

Report on The 24th International Conference on Artificial Intelligence in Education

▼ Table of Contents

 [About the Conference](#)

 [Keynote Speakers](#)

[Prof. Kentaro Inui \(Tohoku University, Japan\)](#)

[Prof. Tamara Clegg \(University of Maryland, USA\)](#)

[Prof. Benedict du Boulay \(University of Sussex, UK\)](#)

 [Selection of Presentations](#)

[“Reducing the Cost: Cross Prompt Pre-finetuning for Short Answer Scoring” Funayama, Hiroaki et. al](#)

[“Trustworthy Academic Risk Prediction with Explainable Boosting Machines” Dsilva, Vegenshanti et. al](#)

[““Why my essay received a 4?” A Natural Language Processing Based Argumentative Essay Structure Analysis” Yang, Bokai et. al](#)

[“Enhancing Stealth Assessment in Collaborative Game-based Learning with Multi-task Learning” Gupta, Anisha et. al](#)

[“Confusion, Conflict, Consensus: Modeling Dialogue Processes during Collaborative Learning with Hidden Markov Models” Earle-Randell, T.V. et al. \(Best Paper Nominee\)](#)

[Prompt-independent Automated Scoring of L2 Oral Fluency by Capturing Prompt Effects” Matsuura, R et. al](#)

[Designing for Student Understanding of Learning Analytics Algorithms” Yeh, C et. al \(check out \[demo!!\]\(#\)\)](#)

[“Improving Automated Evaluation of Student Text Responses using GPT-3 for Text Data Augmentation” Cochran, K et. al](#)

[“Neural Automated Essay Scoring Considering Logical Structure” Yamaura, M et. al](#)

 [Posters](#)

[How Useful are Educational Questions Generated by LLMs? Elkins, Sabina et al.](#)

[Using Decomposed Prompting to Answer Student Questions on a Course Discussion Board](#) Jaipersaud, Jaipersaud, Brandon et al.

["Learning from Auxiliary Sources in Argumentative Revision Classification"](#) Afrin, Taxin et al.

["Visualizing Self-Regulated Learner Profiles in Dashboards: Design Insights from Teachers"](#) Meija-Domenzain, Paola et al.

["Enhancing the Automatic Identification of Common Math Misconceptions Using Natural Language Processing"](#) Gorgun, G et al.

[Impact of Experiencing Misrecognition by Teachable Agents on Learning and Rapport](#)", Asano, Yuya et al.

[Emotionally Adaptive Intelligent Tutoring System to Reduce Foreign Language Anxiety](#)" Ismail, Daneih et al.

["Identifying Usability Challenges in AI-based Essay Grading Tools"](#) Hall, E et al. (virtual)

 [Networking Opportunities](#)

 [References](#)

Driven by the rapid technological advancements and the everlasting pursuit of effective learning methodologies, the educational field is going through a significant transformation. The International Conference on Artificial Intelligence in Education (AIED) has become a well-known forum for researchers, educators, and industry experts to discuss and present the most recent developments in this quickly changing environment. AIED23 was held in the bustling city of Tokyo, Japan and served as a hub for innovative ideas, promising collaborations and inspiring insights that are sure to shape the future of education and our society.

About the Conference

The 24th International Conference on Artificial Intelligence in Education (AIED 2023) was held in Tokyo, Japan and virtually from July 3rd to July 7th, 2023. The conference aimed to explore the latest applications of artificial intelligence in the field of education. This year was the 30th anniversary of the International AIED Society. To celebrate this anniversary, special tracks were held namely "Wide AIED" and "Blue Sky".

This report provides an overview of the keynote speakers, presentations, and networking opportunities that I personally experienced during the conference.

Keynote Speakers

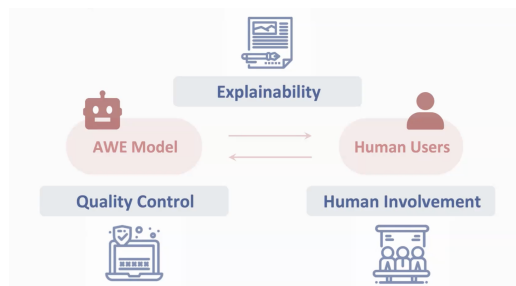
Prof. Kentaro Inui (Tohoku University, Japan)

How Can We Build Trust in Automated Writing Evaluation?

Prof. Kentaro Inui delivered a keynote presentation titled "How Can We Build Trust in Automated Writing Evaluation?" He discussed how trustworthiness is essential for AIED applications and presented Automated Writing Evaluation (AWE) as the ideal application-grounded playground for trustworthiness.

He defined 3 main issues related to trustworthiness in AWE namely:

1. **Quality Control** (focused on AWE model itself)
2. **Explainability** (centered around the interaction between the model and humans) and finally
3. **Human Involvement** (related to human behavior issues).

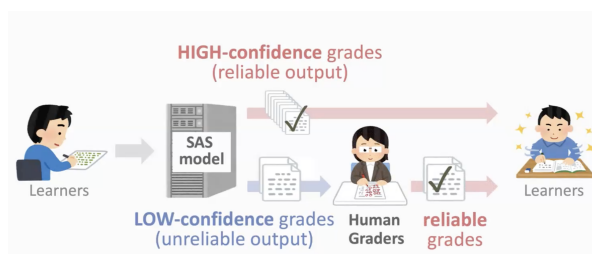


He then introduced studies that address these 3 components and highlighted some of the issues in each.

For explainability in NLP, he emphasized the importance of consulting domain experts to define appropriate explanation forms. The educational field serves as an excellent area for NLP researchers interested in explainability as the pedagogical literature offers plenty of research around corrective and efficient feedback [1, 2].

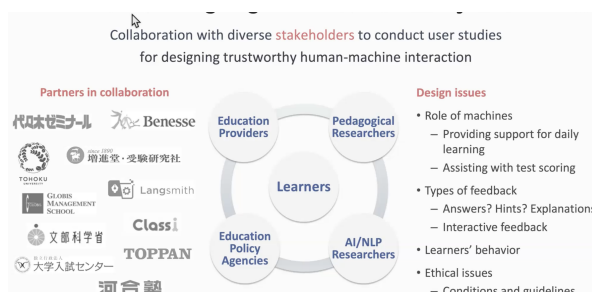
Explorations in Pedagogy	
Prescriptions	Description and references
Focus feedback on the learner	Feedback to the learner should address specific features of his or her work in relation to the task, with suggestions on how to improve (e.g., Butler, 1987; Corbett & Anderson, 2001; Klager & Danesi, 1996; Norris & Hunt, 2000).
Provide elaborated feedback to enhance learning	Feedback should describe the what, how, and why of a given problem. This type of explicit feedback is typically more effective than evaluation of results (e.g., Bangert-Drowns et al., 1991; Gilman, 1969; Moore & Bunting, 2001; Norris & Hunt, 2000).
Present elaborated feedback in manageable units	Provide elaborated feedback in small enough pieces so that it is not overwhelming and does not discourage learners (e.g., Moore & Bunting, 2001; Norris & Hunt, 2000). Presenting too much information may not only result in superficial learning but may also involve cognitive overload (e.g., Moore & Bunting, 2001; Norris & Hunt, 2000). A strategic presentation of feedback allows the possibility to correct for mistakes and give learners sufficient information to correct errors on their own.
Be specific and clear with feedback	If feedback is too specific or clear, it can impede learning and can frustrate learners (e.g., Moore, 2000; Williams, 1997). If possible, try to link feedback clearly and specifically to goals and performance measures (Blöcke, 1992; Norris & Hunt, 2000).

Regarding quality control, Inui suggested that the main open issue is how to control and guarantee the quality of the model's behavior. For example, a straightforward approach could be utilizing model confidence [3], or learning from explanations [4, 5].



Finally, concerning human involvement, the major question NLP researchers should address is “how can we integrate human-centered research to increase trustworthiness?” He highlighted the importance of multidisciplinary collaboration and ethical issues.

To conclude, while there are many challenges to address, there are many experts that are willing to assist with end-user studies



Prof. Tamara Clegg (University of Maryland, USA)

Critical Data Literacy and Why it Matters

A father was surprised to see that his daughter was receiving ads from Target for baby products. Surprisingly, Target figured out that his daughter was pregnant before she did by making use of big data and purchasing habits. That was the example Prof. Tamara Clegg gave to begin her inspiring talk highlighting the fact that AI and big data systems don't exist in a vacuum and how they affect our everyday lives, socially, and politically. While AI and big data systems can help with faster decision making, resource optimization, and performance benefits, they can also lead to complex social, ethical and economic tensions and dilemmas. In addition, since AI is built on learning from datasets, vulnerable populations are often the most negatively impacted by these systems and decisions. For that reason, she highlighted the importance of understanding everyday contexts of Big Data and AI systems to infrastructure them for learner empowerment.

