COLLEGE OF EDUCATION

MARYLAND ASSESSMENT RESEARCH CENTER

# Enhance Large-Scale Educational Assessment Practices with the Advances in Artificial Intelligence

Hong Jiao
Maryland Assessment Research Center
University of Maryland, College Park
USA

November 5, 2024

Presentation at the 7th International Association for Innovation in Educational Assessment

Abuja, Nigeria

# Outline

- Introduction
  - Educational assessment
  - Technology vs. educational assessment
- AI capacity
  - Data analysis of multimodal types of data
  - Data generation/augmentation
- Use cases of AI in educational assessments
- Data augmentation in educational assessments
  - Automated scoring
  - Cheating detection
- Summary

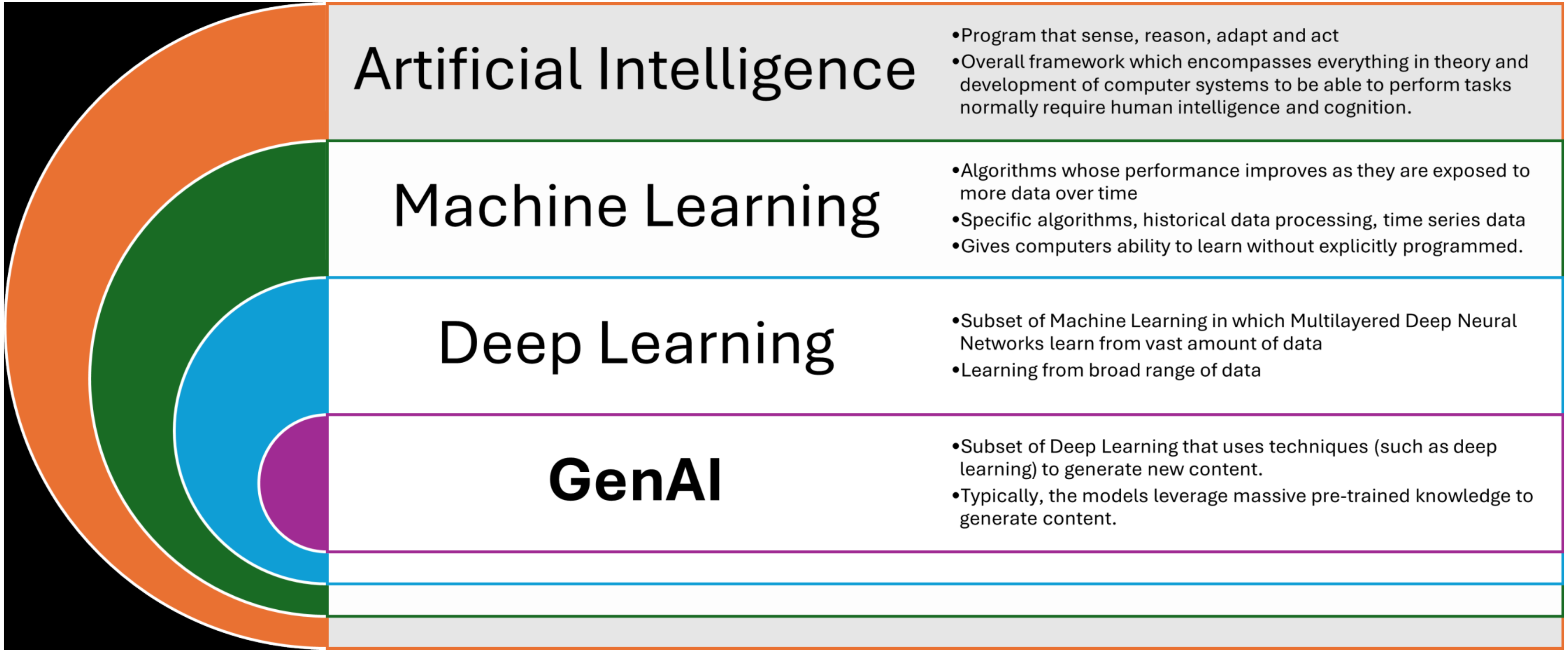
# Educational Assessments

- Needs from different test stakeholders
  - Students: feedback and growths
  - Teachers: areas to improve
  - Parents: are my kids ready?
  - School administrators: where to help?
  - State administrators: what are needed to improve?
- What educational assessments can do
  - Use the shortest tests
  - Provide the fastest and detailed feedback
  - Use technology innovations
  - Track growth

