UMassAmherst
The Commonwealth's Flagship Campus

How are you feeling? Using AI to Improve Human Learning

Beverly Woolf, Danielle Allessio, Ivon Arroyo, Margrit Bretke, Hao Yu,

Dec 2024

Research Questions

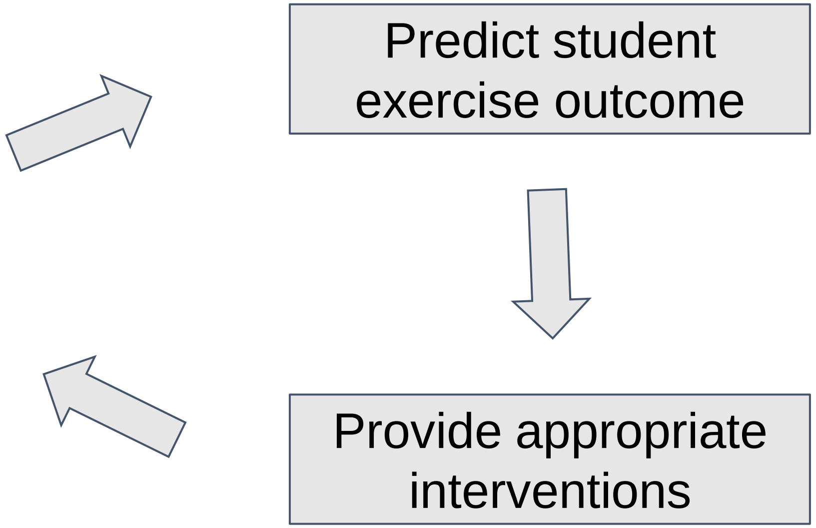
How can computer tutor systems:

- Perform continuous assessment of student outcome, behavior and emotion?
- Create synthetic partners and personalize content/responses?
- Detect and respond to student emotion?
- Create avatars to answer student queries?
- Respond to student behavior, e.g., make mistakes, asks for hints ?
- Tailor messages?

Objective

- Detect and interpret student behavior;
- Respond “just-in-time” to students’ states.

Students solving math exercises online



Agenda

Motivation

Continual Assessment

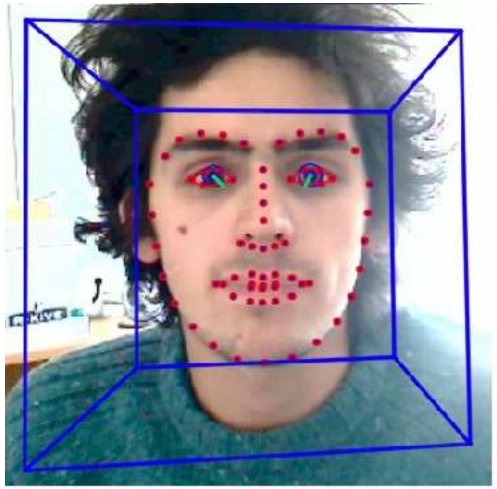
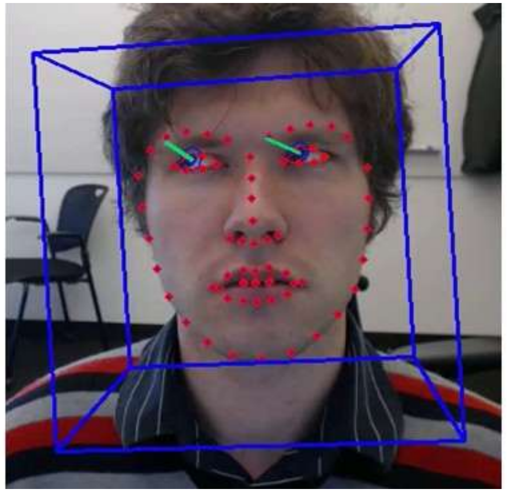
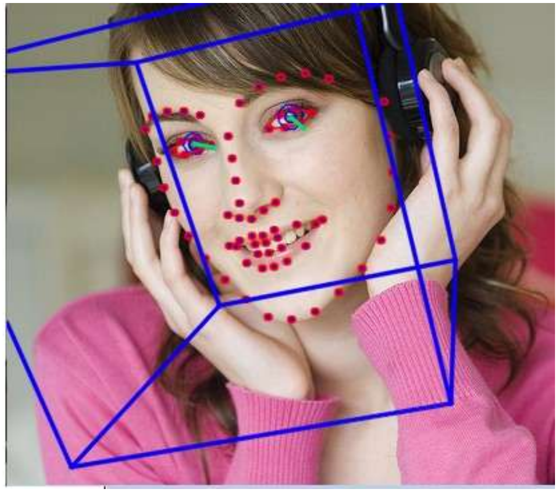
Synthetic and Conversational Partners

Predict Learning

Automate Instructional Content

Conclusions

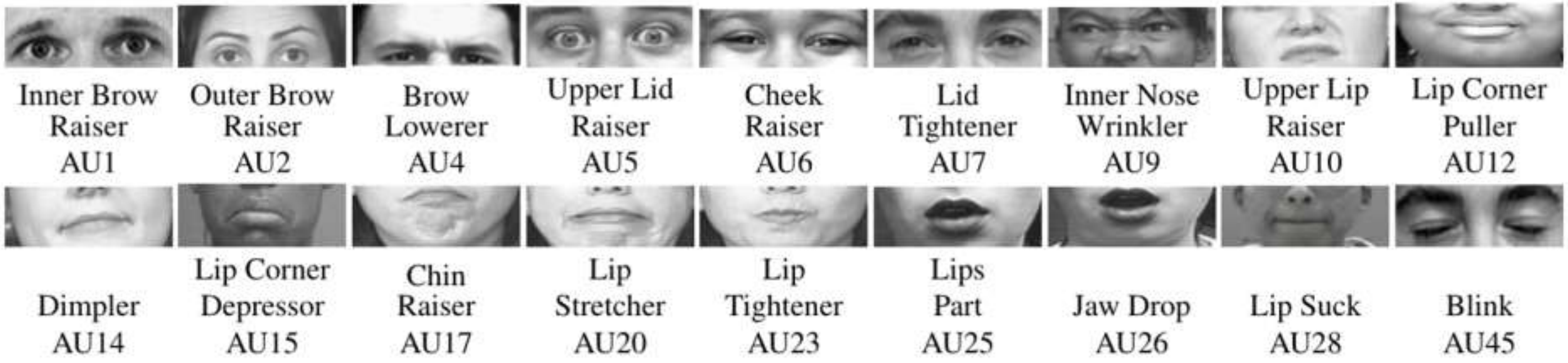
Traditional Facial Features

- Gaze direction
- Head pose
- Facial Action Units
- Facial landmarks

Video Footage Showing the Student's Face

Facial Action Units (from OpenFace 2.0)



MathSpring Children Dataset

- Fiftyone sixth-grade male and female students
- 968 videos of a student working on a single exercise
- We extracted the timing information for each exercise from the log data
- Each video is labeled with an outcome label based on
  - success rate (mistakes made)
  - hints requested
  - time excerpted during problem-solving

Image: [Ruiz, et al., ATL-BP: A Student Engagement Dataset and Model for Affect Transfer Learning for Behavior Prediction, TBIOM 2023]

METHODOLOGY: Affect Model to Predict Outcome

We developed an affect model that predictd student exercise outcomes at an **early** stage by analyzing only the first 5 (or up to 20) seconds of student data.

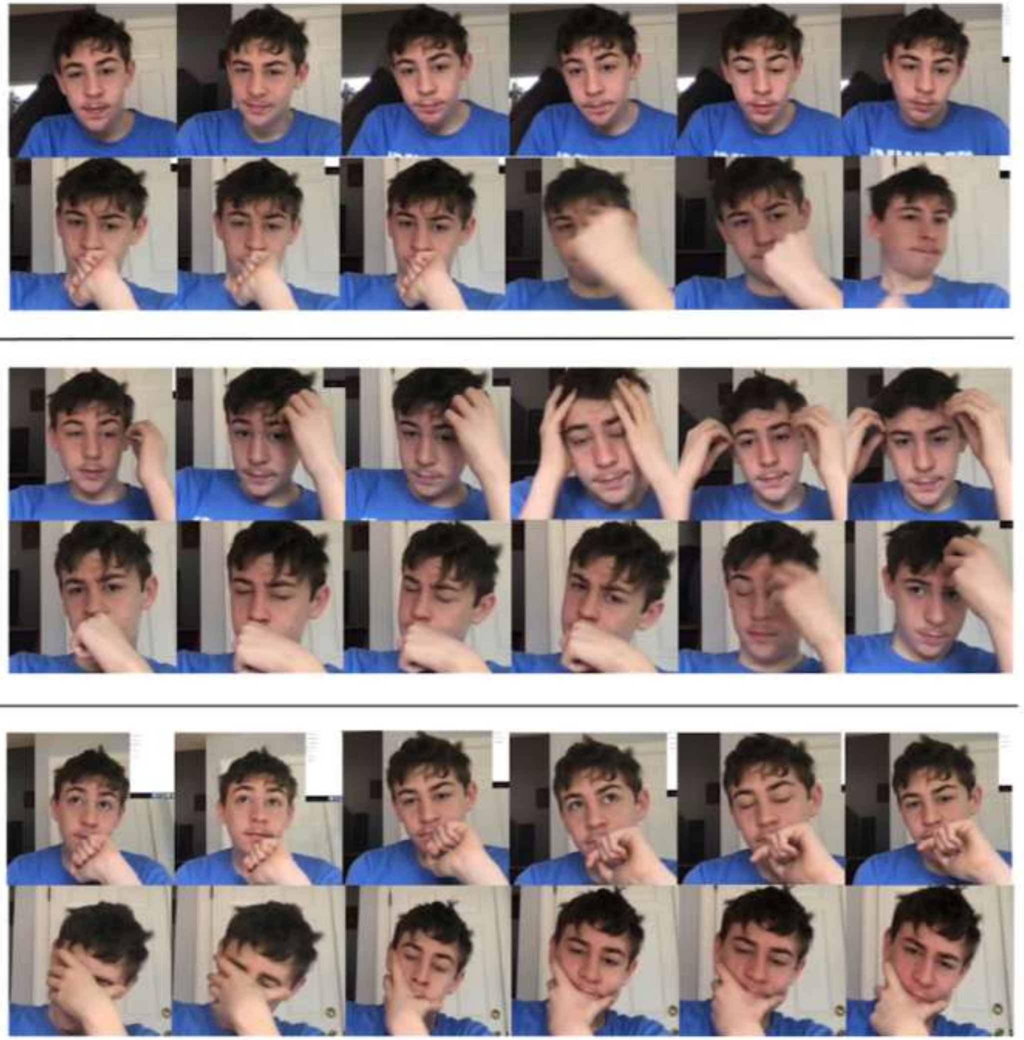
- The **multimodal** system augments video representation with timing information obtained from students' learning log data.
- The model incorporate state-of-the-art facial affective embeddings and an affect-aware Transformer to improve the performance of outcome prediction.

