INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

The 3rd International Conference on Artificial Intelligence in Education Technology (AIET 2022)

TIM SCHLIPPE, QUINTUS STIERSTORFER, MAURICE TEN KOPPEL, PAUL LIBBRECHT

# EXPLAINABILITY IN AUTOMATIC SHORT ANSWER GRADING

Wuhan, China
July 3, 2022

# AGENDA

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

# 1

# INTRODUCTION

# MOTIVATION: UN Sustainable Development Goal 4


SUSTAINABLE DEVELOPMENT GOALS

Image Sources: United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2021); OpenClipart-Vectors/154119/Pixabay.

# MOTIVATION: Challenges in Education

**ACCELERATION**

**COSTS**

SCALING

Image Source: OpenClipart-Vectors/154119/Pixabay.

# AI IN EDUCATION: Potential

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

**AUTOMATION**

GRADING

Image Source: OpenClipart-Vectors/154119/Pixabay.

# AUTO-GRADING

| | |
|---|---|
| **Question** | What is a variable? |
| **Model answer** | A location in memory that can store a value. |
| **Example: Answer 1** | A variable is a location in memory where a value can be stored. |
| **Grading: Answer 1** | **5 of 5 points** |
| **Example: Answer 2** | Variable can be an integer or a string in a program. |
| **Grading: Answer 2** | **2 of 5 points** |

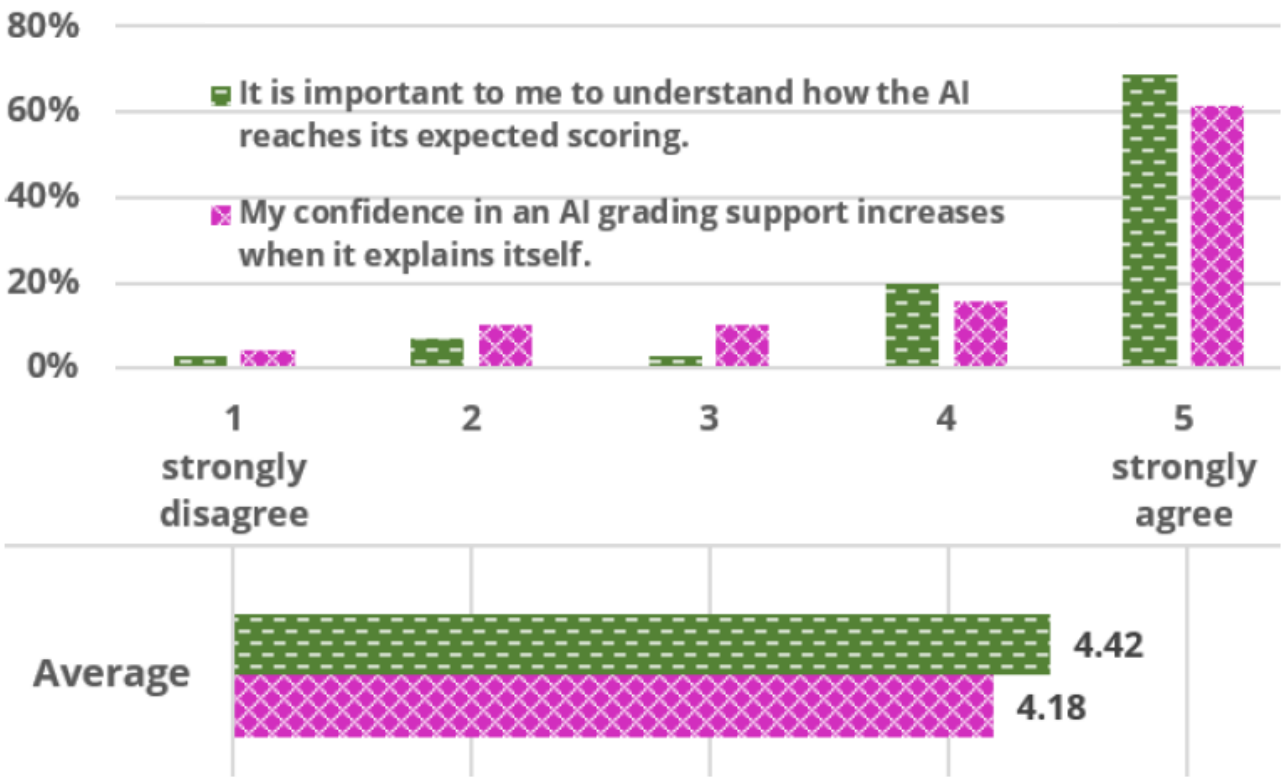
Graphic Source: Custom Depiction.

