THE LEARNING AGENCY LAB

Georgia State University

VANDERBILT UNIVERSITY

Perpetual Baffour, Scott Crossley, Yu Tian, Alex Franklin, Natalie Rambis, Meg Benner, Ulrich Boser.

The Feedback Prize: A Case Study in Assisted Writing Feedback Tools

Executive Summary

Writing is a crucial exercise for lifelong student success. Unfortunately, students are simply not writing enough, and when they do manage to engage in writing activities, they do not receive the feedback necessary to improve. A lack of teacher time and resources to manage the grading burden is one of the main barriers to increased writing opportunities and higher-quality feedback. Resource constraints further compound this issue, as they disproportionately impact Black and Hispanic students, meaning they are more likely to write at the "below basic" level as compared to their white peers.

The Feedback Prize competition series was designed to address this challenge by spurring innovation in the field of assisted writing feedback tools (AWFTs) for students through the use of data science and data science competitions. The regular use of AWFTs in the classroom makes it easier for teachers to grade writing tasks assigned to their students (and assign these tasks more frequently) while also supporting students from all backgrounds to improve their writing skills.

The Feedback Project accomplished these goals through two phases. Starting in 2019, Phase 1 centered around engaging stakeholders in the design process, while Phase 2 focused on the data science competitions. The insight gained from Phase 1 and the models and algorithms created in Phase 2 well position the field to continue investing in the improvement of AWFTs, especially as they pertain to marginalized students.

The work had three main findings:

- **Teachers must be "in the loop" on AI writing tools to impact student outcomes**, and our research highlights the importance of involving teachers at various stages of the AI development process – from data collection to evaluation of performance – in order for these algorithms to have a significant impact. Our research also showed that machines alone are not enough, and AWFTs are best viewed as tools to support teachers rather than a cure-all.

- **High-quality algorithms are able to demonstrate human-comparable accuracy in evaluating student writing,** showing machines can be as effective as humans in evaluating student work and indicating the potential of algorithms to provide valuable feedback.

- **Open data science competitions can facilitate substantial innovation, sparking new research questions and advancements within educational AI.** Several novel algorithms and techniques emerged from the Feedback Prize competitions, and participants dedicated numerous hours to sharing their algorithmic approaches and engaging in reviews of others' solutions on the platform's discussion forums. In the Feedback Prize data science competitions, it is conservatively estimated that competition participants invested time worth more than $240 million across all three competitions, to win a combined prize purse of $270,000.

Background

The Feedback Project was born out of the need to develop more advanced, ethical, and accurate automated writing assessment tools in the classroom. Why do schools need these tools?

Effective writing is critical for success in college and future careers, but few students graduate high school as proficient writers. According to the National Assessment of Educational Progress (NAEP), less than a third of high school seniors are proficient writers. That is especially true within marginalized communities of low-income, Black, and Hispanic students – where less than 15 percent score proficient (NAEP).

To become proficient writers, students need more writing opportunities and more granular feedback to improve. Unfortunately, many teachers struggle to find the time to provide the feedback necessary to help students grow into confident writers. Many educators are burdened by the amount of feedback they need to provide to students.

According to one survey, more than 70 percent of educators say they are overwhelmed with grading, providing feedback, and other administrative tasks. The National Center for Education Statistics reports that educators in low-income schools are almost 20 percent more likely than teachers in other schools to report being overburdened by routine duties like grading essays.

According to research, the most promising solution to this challenge is the use of assisted writing feedback tools (AWFTs) in the classroom. For example, student achievement has dramatically outperformed state averages in districts that use Revision Assistant. Studies have also shown a positive impact for English Language Learners (ELLs) using the program Criterion. In terms of teacher support, the assisted writing feedback tool PEG can drive down the amount of time teachers spend on grading by half, according to one study.

Despite these successes, several factors severely limit the potential impact and scale of AWFTs. Primarily, these tools are expensive and proprietary, making them out of reach for many teachers and students who would benefit.

Secondly, the algorithms of the existing assisted writing feedback tools are also in their infancy, and they need support to improve. In many cases, the tools are not nearly accurate enough, and far more needs to be done to help prepare the field for this potentially groundbreaking technology.

