared basis, this rose to 75%.

50% of teachers stated that it is going to be difficult to moderate another teacher's labelled sentences when those teachers are in different colleges. However, all teachers felt that a moderator role is required to quality assure the training and performance of a subject topic classifier; especially as more teachers are involved. The work of Zheng et al (2019) reiterates this point.

General Findings for Tests 1 and 2: Crowdsourcing

During the first two tests in the pilot, the FirstPass platform demonstrated its ability in supporting teachers and subject specialist data labellers to set up subject topic classifiers that were able to assign correct labels to previously unseen text. Teachers and subject specialist data labellers understood that a classifier's performance is dependent on the volume and quality of training data. Teachers stated that whilst the subject topic classifiers were easy to train, they recognised that support from additional teachers and subject specialist data labellers was required to expand the volume of training data for each of their subject topic classifiers.

The pilot demonstrated the platform's ability to correctly assign labels to subject topic text. As the platform develops, the project team will continue to refine and improve upon the features that support text classification.

The overall message conveyed by teachers in the surveys for the initial two tests in the pilot about FirstPass is that they would like to see additional tools and features that would make

the platform more useful and user-friendly for training subject topic classifiers. They are looking for ways to make the process of training more efficient, effective, and accessible. They would like to see more customisation options, a better way to organise training on the subject topic classifiers, and improved accessibility.

Overall, the suggestions provide valuable information regarding enhancements to the FirstPass platform which can be categorised as features and tools. With regard to features, teachers and data labellers asked for enhancements to be made to accommodate for accessibility requirements such as the ability to alter font sizes and colours. They also requested that AI generative text services should be integrated into the FirstPass platform to ease and speed up the training of subject topic classifiers. For example, if a teacher or a data labeller entered a labelled sentence, they could then trigger the automatic creation of additional sentences for the classifier. With regard to tools, teachers and data labellers asked for the following to be added to the FirstPass platform: display which parts of a sentence the platform labels; the ability to edit, delete or merge labels; and improved search options.

Whilst most participants of the pilot felt that their choice of data labels was manageable, 22% of participants indicated that it was a difficult task to manage. Their insights included the need to edit label names, the ability to delete them or to merge them. This feedback will be incorporated into subsequent versions of the FirstPass platform.

Teachers made use of student submissions to train subject topic classifiers. This was encouraging to see because student submissions will be added to the training data set for each subject topic classifier.

Initially, 23% of participants stated that they found it hard to picture how FirstPass will be used to support their formative assessment practices. As awareness of the platform's functionality and capabilities grew, this percentage fell to zero at the end of Tests 3 and 4. Feedback from participating colleges and students was overwhelmingly positive, with most participants reporting that they would recommend the platform to others.