As developers of AI systems, Clegg highlighted how critical it is to address some questions and teach the younger generation to address them as well:

1. **Who develops AI algorithms? What inferences are they making?**
2. **Who is represented by this data? What access do they have to it?**
3. **Who is benefitting or being harmed by these systems?**

Critical to Address

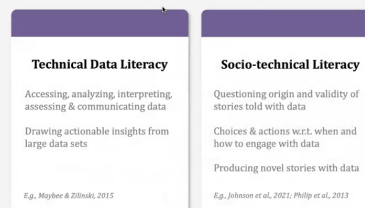


It's imperative for data science and AI education that we facilitate both technical and socio-technical understanding of data for young people

She defined critical data literacy as having 2 main components, **technical** and **socio-technical**. Technical literacy is what comes to mind where the concern is how to deal with the data, whether it's accessing, analyzing, cleaning or drawing actionable insights etc. However, socio-technical literacy is equally important, and it concerns questioning the validity of data-driven decisions and understanding the choices and actions related to data engagement.

Critical Data Literacy

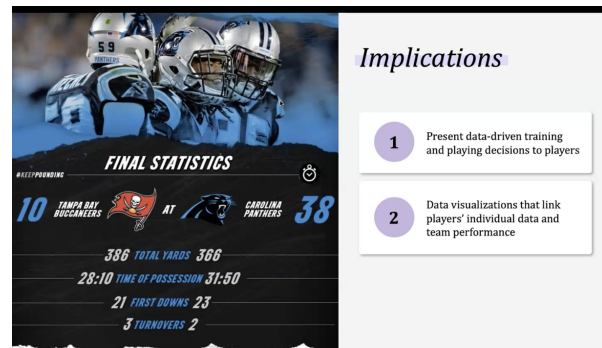
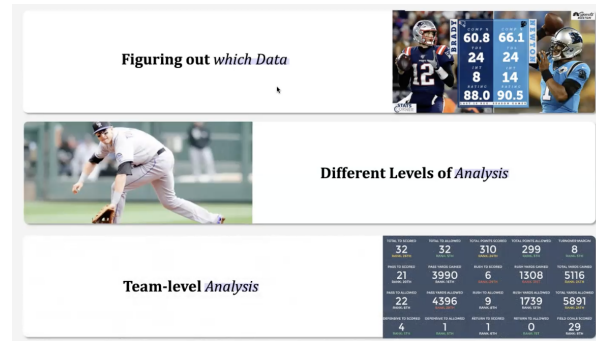
Calderone-Baron et al., 2021; Clegg, et al., 2002



She then began to give examples about both technical and socio-technical data literacy in the field of sports since most major sports team use data analytics to make major decisions such as hiring or firing and where to hold games.

Data Literacy Practices in NCAA Sports (National Collegiate Athletic Association)

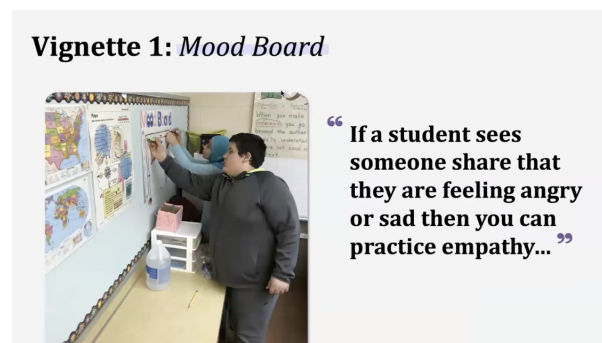
Together with her colleagues, Clegg investigated data literacy practices in NCAA sports. They found that when looking at technical data literacy, it was actually abundant. Coaches talk about how they use data for example, they use gps to collect data about players' speed, collect subjective data about players' wellbeing and emotions etc. Players also talk about how they are attuned to data, for example, film review, tracking weight and nutrition, and reviewing their game stats.



She then offered two vignettes from their work that are closely related to the socio-technical literacy aspect in which they focus on enlivening data for young people.

Vignette 1: Mood Board

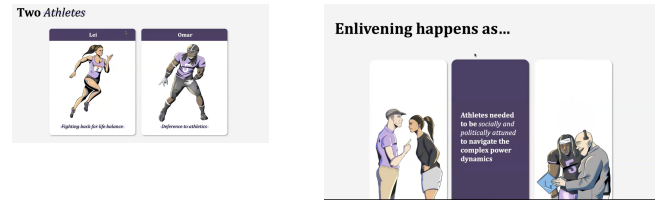
She gave an example from a 6th grade STEM class where students used a mood board to express their feelings. Data was enlivened in this case since youth were able to leverage qualitative and quantitative data about real life socio-emotional issues and were also able to consider how experiences outside of the classroom shaped the moods inside.



Vignette 2: Collegiate Athletes' Enlivened Experience with Data

Clegg then narrated the touching stories of Lei and Omar and their enlivened experiences with data. In the stories she gave, the learners were

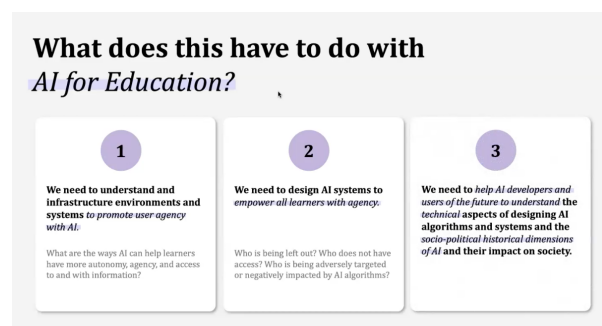
given support for taking agency with data and their applications of it, but in real life contexts, agency is often limited.



Clegg’s work focuses on infra-structuring data literacy practices into communities and institutions. With her colleagues, she explores how community contexts can help (and constrain) people in their day-to-day lives to critically assess data within structures of oppression and how they could utilize data to further their own learning and data-driven actions.

Finally, Clegg ended her talk by stating that as researchers in AIED, it is important to address 3 main needs:

1. Understand and infrastructure environments and systems to **promote user agency with AI**
2. Design AI systems to **empower all learners with agency**
3. Help AI developers and users of the future to understand not only technical aspects but also socio-political historical dimensions of AI and their impact on society



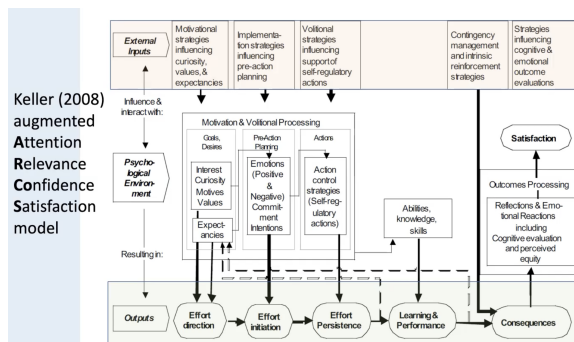
Prof. Benedict du Boulay (University of Sussex, UK)

Prof. Benedict du Boulay began his presentation with a story. A working-class student came up to him in a math class and told him “*Sir, my father is a pig killer and my uncle is a pig killer. And when I leave school, I’m gonna be a pig killer. Why do I have to learn about pythagorus?*”

As motivators for learning, this is one of human teachers’ main roles in the educational system. Other roles that Du Boulay mentioned and supported with research were being orchestrators of the learning system, adapting the system’s use to the educational context and advocating for learner agency. He highlighted the “human” aspects in a human tutor, where a teacher can understand overall context and understand the learner as a person, such as the pig killer student.

Motivational theory was a focus in his talk where he explained how factors such as expectancies, feelings, values and context may vary in emphasis across different learners. While automatically identifying negative motivation either via poor work

rate, body language etc, could be considered an easy classification task, diagnosing the cause and cure of lack of motivation is much more problematic. He introduced the ARCS model of motivation study by Keller [6] which suggests that different strategies for motivation should be used at different stages of learning. Not only should strategies differ by time but they should also be affected by learners' behavior. For instance, appraisal strategy should be different for those learners who experience failure for the first time and for those which failure is a common experience.