# AUTO-GRADING: Trust

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

## STUDY

- **70** professors, lecturers and teachers




 **Explainability in Automatic Short Answer Grading**. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

# 2

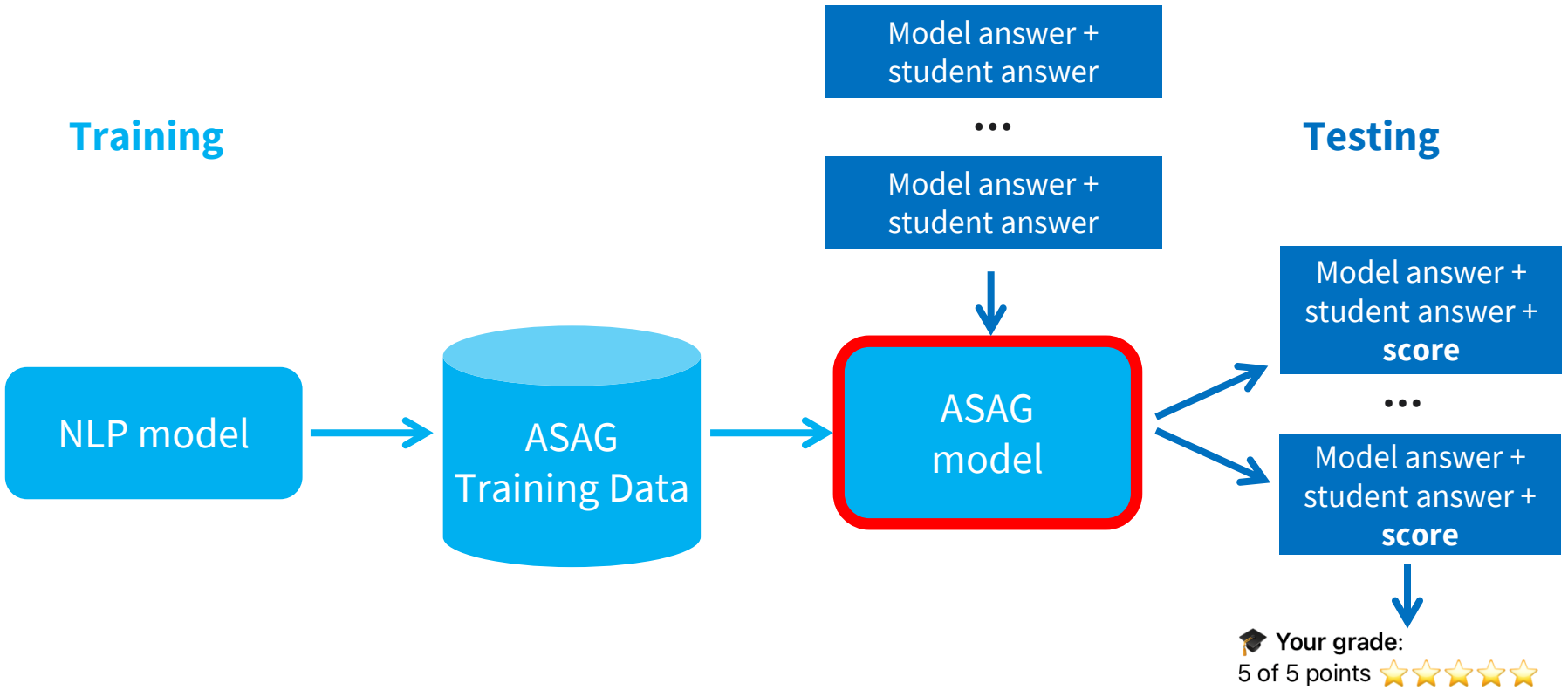
# RELATED WORK

# RELATED WORK: AUTO-GRADING

**AUTOMATIC SHORT ANSWER GRADING**

**Deep learning**

e.g., (Burrows et al., 2014; Camus & Filighera, 2020; Sawatzki et al., 2021; Schlippe & Sawatzki, 2021b)