These less-than-accurate algorithms also mean that struggling students are not well-served. According to many experts and practitioners familiar with this technology, AWFTs work best for middle-of-the road writers rather than students operating far below grade level.

To develop solutions to these challenges, Georgia State University (GSU), Vanderbilt University (Vanderbilt), and the Learning Agency Lab (Lab) teamed up to create the Feedback Prize project. The Feedback Prize project set out to create a set of open-source algorithms within the emerging field of teacher-assisted feedback that helps struggling students, including ELLs, dramatically improve their writing.

To accomplish this admittedly ambitious goal, the project was broken into two phases – Phase 1: Co-Designing with the Field and Phase 2: Data Science Competitions.

Phase 1: Co-Designing with the Field

Phase 1 of the project was designed to help the Lab, GSU, and Vanderbilt establish partnerships with different groups of key stakeholders (teacher development organizations, writing platforms, academics in the field of writing and natural language processing, etc.). These partnerships would eventually create an assisted writing feedback community to highlight the potential for change and enable strong communication between those developing AWFTs and the students and educators who would use the product. Phase 1 was carried out in subphases, detailed below.

Establishing a Community of Interest

In the spring of 2019, the Lab kicked off Phase 1 with a day-long conference of almost 40 thought leaders. The conference discussed the latest developments in assisted writing feedback for struggling writers, and how new technologies can help improve the writing of traditionally underserved populations, while also helping overburdened teachers.

The conference underscored just how invested this community would be in a project designed to provide assisted sentence- and paragraph-level feedback for students who struggle with writing.

Following the conference, the Lab created teacher and research advisory panels that would serve as advisors for designing programs throughout the project. The research advisory board consisted of leaders in writing instruction, natural language processing (NLP), and language development

among English Language Learners (ELLs) and Black and Hispanic students. The teacher advisory board consisted of teacher professional learning organizations, including Teaching Lab and National Writing Project, as well as current middle and high school educators.

The Lab organized regular meetings with the panels to provide updates on the projects and engaged members on critical decision points, including the design of the rubric that would inform the annotation of the student essays.

Deeply Understanding the Teacher/Student Context

Having established the project within the field, the Lab pivoted towards understanding the challenges and possibilities teachers and students face through two projects – teacher/student interviews and The Write Tools Challenge.

Teacher Interviews and Focus Groups.

The team interviewed 70 teachers and 15 students on their experiences teaching and/or learning writing, and using assisted writing feedback tools. The interviews were crucial to understanding the potential limitations and areas of improvement for current tools. These interviews generated a list of the most popular tools, what users liked and disliked, and whether or not teachers or students noticed any bias when using these tools. The findings from these interviews show that current tools could do more to support adaptive learning and reduce bias, especially towards students with dialectic differences.

More specifically, teachers thought that AWFTs exhibited bias towards students with a non-standard English dialect, especially with students from marginalized backgrounds. Teachers also stressed the value of increasing the cultural relevance in instruction and among AWFTs. Many teachers spoke to the importance of students seeing themselves reflected in the curriculum, which AWFTs could support by providing more content that better

represents a range of ethnicities and cultural backgrounds.

Teachers also noted that much of the current work to increase cultural relevance focuses solely on race and ethnicity, even though numerous other identities exist. Many teachers wanted to see more prompts and texts that were more inclusive of different identities like gender, religion, class, etc., as well as prompts and texts that portrayed the intersectionality of identities versus treating these identities as monoliths.

Regarding improving AWFTs ability to support personalized learning, teachers felt that more could be done, especially for low-achieving students. One of the most common suggestions was for AWFTs to include monitoring systems to track individual student progress. Most teachers felt like having long-term data on their students' progress would allow them to better understand the problems that their students are facing and work more effectively with their students.

The findings from these interviews emphasize the need for teacher involvement in the development and evaluation of AI writing tools. By involving teachers in the research process, the Feedback Prize project uncovered the crucial role teachers play in shaping AWFT technologies that are effective, inclusive, and capable of positively impacting student outcomes. The teachers' insights were relevant and meaningful due to their direct experiences and expertise in the classroom.

To complement the teacher interviews, the Feedback Prize project also hosted three focus groups with educators and teacher development organizations to receive feedback on the project focus and design.

The Write Tools Challenge.