Links with Self Regulated Learning	
Prior	Values: Expressing interest; having a positive view about the value of either the interaction itself, or of its successful outcome
	Feelings: Broadly positive feelings about the interaction
	Expectancies: Taking steps to improve the chances of success: clarifying goals; sorting out prerequisites; doing preparatory work; Thinking about what might happen after the interaction
During	Values: Expanding effort; time on task; number and difficulty of problems attempted; going beyond minimum requirement
	Feelings: Self-regulating feelings: being persistent in the face of initial failure; getting started despite anxiety; Maintaining focus despite distractions;
	Expectancies: Choosing problems of appropriate degree of challenge; Using other resources wisely, e.g. the help system; Triangulating understanding
Post	Values: Expressing wish to go further
	Feelings: Expressing satisfaction with either interaction or outcome
	Expectancies: Taking steps to consolidate the new knowledge or skill

He further discussed how motivation is complex as it also interacts with self-regulated learning. Preliminary evidence shows that students that have better meta-affective capability can increase learning gains [7]. (**Meta-affective capability:** those who have the ability to identify their own affective state accurately and further regulate their affective state)

To conclude, I believe the main takeaway from his talk, was that teachers as motivators have a unique role, as they can do things that AIED systems cannot. Teachers have a relationship with students that stands on authenticity and respect. It is one of the goals of AIED to truly understand the skills of human teachers from a scientific point of view so we can develop empathetic and socially intelligent AI systems and Du Boulay's talk was a step in that direction.



Prof. Kentaro Inui: TRUST

Prof. Tamara Clegg: AGENCY

Prof. Benedict du Boulay: AUTHENTICITY & RESPECT



Selection of Presentations

AIED 2023 offered a wide range of research papers and creative initiatives that were presented by academics and professionals from all around the world. We had the opportunity to attend parallel sessions focused on specific themes, such as modeling, evaluation and innovative applications. Unsurprisingly, many

presentations focused on the themes that were presented by the keynote speakers especially trust in AIED systems. Below are some of the notes I took on a selection of presentations I attended:

▼ “Reducing the Cost: Cross Prompt Pre-finetuning for Short Answer Scoring” *Funayama, Hiroaki et. al*

Background

Automated scoring is simply the task of estimating scores for descriptive answers. It has been an active area of research for a long time due its potential for lowering costs and achieving fair evaluation. In this research, the authors focus on Short Answer Scoring (SAS) with rubrics.

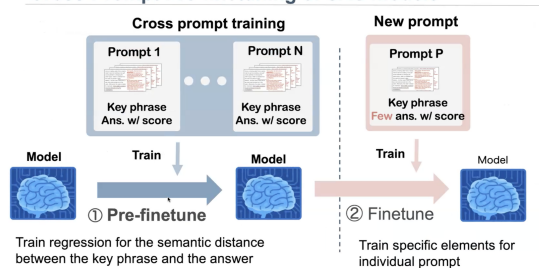
Objective

The main issue addressed is that rubrics differ for each prompt and as a result, individual models have to be trained for each kind which requires a lot of training data that is often unavailable.

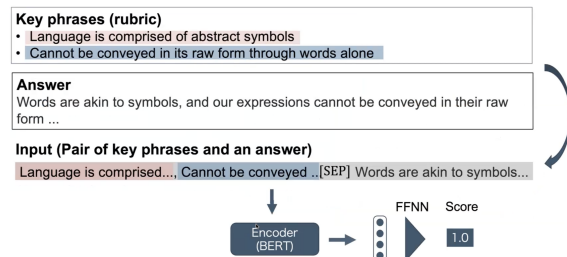
Method

For this reason, the authors explore the potential of cross-prompt training in SAS by incorporating the usage of key-phrases and two-step cross-prompt training. Step one involves pre-finetuning an SAS model on cross-prompt data to predict scores based on the entailment relationship between key phrases and answers. The second step then further finetunes the model on each prompt to create a prompt specific SAS model.

Method: Cross-Prompt Pre-finetuning of SAS models



Model and Input/Output

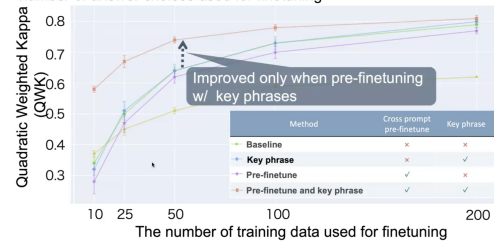


Results & Evaluation

Their experimental results demonstrated that cross-prompt pre-finetuning improves performance and that the number of training data for finetuning can be reduced to 50% without compromising performance.

Effects of cross prompt pre-finetuning

- Quadratic Weighted Kappa (QWK) and standard deviation when varying the number of answer choices used for finetuning



▼ “Trustworthy Academic Risk Prediction with Explainable Boosting Machines” *Dsilva, Vegenshanti et. al*

Background

Many research has been done on academic risk prediction. However, the types of data, the methods and evaluation metrics of course differ. In the context of an online learning scenario with event logs and a binary classification objective, the authors focus their attention on explainability in both features and in evaluation. Can stakeholders trust the recommendations made by the system? What led to the prediction? Is the system fair and reliable?

Objective

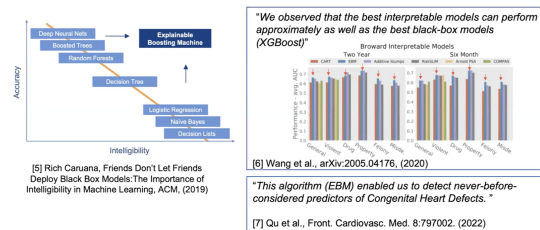
This research focused on three main objectives:

- Motivate use of trustworthy evaluation metrics next to accuracy in model comparison
- Demonstrate the explainable boosting machine (EBM) for an AIED use case
- Show how feature response graphs for online learning behaviors potentially help to gain trust

Method

There is a known tradeoff in machine learning models between accuracy and intelligibility or explainability. The authors make use of a model that offers a balance between these two, namely Explainable Boosting Machines (EBM). EBMs are built on the idea of generative additive models, you have multiple features and you calculate/train a feature response function per feature or pairwise feature interaction, EBM is a fast implementation of the second approach by using trees. EBMs are

Accuracy vs Intelligibility



highly intelligible where feature response functions show the contribution of each function to a final prediction. [Find more here.](#)

$$\text{Additive Model (GAM): } y = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)$$

$$\text{Additive Model with Pairwise Interactions (GA2M) : } y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j)$$

The authors used clickstream data and online learning behavior features to predict failure or dropout. They built baseline models including glass-box models namely Logistic Regression and Decision Trees and black-box models namely Random Forest and Feedforward Neural Network.

Results

In terms of accuracy, EBM was found to outperform the baseline or at least be on par with the baseline methods. The authors used other metrics to evaluate their model for trustworthiness namely:

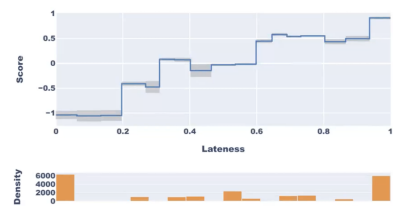
- **Earliness of Prediction:** (EP) mean of first time period when label was correctly predicted
- **Harmonic Mean for Earliness:** EP and taking into account accuracy of the model
- **Temporal Stability of Prediction:** Mean of longest time period where correct predictions were made consecutively
- **Fairness: Statistical Parity Difference (SPD):** measures the difference that the majority and protected groups receive a favorable outcome. Ideal value should be 0.
- **Fairness: Equal Opportunity Difference (EOD):** measures the difference that the majority and protected groups receive a favorable outcome given a correct prediction. Ideal value should be 0
- **Faithfulness: Recall on Important Features:** measures the mean of recall by training the model on important features for explanations generated for the test dataset (for each local explanation in the test dataset: determine k most important features, re-train model only on selected important features and compute recall)

For fairness measures, they found that all models achieved fairness estimates close to zero and that was most probably due to the fact that the dataset was balanced in terms of features. They found that EBM was able to achieve the highest faithfulness score.

All in all, they showed that using trustworthy evaluation metrics allows for a more holistic picture of models, that EBMs may be of potentially high value for AIED tasks (especially tabular data structures) and that feature response

functions allow visualization and interaction with the model especially for domain experts.

Example for Feature Response Graph



Feature function for the feature "lateness" generated by EBM for predicting learners at risk of dropout in week 20 with the gray boxes representing variance around a score. The lower graph represents the density in the different feature value bins.

▼ **““Why my essay received a 4?” A Natural Language Processing Based Argumentative Essay Structure Analysis” Yang, Bokai et. al**

Background

Most Automatic Essay Scoring (AES) tools only provide a holistic score. However, that alone is not enough for the classroom because students need feedback on their essay and teachers demand details of the graded essay.