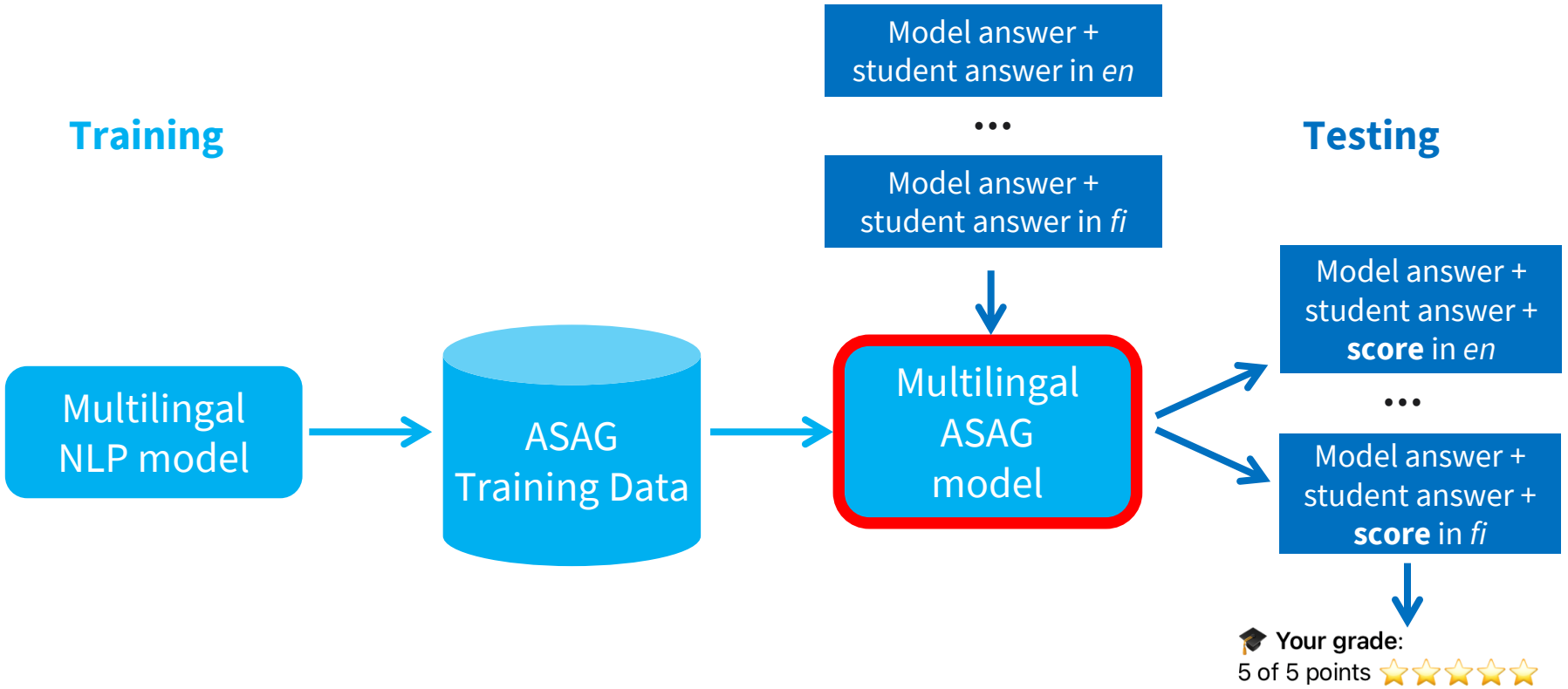

Graphic Source: Custom Depiction.

# RELATED WORK: CROSS-LINGUAL AUTO-GRADING

**AUTOMATIC SHORT ANSWER GRADING**

**Deep learning**

**Cross-lingual Automatic Short Answer Grading**




Graphic Source: Custom Depiction.

# RELATED WORK: AUTO-GRADING

| | multi+ en | multi+ de | multi+ nl | multi+ jp | multi+ zh | multi+ fi | multi+ 6 | mono |
|---|---|---|---|---|---|---|---|---|
| **en** | 0.45 | 0.61 | 0.64 | 0.68 | 0.63 | 0.63 | **0.43** | 0.43 |
| ceb | 0.70 | 0.73 | 0.72 | 0.68 | 0.72 | 0.71 | **0.63** | - |
| sv | 0.63 | 0.67 | 0.68 | 0.73 | 0.72 | 0.68 | **0.48** | - |
| **de** | 0.64 | 0.51 | 0.67 | 0.70 | 0.70 | 0.65 | 0.46 | **0.45** |
| fr | 0.61 | 0.66 | 0.64 | 0.67 | 0.70 | 0.67 | **0.54** | - |
| **nl** | 0.62 | 0.64 | 0.52 | 0.70 | 0.73 | 0.67 | **0.45** | 0.47 |
| ru | 0.68 | 0.73 | 0.83 | 0.74 | 0.75 | 0.78 | **0.52** | - |
| it | 0.62 | 0.65 | 0.72 | 0.71 | 0.73 | 0.70 | **0.52** | - |
| es | 0.61 | 0.68 | 0.76 | 0.68 | 0.72 | 0.65 | **0.49** | - |
| pl | 0.62 | 0.71 | 0.77 | 0.69 | 0.72 | 0.68 | **0.51** | - |
| vi | 0.71 | 0.72 | 0.84 | 0.77 | 0.73 | 0.71 | **0.52** | - |
| **jp** | 0.66 | 0.70 | 0.73 | 0.49 | 0.63 | 0.71 | **0.44** | 0.53 |
| **zh** | 0.63 | 0.71 | 0.77 | 0.69 | 0.50 | 0.79 | **0.41** | 0.44 |
| ar | 0.72 | 0.78 | 0.85 | 0.78 | 0.76 | 0.76 | **0.59** | - |
| uk | 0.65 | 0.70 | 0.82 | 0.73 | 0.73 | 0.75 | **0.54** | - |
| pt | 0.59 | 0.67 | 0.75 | 0.69 | 0.73 | 0.69 | **0.50** | - |
| fa | 0.64 | 0.66 | 0.71 | 0.67 | 0.70 | 0.69 | **0.56** | - |
| ca | 0.64 | 0.70 | 0.74 | 0.70 | 0.76 | 0.67 | **0.53** | - |
| sr | 0.69 | 0.81 | 0.83 | 0.76 | 0.79 | 0.86 | **0.56** | - |
| id | 0.66 | 0.68 | 0.69 | 0.70 | 0.79 | 0.63 | **0.49** | - |
| no | 0.63 | 0.69 | 0.65 | 0.75 | 0.71 | 0.69 | **0.45** | - |
| ko | 0.70 | 0.70 | 0.76 | 0.66 | 0.66 | 0.67 | **0.58** | - |
| **fi** | 0.69 | 0.79 | 0.77 | 0.77 | 0.73 | 0.52 | 0.47 | **0.45** |
| hu | 0.69 | 0.76 | 0.81 | 0.72 | 0.76 | 0.69 | **0.54** | - |
| cs | 0.62 | 0.77 | 0.82 | 0.72 | 0.78 | 0.71 | **0.51** | - |
| sh | 0.66 | 0.77 | 0.79 | 0.74 | 0.78 | 0.79 | **0.53** | - |