Impact of Real-Time Feedback

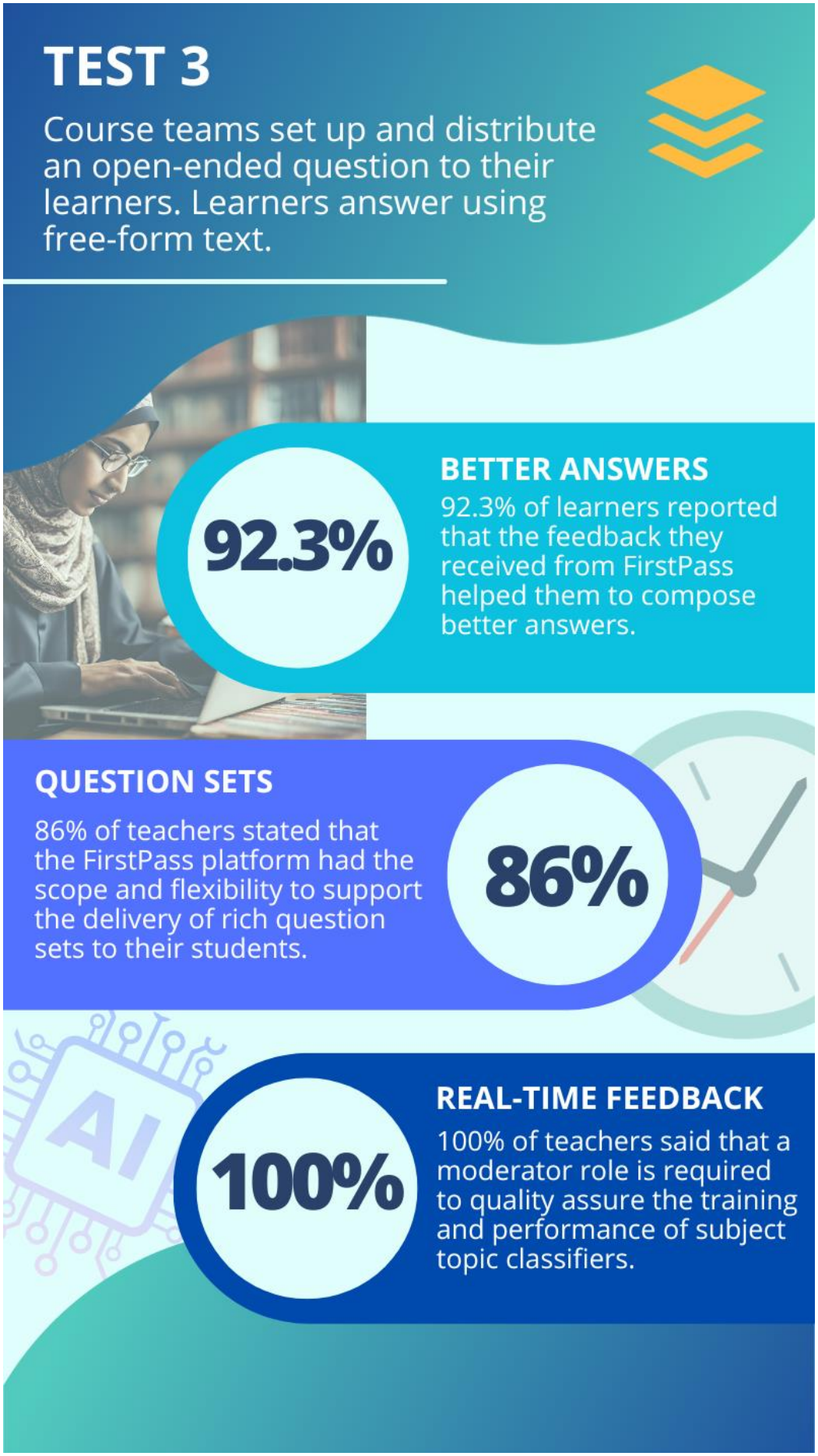
This section details the key findings for Tests 3 and 4 which focused on the real-time feedback that was provided top learners as they responded to open-ended questions on the FirstPass platform.

Key Findings for Test 3

During Test 3 course teams set up one open-ended question for their students which is based on a classifier that has been trained during tests 1 and 2. Students in each course answer one open-ended question.

100% of teachers had no difficulties in setting up an open-ended question on the FirstPass platform. 86% of teachers stated that the assignment simulator allowed them to test the assignment brief before it was released to their students, and it gave teachers the opportunity to engage with the platform as a student. All teachers expressed the opinion that more training data is better than less. During this test, they saw the connection between classification training and the assignment that they were about to present to their students.

71% of teachers stated that the platform allowed them to curate open-ended questions that were wide ranging. Some teachers acknowledged the flexibility but noted that setting up subject classifiers, particularly for outcome-driven qualifications, might require significant effort due to the need for a range of answers for each criterion. Others found it relatively easy to add and edit multiple questions, while one respondent suggested the importance of distinguishing between different question types. Some participants felt that the page currently provided such flexibility, while one respondent expressed the need for more guidance for tutors to fill out key information and requirements like duration, deadline dates, and response length. Additionally, one participant pointed out that the assignment set-up page seemed more suitable for longer narrative writing tasks but still found it easy to add questions and link them to different labels on the classifier.


TEST 3
Course teams set up and distribute an open-ended question to their learners. Learners answer using free-form text.
92.3%
BETTER ANSWERS
92.3% of learners reported that the feedback they received from FirstPass helped them to compose better answers.
QUESTION SETS
86% of teachers stated that the FirstPass platform had the scope and flexibility to support the delivery of rich question sets to their students.
86%
AI
100%
REAL-TIME FEEDBACK
100% of teachers said that a moderator role is required to quality assure the training and performance of subject topic classifiers.