Student Effort Outcome Goal: Label Frames

SOF

SHINT

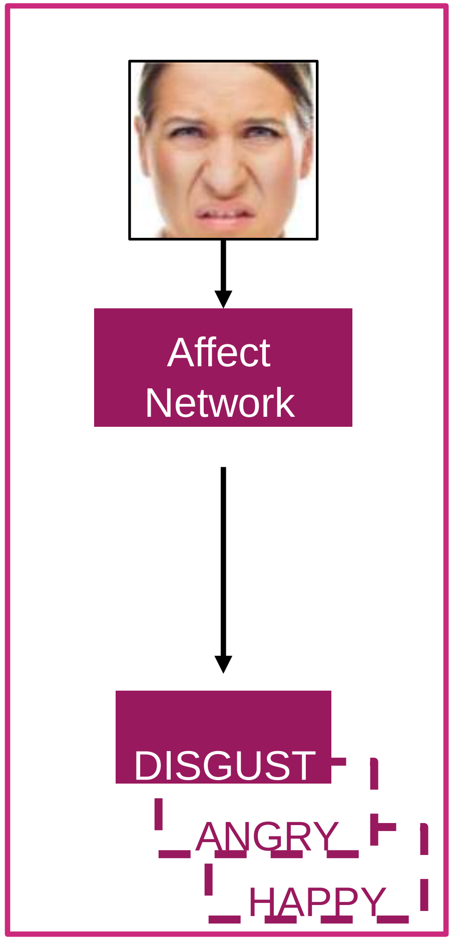
GIVEUP



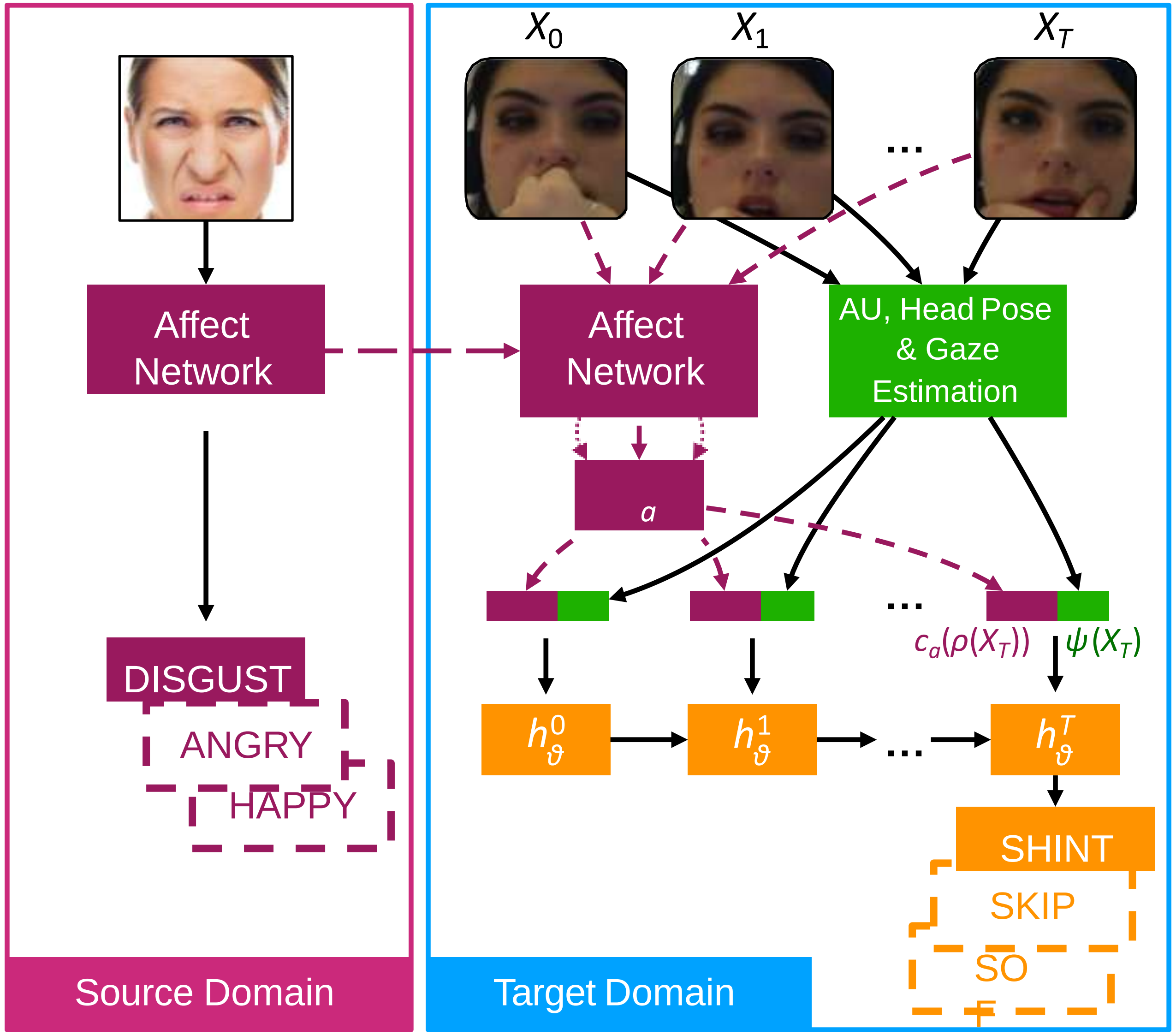
EFFORT CHART

POSITIVE	SOF:The student SOLVED the problem correctly on the FIRST attempt, without any help.
	ATT:The student ATTEMPTED once incorrectly, but self-corrected (answered correctly) in the second attempt, no help request.
	SHINT:Student SOLVED problem correctly after seeing HINTS.
	SHELP:The student got the problem correct but saw at least one video.
NEGATIVE	GUESS:The student apparently GUESSED, clicked through 3-5 answers until getting the right one, and did not ask for hints/videos etc.
	NOTR:NOT even READING the problem --The student answered too fast, in less than 4 seconds
	SKIP:The student SKIPPED the problem (did not do anything on the problem).
	GIVEUP:The student started working on the problem, but then GAVE UP and moved on without solving it correctly.

Affect Features



Affect Transfer Learning



Emotion Influences Learning

Literature shows that:

- Positive emotions (e.g., confidence) improve performance;
- Negative emotions (e.g., low self-confidence) can reduce performance;
- Boredom reduces performance;
- Student interest supports learning in general.

D. Goleman. Emotional intelligence. why it can matter more than fq. *Learning*, 24(6):49–50, 1996.

R. Pekrun, T. Goetz, L. M. Daniels, R. H. Stupnisky, and R. P. Perry. Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, 102(3):531, 2010.

Methodology: Detect Head Direction

A deep neural network recognized the <u>direction of students</u>' <u>heads</u> as they use a tutor, student's gaze;

If students are not looking directly at the screen, they might be distracted and the tutor can bring their attention back.

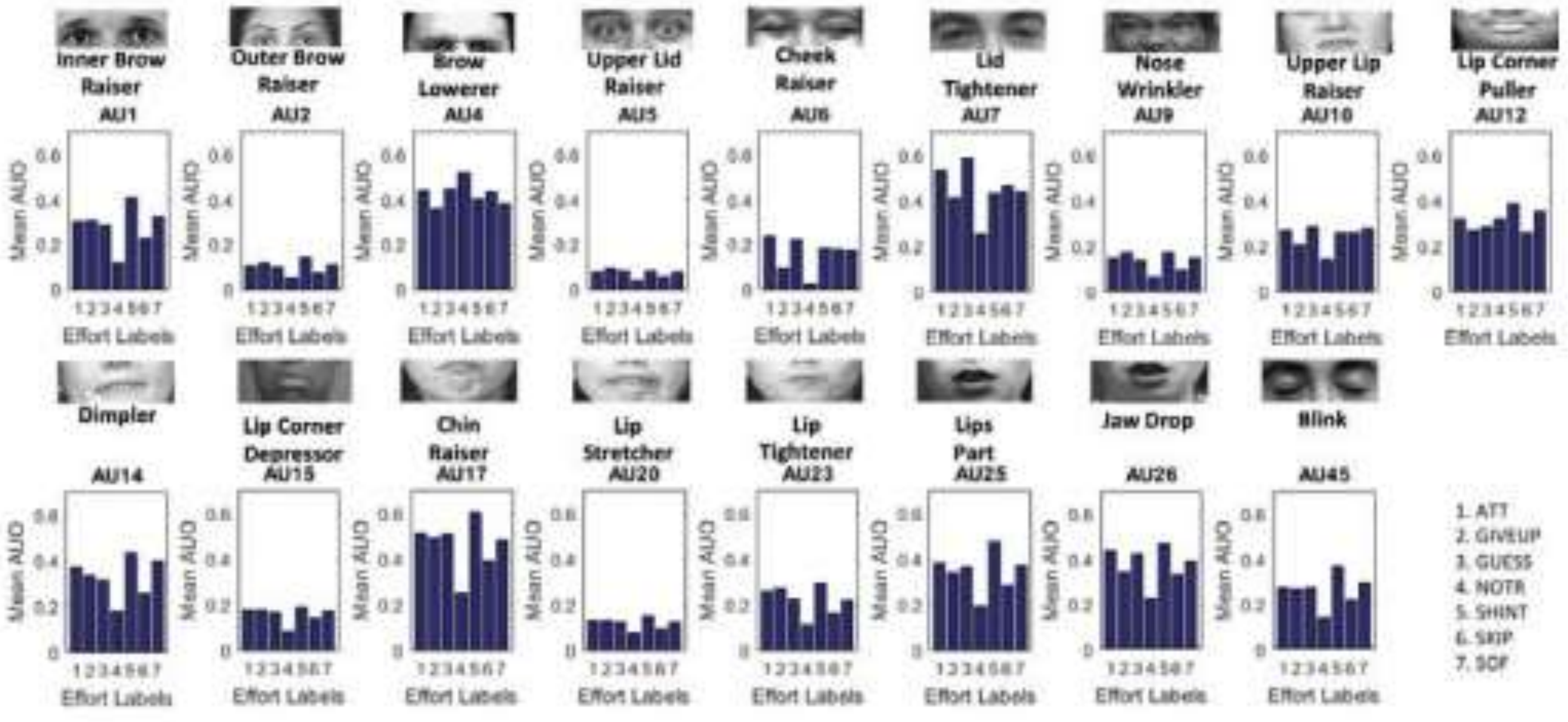
Detect Head Movement

A fully connected neural network was trained to detect the presence and intensity of three action units for smile, nose wrinkle and frown (AU4, AU9, and AU12, respectively);

Face detection, face extraction, and facial landmark detection are all performed using face-api.js with Tensorflow.js;

The model was trained on a dataset of 35K face images labeled with 68 face landmark points. The input of the model consists of facial landmarks of eyebrows, eyes, nose, and mouth. The output contains the confidence predictions of the 3 facial action units.