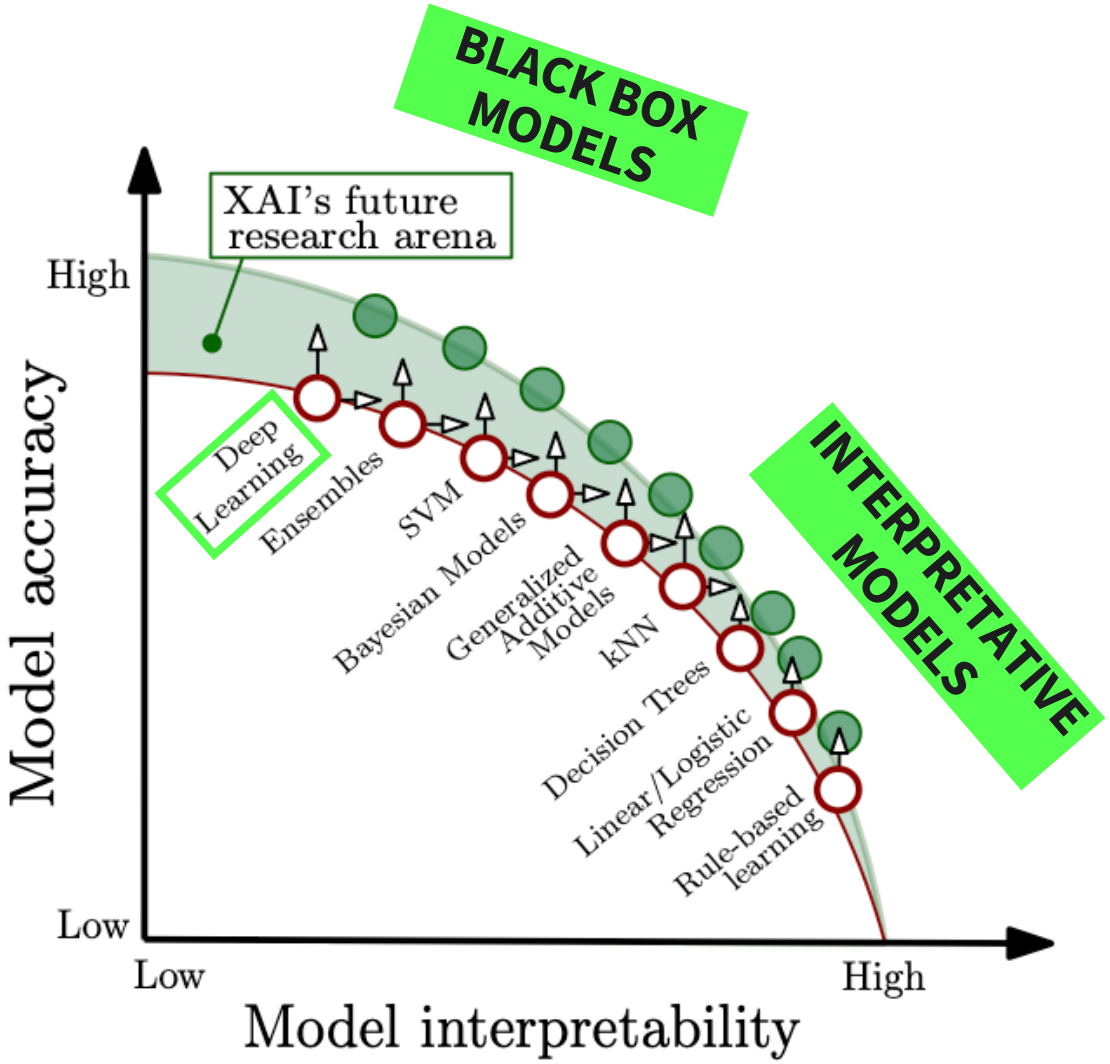
**<0.75 POINTS**
**MEAN ABSOLUTE ERROR**