Student Feedback on answering open-ended questions on the FirstPass platform

92.4% of learners stated that FirstPass correctly detected what they entered on the screen as they composed their responses to open-ended questions. 92.3% of learners said that the feedback from FirstPass helped them to compose better answers. All learners welcomed the immediacy of the feedback from the platform.

Learners were asked to indicate what other types of feedback they wanted from FirstPass. These are listed below.

Type of feedback	
Video feedback from teacher	20%
Badges, medals, and emoticons	30%
Audio feedback from teacher	30%
Feedback summary from a LLM	40%
A point score	40%
A progress bar	80%
Text summary from the teacher	90%
Spelling and grammar checker	90%
Hints	90%

75% of students stated that the FirstPass platform delivered feedback that was on par with the feedback that they received from their teachers; especially regarding omissions or where additional information was required in their work. Effective feedback at the point of action was particularly referenced by students as they used the FirstPass platform, with all students commenting on the benefits of receiving feedback straight away. Students stated that they could amend and improve their work straight away which was very different from making changes to work after a few days or more before they traditionally received feedback from their teachers.

Summary of Test 3 findings: Course teams set up one open-ended question for their students which is based on a classifier that has been trained during tests 1 and 2. Students in each course answer the open-ended question.

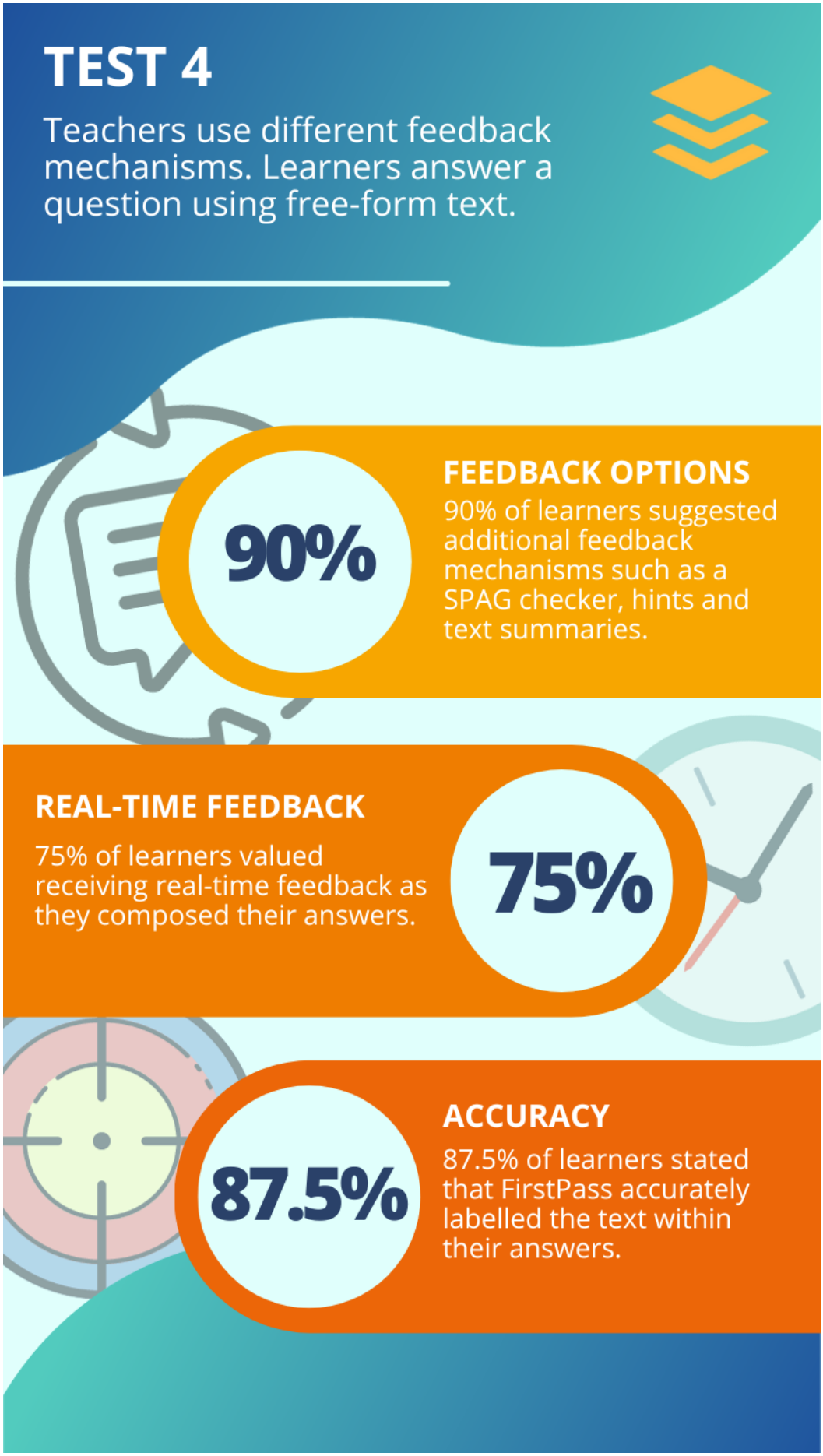
86% of teachers stated that their subject topic classifiers needed to have more training data if they were going to be used to underpin the delivery of open-ended questions. This was reiterated by 43% of teachers saying that they did not have enough labelled sentences (training data) for the assignment simulator to provide accurate feedback when they tested their assignments before they were presented to learners. 86% of teachers stated that the FirstPass platform had the scope and flexibility to support the delivery of rich question sets to their students.