Objective

- 1. Extract and identify argumentative elements and structures from the text (identify claims, evidence in students’ essays to improve performance of the AES model.
- 2. Explore association between argumentative structure and essay scores
- 3. Provide a design mockup for a future feedback tool

Method

The datasets used were (Feedback Prize Dataset from Kaggle, their own ACT Writing Test). For argumentative element classification, they used DeBERTa and DistilBERT pretrained models. They trained the model to classify only (Positions, Claims and Evidence) since they held the majority of the dataset and considered features such as counter-claim to be part of claims. The model was then trained with a cross-prompt, sentence level approach, a single ensemble model covering multiple prompts. After classification, argumentative structures are then extracted via sequential pattern mining using SPMF package to provide essay structure feedback.

ACT Writing: Higher scored essays use a different structure

- Sequential pattern mining results
- Using more Evidence-Claims-Position structure

Scores	Top Three Frequent Sequences	Counts and Proportions
12	evidence claim claim claim claim claim position	59 (70.24%)
	claim evidence claim claim claim claim position	58 (69.05%)
	evidence claim claim claim claim claim position	58 (69.05%)
10-12	claim evidence claim claim claim claim position	807 (56.91%)
	claim claim evidence claim claim claim position	796 (56.14%)
	evidence claim claim claim claim claim position	786 (55.43%)
7-9	position claim claim claim claim evidence claim claim	2982 (40.38%)
	position claim claim claim evidence claim claim claim	2978 (40.33%)
	position claim claim claim claim evidence claim	2938 (39.79%)
4-6	position claim claim claim evidence	1348 (28.37%)
	position claim claim evidence claim	1344 (28.28%)
	position claim evidence claim claim	1318 (27.74%)
2-3	position claim claim claim evidence	14 (3.21%)
	position claim claim evidence claim	14 (3.21%)
	position claim evidence claim claim	13 (2.98%)

Results

Using the ensemble model they achieved overall accuracy of 0.85 when classifying argument structures. They found that the most frequently used structure in the essays (accounting to 40% of the ACT Writing Essays) was: (Position - Claims - Evidence - Claims)

When analyzing essays by score points they found that:

- The longer the essay, the higher the score
- Higher scored essays include more positions and evidences (Popular structure: Evidence - Claims - Position)
- Lower scored essays had less argumentative structures (Popular structure: Position - Claims - Evidence)

Design mockup for a future feedback tool:

- Length
- Argumentative elements
 - How the Position and Evidence sentences should be used.
- Argumentative sequences
 - How the Evidence, Claim, and Position sentences should be placed.



▼ “Enhancing Stealth Assessment in Collaborative Game-based Learning with Multi-task Learning” *Gupta, Anisha et. al*

Background

Collaborative game-based learning environments combine the benefits of collaborative learning and game-based learning environments. It can help support collective brainstorming towards a shared goal and help guide peers. These learning environments often come with a chat interface to offer a platform for students to exchange their ideas. But what happens when you leave middle school students to talk amongst themselves? They often engage in out-of-domain chat. While some of it may be constructive, it can lead to students transiting to negative affective states and lead to unproductive learning interactions.

Objective

- 1- Predict out-of-domain chat behavior beforehand and students’ learning outcomes while interacting with the game.
- 2- Create a multi-modal multi-task stealth assessment to model students’ learning without disrupting their engagement.

Method

Their study was conducted using the “Crystal Island: Ecojourneys” dataset where students have to investigate an outbreak on a remote island that is affecting the fish population. Their task is to identify the illness while interacting with the game. The collaborative aspect in the game exists where students gather at a virtual whiteboard to discuss and vote on ideas and also engage in a group chat.


CRYSTAL ISLAND: ECOJOURNEYS

Narrative

- Investigate outbreak affecting the fish on a remote island
- Identify illness

In-Game Activities

- Talk to NPCs
- Explore the map
- Read in-game content

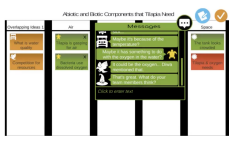


Data Collection

- Played for 6 classroom periods
- 72 students divided into 18 groups
 - 43% female
 - 11-12 years old
- Pre-test and Post-test conducted
- Chat messages
 - 8,350 chat messages
 - 2,000 messages sent by facilitator
 - Average of 463.89 messages in each group
 - Average of 92.78 messages sent by each student

Collaboration in CRYSTAL ISLAND: ECOJOURNEYS

- Gather at virtual whiteboard to discuss and vote on ideas
- In-game group chat
 - Regulated by a facilitator



Out-of-Domain Rubric

Label	Rubric	Example
On-task	Text that contributes to the science discussion in the group, demonstrates or addresses a relevant affective state, fosters collaboration, or asks a relevant question	"Im at the whiteboard now, what do I do?" "Tilapia need warm temperatures in their water, dissolved oxygen, and food (Cant remember what food)"
Out-of-domain	Text that is meaningless or unintelligible, or off-topic conversation that is unrelated to learning outcomes and fails to address affective states of students in the group	"REEEEEEEEE" "are u boy or girl" "WHATS A TURTLE" "gonna be as mean as possible"

Out-of-domain Contribution

Out of domain contribution

$$= \frac{\text{\# student's out-of-domain messages}}{\text{\# out-of-domain messages in group chat}}$$

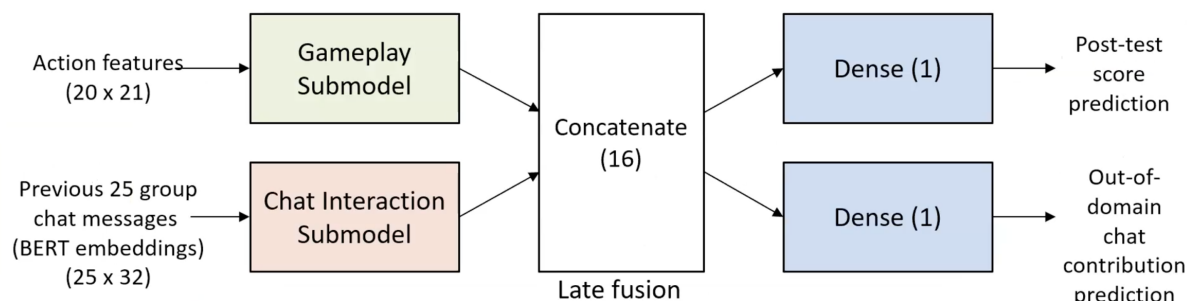


The multimodal features they extracted from the dataset were as follows:

- Game Action Features
 - 21 distinct action types
 - talking to an NPC, moving to a different location, voting on a sticky note
 - count vector to represent volume of each action
- Chat Features
 - Average word-level pre-trained DistilBERT embeddings to represent a chat message (768 dimensions)
 - Dimensionality reduction using PCA (32 dimensions)

They converted the task of predicting post-test score and out-of-domain contribution from originally being regression tasks to classification tasks to make the feedback more actionable.

Multimodal Multi-Task Stealth Assessment Model



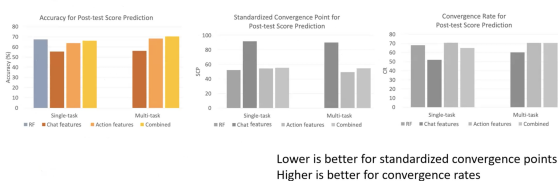
Regarding the model's architecture, they used 2 separate sub-models to process each of the 2 different modalities or features. For each game action that the student takes, the model look at the past 20 actions, and feeds them through an LSTM model. The output is then combined with the output of the other sub-model which works similarly except with a GRU and previous 25 group chat messages.

Results

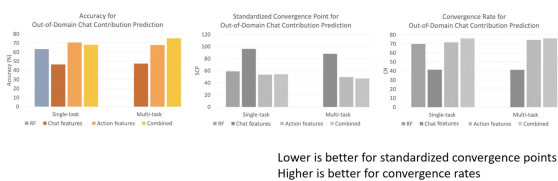
The authors evaluated their experiments using group-level nested cross validation with 5 inner folds and 3 outer folds where in each fold, 20% of training set is used for validation. They maintained a random forest baseline and evaluated across 3 main metrics, namely:

1. Accuracy
2. Standardized convergence point (Min et al. 2016)
3. Convergence rate (for what percent of students are we able to make correct predictions)

Post-test Score Prediction Results



Out-of-Domain Chat Contribution Prediction Results



The deep learning based multimodal stealth assessment framework that they proposed was able to outperform a non-neural random forest baseline for both prediction tasks. The multi-task learning models outperformed the single-task learning models across all 3 evaluation metrics, especially for models predicting post-test performance using only gameplay features which suggests that both tasks look at

similar features and make same inferences from gameplay features. They obtained best early prediction results for post-test performance prediction using both game-trace logs and chat interaction evidence which shows that group chat could be beneficial for post-test performance prediction.