Mean Absolute Error out of 5 points

**0.75 POINTS**
**HUMAN GRADER VARIABILITY**

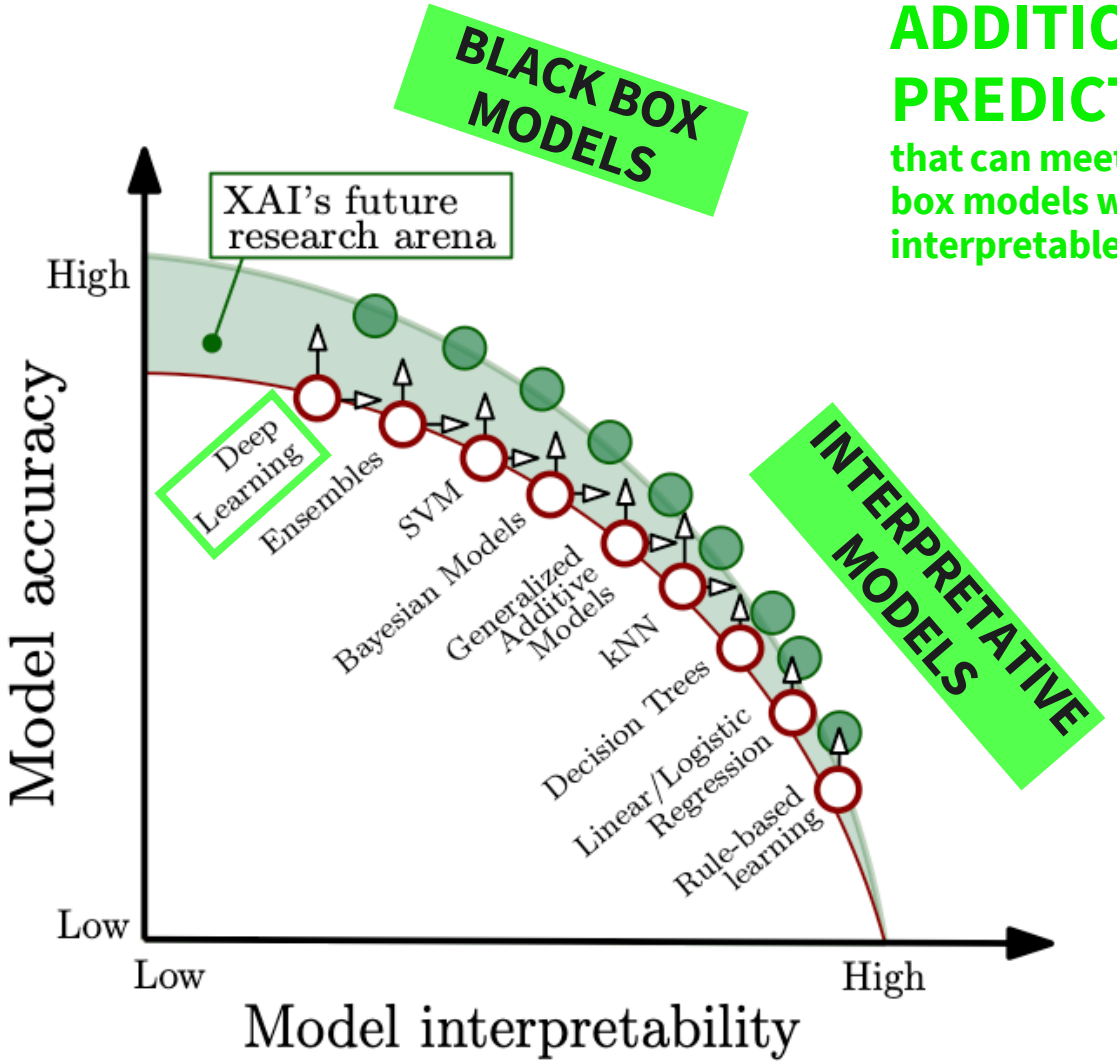
Graphic Sources: Schlippe & Sawatzki (2021b).

# RELATED WORK: EXPLAINABILITY




Source: Barredo Arrieta et al. (2020).

# RELATED WORK: EXPLAINABILITY




**ADDITIONAL PREDICTIVE MODELS**

**that can meet the accuracy of black box models while being intrinsically interpretable**

Source: Barredo Arrieta et al. (2020).

# RELATED WORK: EXPLAINABILITY

| XAI method class | Description |
|---|---|
| *confidence score* | Certainty of a model's prediction is made interpretable and inspectable (van der Waa et al., 2020) |
| *word highlighting* | Words are color marked to indicate their relevance towards the classification (Ribeiro, Singh & Guestrin, 2016) |
| *concept activation* | High level human concepts are used to explain a classification (Kim at al., 2018) |

# RELATED WORK: EXPLAINABILITY

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

| XAI method class | Description |
|---|---|
| *confidence score* | Certainty of a model's prediction is made interpretable and inspectable (van der Waa et al., 2020) |
| *word highlighting* | Words are color marked to indicate their relevance towards the classification (Ribeiro, Singh & Guestrin, 2016) |
| *concept activation* | High level human concepts are used to explain a classification (Kim at al., 2018) |

| XAI method for ASAG | XAI method class |
|---|---|
| *Predicted Points* | — |
| *Predicted Points with Confidence Scores* | *confidence score* |
| *Predicted Points with Confidence Scores and Similar Answers* | *confidence score* |
| *Predicted Points with Relevance of Words in the Answer* | *word highlighting* |
| *Predicted Points with Matching Positions* | *concept activation* |