Building on the knowledge gleaned from the teacher interviews and focus groups, The Write Tools Challenge was launched to solicit ideas for assisted writing tools from 6th-12th grade educators nationwide. Teachers were asked to submit a proposal outlining a tool that would enhance instruction and support student learning in an inclusive classroom.

After receiving 136 proposals, the Lab reviewed all the submissions and assembled a panel of educators to evaluate the finalists' proposals. The submissions were full of innovative ideas, including using students' work as exemplars or tracking student productivity to identify a need for teacher support.

The judges selected Dan Pier, a Spanish teacher at St. John's College High School in Washington, D.C., as the winner of The Write Tools Challenge. At a high level, Pier's tool would move beyond grammar to support students in the writing process from start to finish.

This tool would provide a platform that facilitates group and/or individual brainstorming and idea organizing, allowing students to input, arrange, and rearrange ideas similar to a mind map. Once students have completed and submitted their writing, the tool would provide both positive and negative feedback, allowing the teacher to specify which aspects of the piece should receive which type of feedback.

The tool would extend the customization options, both at the class and individual student level, by allowing teachers to select what writing features are analyzed. This would help mitigate overwhelming students with feedback, while also allowing the teacher to tailor their feedback based on the student's writing level.

Pier's tool also addressed an issue raised in the teacher interviews – providing prompts that address a variety of identities and interests.

While there were many takeaways from this challenge, it solidified the Lab and GSU's understanding of what teachers need to help their students become more confident writers. The top themes that surfaced revolved around creating more support for students as writers, and ensuring that support is inclusive and accessible. The educator panel also further highlights the role of teachers in the AI decision-making process and their expertise in evaluating the feasibility and potential effectiveness of AWFTs.

Similar to the teacher interviews, challenge submissions also highlighted the need for personalized feedback over time. The data generated from tracking this feedback can be used in diagnostic assessments to isolate areas for growth, as well as highlight students' strengths. That would allow teachers to better support their students, and for students to set and track milestones.

Additionally, the need for inclusive, accessible content was also raised. Echoing themes from the teacher interview, submissions to the challenge focused on including mentor or exemplar texts from writers with diverse cultural, linguistic, or socioeconomic backgrounds. Submissions also underscored that teachers need tool features that support ELLs to develop writing in their primary language and provide scaffolding for English writing proficiency.

Finally, challenge submissions made it clear that students need support in all phases of the writing process – not just feedback on a completed piece of writing. Pre-writing feedback, graphic organizers, and prompts catering to scaffolding and planning were all mentioned as helpful features to ensure students receive assistance along the way.

Understanding the Assisted Writing Feedback Tool (AWFT) Market

To ensure that the Lab and GSU could spur technology that would more broadly add significant value to the existing AWFT market and education ecosystem, it hired EdSolutions to conduct a landscape analysis.

EdSolutions carefully considered the use and potential for AWFT in school districts predominantly serving marginalized and underrepresented students. The analysis confirmed that current AWFTs tend to hyperfocus on surface-level elements of writing, such as grammar and spelling, as opposed to feedback points that would support a student to improve their writing more generally. More specifically, there are three common categories of AWFTs.

The simplest of the three types is Embedded Editors. They provide surface-level feedback and are typically embedded into browsers. Similar to Embedded Editors are In-Site Supporters. With these tools, students have to go to the product's webpage to type into a designated text box. However, these tools typically only offer the same proofreading functionality as the Embedded Editors. Finally, there are Actionable Growth Tools, which provide the most in-depth and personalized feedback of the three. However, they are often paid subscriptions, making them inaccessible to many teachers.

The biggest takeaway from this analysis is that the most comprehensive tools – the ones that provide the most personalized and meaningful feedback – are the least popular district-wide, although these tools support writing instruction better than surface-level feedback tools. These findings only served to amplify the need for improved AWFTs.

Alongside this market analysis, the Lab and GSU also analyzed available open-source NLP tools to determine what elements of writing were already captured within high-quality open source tools and which components of argumentative writing could not yet be identified by open source tools. The research surfaced that features related to argumentation and the presentation of evidence were in the greatest need of additional innovation.