▼ **“Confusion, Conflict, Consensus: Modeling Dialogue Processes during Collaborative Learning with Hidden Markov Models” Earle-Randell, T.V. et al. (Best Paper Nominee)**

Background

There is growing recognition that AI technologies can and should support collaborative learning. While great progress has been made in modeling dialogs for middle and high school students, the dialogue processes that characterize collaborative talk between elementary-aged learners are not currently well understood. A promising approach to moving towards AI-augmented support of collaboration for young learners is to build a bottom-up model of dialogue and explore ways in which these learners move between the states or moments of collaborative talk.

To analyze collaborative talk, they leverage a leading discourse theory by Neil Mercer that deconstructs collaborative talk into 3 different components:

1. **Exploratory Talk:** embodiment of collaborative, critical thinking where learners express incomplete thoughts as they forge their own understanding of the content
2. **Cumulative Talk:** characterized by learners constructing a shared knowledge base by positively and uncritically building on each others contributions
3. **Disputational Talk:** learners disagree with each other and make their own decisions instead of collaborating with their partner

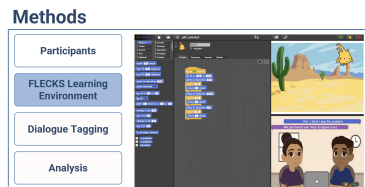
In this research they focus primarily on exploratory and disputational talk.

Objective

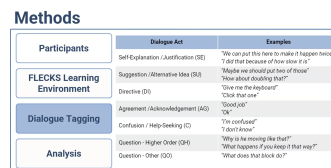
What are the dialogue states that characterize collaborative talk between elementary school learner dyads, and how do these dyads transition between these dialogue states?

Method

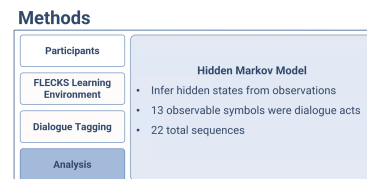
The data used was gathered as a part of a larger study where students participated in a 5 day classroom study in which they completed 40 min pair programming activities using the FLECKS learning environment.



FLECKS learning environment, contains 2 virtual agents designed to foster collaborative talk by modeling collaborative behavior (not focus of this study)



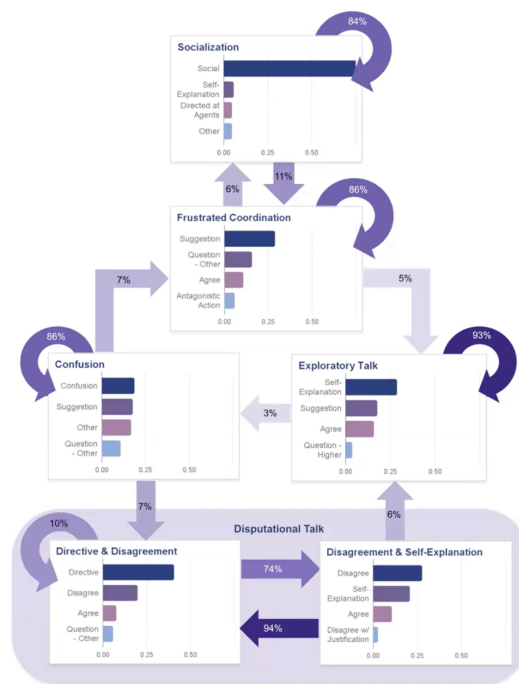
Tagging all dialog transcripts between students



To model learners' collaborative talk, this study implemented a Hidden Markov Model. (*HMM is a probabilistic graphical model that can be used to describe the hidden processes or states that influence a sequence of observable symbols*). The constructed observable symbols were the 13 dialog acts that were extracted and the sequences were the 22 pair programming sessions. They experimented with 3-12 hidden states and a leave one out validation method to finally determine that 6 states performed the best.

Results

Results



Resulting HMM

LearnDialogue

Exploratory Talk while the most common, only had 1 primary way for students to leave it: Confusion. It occurs when students reach an impasse where it is hard to move forward which can lead to Frustration and Disengagement. Once students entered the Confusion state, there was an 86% chance they would dwell

there until they transitioned to Frustrated Coordination or Disputational Talk. The only way out of Disputational Talk is to cycle back to Exploratory Talk, by justifying and working cooperatively.

To summarize, supporting collaborative talk between partners is an important direction for AI augmented technologies to move into. The paper presented an analysis of states and flow of dialog between elementary school students dyads who are collaboratively learning and found that disputational dialogue appears to be an important component of larger cycles of exploratory talk between young learners.

▼ **Prompt-independent Automated Scoring of L2 Oral Fluency by Capturing Prompt Effects” Matsuura, R et. al**

Background

In context of foreign language learning, speaking ability is evaluated in terms of fluency or “the speaking ability to produce smooth utterances”. A good speaker tends to speak fast with few pauses and hesitation. To evaluate fluency, it is necessary to prompt the learners to score their performance.

This paper proposes an automated fluency scoring system which is compatible with multiple-prompt settings. Different prompts impose different cognitive demands on speaking and it is important to evaluate fluency consistently across various prompts. Human experts estimate prompt effects from non-temporal features such as content, lexico-grammar and pronunciation.

Objective

→ Develop an automated fluency scoring system which is compatible with multiple-prompt setting

- Do content, lexico-grammar and pronunciation features improve performance on automated fluency scoring system in multiple-prompt setting?
- Which features contribute to improved performance of automated scoring?

Method

Architecture of proposed system:

Use **prompt-id**, **sentence embedding** & **speech representation** to capture prompt effects

- **Prompt-id:**

- One-hot-vector of prompts

- **Sentence embedding:**

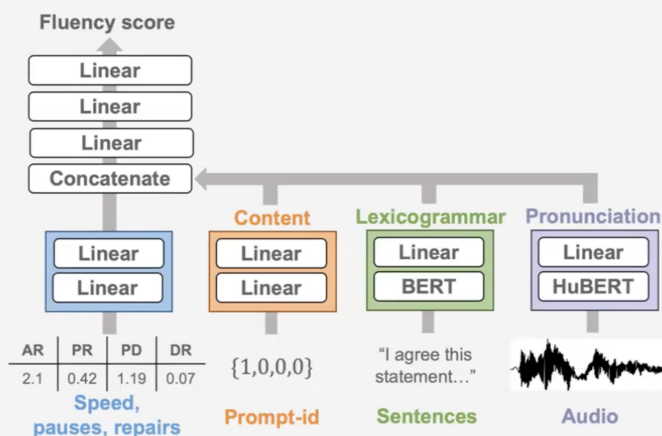
- Extract from BERT [Devlin+, 2019]

- **Speech representation:**

- Extract from HuBERT [Hsu+, 2021]

- **Temporal feature:**

- 12 speed, pause & repair features [Matsuura+, 2022]
e.g.) articulation rate, pause ratio & duration, disfluency ratio



To capture prompt effects, the model is conditioned by prompt ids, sentence embeddings from BERT and speech representations extracted from HuBERT.

In the experiments, they used English monologue speech data belonging to 128 Japanese learners of English, 4 main prompts and ground truth labels of a 6-point scale fluency score. The speech samples were all manually evaluated by PhD students in Applied Linguistics.

Results

They compare their model with a baseline of a conventional method where the system only uses speed, pause and repair features.

Evaluation metrics employed:

- Quadratic weighted kappa (QWK)
- Pearson's correlation coefficient (PCC)

Result: Examine agreement btw ground truth & predicted fluency score

Content, lexicogrammar & pronunciation features improved performance of automated scoring in multiple-prompt setting

- Conditioning by the non-temporal features may work well in the multiple-prompt setting

Model	QWK↑	PCC↑
Conventional	0.797	0.904
Proposed	0.863	0.932

Their method was able to outperform the conventional method baseline which indicates the effectiveness of the employed features for creating an automated scoring system in a multi-prompt setting.

They performed an ablation study to test which features contributed to the improved performance of automated scoring and found that the pronunciation makes a strong contribution, while lexico-grammar may not be necessarily that strong. They found a negative impact by just inputting sentence embeddings. However, a limitation of their study is that the proposed method is not entirely prompt-independent since a one-hot vector is used for the prompt-ids.

▼ Designing for Student Understanding of Learning Analytics Algorithms" Yeh, C et. al (check out [demo!!](#))

Background

It is well-known how there is an increasing prevalence of artificial intelligence enhanced decision-making systems in educational settings and our daily lives. Therefore it is equally important to understand the limitations of these systems. Explainable AI is often spoken about as a way of increasing algorithmic transparency and user understanding. A possible way of achieving this is post-hoc XAI which often involve interactive visualizations. AIED has a lot to contribute to the XAI field.