92% of learners stated that FirstPass correctly detected what they entered on the screen as they composed their responses to open-ended questions. 92% of learners said that the feedback from FirstPass helped them to compose better answers. All learners welcomed the immediacy of the feedback from the platform.

Key Findings for Test 4

Test 4 entailed course teams set up one open-ended question which uses an alternative feedback mechanism. Students in each course answer the open-ended question.

The findings regarding the use of FirstPass to deliver real-time feedback to support formative assessment underscore its potential as a valuable tool in enhancing the educational experience. 75% of students welcomed the opportunity of receiving real-time feedback which indicates that FirstPass effectively addresses a crucial need in the formative assessment process. This real-time feedback mechanism not only meets students' expectations but also significantly contributes to their ability to produce improved answers, as acknowledged by 92.3% of learners. The desire for specific feedback mechanisms, such as text summaries, spelling and grammar checks, and hints, by 90% of students reflects their engagement with


TEST 4
Teachers use different feedback mechanisms. Learners answer a question using free-form text.
90%
FEEDBACK OPTIONS
90% of learners suggested additional feedback mechanisms such as a SPAG checker, hints and text summaries.
REAL-TIME FEEDBACK
75% of learners valued receiving real-time feedback as they composed their answers.
75%
87.5%
ACCURACY
87.5% of learners stated that FirstPass accurately labelled the text within their answers.

the platform and the importance of tailored feedback to support their writing and learning. Therefore, these findings suggest that FirstPass has the potential to play a pivotal role in enhancing formative assessment practices by providing timely and personalised feedback, which in turn shapes and improves student performance and engagement in the learning process.

During an open exchange of ideas with teachers they stated that providing diverse types of formative assessment feedback on FirstPass platform offers them the opportunity to tailor guidance to individual student needs, employing targeted support strategies, and enhancing the efficiency and timeliness of feedback delivery. This approach would lead to improved learning outcomes, increased student engagement, and valuable data for data-driven instruction. Teachers stated that the availability of multiple feedback mechanisms would enhance the overall quality of feedback and it would promote a more engaging and effective learning experience for their students.

Test 4: Course teams set up one open-ended question which uses an alternative feedback mechanism. Students in each course answer the open-ended question.

Summary of findings

75% of learners stated that they valued receiving real-time feedback from the FirstPass platform. When asked about the different feedback mechanisms that they wished to receive, 90% of learners asked for a text summary from teachers, a spelling and grammar checker and they asked for hints to be displayed to support their writing.