Action Unit According to Effort



X-axis => Student learning outcome classes (ATT, GIVEUP, GUESS, NOTR, SHINT, SKIP, SOF)
Y-axis => Mean AUO score.

Grayscale images depict the AUs, https://www.cs.cmu.edu/?face/facs.htm

Children Datasets

Row #1: Dartmouth database.
Row #2: Child Affective Facial Expression (CAFE).
Row #3: LIRIS-CSE.
Column #1 "happiness"; #2 "surprise"; #3 "sadness"; #4 "anger"; #5 "fear"; and #6 "disgust".

CAFE Data Set

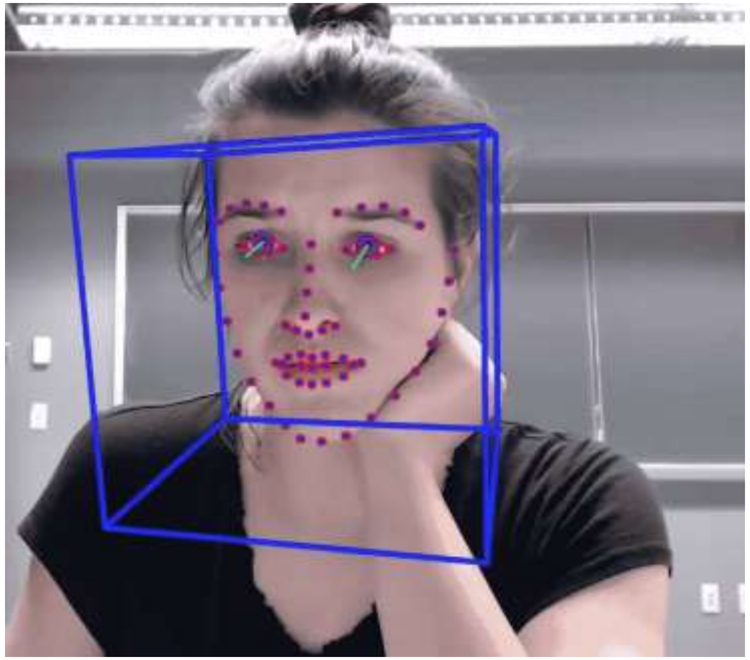
Khan, R. A., Crenn, A., Meyer, A., & Bouakaz, S. (2019). A novel database of children's spontaneous facial expressions (LIRIS-CSE). *Image and Vision Computing*, *83*, 61-69.

K. A. Dalrymple, J. Gomez, B. Duchaine, The dartmouth database of childrens faces: Acquisition and validation of a new face stimulus set, PLOS ONE 8 (11) (2013) 1–7. doi:10.1371/journal.pone.0079131. URL https://doi.org/10.1371/journal.pone.0079131

L. V., T. C., The child affective facial expression (CAFE) set: validity and reliability from untrained adults, Frontiers in Psychology 5 (1532). doi:http://doi.org/10.3389/fpsyg.2014.01532.

Feature Extraction using OpenFace 2.0

- OpenFace 2.0 is an opensource tool for facial analysis
  - 68 facial landmarks;
  - Head orientation;
  - Gaze direction;
  - Eye detection;
  - 18 Facial action units presence;
  - 17 facial action units intensities.



We concatenate all of the above as a feature vector of length 8241 and pass it to an LSTM-based RNN.