At the moment, there is no established evidence-based way for XAI designers to determine which concepts are necessary to teach. Many AI explanations still require extensive prior knowledge of machine learning.

In addition, the relationship between algorithmic transparency and user outcomes in learning analytics systems is unclear. Even if we reach 100% algorithmic transparency, it is not guaranteed that we will reach the outcomes that we want, it may or may not lead to increased use of AI systems or trust or even

user understanding. For example, there is evidence where students do not necessarily have positive feelings towards transparent grading schemes.

Objective

How to build good explainables and evaluate their effectiveness?

Research Questions

RQ1: What are the knowledge components of algorithmic understanding for BKT?

RQ2: How does decreasing the transparency of BKT's limitations affect algorithmic *understanding*?

RQ3: How does decreasing the transparency of BKT's limitations affect perceptions of algorithmic *fairness* and *trust*?

Method

They used Bayesian Knowledge Tracing as the base example for an AI algorithm to explain. Any other model could also be used by following same process to achieve an evidence based effective explainable.

Bayesian Knowledge Tracing

We focus on studying [Bayesian Knowledge Tracing \(BKT\)](#), an AI algorithm that predicts skill mastery

There are four parameters involved in BKT (each with a value between 0 and 1, inclusive):

- **P(init):** the probability that the student already knew a skill.
- **P(transit):** the probability that the student will learn a skill on the next practice opportunity.
- **P(slip):** the probability that the student will answer incorrectly despite knowing a skill.
- **P(guess):** the probability that the student will answer correctly despite not knowing a skill.

$$p(L_t | obs = correct)_u^k = \frac{p(L_t)_u^k \cdot (1 - p(S)^k)}{p(L_t)_u^k \cdot (1 - p(S)^k) + (1 - p(L_t)_u^k) \cdot p(G)^k}$$

$$p(L_t | obs = wrong)_u^k = \frac{p(L_t)_u^k \cdot p(S)^k}{p(L_t)_u^k \cdot p(S)^k + (1 - p(L_t)_u^k) \cdot (1 - p(G)^k)}$$

$$p(L_{t+1})_u^k = p(L_t | obs)_u^k + (1 - p(L_t | obs)_u^k) \cdot p(T)^k$$

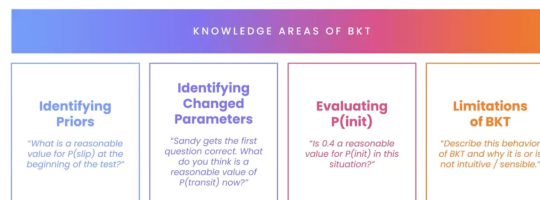
While BKT is transparent, it is complex to understand on the fly. It also has flaws like model degeneracy, and the fact that it does not incorporate forgetting or passage of time.

To answer the research questions, they conducted the following process:

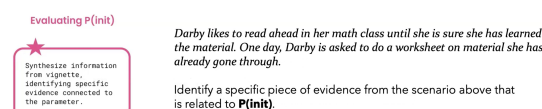
1. Identify Knowledge Components (KC) of BKT and build assessments
2. Build post-hoc explainable, a learning activity that incorporates and support each of the components
3. Evaluate user outcomes

To identify KCs, they perform cognitive task analysis which is a think aloud method for systematically identifying the necessary concepts to teach stakeholders about a topic (e.g. an AI algorithm) and with CTA, they identified four broader knowledge areas of BKT, which can be broken down into 19 smaller KCs. To assess KCs, they developed questions specifically targeting each concept.

Then to build an explainable for BKT, they built it alongside with another skill that they were trying to teach so they could use BKT's inputs and outputs as the participant learns something. (ex. a skill that is useful and not everybody knows is sign language). They developed an interactive web app that also incorporates best pedagogical practices like prompts, multiple choice etc .



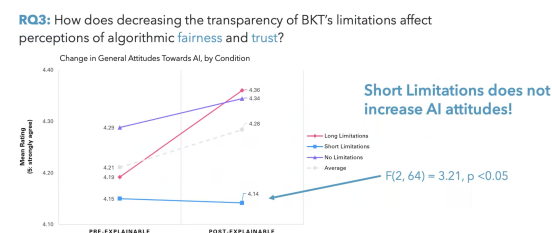
knowledge components of BKT



vignette type questions for assessment

Results

To assess how transparency decrease affects understanding and perceptions of fairness and trust, they carry out a randomized and controlled experiment where they vary levels of transparency and measure the impact on user perception of trust and fairness. Their findings showed the effects of varying transparency on algorithmic understanding and user perceptions which surprisingly showed that less information is not necessarily always worse.



▼ “Improving Automated Evaluation of Student Text Responses using GPT-3 for Text Data Augmentation” Cochran, K et. al

Background

Teachers use formative assessments to evaluate progress on textual responses for many advantages:

- students can build components of knowledge along the way to solve larger problems
- students can communicate conceptual understanding of the topic, have self-assessment skills and self-regulated learning skills
- teachers can become aware of difficulties along learning path

Formative assessments of student writing takes time and the sooner the feedback the more opportunity for feedback to be received while the brain is in a receptive state.

Objective

Deep learning models need lots of data to perform well and many educational studies have very few student responses. Smaller educational data sets are also often imbalanced.

Solution: Balance dataset or expand it. Easy with image data. With text? Text augmentation techniques:

- back translation
- masking
- add noise with typos for example
- use hyponym (more specific term) or hypernym (broader meaning)
- self augmentation, using copies of given data (may overfit however)

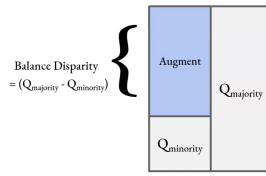
Research Questions

1. Can artificially generated responses improve base model classification performance?
2. Can artificially generated responses outperform self-augmentation when used for training models for sentence classification?
3. Does temperature sampling of the artificially generated student responses affect model performance?
4. Does performance ultimately degrade when model reaches a sufficient level of augmentation?

Method

Use GPT 3.5 for robust text generation, use student responses in a prompt for text generation and compare different temperatures when generating.

First, create a pool of augmented data.



Next, Increase Minority To Balance Dataset



Finally, Augment After Balancing

Models created for each of the 7 concepts.

Used three (3) temperatures (0.1, 0.5, 0.9) plus a self-augmented version of each 9 Augmentation levels: 0, 1x, 3x, 8x, 21x, 34x, 55x, 89x, and 100x

This resulted in 252 individual models

80% Training / 20% Test

Multilingual Model: Microsoft Multilingual L12 H384

Results

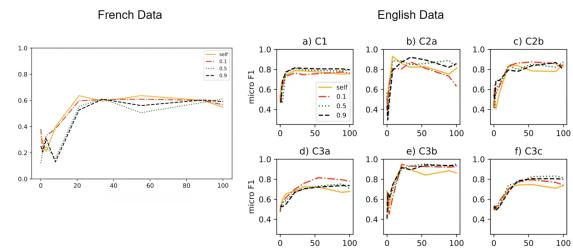
Generative AI augmentation techniques can produce responses that are semantically different than the prompted sentence.

1. Balanced data sets improve in performance with the addition of generated augmentation
2. Generated responses improve performance over self-augmentation
3. Temperature variation didn't affect performance as much as expected
4. Peak performance varied but didn't exceed 89x

Performance of baseline vs. all augmented methods (micro-F₁)

Concept	% Maj. Label	Baseline		Max Performance		
		<i>a priori</i>	Unaug.	Self	GPT-3.5	Aug.
French	73	0.720	0.575	0.636	0.612	21x
C1	89	0.940	0.735	0.789	0.815	0.6x
C2a	73	0.850	0.757	0.931	0.921	8x
C2b	67	0.670	0.547	0.852	0.874	55x
C3a	54	0.700	0.532	0.726	0.815	55x
C3b	77	0.770	0.684	0.926	0.952	55x
C3c	60	0.600	0.568	0.747	0.832	89x

Results: Temperatures vs. Self-Augmentation



▼ "Neural Automated Essay Scoring Considering Logical Structure" Yamaura, M et. al

Objective

Neural Models cannot directly consider logical structure in student essays. The main objective is to incorporate information about the logical structure of the essay in hope of improving the performance of AES models. A method that can be used for this is Argument Mining. H. Nguyen and D. Litman (2018) handcrafted features such as the number of premises and used those as inputs. Ultimately, the improvement in accuracy was limited. The objective of this research is to incorporate the features in a way that they do not need to be handcrafted as this could be the reason as to why logical structure was not properly captured.