Teacher Surveys

The final component of Phase 1 was a round of teacher surveys to gain insights into teachers' opinions on AWFTs and a better understanding of writing instruction in the classroom. After speaking with 200 teachers nationwide, three high-level takeaways emerged.

Exactly as the research indicates, teachers confirmed that students are not writing enough, whether inside or outside of the classroom. A majority of teachers reported that they assign up to two pages of writing a week for homework. However, only a quarter of teachers report assigning in-class writing daily.

Secondly, teachers are mostly unaware of the AWFTs that are currently on the market, or their functionality. As a result, over 70 percent of the teachers surveyed did not use AWFTs in their classrooms.

Finally, despite limited knowledge of AWFTs, the teachers surveyed were willing to try new tools if they were free and accessible. Even without having used AWFTs personally, most teachers believe that AWFTs can help all students and help to promote equity in the classroom by supporting personalized learning.

For example, teachers suggested that these types of tools can level the playing field by providing the same amount of resources to students. Students in better-resourced schools have more access to collaborative learning experiences and meaningful writing instruction. However, students in low-resource schools tend to have larger classroom sizes, meaning teachers have less time to provide feedback in class. AWFTs can help increase the amount of meaningful writing instruction students receive by removing the burden on teachers to give feedback, therefore giving students more opportunities to practice their writing.

That is especially true for students learning English as a second language (English Language Learners or ELLs). ELLs often spend more time than their peers trying to find the right word or constructing sentences in English. Providing exemplars to ELLs, like sentence starters or predictive text, can help students navigate the writing process while simultaneously learning English.

Although many teachers believe in the potential of AWFTs to help students to become better writers, they did express concerns about students becoming dependent on the software. Teachers argued that the best way to address these concerns is to include educators in the design process of new tools, to create opportunities for scaffolded revision.

These findings reinforce the insights obtained from the teacher interviews, focus groups, and The Write Tools Challenge, which highlighted the significance of involving teachers in the development of AWFT technologies. The surveys specifically revealed that teachers acknowledged the inadequate amount of writing being done by students, had limited awareness of existing AWFTs, but were open to exploring new tools to support student learning and equity in the classroom.

Overall, these findings suggest that there is huge potential to reshape the writing classroom with new and improved AWFTs.

Phase 2: Data Science Competitions

Having worked closely with project stakeholders, the Lab and GSU began to collect data that could be annotated and then used for a data science competition. Approximately 600,000 essays from eight different organizations and states were collected over two years. Sources included states, districts, national educational providers, and online writing platforms.

Approximately 32,000 essays were then selected to build two datasets as the basis of the data science competitions: the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus, consisting of over 25,000 argumentative essays written by students in grades 6-12; and the ELL Insight, Proficiency, and Skills Evaluation (ELLIPSE) Corpus, an English language development corpus consisting of over 6,000 essays written by ELLs in grades 8-12. These corpora were developed based on input from teacher advisory boards in which the boards indicated that algorithms that could provide feedback on argumentation and ELL proficiency would be of the most value.

Annotation Scheme: PERSUADE Corpus

The essays in the PERSUADE corpus were annotated for elements commonly found in argumentative writing, and the rubric served as the basis for the first two competitions in the Feedback Prize series: Evaluating Student Writing and Predicting Effective Arguments. The rubric was developed in-house and went through multiple revisions based on feedback from two teacher panels and a research advisory board comprising of experts in writing, discourse processing, linguistics, and machine learning. The advisory boards ensured that the data labeled would appropriately capture the relevant aspects of argumentation and writing proficiency.

The discourse elements chosen for the annotation scheme also come from adapted or simplified versions of the Toulmin argumentative framework. Labels and brief descriptions for the elements are provided below.

- **Lead.** An introduction begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.
- **Position.** An opinion or conclusion on the main question.
- **Claim.** A claim that supports the position. Counterclaim. A claim that refutes another claim or gives an opposing reason to the position.
- **Rebuttal.** A claim that refutes a counterclaim. Evidence. Ideas or examples that support claims, counterclaims, rebuttals, or the position.
- **Concluding Statement.** A concluding statement that restates the position and claims.
- **Counterclaim.** A claim that refutes another claim or gives an opposing reason to the position.
- **Evidence.** Ideas or examples that support claims, counterclaims, rebuttals, or the position.