Agenda

Motivation

Continual Assessment

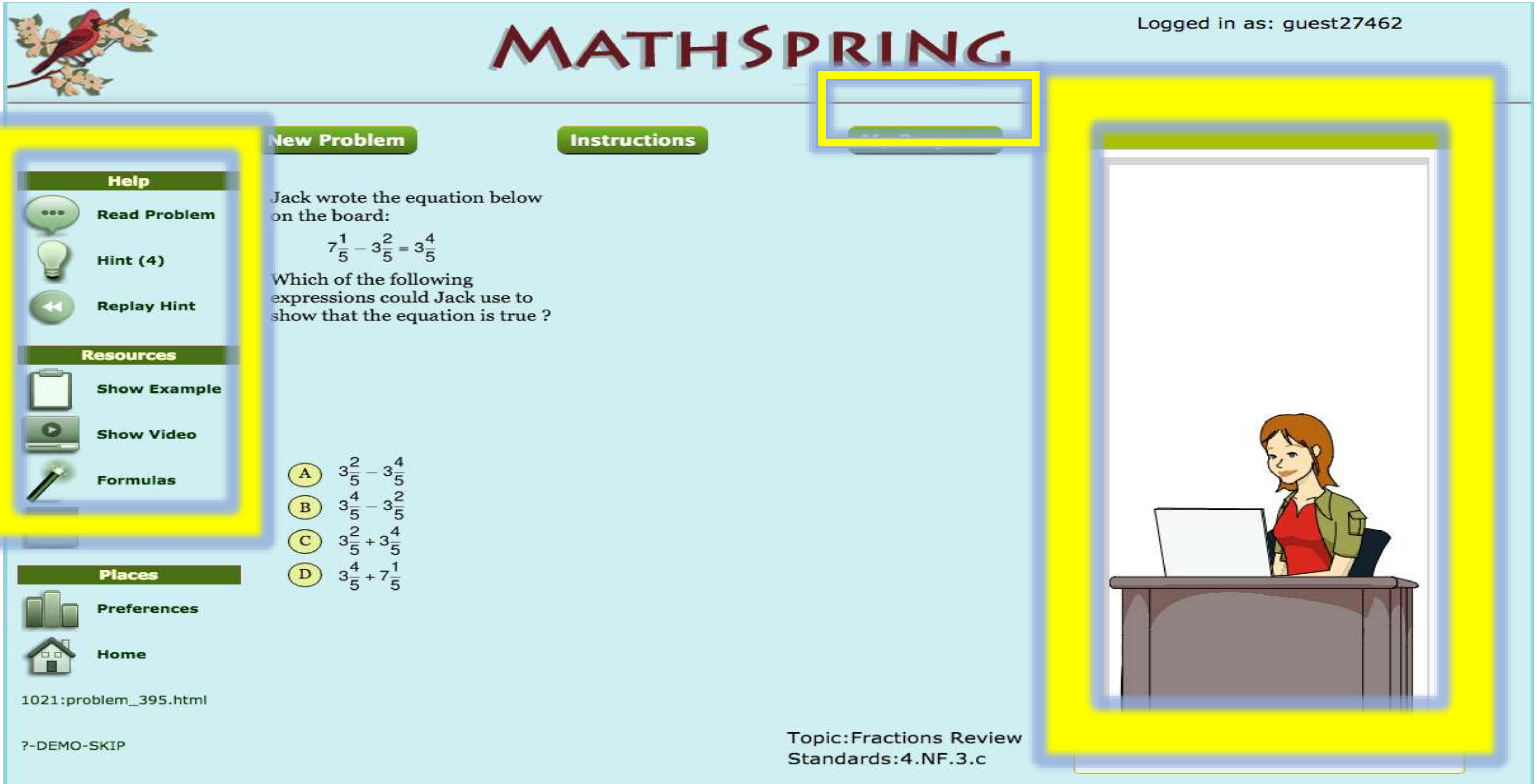
Synthetic and Conversational Partners

Predict Learning

Automate Instructional Content

Conclusions

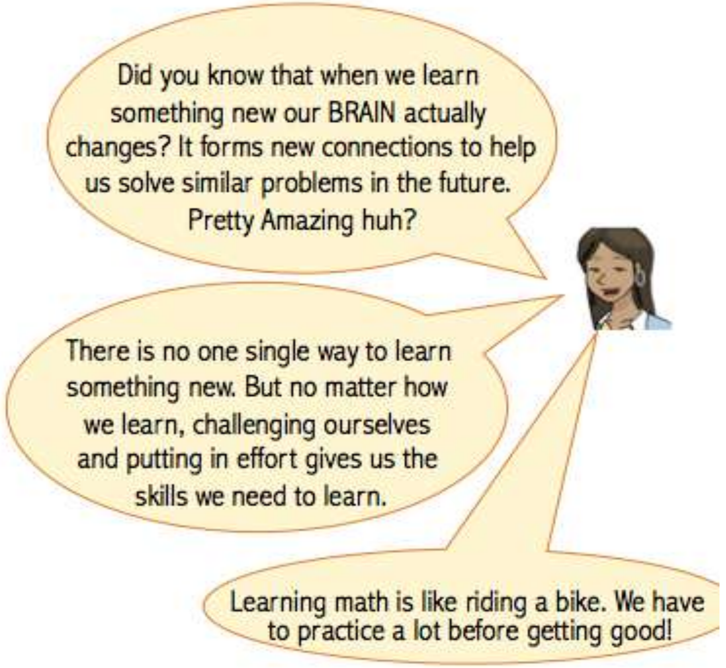
The MathSpring Tutor



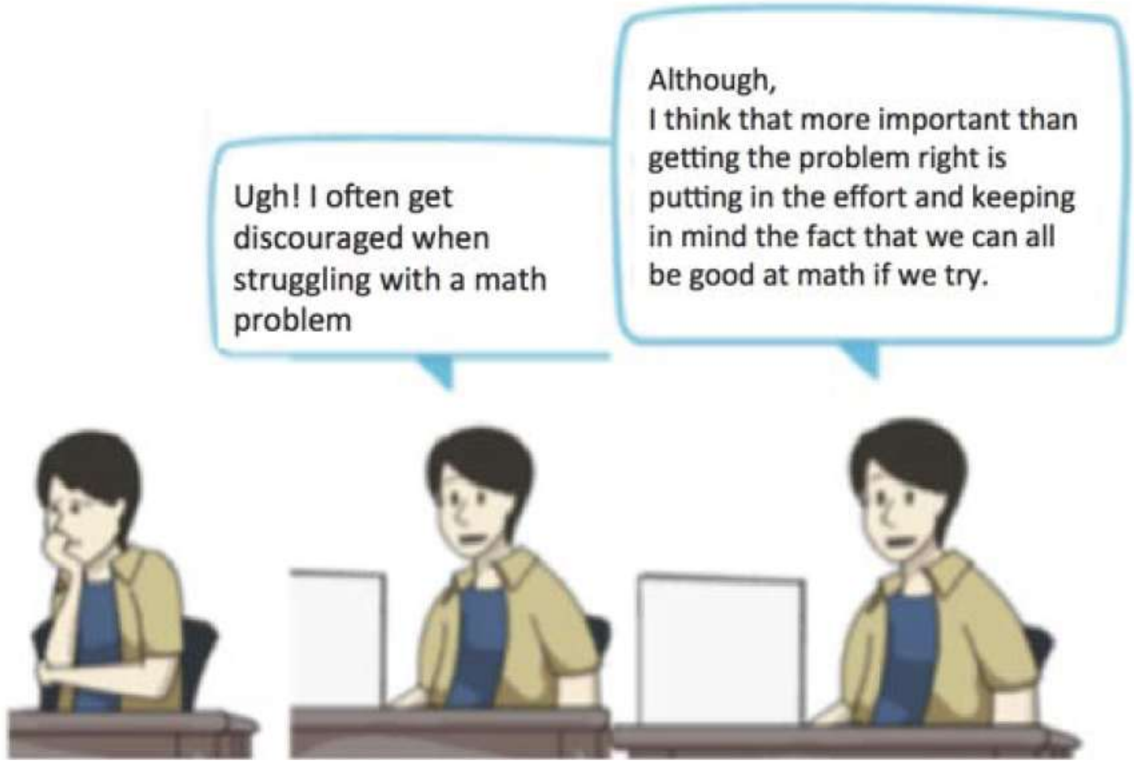
Growth Mindset

- Students who believe that intelligence can be increased tend to seek out academic challenges;
- Growth Mindset messages led to gender differences and more problems right on students' first attempt.
- D'Mello found successful results using Empathy,
  (D'Mello et al., 2010).

Animated Characters



Companions provide growth mindset messages, encouraging students to put in effort to become good at math.



(Left): Visual acknowledgement of anxiety
(Middle): verbal acknowledgement
(Right): connector and resolution via growth mindset message

Three categories of responses

Empathy	"Don't you sometimes get frustrated trying to solve math problems? I do."
Growth Mindset	"Did you know that when we practice to learn new math skills our brain grows and gets stronger?"
Success / Failure	"Very good, we got another one right!"

Data

- N = 61 students (grade 6);
- All students completed a pretest and posttest;
- 21066 event logs in total;
- Recorded for 3 classes over 4 separate days;
- Total messages received per condition – 978 (empathy), 763 (growth mindset) and 882 (success/failure).

Results

Interest - Students' interest in math. "Are you interested when solving math problems?"
Excitement - How exciting students find math. "Do you feel that solving math is exciting?"
Confusion - How confused students feel while solving math problems. "Do you feel confid
eventually be able to understand the Mathematics material?"
Frustration - How frustrating students find math. Average of "Do you get frustrated when s
lems?" and "Does solving math problems make your feel frustrated?"
Learning Orientation - How much students focus on learning as opposed to performance.
you are doing math exercises, is your goal to learn as much as you can?" and "Do you prefer lea
that make you curious even if that means you have to work harder?"
Performance Approach Goals - "Do you want to show that you are better at math than y
Math Value - How important do students think math is. "Compared to most other activities,
it or you to be good at math?"
Math Liking - Measure of how much students like math. "Do you like your math class?"
Math Test Performance - Student's score on math questions that are representative of the
MathSpring.

Partial Correlations (Posttest) Result

Variable	Empathy Messages		Growth Mindset Messages		Success/Failure Messages	
	corr	p	corr	p	corr	p
Interest	**0.28***	0.03	0.19	0.15	-0.20	0.14
Confusion	-0.05	0.74	-0.05	0.74	**0.32***	0.02
Learning Orientation	0.02	0.85	0.02	0.88	**-0.24**$^{+}$	0.06
Math Value	**0.25***	0.05	**-0.22**$^{+}$	0.09	-0.10	0.45
Performance	-0.01	0.93	**-0.23**$^{+}$	0.07	-0.13	0.33

$^{+}\ p \leq 0.10$, $^{*}\ p \leq 0.05$

Partial correlations between different types of messages seen and posttest variables, accounting for the corresponding pretest value, time spent in tutor and message frequency.

Observations

- With exposure to more empathic messages, students exhibited higher levels of interest and valued math knowledge more;
- With Growth Mindset messages, students got more problems correct on their first attempt, yet valued math knowledge less and had lower post test performance scores;
- With success/failure messages, students were less learning-oriented and more confused.

Animated Characters

- Empathic responses (frustration and anxiety) are associated with higher student learning;
- Empathic responses reduce student anxiety and boredom;
- Success/failure messages produce boredom and anxiety;
- Growth Mindset messages induce learning/mastery/growth orientation goals.

Mediating factors: changes in behavior (slow down problem solving, become more careful, seek hints).

Agenda

Predict Student Outcome and Emotion

- As students work online, they demonstrate engagement and emotions (e.g., confusion, boredom, excitement);
- Information about engagement and emotion aids in understanding students’ progress, suggesting when and who needs further assistance.

Predict Problem Solving Outcome

- Students solve math problems online;
- For each problem their outcome is predicted
- **Eight** possible behaviors/efforts:
    - Attempted
    - Gave up
    - Guessed
    - Not read
    - Solved with hint
    - Skipped
    - Solved on first attempt
    - Solved with help

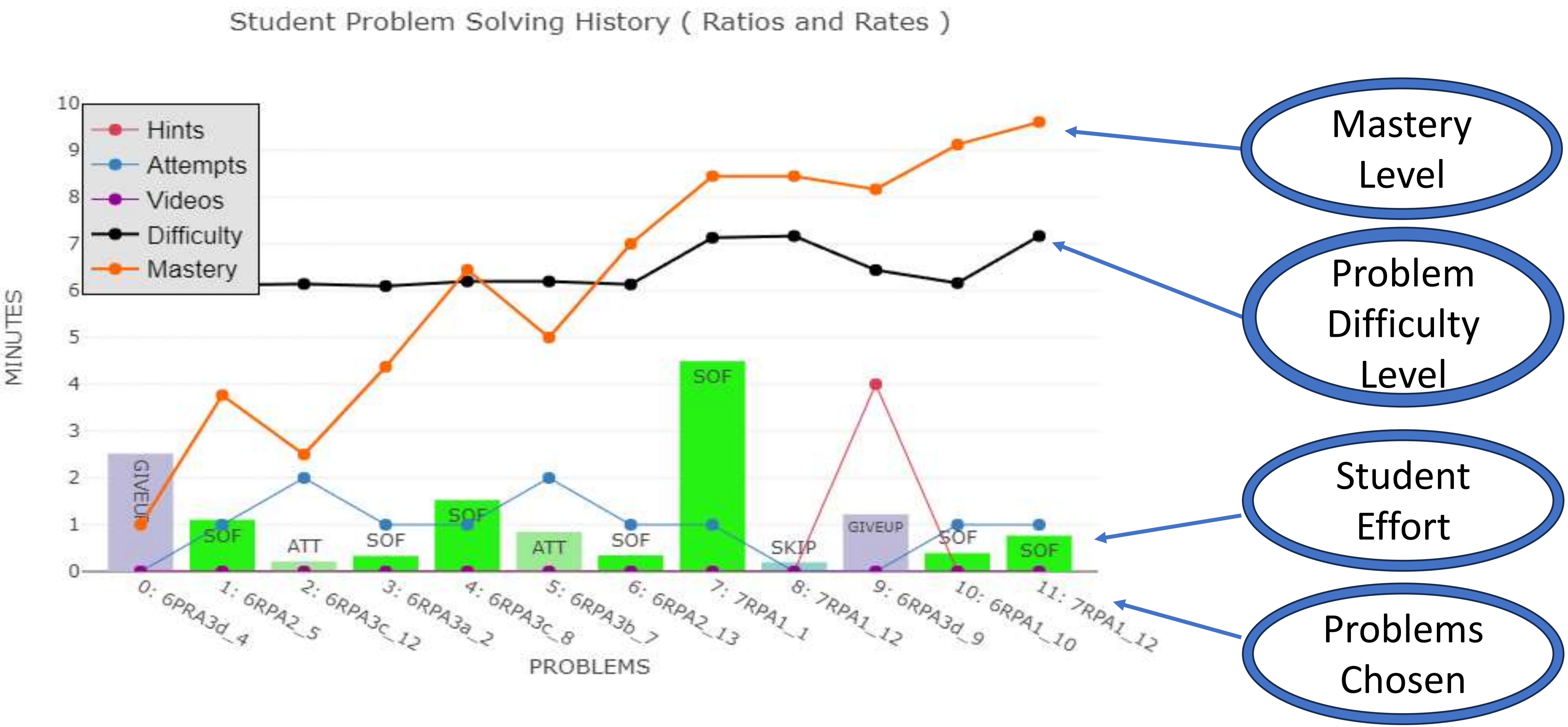
- Use of deep learning for facial features to **predict** the student's outcome

Use AI to Select Problem Difficulty

- This model predicts the problem outcome of a student given any difficulty level;

- This model is used to find a optimal problem difficulty level that leads to a positive learning outcome for the next problem.