Method

1. Extract logical structures using argument mining. (**Argument mining**: a machine learning method that inputs sentences and estimates the logical structure)
2. Convert extracted logical structure into numerical form and incorporate it into the model. To numerate the logical structure, they create what they call a “**visible matrix**”. The goal is to simply make words visible if they are logically connected and invisible if they are not. The dots represent relevance between two words in a sentence. Calculate t -th self attention vector by incorporating the visible matrix M and call this BERT based model “BERT-LS”.
3. Concatenate outputs from BERT-LS and conventional model to utilize both logical structure from BERT-LS and retain some of the linguistic information from the conventional model. Use concatenation as input to a linear layer to predict score.

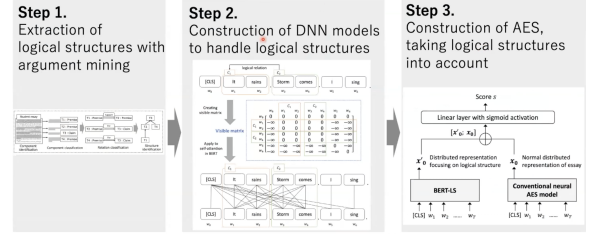
In their experiments, they used ASAP dataset which consists of essays written by English native speakers in 8 different tasks along with their scores.

Results

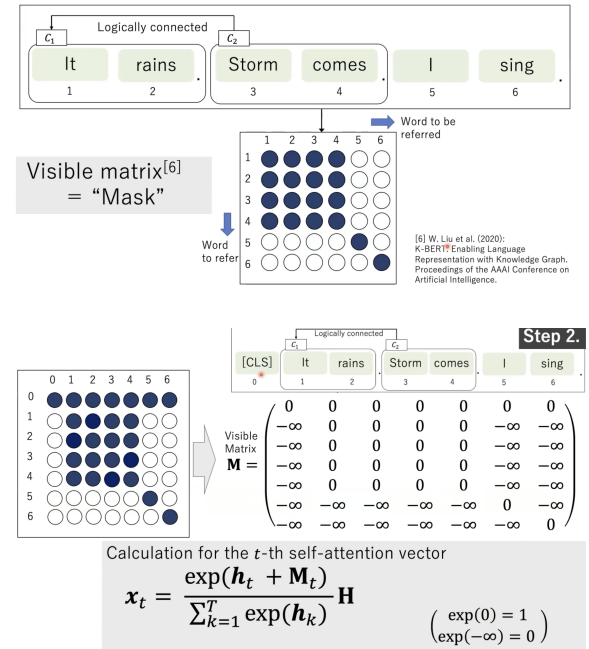
They evaluated their experiments with 5-fold cross validation and QWK (Quadratic Weighted Kappas) as the main metric. Results showed that incorporating logical information of essays can improve the accuracy of neural AES models, especially with persuasive type prompts.

Proposed Method

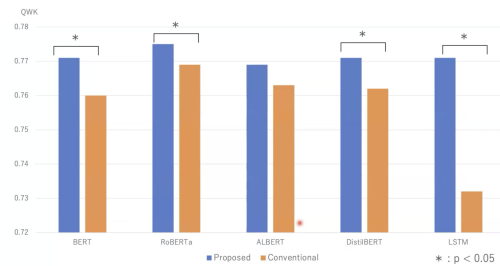
Process for incorporating logical structure



Process Behind Visible Matrix Creation



Experimental Results (Compared by Average)



Posters

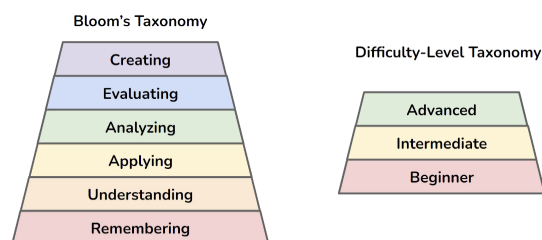
Poster sessions were held on July 4th, 5th and 6th. On July 4th, I had the chance to present my work as a poster and have meaningful discussions with other researchers. Below are some notes I took on other poster sessions I attended as well and links to interesting ones.

July 4th

▼ How Useful are Educational Questions Generated by LLMs? Elkins, Sabina et al.

Use LLMs (InstructGPT), controllable text generation (specific types of questions), and few-shot learning to generate educational questions. Collect opinions / human evaluation from real-world teachers (biology, ML domains) to show that generations are both high-quality and useful.

Controlled Generation of Question Types

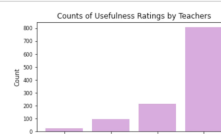


Results

Metric	Definition	Mean Across Domains
Relevance	Is the question related to the context provided?	97%
Grammar	Is the question grammatically correct?	96%
Adherence	Is the question an instance of the question type provided?	68%
Answerable	Is the answer in the input context?	92%

Results

Metric	Definition
Useful	Assume you wanted to teach about context X. Would you use candidate Y in a lesson, home work, quiz, etc? Rated as a scale from not useful (1) to useful as is (4).



▼ Using Decomposed Prompting to Answer Student Questions on a Course Discussion Board Jaipersaud, Jaipersaud, Brandon et al.

Responding to student questions on discussion forums can be a very time consuming process. This paper presents an approach to answer student questions by prompting GPT-3 using a prompting strategy known as **decomposed prompting**.

Different questions on discussion board may require different prompting strategies. They categorize questions into 4 categories (Conceptual, Homework, Logistics and Non-answerable). Using GPT-3, they first categorize the questions, then use a specialized prompt or ignore altogether. Afterwards, they manually annotate the GPT answers.

Their classification results showed an accuracy of 81%. The authors then manually annotate the model's answers to the question and label them either "bad" or "good" with "bad" being 20 answers and "good" being 8 answers. They further categorized the reasons for the "bad" answers to be belonging to

(Misclassified, Factually Incorrect, Outside of Course Scope, Other (such as misunderstanding, incoherence, incorrect assumptions, adding irrelevant information))

Investigating Patterns of Tone and Sentiment in Teacher Written Feedback Messages” Baral, S et al.

July 5th

▼ **“Learning from Auxiliary Sources in Argumentative Revision Classification”** **Afrin, Taxin et al.**

The objective of this research is to classify desirable reasoning revisions in argumentative writing. However, hand-annotated dataset are small in size and therefore difficult to use state-of-the-art models. Therefore, they take advantage of auxiliary data sources of revisions in solving this problem by utilizing multi-task learning for training on different sources of data and transfer learning for capturing the relationships between the data.

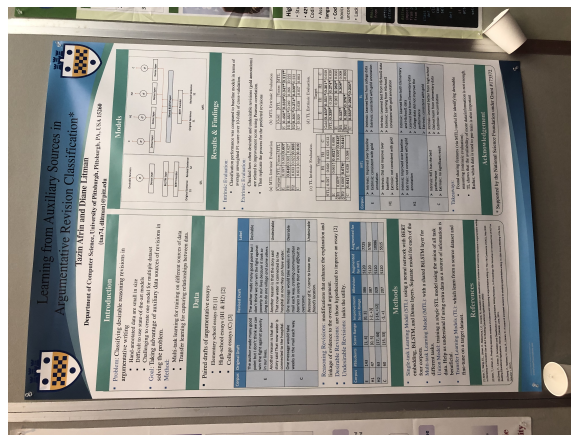
The data used in this study was paired drafts of argumentative essays from elementary, high school and college. They are written by students in response to a prompt and revised in response to feedback and finally graded with respect to a rubric.

They focus on reasoning revisions or those modifications that enhance the explanations and linkage of evidence to the overall argument. Desirable revisions are those that are hypothesized to improve an essay and undesirable ones are those that lack the utility.

They compare different types of models namely, Single-task Learning Model, Multi-task Learning Model, Union Model and Transfer Learning Models. For evaluation, they employ both intrinsic and extrinsic evaluation. In intrinsic evaluation, the classification performance was compared to baseline models in terms of average unweighted F1 score. In extrinsic evaluation, they checked how often desirable and undesirable revisions are related to improvement score and used Pearson correlation. The process is then replicated for the predicted revisions.

Takeaways: they found that sharing features via multi-task learning was useful for identifying desirable reasoning or intrinsic evaluation. Transfer learning showed that the availability of more data was not enough, rather, which data was used for pre-training was also important.

▼ **“Visualizing Self-Regulated Learner Profiles in Dashboards: Design Insights from Teachers”** Meija-Domenzain, Paola et al.



In flipped classrooms (FC), students do pre-class activities and then attend to interactive face-to-face sessions, these pre-class activities have been found to be essential for students' success. Therefore, it is essential to best assist and support them during these activities.