# Technology vs. Educational Assessments

- Assessment is closely tied with technology.
    - Computer-based assessment programs (about 15 year ago, RTTT program, technology-enhanced innovative items)
    - Computerized adaptive testing (about 10 years ago, Consortium tests, SBAC, PARCC)
- Instant advantages of previous successes of computer technology in assessments
    - Efficiency in test administration
        - no shipping test booklets
        - no scanning of answer sheets
        - reduce potential errors due to logistics
    - Fast scoring reporting and feedback
    - More accurate estimation of learning outcomes with shorter test lengths (CAT)
    - More authentic assessment format: interactive item types
    - More data that can be collected: item responses and response process data

# AI, Machine Learning and Natural Language Processing



**Artificial Intelligence**

- Program that sense, reason, adapt and act
- Overall framework which encompasses everything in theory and development of computer systems to be able to perform tasks normally require human intelligence and cognition.

**Machine Learning**

- Algorithms whose performance improves as they are exposed to more data over time
- Specific algorithms, historical data processing, time series data
- Gives computers ability to learn without explicitly programmed.

**Deep Learning**

- Subset of Machine Learning in which Multilayered Deep Neural Networks learn from vast amount of data
- Learning from broad range of data

**GenAI**

- Subset of Deep Learning that uses techniques (such as deep learning) to generate new content.
- Typically, the models leverage massive pre-trained knowledge to generate content.

https://www.linkedin.com/pulse/evolution-ai-key-concepts-gen-ai-mandar-kulkarni-63fbf/

# AI Capacity

## Data Analysis

- Machine learning
  - Supervised learning
    - Prediction
    - Classification
  - Unsupervised learning
    - Clustering
    - Association
    - Dimensionality reduction
  - Reinforcement learning
- Natural language processing (NLP)
  - Analysis and understand spoken and written data
  - Extract linguistic features by processing text/speech data

## Data Generation/Augmentation

- Large Language Models (LLMs)
  - Generate text data
  - Respond to prompting requests

# Why Need AI in Assessment

## Traditional Psychometric Analysis

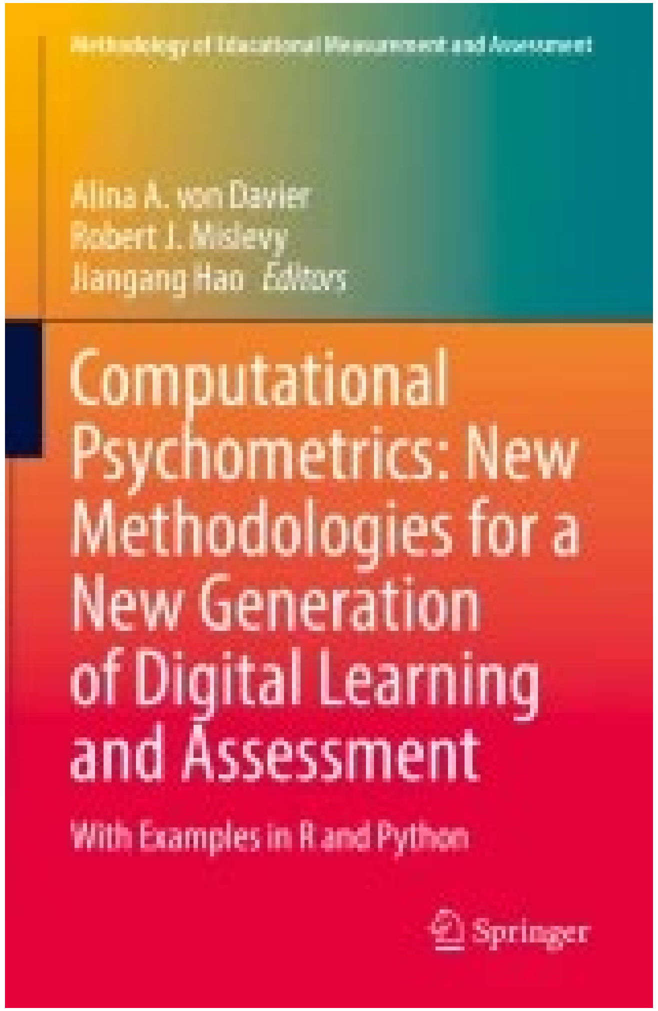
- Data types
  - Structured item product data
    - Item response data
      - Dichotomously scored
      - Polytomously scored
  - Limited unstructured item process data
    - Item response time
    - Answer change counts and patterns
    - Item revisit counts and patterns