The essays in the PERSUADE corpus also received a holistic score for essay quality and effectiveness ratings for the individual discourse elements. Two rubrics were finalized for the argumentative elements and holistic essay scoring.

Annotation Scheme: ELLIPSE Corpus

The essays in the ELLIPSE corpus were annotated for English language development and acquisition, and the rubric served as the basis for the third competition in the Feedback Prize series, English Language Learning. The rubric is based on a literature review of the components that comprise language proficiency.

The final rubric comprises a single holistic score of overall language proficiency and six analytic scores related to specific features of the language. They are cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The holistic and analytic scores are based on a 5-point Likert scale. A score of 5 relates to a native-like facility in English language proficiency, while a score of 1 relates to limited ability in English language proficiency.

Feedback on the rubric was first provided by a teacher advisory board that consisted of ten English teachers who taught ELLs. By actively engaging ELL teachers in this process, the algorithms would be trained on data annotations that reflect the objectives of ELL writing instruction. The rubric was next reviewed by a research advisory board composed of experts in second language acquisition, ELLs, and composition. The rubric was then modified to account for the feedback provided by the teacher and research advisory boards.

Selecting Essays for Annotation

Essays for the PERSUADE corpus were selected to reflect a range of writing from diverse student populations that are representative of the writing population in the United States. All essays included information on the student's gender, race/ethnicity, and grade level. A subset of the corpus also contains data on student eligibility for federal assistance programs such as free or reduced-price school lunch, Temporary Assistance for Needy Families, and the Supplemental Nutrition Assistance Programs, which the Lab broadly defines as markers of economic disadvantage. A large sub-sample of the essays in the corpus also includes information on ELL status and disability status. The goal was to build a corpus collection that closely resembled the U.S. secondary public school population in racial, gender, and economic composition, using data from the National Center for Education Statistics as a benchmark.

The demographic data also allow for the identification and analysis of differences within subgroups of populations, which will help to reduce bias in the resulting algorithms.

It was also important to consider the original writing assignment (i.e., prompt) and the nature of the writing task (i.e., independent or source-based writing) for the curation of the PERSUADE corpus. Source-based writing requires the student to refer to a text, while independent writing excludes this requirement. Essays selected for the PERSUADE corpus were an even split between source-based (based on 15 writing prompts) and independent writing.

The ELLIPSE corpus was selected to reflect writing from ELLs in the United States. The 9,000 essays selected for annotation contained demographic and individual difference measures, including gender, race/ethnicity, grade level, and economic disadvantage. All essays were independent writing for which no background knowledge of the topic was required, and students were not provided source texts. The essays selected for annotation were based on 44 different prompts.

In summary, the PERSUADE and ELLIPSE corpora underwent careful curation processes to ensure they could train algorithms effective at evaluating student writing. The corpora provided a representative sample of student writing, considering various demographics, writing tasks, and quality metrics.

Human Rater Training and Annotations Based on Rubrics

The Lab and GSU ran a Request for Proposals to select a firm that would use the rubrics to annotate the essays and prepare them for the competitions. The Lab selected Georgia Center for Assessment (GCA) to manage the annotation and scoring for the PERSUADE corpus and determined that GSU was best suited to internally manage the annotation process for the ELLIPSE dataset.

GCA had significant experience annotating essays to build a dataset for machine learning algorithms and had a team of raters with experience in classroom instruction. Given this, GCA gathered a group of experienced writing teachers who taught in diverse school communities to provide feedback on the argumentative and discourse elements rubric, and advised during the range finding and exemplary development before scoring began. Each essay rater also had at least two years of experience in data annotation.

For the PERSUADE corpus, GCA enlisted 24 raters. Of the 24 raters, 15 identified as female, eight identified as male, and one identified as non-binary. Half of the raters had a bachelor's degree as their highest level of education, while the other half attained a graduate degree. Most raters were above the age of 35, and fewer than half were white, while the remaining were Asian or Black.

For the ELLIPSE corpus, GSU recruited 26 raters. Of the 26 raters, 21 identified as female, three identified as male, and two identified as other. Seven of the raters were undergraduate students (seniors), 12 were master's students, two had completed a master's degree, and five were Ph.D. students. Most raters were in an applied linguistics department, and all raters had experience teaching English as a second language. Many were between the ages of 20-30, and half of the raters were white.