Keep Students Motivated by Choosing Good Problems



Problem Difficulty

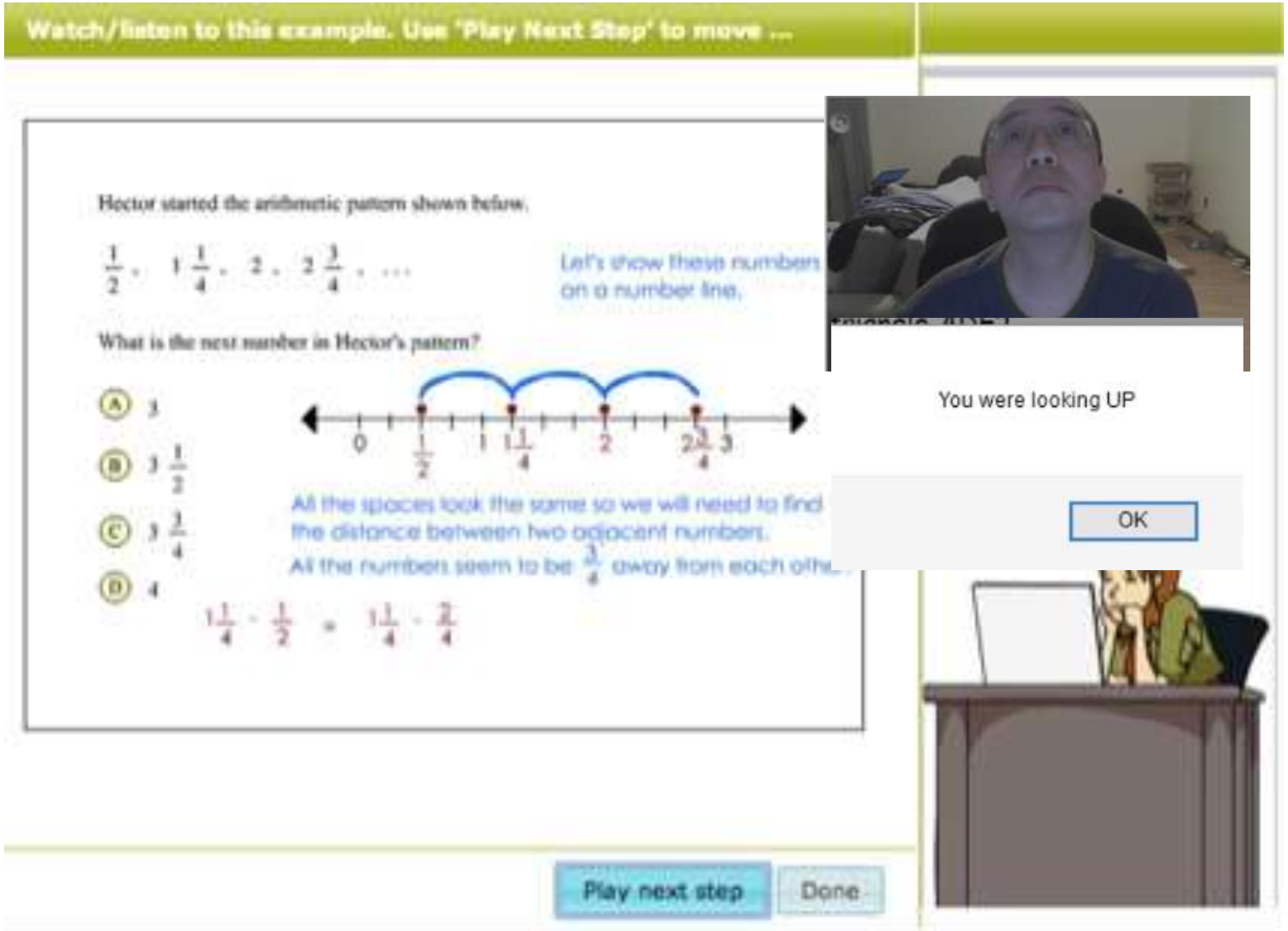
- Given candidate difficulty levels for the next problem, run the model for each difficulty to obtain the corresponding problem outcome and select one with a positive problem outcome.

- Select the level that yields a positive problem outcome based on prediction results and a problem with a specific difficulty level presented as the next problem.

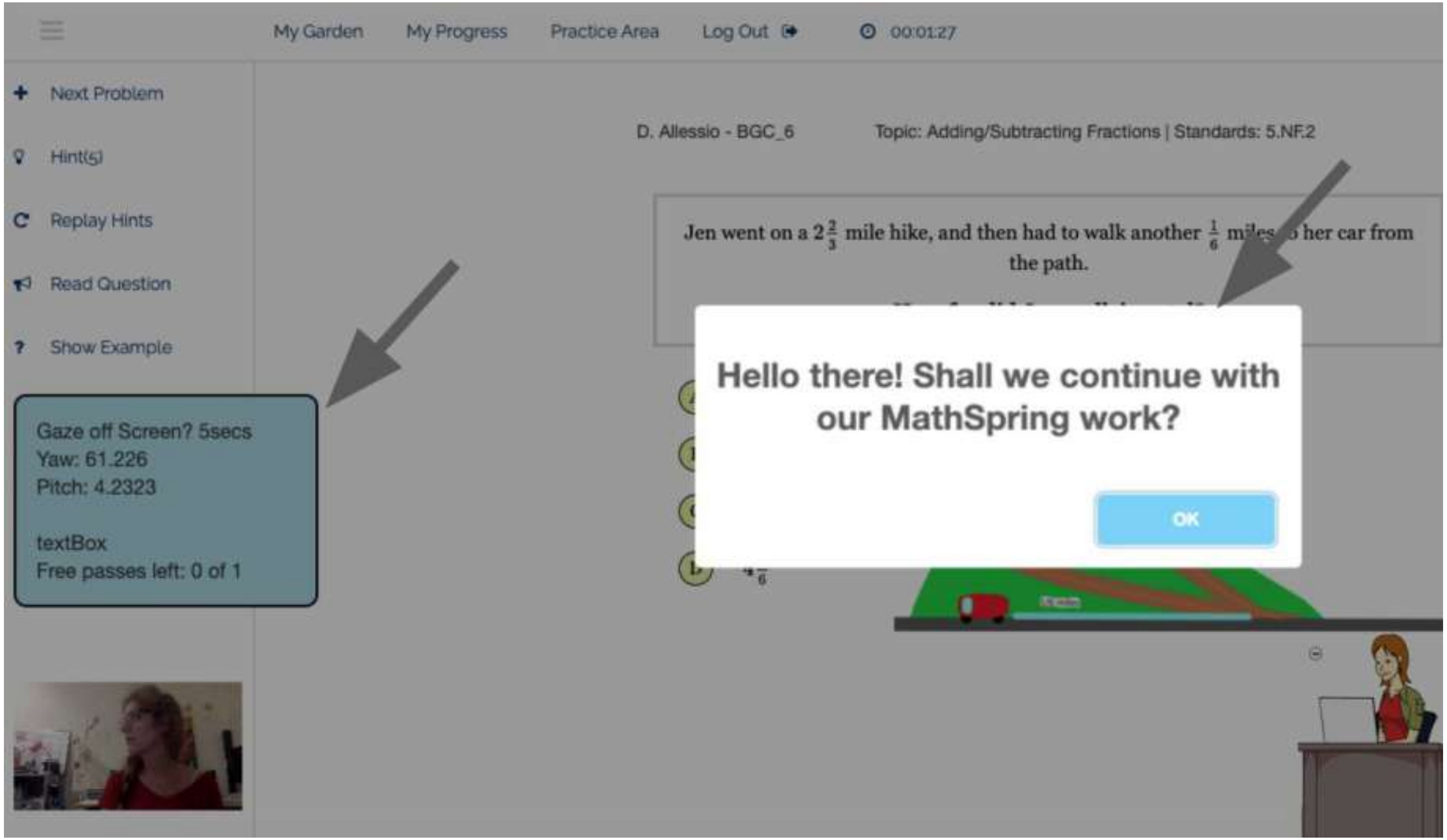
Predict Student Attention & Head Movement

- Recognize the direction of students' heads as they use a tutor, student's gaze;
- If students are not looking directly at the screen, they might be distracted and the tutor can bring their attention back.

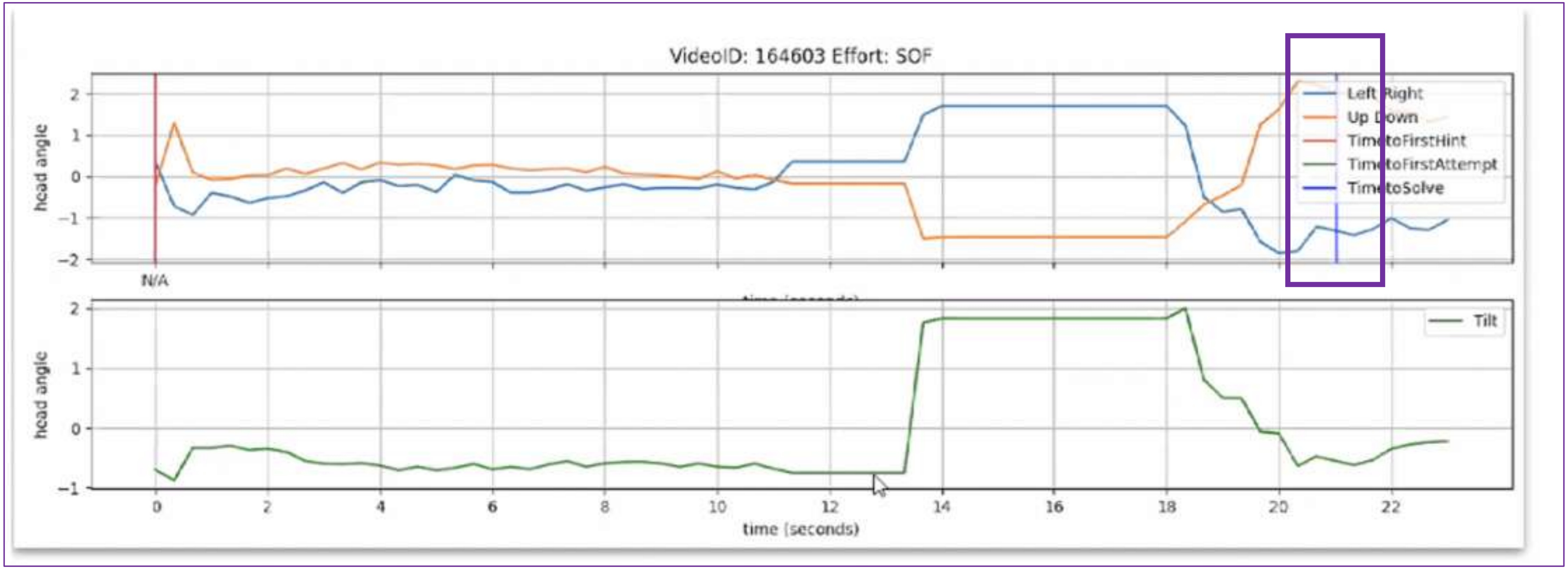
Detect Head Movement



Recognize Student Wandering



Record the Location of a Student’s Head



<u>Student solved this problem on first attempt</u>: Lines indicate left/right head position (blue), up/down (orange). The second chart shows head tilt (green).