## AI-Enhanced Computational Psychometric Analysis (von Davier, Mislevy, & Hao, 2021)

- Data types
  - Unstructured item response data
    - Text data
    - Speech data
  - Unstructured item process data
    - Item response time
    - Item revision counts and patterns
    - Item revisit count and patterns
    - Action sequence data

# Computational Psychometrics

- Computational Psychometrics (von Davier, Mislevy, & Hao, 2021) provides a new framework to re-conceptualize assessment theory and practices in the era of digital assessment with the advances in machine learning, natural language processing, and generative AI.
- It integrates principled traditional psychometric theory/methods and machine learning algorithms to enhance the measurement theory and practices in digital assessment when assessment data may become available in both structured item response data and unstructured item process and multimodal data (text/speech/image/biometric).



Methodology of Educational Measurement and Assessment

Alina A. von Davier
Robert J. Mislevy
Jiangang Hao *Editors*

Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment

With Examples in R and Python

Springer

https://link.springer.com/book/10.1007/978-3-030-74394-9

# Test Development Process

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications/ test blueprints
- Item development: item reviews
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores: standard setting
- Reporting test results
- Item banking
- Test technical report
    - Validity
    - Reliability
    - Fairness

## Responsible Parties

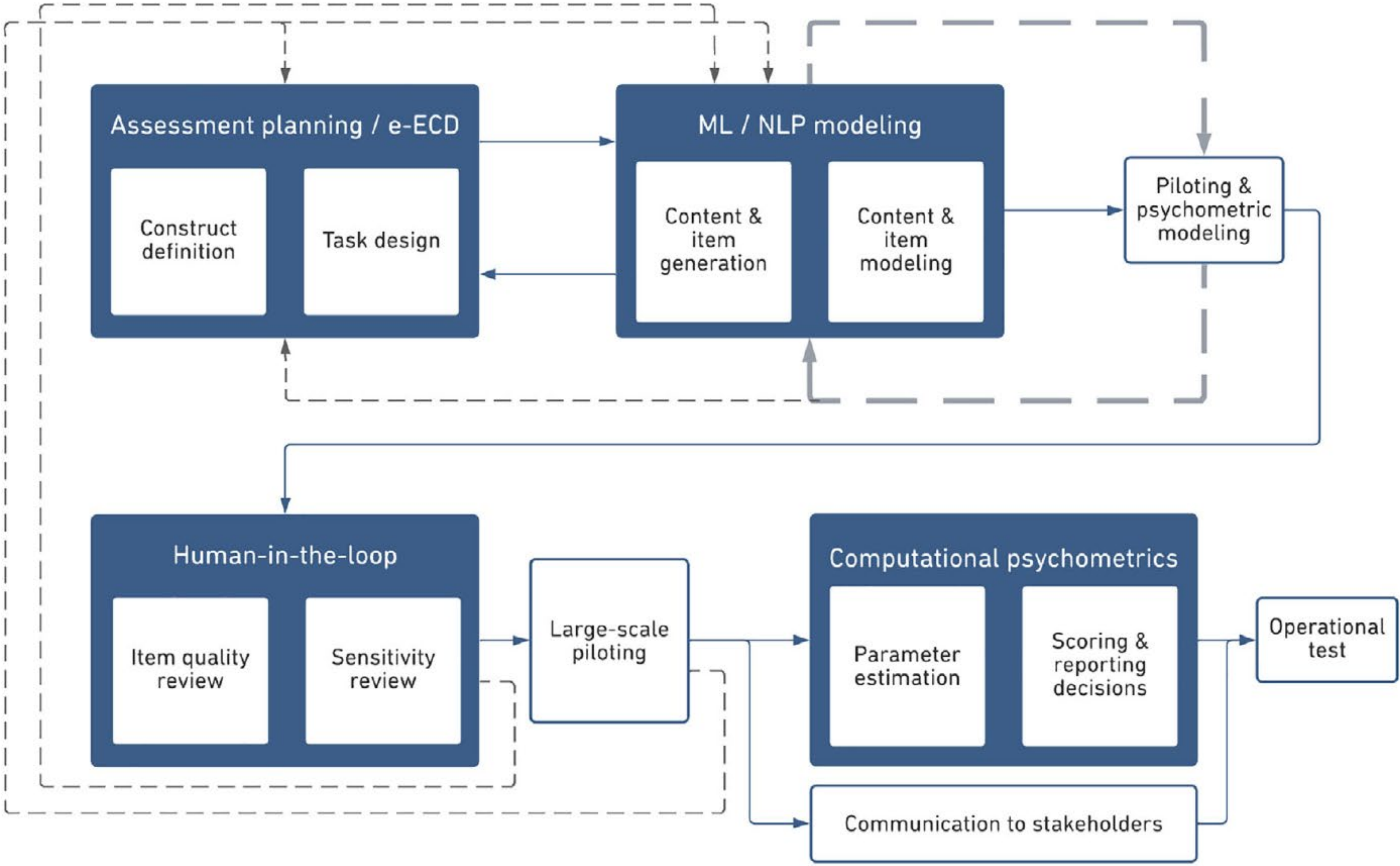
- Test stakeholders
- Policy makers
- Content experts
- Psychometricians
- IT/AI technology experts
- Project managers