Before norming, all raters for the PERSUADE and ELLIPSE corpora received anti-bias training and anti-bias strategy instruction designed to address issues of bias that occur during scoring and are inherent to the use of standardized rubrics.

All raters took the Implicit Bias Module Series developed by the Kirwan Institute at The Ohio State University to mitigate potentially harmful unconscious biases held by raters. The series covers a wide range of topics, including the formation of implicit bias and feasible ways to prevent and intervene against the bias. The raters spent around 50 minutes on the online bias training and obtained a certification of completion. After the bias training, all raters were trained and normed on similar writing samples not included in the original corpora. This involved familiarity with the rubric scales, the wording within the rubric, group scoring of essays, and independent practice scoring.

The PERSUADE and ELLIPSE corpus were both annotated using a double-blind rating process with 100 percent adjudication such that each essay was independently reviewed by two expert raters and adjudicated by a third expert rater. A Many-Facet Rasch Measurement (MFRM) analysis for the raters and texts was also conducted for the ELLIPSE corpus to check additional aspects of reliability.

Overall, the careful selection of human raters, comprehensive training, anti-bias measures, and the adoption of a double-blind rating process ensured the Feedback Prize could produce algorithms on par with humans in evaluating student work. The rigorous rating process resulted in well-developed annotations within the PERSUADE and ELLIPSE corpora, providing a reliable and accurate basis for machines to assess student writing.

Selecting Essays for Competitions and Publication

The final PERSUADE corpus comprises 25,996 essays annotated for argumentative and discourse elements, relationships between these elements, effectiveness ratings for the elements, and holistic essay scores. The public dataset also includes detailed demographic information for the writers.

The final ELLIPSE corpus comprises 6,482 essays that showed reliability in the MFRM analysis. The final dataset includes the ELL essays and information about the essays, including file names, prompts, and simple descriptive data for each essay. Such as word count, sentence count, and paragraph count. The data frame contains the holistic and analytic scores for each essay and demographic information about the writer, including gender, race/ethnicity, grade level, and economic status.

Additionally, the Lab used automatic and manual processes to scrub the PERSUADE and ELLIPSE data for personally identifying information (PII). It was a multi-stage procedure requiring a combination of named entity recognition (NER) algorithms to search student text and human reviewers to validate the NER results. Human raters for the PERSUADE and ELLIPSE corpora also identified additional cases of PII in PERSUADE, which were not flagged by NER.

The PERSUADE and ELLIPSE corpora are the largest, public, annotated datasets on student writing.

Competitions

Utilizing the PERSUADE and ELLIPSE corpora, three competitions were launched on Kaggle, a popular platform for hosting data science and machine learning competitions. Based on competitor reports, it's conservatively estimated that competition participants invested time worth more than $240 million across all three competitions, to win a combined prize purse of $270,000[1]. The details on each competition can be found below.

Feedback Prize - Evaluating Student Writing (Feedback 1.0)

Feedback 1.0 launched on December 14, 2021, and ran through March 15, 2022. More than 2,000 teams and over 2,500 competitors participated in this challenge, generating nearly 34,000 submissions. In this competition, participants were tasked with developing algorithms that can identify and segment 6th to 12th-grade student essays (from the PERSUADE corpus) into elements of an argument (e.g., claim, evidence, etc.) and label each one.

This competition was successful in terms of its rich participation and getting Kagglers (those who compete on Kaggle) interested in future Feedback competitions. It was also successful in terms of the models generated. Of particular note is the accuracy rate of the winning models. The first-place team's model has a 75% accuracy rate, similar to the rate human readers agreed on the annotations of the Feedback Prize data (73%).

To better understand those winning models, the Lab and Vanderbilt are currently working with experts and specialists, to allow for those models to be accurately and efficiently utilized for open-source automated writing assessment tools.

Feedback Prize - Predicting Effective Arguments (Feedback 2.0)

Feedback 2.0 ran from May 24, 2022, through August 23, 2022. The competition built on its predecessor by tasking participants with developing models that predict the quality of argumentative elements (i.e. effective, adequate, or ineffective).