Student solved the problem (vertical blue line, right).

Predict Student Attention: Head Movement

- A fully connected neural network was trained on a dataset of 35K face images labeled with 68 face landmark points. The input of the model consists of facial landmarks of eyebrows, eyes, nose, and mouth. The output contains the confidence predictions of the 3 facial action units.

- Face detection, face extraction, and facial landmark detection are all performed using face-api.js with Tensorflow.js;

Methodology: Detect Head Wandering

- Sixty-eight (68) undergraduates used MathSpring.org, and their faces were videotaped via laptop webcam, resulting in 2822 recorded problem-solving interaction samples;
- To assess model detection accuracy, we ran head pose estimation on the MathSpring videos with the student facing up, down, left, right or facing straight at the screen (340 images);
- Two human coders labelled the images and achieved 97.4% accuracy when comparing the human annotations and system predictions.

Summary

In sum, facial and gesture recognition is combined with self-reports to detect student emotion and expression;

The combination of head position and facial expressions and logged problem-solving activity provides invaluable data to assess and address student performance.

Agenda

Prompt for ChatGPT

"You are a math teacher with 10 years of experience teaching math to K12 learners from diverse backgrounds including a variety of socio-economic, cultural, and languages in the US. I will give you a math word problem, a multiple-choice question or a short answer type, along with the hints needed to solve the problem. I need you to generate the same number of hints as given in the problem. How would you improve each hint from the perspective of the math teacher? I will also give you what grade each problem is from. My first math word problem is {}"

Gattupalli, S.S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B.P. and Woolf, B., 2023. Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning. In *LLM@ AIED* (pp. 151-162).

Source: GitHub Prompt Engineering repo

Generate Hints

Transformer-generated hints produced via prompt engineering techniques.

Can significantly benefit students grappling with mathematical word problems.

Teachers must be involved in reviewing hints, whether human-crafted or transformer-generated.

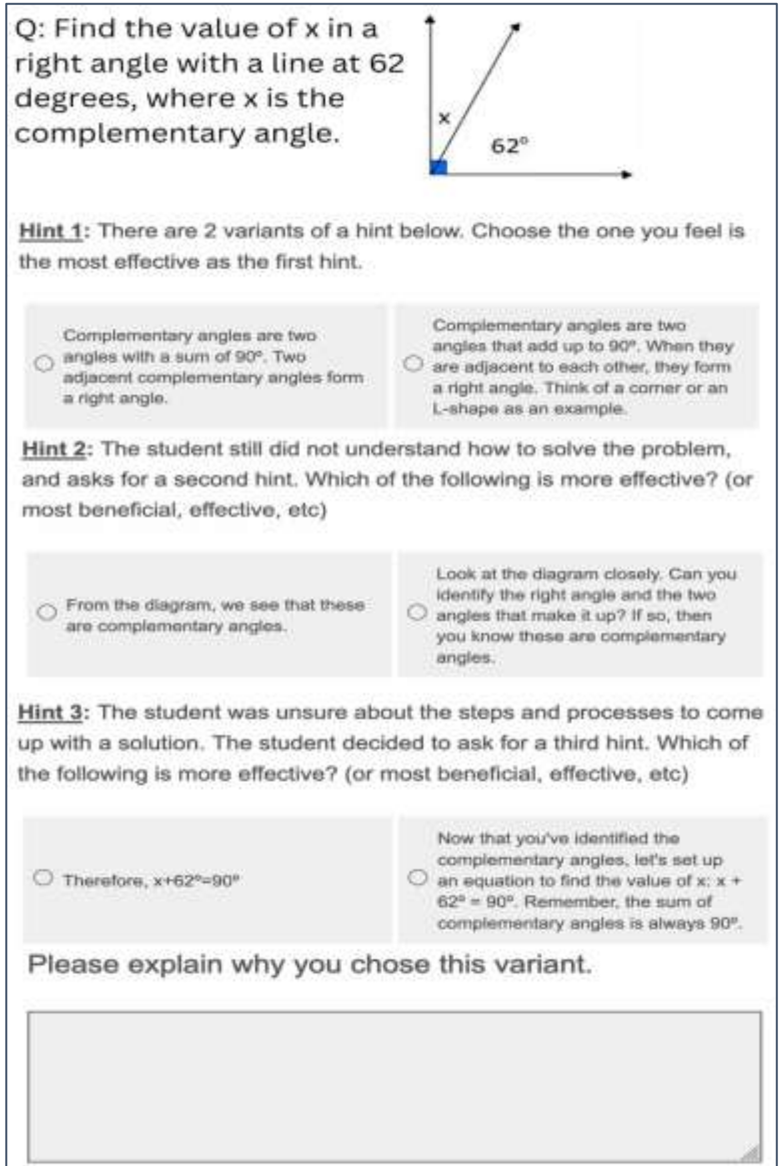
Gattupalli, S.S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B.P. and Woolf, B., 2023. Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning. In *LLM@ AIED* (pp. 151-162).

Methodology: Evaluate AI Generated Hints

- N= 33; pre-service teachers
  - Survey collected perceptions and responses for 5 random MS math word problems
  - Human-crafted vs. transformer-generated hints; crafted by real human teachers, the other was generated by GPT-4
  - Also answered "Why did you choose this hint variant?"
- To eliminate any potential biases, the participants were not aware of whether the hints were human-created or transformer-generated

Gattupalli, S.S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B.P. and Woolf, B., 2023. Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning. In *LLM@ AIED* (pp. 151-162).

https://osf.io/t84v7/



Q: Find the value of x in a right angle with a line at 62 degrees, where x is the complementary angle.

Hint 1: There are 2 variants of a hint below. Choose the one you feel is the most effective as the first hint.

- Complementary angles are two angles with a sum of 90°. Two adjacent complementary angles form a right angle.
- Complementary angles are two angles that add up to 90°. When they are adjacent to each other, they form a right angle. Think of a corner or an L-shape as an example.

Hint 2: The student still did not understand how to solve the problem, and asks for a second hint. Which of the following is more effective? (or most beneficial, effective, etc)

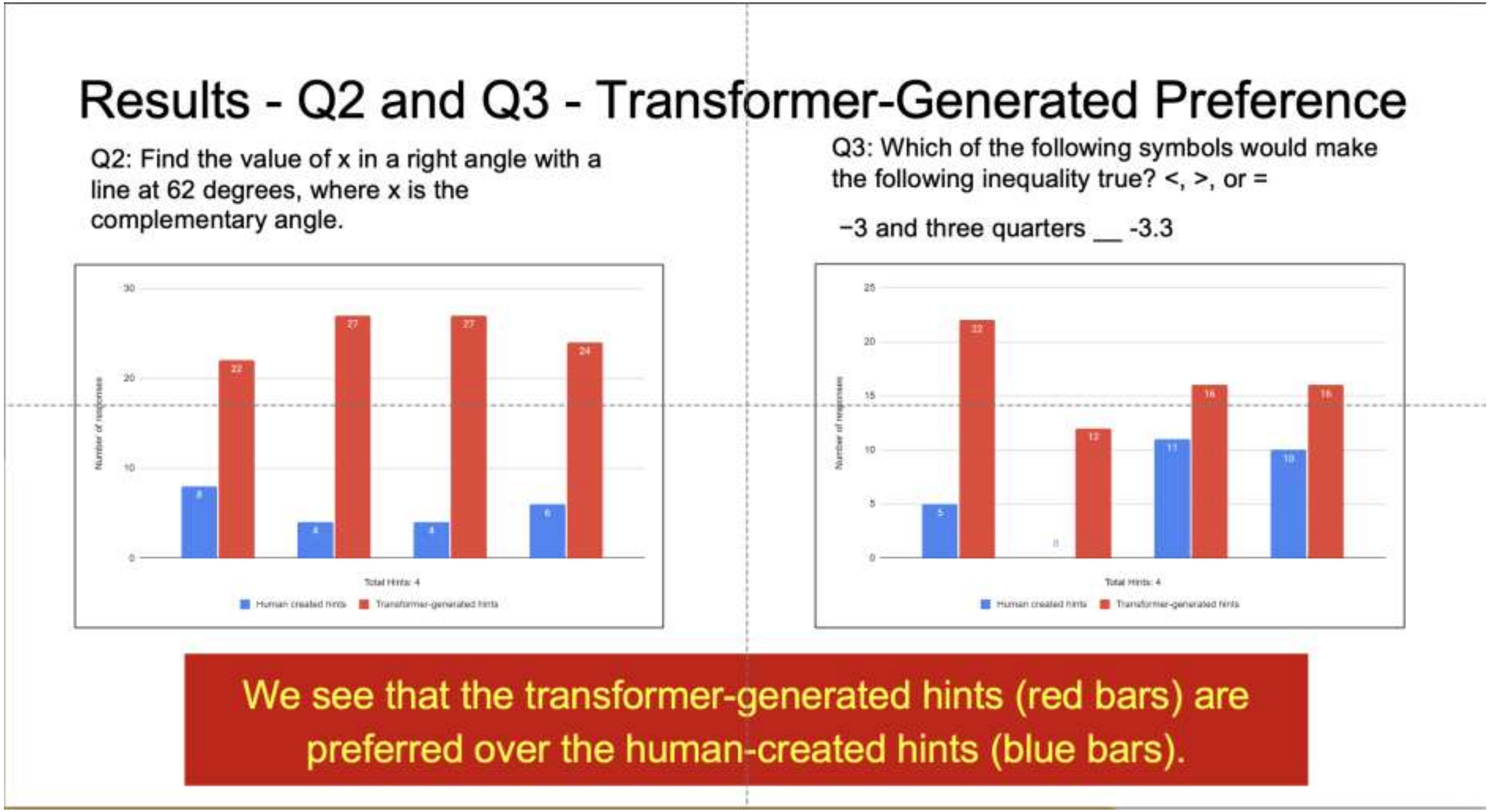
- From the diagram, we see that these are complementary angles.
- Look at the diagram closely. Can you identify the right angle and the two angles that make it up? If so, then you know these are complementary angles.