*A system of systems based on principled design*

# Duolingo Example

**Figure 1**

*Content Creation at Scale Using Human-in-the-Loop AI from the Duolingo English Test*



Hao et al. (2024)

# Use Case: **Overall Plan**

## Test Development Steps (Downing, 2006)

- **Overall plan**
- Content/construct definition
- Test specifications
- Item development
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - Fairness

## Assessment Design

- Assessment design
  - Without investing in too much resources, assessment designers can create prototype of an assessment program.
  - An iterative process
  - Current practice follows a sequence.

# Use Case: **Automated Item Generation**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- **Item development**
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - Fairness

## Item writing

- Item development
  - Dramatically reduce the time, financial resources, human resources in item development
  - Create passages without copyright issues
  - Create images without copyright permission
  - Create sample answers

**von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika* 83, 847-857.** https://doi.org/10.1007/s11336-018-9608-y

**von Davier, M. (2019). Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model.** https://doi.org/10.48550/arXiv.1908.08594

# Use Case: **Item Alignment**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- **Item development**
- **Test** design and **assembly**
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - **Validity**
  - Reliability
  - Fairness

## Automated Item Alignment

- Content standards
  - Reading
  - Math
- Cognitive complexity
  - Bloom's taxonomy
- Many studies in educational technology & learning analytics

Wang et al. (2023): https://doi.org/10.21203/rs.3.rs-3740769/v1

# Use Case: **Field Testing**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- **Item development**
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- **Calibration**, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
    - Validity
    - Reliability
    - Fairness

## Generate item responses for IRT calibration

- Hotaka Maeda, Smarter Balanced NCME 2024 paper Training AI to Generate Human-Like Item Responses for Field-Testing
- Proposed a method of using large-language models to generate human-like item responses by training the model on simulated item response data.
- Results showed moderate levels of success in using AI-generated responses to recover item parameters.
- If refined further, AI could potentially replace human test-taker responses for field-testing new items.

# Use Case: **DIF Detection**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- **Item development**
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - **Fairness**

## Detection of DIF

- Mangino, Finch, French, & Demir. (2023, April). *Identification of differential item functioning using machine learning.* Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.