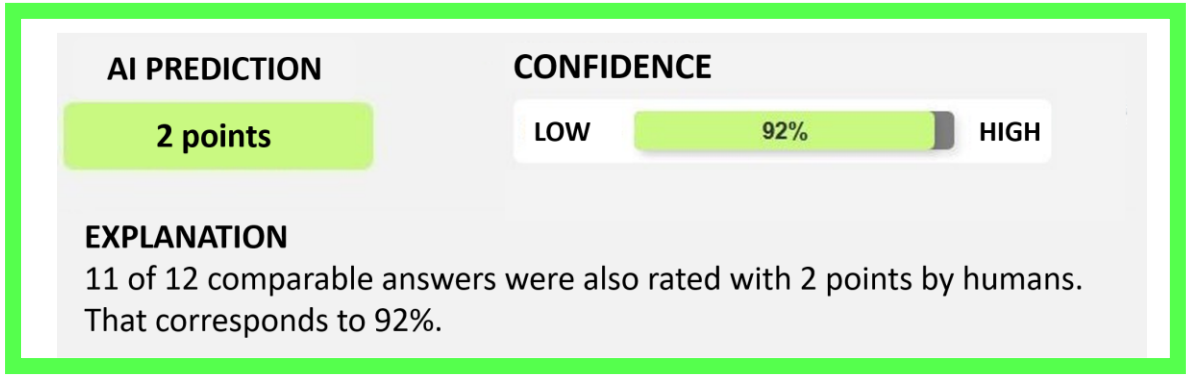
3

# EXPLAINABILITY IN AUTOMATIC SHORT ANSWER GRADING

# AUTO-GRADING: Explainability: XAI methods


points

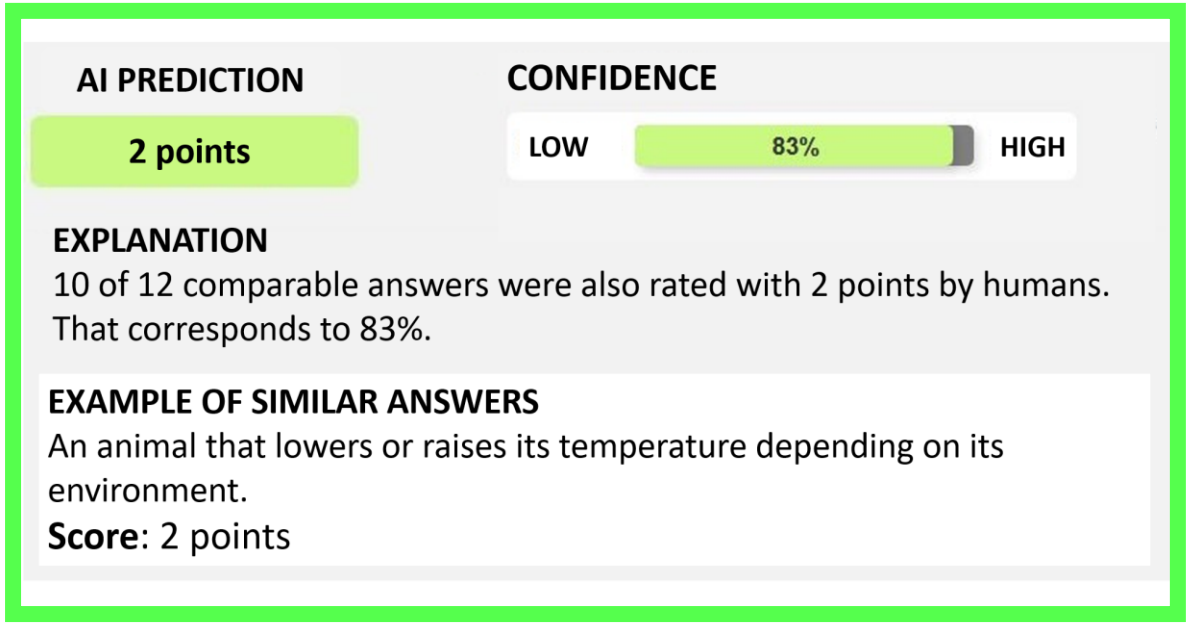
# AUTO-GRADING: Explainability: XAI methods




**confidence**

 **Explainability in Automatic Short Answer Grading**. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

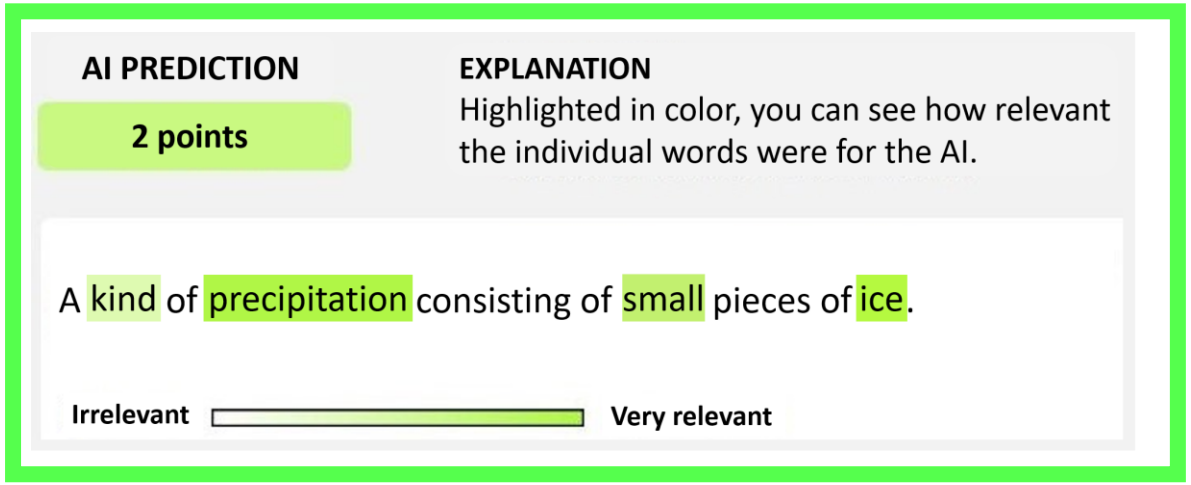
# AUTO-GRADING: Explainability: XAI methods




**confidence + similar answers**

 **Explainability in Automatic Short Answer Grading**. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

# AUTO-GRADING: Explainability: XAI methods




**relevance of words**

# AUTO-GRADING: Explainability: XAI methods

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

**AI PREDICTION**

**2 points**

**EXPLANATION**

The AI assistant compared with the model answer and found 2 content matches.