The competition ran separately from Feedback 1.0 because it used a smaller subset of the data from the first competition (around 6,900 out of the 26,000 essays, or a little over one quarter), which had a stronger balance of effectiveness scores. It also ran separately because the competition task for Feedback 1.0 was already at a high level of difficulty. For Feedback 1.0, teams had to build models that could segment and label argumentative elements in an essay. For Feedback 2.0, teams focused solely on predicting the quality of these elements, ensuring winning models from both Feedback 1.0 and Feedback 2.0 could effectively handle both tasks.

Feedback 2.0 also prioritized computationally efficient algorithms. In other words, models that are simple and fast but still achieve high accuracy. Heavy and complex models can negatively impact the environment with their energy consumption, and they are also less suitable for use in a real-world software tool. This was the first competition on Kaggle to offer a prize-incentivized, "efficiency" track.

Once again, this competition was incredibly successful. In the main track, not concerned with computational efficiency, the first-place team's model reported a Root Mean Square Error (RMSE) score of 0.554. The same team, who also placed first in the efficiency track, reported a RMSE score of 0.558 in the efficiency track. The scores are virtually the same. However, the run time for the winning model on the main track was almost six hours versus six minutes on the efficiency track. This is a monumental development, not only for this project but for the field as well.

[1] $150 per hour at 20 hours per week, over a period of 13 weeks per competition, totalling $39,000 per team

Feedback Prize - English Language Learning (Feedback 3.0)

The third and final Feedback competition ran from August 30, 2022, through November 29, 2022. This competition received the most participation of all Feedback competitions with over 3,000 competitors and over 2,600 teams generating nearly 50,000 models.

This competition asked participants to develop models that score essays based on language proficiency, utilizing a similar dataset of student writing that was used in the first and second Feedback Prize competitions, but containing essays exclusively from English Language Learners. Available essay scoring tools are currently unable to provide feedback based on the language proficiency of the student, meaning the final evaluations are often skewed. The primary goal of this competition is to sensitize these tools to differences in language proficiency.

Like the previous Feedback Prize competition, Feedback 3.0 garnered impressive, winning models utilizing state-of-the-art techniques in NLP. And similar to Feedback 1.0 and 2.0, the winning algorithms for Feedback 3.0 achieved a level of accuracy comparable to the human raters during the annotation phase.

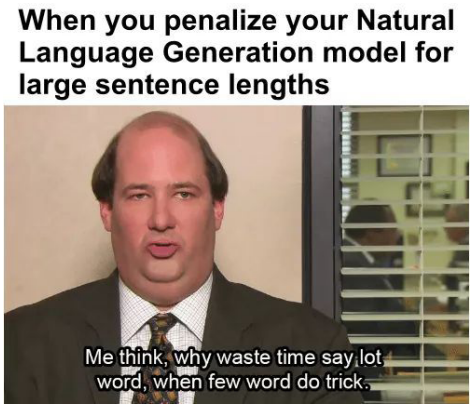
Due to the enthusiasm and effectiveness around the efficiency track in Feedback 2.0, an efficiency track was also created for Feedback 3.0. The efficiency track has been such a successful experiment that Kaggle is looking into making the efficiency track a first-class part of its site.

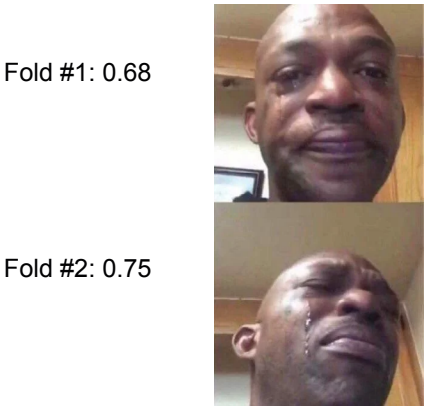
Throughout all competitions, Kagglers expressed their satisfaction with the data, the individual competition tasks, and the competition series. Only a few days after launching Feedback 3.0, a Kaggler commented that the Feedback series was one of the best Kaggle competition series.

In addition to the praise, Kagglers also enthusiastically shared memes throughout the competition series. They ranged in themes from data science and Kaggle-specific jokes to more direct references to developments in the Feedback series. The entirety of the meme discussion threads can be found here, here, and here.