# Use Case: **Item Parameter Modeling**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- **Item development**
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- **Calibration**, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - Fairness

## Item parameter prediction

- Use item text features to predict item parameters
  - item difficulty: p-values and IRT difficulty parameters
  - item discrimination parameters: item-total correlations and a-parameters in an IRT model
- Screening item quality before putting them in field-testing or skipping field-testing

# Use Case: **Cheating Detection**

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- Item development
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - **Validity**
  - Reliability
  - Fairness

## Types of Cheating Detection

- Enhance cheating detection using multiple data types
- Cheating detection of AI generated essays in state tests

# Automated Scoring

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- Item development
- Test design and assembly
- Test production
- Test administration
- **Scoring test responses**
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - Fairness

## Scoring in state assessment programs

- Essays
  - Written communication skills
- Constructed-response items
  - Constructed-response (CR) items are often expected to better assess higher-order thinking skills
    - Reasoning and argumentation (Downing & Haladyna, 2006).
    - Synthesizing information from multiple sources like paired passages, paired reading materials and numerical data presented in complex data system.
  - More authentic
  - Remove possible guessing in multiple-choice items

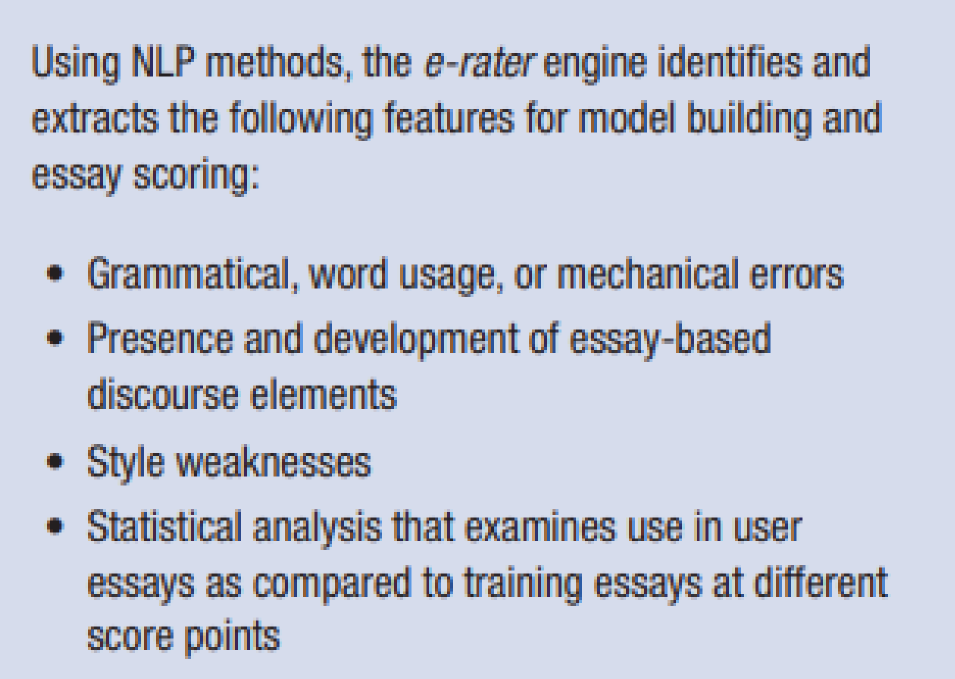
https://cambiumassessment.com/-/media/project/cambium/corporate/pdfs/cai-cambium-comparing-robustness-automated-scoring-approaches.pdf

# Automated Scoring

## Historical Explorations

- 1st automated essay scoring system:
  - Page (1966, 1968): Project Essay Grader (PEG)-Measurement, Inc.
- e-rater in 1999-ETS (Attali & Burstein, 2006; Burstein, Chodorow, and Leacock, 2003)
- Intelligent Essay Assessor-Pearson (IEA; Zupanc & Bosnic, 2015)
- Autoscore (AIR, Cambium Assessment, Inc.)
- LightSIDE-Carnegie Mellon
- IntelliMetric (Vantage Learning)
- Lexile Writing Analyzer-Meta Metrics
- Bookette-DRC
- CRASE-ACT
- Criterion (Burstein et al., 2004) for analytical scoring of essays
- c-rater(Leacock & Chodorow 2003)
- m-rater (Bennett, Morley, & Quardt, 2000; Bennett, Morley, Quardt, & Rock, 2000; Bennett et al., 1999; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997)