General Findings for Tests 3 and 4: Impact of Real-Time Feedback

The findings from Tests 3 and 4 on the impact of real-time feedback through the FirstPass platform provide valuable insights into the dynamics of formative assessment practices. Test 3 revealed a high degree of teacher satisfaction with the platform, with 100% of teachers reporting no difficulties in setting up open-ended questions. Additionally, the positive feedback that was received about the assignment simulator's role in allowing teachers to test

assignment briefs before student distribution indicates that teachers value the opportunity to engage with the platform as both teachers and students.

One of the noticeable findings in Test 3 was the repeated emphasis on the importance of training data for subject topic classifiers. Teachers expressed a desire for more training data, recognising the important connection between data quality and the accuracy of feedback and classification. This underscores the critical role that data plays in machine learning-based educational platforms like FirstPass, particularly with regards to formative assessment practice. Adequate training data is also seen vital component for ensuring that the platform can effectively support both teachers and students during formative assessment activities.

Student feedback on FirstPass was overwhelmingly positive, with 92.4% of learners stating that the platform correctly detected their responses and 92.3% acknowledging that the feedback from FirstPass helped them compose better answers. What's noteworthy is that students appreciated the immediacy of the feedback. They valued the ability to receive guidance and make improvements promptly, which stands in contrast to traditional feedback methods that often involve longer waiting times. This aligns with the current educational strategies that prioritise timely interventions and actionable feedback to enhance student engagement and performance.

Moreover, the specific feedback mechanisms requested by students, including text summaries, spelling and grammar checks, and hints, highlight their desire for comprehensive support in their learning journey. These requests reflect the recognition that students benefit from diverse forms of feedback tailored to their individual needs. It underscores the importance of personalisation and adaptability in educational technology.

DISCUSSION

This section of the report returns to the two hypotheses that were made prior to the start of the pilot. The first stated that the utilisation of crowdsourcing to train subject topic classifiers on the FirstPass platform will lead to the creation of higher-quality open-ended questions compared to questions generated solely by individual educators. And the second stated that

the provision of real-time feedback to students through the FirstPass platform will lead to improved student performance in composing responses to open-ended questions compared to traditional feedback delivery methods with delayed feedback.

The pilot demonstrated that it is possible for teachers and other subject specialist teachers in different education settings could work in a co-ordinated manner to train AI models which underpin formative assessment practices. Key findings highlight the positive impact of real-time feedback on both students and teachers. Nearly 94% of learners reported submitting improved answers with the aid of real-time feedback. This indicates that students benefited from real-time effective via the FirstPass platform as they composed their free-form responses to open-ended questions. And in broader terms, the results of the FirstPass pilot study show that the computer mediation of open-ended questions can be conducted as part of formative assessment practice in education settings.

In the broader context of educational research, the FirstPass pilot represents a significant step toward addressing the long-standing challenges of conducting the formative assessment of open-ended questions. In addition, the use of crowdsourcing to train subject topic classifiers and the ability to offer real-time feedback to students sets FirstPass apart as an innovative tool for enhancing formative assessment practices.

Feedback can be offered to learners in a variety of forms as they compose their responses to open-ended questions. These include marks, automated textual summaries that detail what has been completed and yet to be completed, graphical and textual feedback about what topics have been covered in the body of an answer, word counts, sentence complexity and SPAG. The feedback mechanisms will be refined and improved upon as the platform develops during the next stage of development.

Acknowledging the strengths and limitations of the study is crucial. The pilot study provides valuable insights, but it's important to consider that the sample size was relatively small. As such, while the findings are promising, further research with larger and more diverse samples would be valuable to validate the platform's effectiveness across a wider range of educational settings. The lessons learnt during the pilot highlighted the need for further work to be

undertaken to reduce some of pinch points associated with model training. And the pilot identified the need for additional toolsets to support collaboration between teachers and data labellers. An ongoing dialogue with teachers and students will inform the continued development, improvement and delivery of real-time feedback to students. All this sets the stage for future research directions.

The pilot study suggests several avenues for future research. One essential area is the scalability of the FirstPass platform. Can it effectively support formative assessment in larger educational institutions and across diverse subject areas? Additionally, further investigation into the platform's impact on specific student groups, such as autistic learners, could yield valuable insights.