Everything **we consciously perceive** +1 such as vision, is **processed and transmitted through body parts** +1.

**points + matching positions**

# 4

# USER STUDY

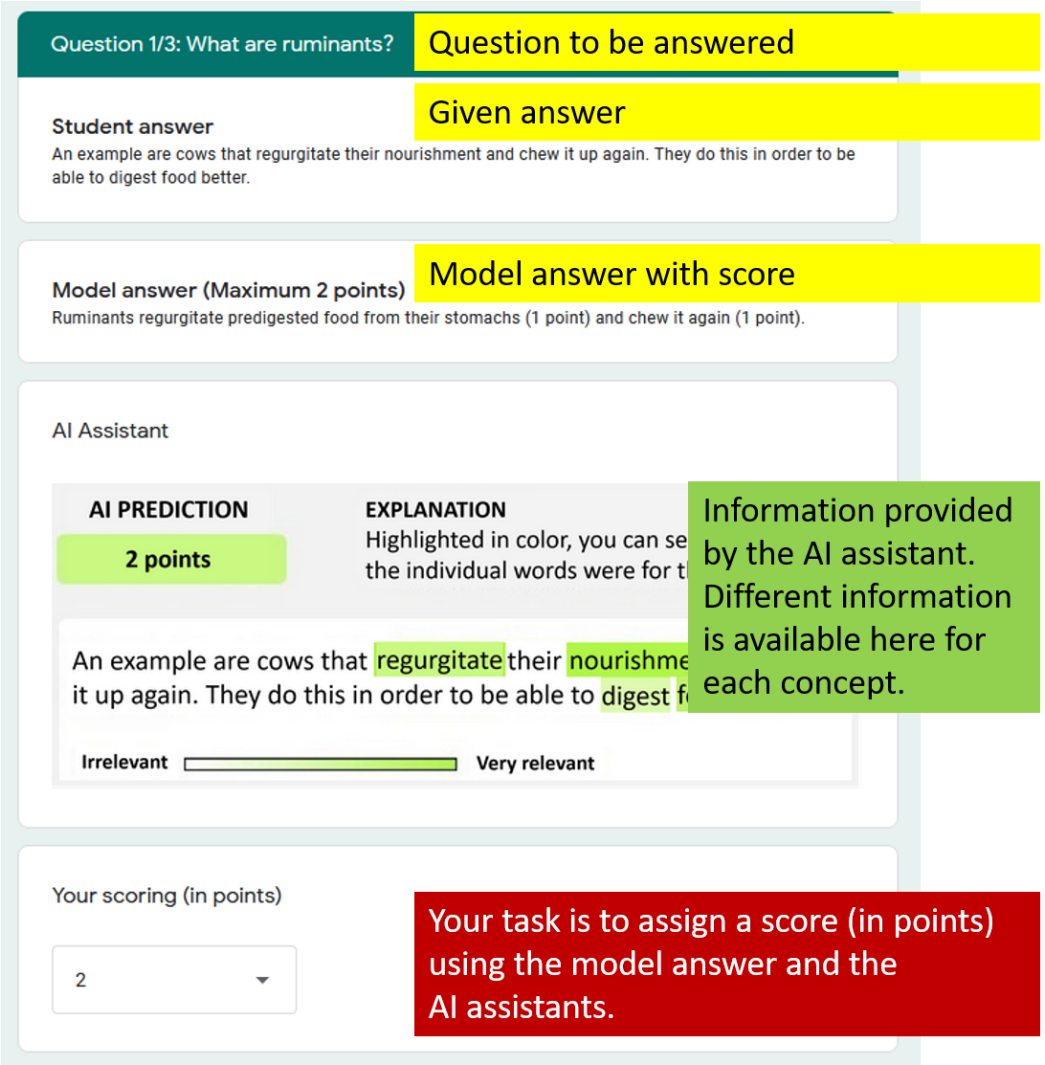
# AUTO-GRADING: Explainability

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

## STUDY

- **5 XAI methods**
- **70** professors, lecturers and teachers
- **9** aspects evaluated

  ✓ trust, ✓ informative content, ✓ speed, ✓ consistency & fairness, ✓ fun, ✓ comprehensibility, ✓ applicability, ✓ use in exam preparation, ✓ in general




 **Explainability in Automatic Short Answer Grading**. Tim Schlippe, Quintus Stierstorfer, Maurice ten Koppel, Paul Libbrecht. AIET 2022.

# AUTO-GRADING: Explainability: XAI methods

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

**AI PREDICTION**

**2 points**

**points**

**AI PREDICTION**

**2 points**

**CONFIDENCE**

LOW 92% HIGH

**EXPLANATION**

11 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 92%.

**confidence**

**AI PREDICTION**

**2 points**

**EXPLANATION**

Highlighted in color, you can see how relevant the individual words were for the AI.

A kind of precipitation consisting of small pieces of ice.

Irrelevant — Very relevant

**Relevance of words**

**AI PREDICTION**

**2 points**

**CONFIDENCE**

LOW 83% HIGH

**EXPLANATION**

10 of 12 comparable answers were also rated with 2 points by humans. That corresponds to 83%.

**EXAMPLE OF SIMILAR ANSWERS**

An animal that lowers or raises its temperature depending on its environment.