Some of the memes that appeared in the Kaggle competitions

Many of the memes surfaced technical issues in natural language processing.





Computationally Efficient Models

As mentioned above in Feedback 2.0 and Feedback 3.0, incorporating a track dedicated to computationally efficient models proved to be a success in both engaging competitors, and generating more usable models.

Competitors reported that balancing efficiency and performance was an engaging aspect of the competition and used creative techniques to decrease the runtime of models in some cases by up to tenfold, while still maintaining similar levels of performance when compared to larger models. In Feedback 2.0, the number-one team decreased their model's runtime by sevenfold, while only losing 0.6% in overall accuracy. These results demonstrate that it is possible to significantly improve the speed of these algorithms with only negligible decreases in accuracy.

Knowing that efficient models can perform at the same level as more complex models is encouraging since developing efficient models is vital for producing practical algorithms in educational settings where computational resources are often limited. Additionally, these efficient models save resources, which can have a significant impact both financially and environmentally.

Conclusion

Three years ago, the Feedback Prize project set out to achieve the incredibly ambitious goal of spurring innovation in the field of AWFTs to improve the writing outcomes for students and to support their teachers in the grading and feedback process. The Feedback Project was incredibly successful in meeting this goal without losing sight of the equity issues at the core of this challenge.

Three key findings come out of the work:

- **Teachers must be kept in the loop when it comes to AI in classrooms in order to raise outcomes,** and Phase 1 of the Feedback Prize project demonstrated the value of teachers' insights in guiding the design and improvement of AWFTs. By soliciting their input through interviews, focus groups, and surveys, the Feedback Prize project derived meaningful insights from teachers on how AWFTs can address the specific needs and challenges of students in the classroom.

  For instance, most teachers wanted to be involved in the design process of new AWFTs. They advocated for opportunities for scaffolded revision and the inclusion of monitoring systems to track individual student progress. These suggestions reflect teachers' desire for AWFTs to support adaptive learning, reduce bias, and cater to the needs of low-achieving students. Additionally, while most teachers recognized the potential of AWFTs to support personalized learning and equalize access to resources, many were unaware of the existing AWFTs on the market and their functionality.

  More importantly, our research makes clear that machines alone will not be enough. Machines cannot fulfill all the complex needs of students, and while AWFTs can offer valuable support, teachers must be at the forefront of engaging with and implementing these tools. They possess a deep understanding of effective practice in writing instruction. This makes teachers central to both engaging and deploying technology and designing AWFTs that meet educators' requirements for instruction will ensure these tools have a significant impact and relevance in the classroom.

- **The data science competition series demonstrate that machines can be equally as effective as humans in evaluating student writing.** Over the course of three years, after closely working with educators and other stakeholders in the field, two groundbreaking

data corpora were developed and three data science competitions were hosted – generating just shy of 100,000 submissions. The winning submissions showcased advanced large language models (LLMs) that possess the capability to analyze different aspects of a writer's arguments and assess their effectiveness.

Notably, the winning models in Feedback 1.0 achieved an accuracy rate of 75%, comparable to the agreement between human readers in annotating the data. As a result, students and teachers are much closer to using quality, accessible, and equitable AWFTs.

- **Open data science competitions can play a vital role in driving innovation and advancements within the field.** The competitions hosted on Kaggle attracted thousands of participants who collectively invested a significant amount of time and effort, estimated to be worth more than $240 million, to compete for a combined prize purse of $270,000. The high level of participation demonstrates the enthusiasm and interest in these competitions, indicating the valuable impact they have had on the data science community. The success of the competitions is further evident in the quality of the models generated, demonstrating the potential of collaborative efforts in leveraging artificial intelligence to enhance writing assessment.

  In the end, the success of the Feedback Prize competition series, coupled with the engagement, enthusiasm, and positive feedback from participants, solidifies the notion that collaborative and open competitions drive innovation in the field of data science for education. These competitions underscore the advantage of collaborative efforts in fostering advancements in automated writing assessment, promoting the development of efficient models, and showcasing the power of artificial intelligence and machine learning in educational contexts. The Feedback Prize's success proves that similar initiatives have the potential to dramatically impact the educational landscape for good.