The practical implications of the FirstPass platform are significant. It has the potential to reduce the time and effort required for teachers to assess open-ended responses, thus improving efficiency. Moreover, the platform's real-time feedback empowers students to enhance the quality of their work. In a theoretical context, FirstPass contributes to the evolving field of AI in education, demonstrating the feasibility and benefits of using AI-powered tools for formative assessment. Bringing the findings together, the FirstPass pilot study underscores the potential of AI-driven solutions in education. It highlights the importance of real-time feedback and collaborative crowdsourcing in improving formative assessment practices. This synthesis emphasises the transformative role that technology can play in enhancing both the teaching and learning experience.

During the FirstPass pilot, it became evident that the choice of subjects and topics was a pivotal factor in shaping the study's execution. The broad spectrum of subjects initially chosen by pilot colleges posed challenges in terms of managing and assessing the effectiveness of the platform across diverse educational settings. In retrospect, it is evident that a more focused approach, where multiple colleges train, test, and use the same set of subject classifiers, could offer enhanced manageability and comparability of results. This lesson underscores the importance of refining the scope of subjects in future implementations of the FirstPass platform, aligning with the necessity for a structured and controlled research environment.

The data labelling process emerged as a critical component in preparing subject topic classifiers for the FirstPass platform. Insights gathered from 64 teachers involved in the pilot underscored the importance of structured data labelling activities. Lessons learned include the need for meticulous planning of the data labelling process, the provision of user-friendly data labelling tools, and the definition of clear and consistent labelling criteria. These lessons emphasise the significance of refining the data labelling process to enhance the quality and efficiency of classifier preparation.

Quality control and training of data labellers were identified as pivotal aspects of the data labelling process. The FirstPass pilot shed light on the importance of having a quality control mechanism in place, which involves multiple labellers, label comparison, and resolution of discrepancies. Additionally, the need for providing adequate training to labellers became evident to ensure their understanding of labelling criteria and consistency in their work. These lessons highlight the significance of investing in quality control measures and comprehensive training programs to maintain data accuracy and reliability.

The FirstPass pilot emphasised the value of incorporating feedback and iteration into the data labelling process. Insights from teachers and stakeholders indicated that continuous feedback loops allowed for process improvement and enhanced efficiency. Lessons learned from this experience underscore the importance of an adaptable approach that welcomes feedback from participants, promoting ongoing refinement and optimisation of data labelling procedures.

A critical lesson drawn from the pilot was the need to be vigilant about potential sources of bias in the data labelling process. The FirstPass pilot revealed that addressing bias requires careful consideration, such as involving multiple labellers and conducting regular reviews for label consistency. This lesson underscores the importance of proactive measures to ensure unbiased data labelling practices.

In summary, the FirstPass pilot study offers a compelling case for the adoption of AI-powered formative assessment tools. The platform's real-time feedback and crowdsourcing model

show promise in addressing longstanding challenges in open-ended question assessment. While the study is not without limitations, its findings contribute to the ongoing conversation about the role of technology in education and the potential for AI to support teachers and learners. Overall, the FirstPass pilot study represents a significant step toward more efficient and effective formative assessment practices, with implications for educators and students alike. It paves the way for further exploration and development of AI-driven solutions in the field of education.

Next Steps

The research project has demonstrated that the computer mediation of open-ended questions is a fertile field to explore. Whilst the project has highlighted the gains to be had in using FirstPass to support formative assessment practices, it has highlighted further gaps in knowledge and some of these are expressed in the following questions. Firstly, the research project spanned a broad range of vocational subjects, so how may FirstPass support formative assessment practices in a narrower and more focused context? Secondly, the research project highlighted inefficiencies and bottle necks in model training, so what improvements could be made to model training and model accuracy? And thirdly, whilst the FirstPass platform demonstrated tangible benefits to learners through real-time feedback, what additional feedback could be offered to learners and teachers as part of formative assessment practices?

These questions will be explored during 2023-24 when Bolton College will continue to work with the NCFE on further development and testing of FirstPass, again supported by the Assessment Innovation Fund. The research will offer the opportunity to carry out further tests to assess the efficacy of the FirstPass platform to support formative assessment practices in two specific courses, the Level 2 Diploma in Early Years Practitioner and the Technical Award in Child Development and Care at Levels 1 and 2.