Hint 3: The student was unsure about the steps and processes to come up with a solution. The student decided to ask for a third hint. Which of the following is more effective? (or most beneficial, effective, etc)

- Therefore, x+62°=90°
- Now that you've identified the complementary angles, let's set up an equation to find the value of x: x + 62° = 90°. Remember, the sum of complementary angles is always 90°.

Please explain why you chose this variant.

Pre-Service teachers preferred Transformer hints when the hint provided details about the problem.

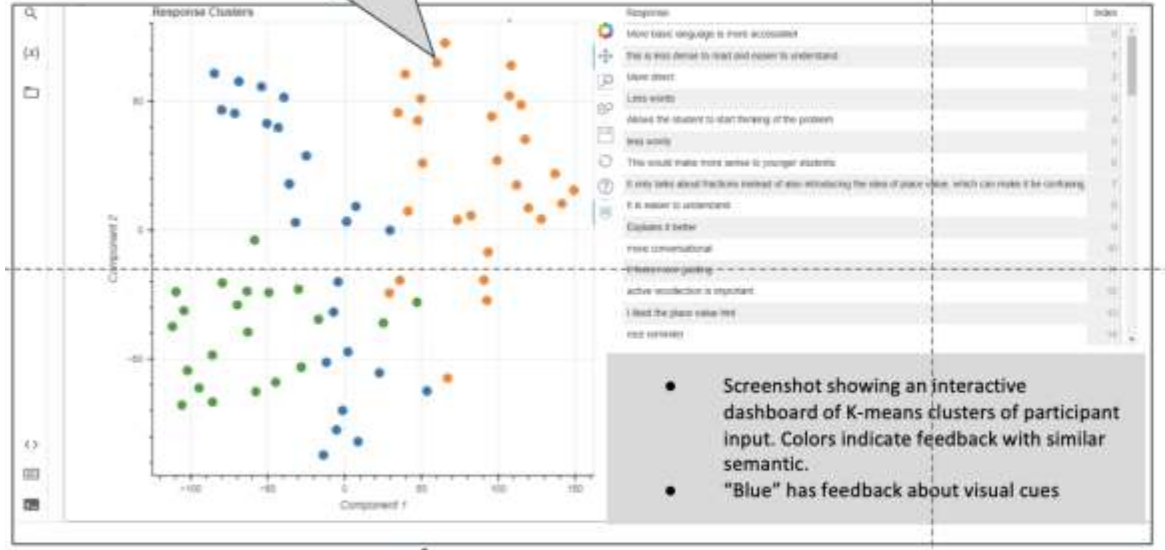


Gattupalli, S.S., Lee, W., Allessio, D., Crabtree, D., Arroyo, I., Woolf, B.P. and Woolf, B., 2023. Exploring Pre-Service Teachers' Perceptions of Large Language Models-Generated Hints in Online Mathematics Learning. In *LLM@ AIED* (pp. 151-162).

Identify Themes in Hints

Results - Thematic Analysis

Each point represents a participant's feedback



- Screenshot showing an interactive dashboard of K-means clusters of participant input. Colors indicate feedback with similar semantic.
- "Blue" has feedback about visual cues

BART and **K-means clustering** with visualization of embeddings

BERT:
Top themes and related feedback for transformer hint preference

Topic Cluster	Theme from Transformer-Generated Hint
1: Explanation Through Connection	• "The language is more descriptive ie the explanation that a right angle forms an L shape." • "I feel like when it comes to applying concepts in an equation format it becomes more confusing to not elaborate on how equations are formed or why they appear the way they do. Even if the previous two hints lead up to the equation itself, having this hint appear like this might draw the two hints into a full circle with this elaboration, because if the child struggles with just the first two hints, they might need more than just given steps, and rather, scaffolding alongside with the information provided."
2: Guidance	• "This coincides with the first question and is closer to what they have most likely been taught!" • "[T]his guides students through the problem."
3: Simple Walkthrough	• "[L]eads the student into the correct 1st step." • "[M]akes it very easy for a student to understand what they should be doing."

Transformer Hints Themes

BERT:
Top themes and related feedback for human hint preference

Topic Cluster	Theme from Human-Created Hint - Visual Cues
1: Explanation with Example	• "I like being able to see the explanation in pictures/numbers more than words. I get a bit lost in the words version." • "I like how in this one the expression is written out and how to find the multiples of each number."
2: Visualization	• "The visual! Much more helpful than the words." • "[T]he picture helps."
3: Explanation in Detail	• "Explains why it works and substitute the original equation." • "If a student needs an additional hint, they may need a little more information, which is why I chose the first option."

Human Hints Themes

Generate and Assess Exam Content

Produce Questions

Load large text into system;
Upload a document (PDF)

Receive a singe paragraph with bulleted items;
Or ask for a summary;
Or to produce test question

Chatbot will highlight relevant portions of text

Fernandez, N., Scarlatos, A. and Lan, A., 2024. SyllabusQA: A Course Logistics Question Answering Dataset. *arXiv preprint arXiv:2403.14666*.

Automatic Text Grading:
Text snippets from an example grade 8 reading comprehension item

Human-generated question

Passage	Long ago, a poor country boy left home to seek his fortune. Day and night he traveled, stopping to eat at inns along ...	
Question	Describe what kind of person the merchant is. Give one detail from the story to support your answer.	
	Student Responses	**Scores**
Dataset	"The merchant is a very optimistic and persevering person. For example, it states how the he kept walking on even with an empty stomach. This shows that he is hopeful and not willing to give up after he had come so far."	3
	"the merchant is a determined and honest person"	2
	"Dishonest because he didn't want to pay for the eggs."	1

Student Responses

Graded by LLM

- Fernandez, Scarlatos, Lan 2024: Syllabus QA: A Course Logistics Question Answering Data Set, arXiv:2023.14666v@

Experimental Results

Approach	Avg. QWK	p-value
Human	0.878	-
Majority	0.527	-
Feature Engineering + Random Forest	0.443	-
Stacked LSTM	0.657	-
Clustering + Classification	0.709	-
BERT (response)	0.828	-
BERT (passage+question+response)	0.828	0.414
BERT in-context	0.833	0.001
BERT multi-task	0.833	1.6×10^{-4}
Meta-trained BERT in-context	**0.841**	9.6×10^{-5}

Quadratic Weighted Kappa (QWK) is a metric used to compare the consistency and fairness of grades assigned by different teachers in educational assessments.

Fernandez, N., Scarlatos, A. and Lan, A., 2024. SyllabusQA: A Course Logistics Question Answering Dataset. *arXiv preprint arXiv:2403.14666*.

LLM Prediction

Human Teacher Grading

Example Student Errors

Error Type	Example Student Response	Predicted Score	Ground Truth
Spelling/grammar	"*mearchant are* a good man because he *played* the innkeeper and kept his . . . "	2	3
Human error & subjectivity	"Long ago a poor country boy left home to seek his fortune. Day and night he . . . "	1	3
Infrequent correct answers	"merchant is described as *brave* as he got on a ship and visited multiple ports . . . "	2	4
No reasoning	"The merchant is dishonest because he *doesn't want to pay* for the eggs . . . "	3	1
Character coreference	"The *merchant* is greedy because he gives the *innkeeper* eggs but when the boy . . . "	4	1

Grading Student Explanations

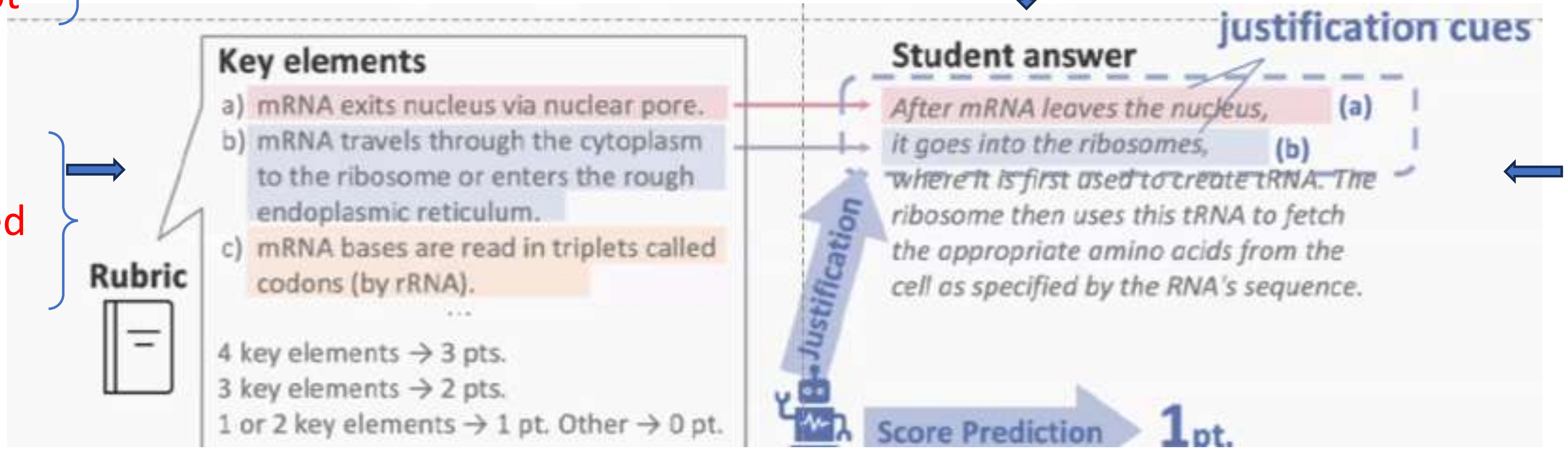
Teacher-generated LLM Prompt

Prompt: Starting with mRNA leaving the nucleus, list and describe four major steps involved in protein synthesis.

Student Explanation

LLM generated answer

Machine-generated grading



Inui, K., Ishii, Y., Matsubayashi, Y., Inoue, N., Naito, S., Isobe, Y., Funayama, H. and Kikuchi, S., 2023. Frontiers in Explainable Automated Writing Evaluation. *IEICE ESS Fundamentals Review*, *16*(4), pp.289-300.