**Score**: 2 points

**confidence + similar answers**

**AI PREDICTION**

**2 points**

**EXPLANATION**

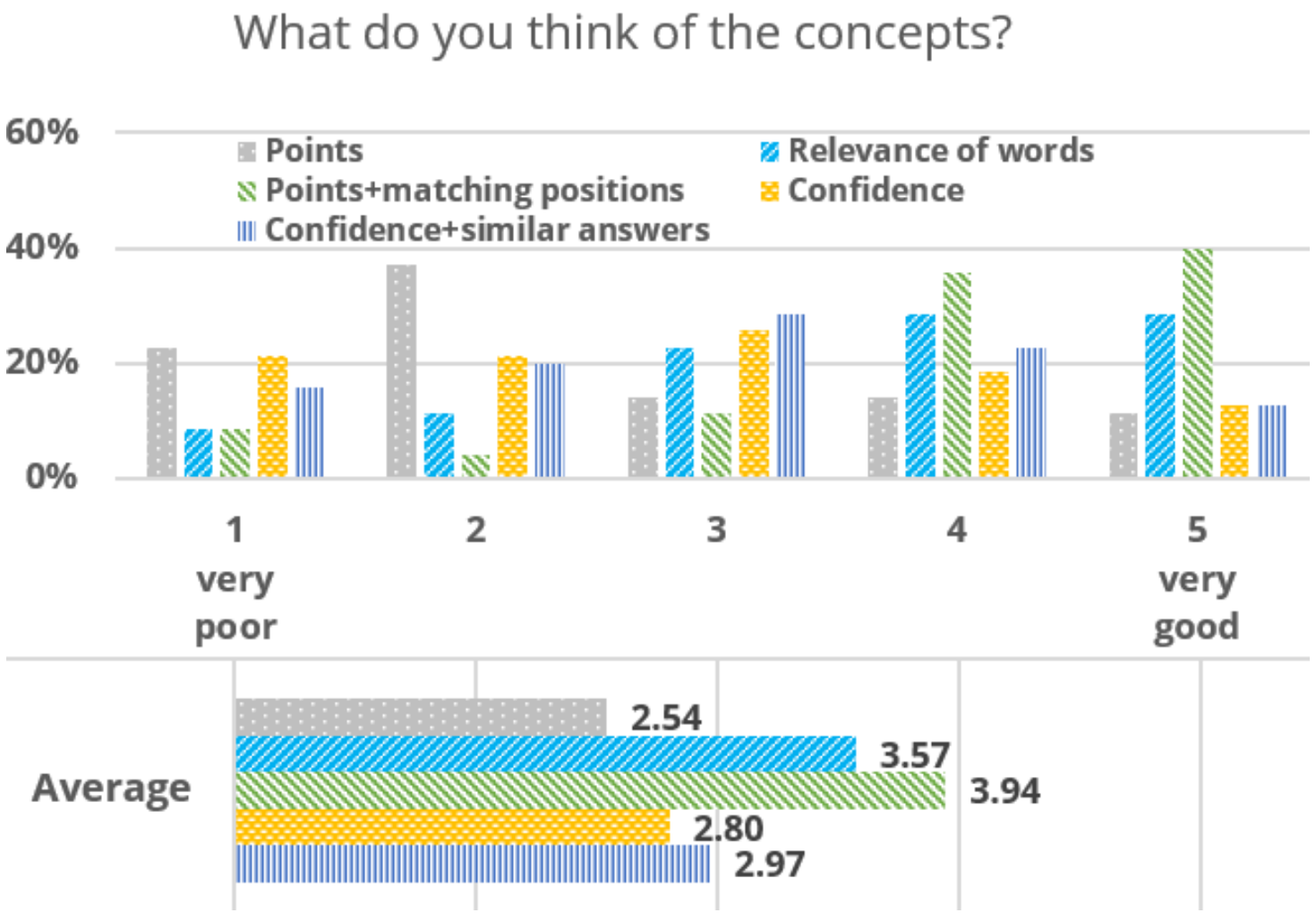
The AI assistant compared with the model answer and found 2 content matches.

Everything **we consciously perceive** +1 such as vision, is **processed and transmitted through body parts** +1.

**points + matching positions**

✓ trust, ✓ informative content, ✓ speed, ✓ consistency & fairness, ✓ fun, ✓ comprehensibility, ✓ applicability, ✓ use in exam preparation, ✓ in general

# AUTO-GRADING: Explainability: XAI methods




 **Explainability in Automatic Short Answer Grading**. T. Schlippe, Q. Stierstorfer, M. ten Koppel, P. Libbrecht. AIET 2022.

# 5

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

# CONCLUSION & FUTURE WORK

# CONCLUSION AND FUTURE WORK

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

**Conclusion**

- Investigated and evaluated different methods for explainability in automatic short answer grading
- Survey of over 70 professors, lecturers and teachers
- Important for them to understand how the AI reaches its scoring and their confidence in an AI grading support increases when it explains itself
- Displaying the predicted points together with matches between student answer and model answer is rated better than the other tested XAI methods.
- Evaluated aspects: trust, informative content, speed, consistency and fairness, fun, comprehensibility, applicability, use in exam preparation, and in general.

**Furture Work**

- Analyze the use of our XAI methods in interactive training programs to prepare students optimally for exams
- Direct interpretation of the complex ASAG models could be also investigated

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

# THANK YOU

Tim Schlippe
tim.schlippe@iu.org