Additionally, Bolton College will be funded by the Education Endowment Foundation (EEF) to undertake a small-scale pilot with 8 – 12 further education colleges. The pilot will assess how FirstPass can support formative assessment practices to aid learners within GCSE English

Language GCSE resit cohorts. If the project is successful, the EEF will increase its funding to support a larger scale project during 2024-25.

Conclusion

The FirstPass pilot project, funded by the NCFE and its Assessment Innovation Fund, has demonstrated promising results. The platform has shown its potential in supporting teachers and students with the formative assessment of open-ended questions. The pilot has involved six participating colleges, and the progress made thus far has been in line with the project milestones and payment schedule.

Key findings from the pilot include the effectiveness of the platform in training subject topic classifiers, positive feedback on its functionality and usability, and valuable insights from participating colleges and students. Some challenges and areas for improvement have also been identified, such as the need to improve the training toolkit, better management of data labelling activities, and addressing potential biases in the labelling process.

In response to the feedback, the project team has outlined several next steps to improve the platform, including adding new features to the classification training toolset, improving the single sign-on process, and incorporating external AI generative text services to aid in training subject topic classifiers.

Whilst the funded programme is designed to assess the efficacy of using FirstPass as a formative assessment platform for open-ended questions, the feedback from the pilot colleges is being used to further refine and enhance the platform. This will ensure that the FirstPass service continues to evolve and improve as it is used to support the formative assessment of open-ended questions.

References

Bieber, M., Engelbart, D., Furuta, R., Hiltz, S.R., Noll, J., Preece, J., Stohr, E.A., Turoff, M. and Van De Walle, B. (2002). Toward Virtual Community Knowledge Evolution. Journal of Management Information Systems, 18(4), pp.11–35. doi:10.1080/07421222.2002.11045707.

Boons, M., Stam, D. and Barkema, H.G. (2015). Feelings of Pride and Respect as Drivers of Ongoing Member Activity on Crowdsourcing Platforms. Journal of Management Studies, 52(6), pp.717–741. doi:10.1111/joms.12140.

EskénaziM., Gina-Anne Levow, Meng, H., Parent, G. and Suendermann, D. (2013). Crowdsourcing for Speech Processing Applications to Data Collection, Transcription and Assessment. Oxford, UK John Wiley & Sons, Ltd.

Martins, J., Branco, F., Gonçalves, R., Au-Yong-Oliveira, M., Oliveira, T., Naranjo-Zolotov, M. and Cruz-Jesus, F. (2019). Assessing the success behind the use of education management information systems in higher education. Telematics and Informatics, [online] 38, pp.182–193. doi:10.1016/j.tele.2018.10.001.

Owusu-Agyeman, Y. 2021, "Transformational leadership and innovation in higher education: a participative process approach", International journal of leadership in education, vol. 24, no. 5, pp. 694-716.

Schonlau, M., Gweon, H. and Wenemark, M. (2019). Automatic Classification of Open-Ended Questions: Check-All-That-Apply Questions. Social Science Computer Review, 39(4), pp.562–572. doi:10.1177/0894439319869210.

Smith, G.G., Haworth, R. and Žitnik, S. (2020). Computer Science Meets Education: Natural Language Processing for Automatic Grading of Open-Ended Questions in eBooks. Journal of Educational Computing Research, 58(7), pp.1227–1255. doi:10.1177/0735633120927486.

Wang, A., Hoang, C.D.V. and Kan, M.-Y. (2012). Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation, 47(1), pp.9–31. doi:10.1007/s10579-012-9176-1.

Ye, H. (Jonathan) and Kankanhalli, A. (2017). Solvers' participation in crowdsourcing platforms: Examining the impacts of trust, and benefit and cost factors. The Journal of Strategic Information Systems, [online] 26(2), pp.101–117. doi:10.1016/j.jsis.2017.02.001.

Zheng, L. and Chen, L. (2019). DLTA: A Framework for Dynamic Crowdsourcing Classification Tasks. IEEE Transactions on Knowledge and Data Engineering, 31(5), pp.867–879. doi:10.1109/tkde.2018.2849385.