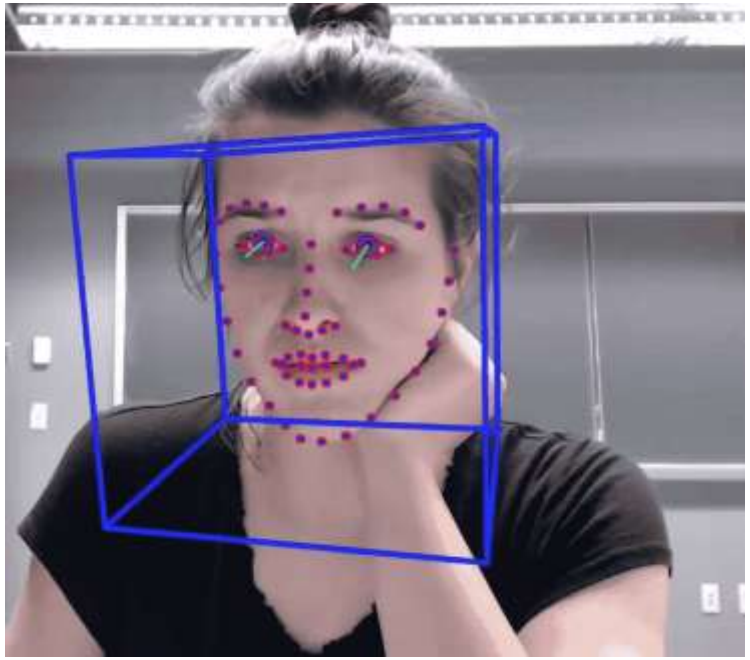
Mizumoto, Bea.
Sato, AIED 2023

LLM Limitations and Considerations

- The effectiveness of personalized feedback systems can vary significantly based on factors including: student preparation and prior knowledge; subject matter complexity; implementation quality; available computational resources; and teacher training;
- LLM hallucinations and incorrect statements;
- Data is subject to algorithmic bias (in terms of gender, race, and culture);
- Quantity and quality of neural network training data are critical to ensure diverse, ethical and inclusive instruction;
- Models can exacerbate falsehoods, inaccuracies and biases of gender, race, or culture at scale;
- Text (and graphics) produced by LLMs are difficult to explain (interpret), placing their accuracy and truthfulness in doubt.

Feature Extraction using OpenFace 2.0

- OpenFace 2.0 is an opensource tool for facial analysis
  - 68 facial landmarks;
  - Head orientation;
  - Gaze direction;
  - Eye detection;
  - 18 Facial action units presence;
  - 17 facial action units intensities.



We concatenate all of the above as a feature vector of length 8241 and pass it to an LSTM-based RNN.

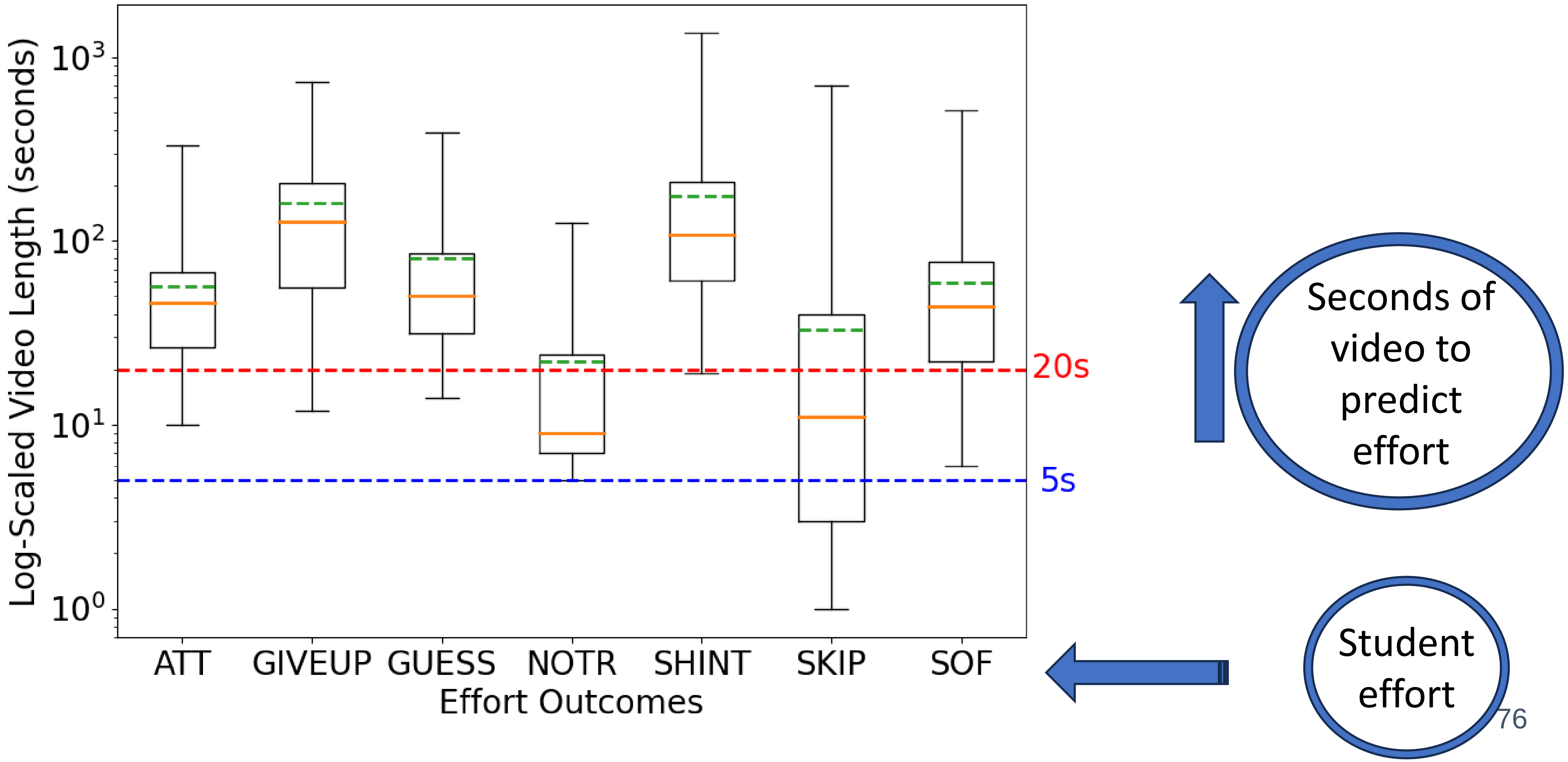
Predict Student Engagement

Method	Accuracy (%)
MobileNet (pretrained ImageNet)	94
Xception (pretrained ImageNet)	88
VGG16 (pretrained ImageNet)	85
Head pose Estimator (Logistic Reg.)	60
Head pose Estimator (Conditional)	55

Predict student engagement using different deep learning and head pose estimate approaches. Deep learning models are more suitable for this type of classification task compared to the head pose estimation. Best results obtained by the model with less complexity, MobileNet.

Distribution of Video Length per Effort Outcome Class

Our models attempt to make an outcome prediction much earlier than the average time a student spends on an exercise for 5 of the 7 student effort outcome classes.



Results: 2-Behavior Outcomes (positive or negative)

Data	Method	F-Score	Balanced Acc.	κ
5 seconds	Majority Vote	0.34±0.04	0.50±0.00	0.00±0.00
	Timing Module (timing info only)	0.39±0.01	0.52±0.01	0.03±0.02
	ATL-BP [27]	0.59±0.04	0.60±0.03	0.19±0.07
	EPAT (*Ours* visual info only)	**0.59±0.04**	**0.60±0.05**	**0.19±0.09**
	EPATT (*Ours* visual + timing info)	0.55±0.08	0.58±0.04	0.15±0.08
10 seconds	Majority Vote	0.35±0.04	0.50±0.00	0.00±0.00
	Timing Module (timing info only)	0.44±0.05	0.55±0.03	0.09±0.05
	ATL-BP [27]	0.61±0.04	0.61±0.04	0.22±0.08
	EPAT (*Ours* visual info only)	0.60±0.03	0.60±0.03	0.21±0.06
	EPATT (*Ours* visual + timing info)	**0.61±0.04**	**0.62±0.03**	**0.24±0.07**
15 seconds	Majority Vote	0.35±0.02	0.50±0.00	0.00±0.00
	Timing Module (timing info only)	0.48±0.04	0.56±0.03	0.13±0.05
	ATL-BP [27]	0.59±0.03	0.59±0.04	0.18±0.07
	EPAT (*Ours* visual info only)	0.57±0.02	0.57±0.02	0.15±0.03
	EPATT (*Ours* visual + timing info)	**0.61±0.06**	**0.63±0.04**	**0.26±0.07**
20 seconds	Majority Vote	0.35±0.01	0.50±0.00	0.00±0.00
	Timing Module (timing info only)	0.52±0.03	0.59±0.02	0.19±0.04
	ATL-BP [27]	0.62±0.05	0.63±0.04	0.25±0.09
	EPAT (*Ours* visual info only)	0.61±0.02	0.62±0.02	0.23±0.04
	EPATT (*Ours* visual + timing info)	**0.67±0.04**	**0.68±0.04**	**0.35±0.09**

Agenda

Motivation

Continual Assessment

Synthetic and Conversational Partners

Predict Learning

Automate Grading

Conclusions

Conclusions

- Early outcome prediction provided by combining timing information with visual affect analysis;
- Visual information is available as early as 5 seconds. Log data becomes available with time;
- Superior performance of facial recognition over baselines and previous state-of-the-art;
- Affective messages build student-tutor rapport and influence students;
- Empathic messages result in improved student interaction and experience;
- Empathic messages result in more confidence, patience, higher levels of interest, and more valued math knowledge;
- Success/failure messages result in more mistakes, less learning-orientation, and more confusion.

Contributions

- Proposed a new student engagement video dataset in an online learning environment;
- Used advanced conversational AI technologies to personalize instruction;
- Used deep learning approach that out-performs traditional head pose classifiers;
- Explored the influence of emerging large language models, such as ChatGPT and Bard, in education;
- Compared LLM-generated content to human-created hints in online math tutoring;
- Interventions from virtual companions are presented when wandering is detected.

UMassAmherst
The Commonwealth's Flagship Campus

How are you feeling? Using AI to Improve Human Learning

Beverly Woolf, Danielle Allessio, Ivon Arroyo,
Margrit Bretke, Hao Yu,

Dec 2024