Their research questions were:

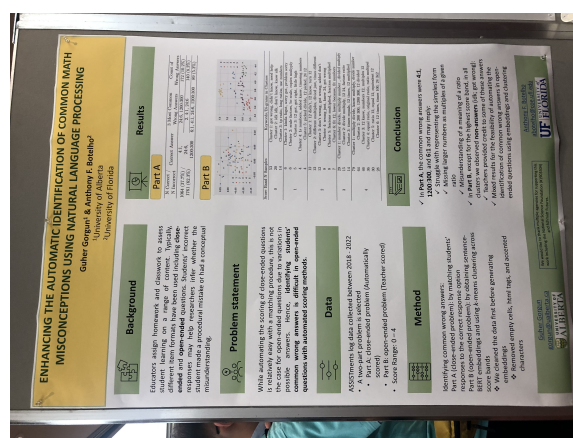
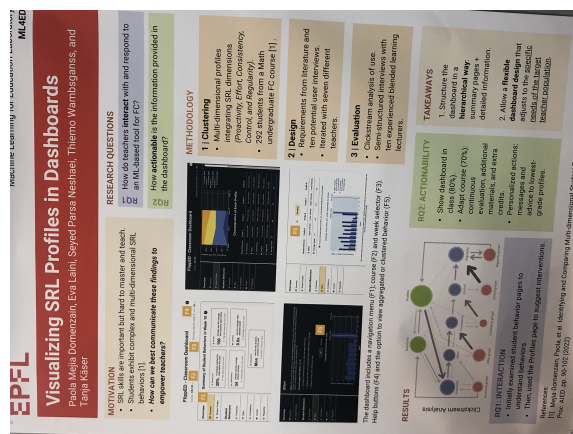
1. How do teachers **interact** with and respond to an ML-based tool for FC?
2. How **actionable** is the information provided in the dashboard?

To answer these questions, they cluster profiles of students that integrate SRL dimensions (proactivity, effort, consistency, control and regularity). Then, they design the dashboard with a teacher-centered approach and an iterative process. Finally, they evaluate using clickstream analysis of use in addition to semi-structure interviews with ten experienced blended learning lecturers. They found that teachers initially explore behavior pages to understand and then use summary and overview pages to suggest interventions. Among the interventions were:

- showing dashboard in class
- adapting course
- or personalized actions depending on the profiles

▼ “Enhancing the Automatic Identification of Common Math Misconceptions Using Natural Language Processing” Gorgun, G et al.

The objective of this research was to identify common wrong answers given in a two-part problem. The problem consists of a close-ended question followed by an open-ended one where students provide reasons or explanations for the close-ended part. Identifying wrong answers in the close-ended question is relatively easy where we can use a matching algorithm for example. However, for open-ended questions, it is a difficult problem due to the variations in possible answers. In this research, they obtained sentence BERT embeddings and then followed that by clustering using k-means. Their experimental results show that for the close-ended question,



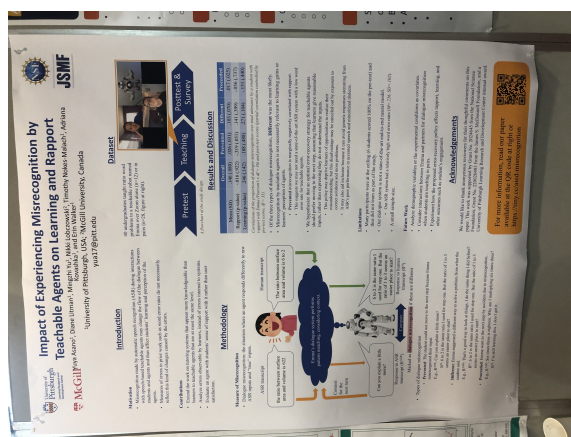
students may be struggling with representing simplest fraction form and misunderstanding of the meaning of a ratio. However, for the open-ended question, they found mixed results for the feasibility of automating the identification of common wrong answers using their method of embeddings and clustering.



July 6th

▼ **Impact of Experiencing Misrecognition by Teachable Agents on Learning and Rapport**, Asano, Yuya et al.

When you're talking to AI (Siri, Alexa etc) sometimes your speech gets mis-recognized which can be annoying and can affect your user experience. In this research they investigate the impact of mis-recognition in the context of tutoring dialogue. They automatically quantify the mis-recognitions observed by learners instead of the ones internal to systems and evaluate agents with students' sense of rapport with it rather than user satisfaction. To measure mis-recognition, they feed both the ASR transcript and the human transcript to the same dialog system given same context and compare the output between both. If responses are different, it is marked as a dialog mis-recognition. Their results show that dialogue mis-recognition by teachable agents is not necessarily relevant to learning gains or learners' rapport with agents.



▼ **Emotionally Adaptive Intelligent Tutoring System to Reduce Foreign Language Anxiety** Ismail, Daneih et al.

When learning a foreign language, feelings of stress, worry and apprehension can accumulate and negatively impact learners cognitively, psychologically and physiologically. For this reason, they build an emotionally adaptive intelligent tutoring system that is able to detect anxiety and provide intervention as needed. They use a random forest chain regressor, and sensor free human behavior metrics to predict

anxiety and change. The interventions include motivational supportive feedback presented by an agent via voice or text and explanatory feedback presented via the same modalities. They found that adaptive feedback reduces anxiety more than fixed strategies. However, they didn't find significant differences in learning gains.

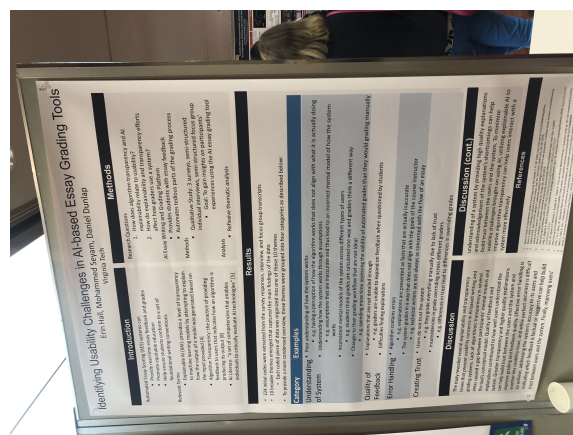
▼ **“Identifying Usability Challenges in AI-based Essay Grading Tools” Hall, E et al.(virtual)**

Automated essay scoring systems (AES) has made it possible to provide real-time feedback and grades for student essays, thus promoting equitable education and creating a level playing field by ensuring that all students conform to a set of foundational writing conventions.

However, two main questions arise in this context:

1. How does explainability and transparency relate to usability?
2. How do explainability and transparency efforts affect how graders use a system?

They conducted a qualitative study to gain insights on participants' experiences using an AI essay grading tool. 10 broad themes emerged that captured the findings of their study. These findings revealed several transparency and explainability issues that present usability concerns.



Networking Opportunities

From distinguished professors with years of experience in the field, to passionate students pursuing their PhD, AIED23 served as a golden opportunity to meet great people, make deep connections and share insightful conversations. I established connections with several researchers who expressed interest in potential collaborations and knowledge-sharing opportunities in the future. The conference utilized the Whova platform for virtual participation which further facilitated establishing connections.

In conclusion, the conference successfully achieved its goal of fostering discussions and sharing knowledge about the integration of AI in education. It deepened my understanding of the challenges, opportunities, and ethical considerations associated with this emerging field.



References

1. Wiegrefe, Sarah, and Ana Marasović. "Teach me to explain: A review of datasets for explainable natural language processing." *arXiv preprint arXiv:2102.12060* (2021).
2. Zhang, Zhe Victor, and Ken Hyland. "Fostering student engagement with feedback: An integrated approach." *Assessing Writing* 51 (2022): 100586.
3. Funayama, Hiroaki, et al. "Balancing cost and quality: an exploration of human-in-the-loop frameworks for automated short answer scoring." *International Conference on Artificial Intelligence in Education*. Cham: Springer International Publishing, 2022.
4. Sato, Tasuku, et al. "Plausibility and faithfulness of feature attribution-based explanations in automated short answer scoring." *International Conference on Artificial Intelligence in Education*. Cham: Springer International Publishing, 2022.
5. Hartmann, Mareike, and Daniel Sonntag. "A survey on improving NLP models with human explanations." *arXiv preprint arXiv:2204.08892* (2022).APA
6. Keller, J.M. (2012). ARCS Model of Motivation. In: Seel, N.M. (eds) Encyclopedia of the Sciences of Learning. Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_217
7. Rebolledo-Mendez, Genaro, et al. "Meta-affective behaviour within an intelligent tutoring system for mathematics." *International Journal of Artificial Intelligence in Education* (2022): 1-22.