## e-rater features



Using NLP methods, the *e-rater* engine identifies and extracts the following features for model building and essay scoring:

- Grammatical, word usage, or mechanical errors
- Presence and development of essay-based discourse elements
- Style weaknesses
- Statistical analysis that examines use in user essays as compared to training essays at different score points

https://www.researchgate.net/profile/Mohamed_Mourad_Lafifi/post/Methods_that_can_use_for_generate_automated_essay_type_questions_for_a_given_text/attachment/5a34ebb6b53d2f0bba44bf17/AS:572129265164289@1513417654657/download/AutomatedScoring.pdf

# Automated Scoring

## Feature-based automated scoring

- Features (Ke & Ng, 2019; Shermis, 2014; Uto et al., 2020)
    - Syntactic features
        - Part-of-speech of words: nouns, adjectives
        - Grammar errors
        - Spelling errors
    - Semantic features
        - Semantic similarity
        - Histogram-based features
    - Lexical features
        - Usage: sentence variation,
        - Single words, stemmed or lemmatized words
        - Prefix and suffix
        - Overlapping between sentences
    - Length-based features
        - Word count
        - Sentence count
        - Average word length
    - Word-based features
        - Number of useful n-grams
        - Spelling errors
        - Sentiment words
    - Readability
        - Number of difficult words
        - Readability index: Flesch-Kincaid reading ease, Gunning fog, SMOG index
    - Argumentation features
        - Number of claims and premises
        - Argument tree depth
    - Similarity measures
        - to training essays at different score points
        - to other reference responses

## Deep learning-based automated scoring (Lottridge et al., 2020)

- Deep learning (multilayered neural networks)
    - Long short-term memory (LSTM)networks
    - Recurrent Neural Networks (RNN)
    - Convolution Neural Networks (CNN)
- Transformer networks
- Bookette, developed by CTB, is one of the first AES systems developed using neural networks (Rich, Schneider, & D’Brot, 2013)

https://cambiumassessment.com/-/media/project/cambium/corporate/pdfs/cai-cambium-comparing-robustness-automated-scoring-approaches.pdf

# Automated Scoring

## Hybrid approach to automated scoring (Whitmer et al., 2023)

- NAEP automated scoring challenges exemplified the success of the hybrid approach
  - Ensemble of features and LLM embeddings
- 2021 NAEP Automated Scoring Challenge for CR items in reading assessment
  - to extract features adding N-gram vectors, passage similarity measures, and deep neural network embeddings to train the best ensemble model from multiple classifier and regression models.
  - fine-tuned a BERT model for all items with in-context learning adding the passage and response to be scored as well as other scored example responses to develop a final linear classifier to predict the score.
  - used an ensemble model including Convolutional BERT model and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA; Clark et al., 2020).
  - The hybrid models integrating the beddings from BERT or BERT extended models and the handcrafted features seem to perform slightly better compared with the pure ensemble models from BERT and BERT extended models (Lottridge, 2023; Lottridge et al., 2023; Ormerod, 2022a; Ormerod et al., 2022; Zhou et al., 2002).

## Comparison

- In general, LLM-based models performed better.
- The ensemble learning of either multiple deep learning models or a hybrid of handcrafted features plus deep learning embeddings with classical machine learning models may work better (Zhou et al., 2002).

# Challenges in Automated Scoring

- Scorability: not all items are scorable
- Explainability of automated scores
- Bias in automated scoring engine
- Small sample size in some score categories: class imbalance issues
- Cheating or gaming the automated scoring system

# Class Imbalance in Automated Scoring

- Small sample sizes in some score categories are frequently observed in automated scoring.
- Table 1 summarizes the n-counts for each score category for six prompts in the Kaggle ASAP dataset.
- The worst case is in Prompt 1 with score sample sizes ranging from 0.1% to 38.5% out of the total sample size for this prompt while the best case is in Prompt 4 with score sample sizes ranging from 14.3% to 35.9%.

# Class Imbalance

- Sample sizes for each score category for six prompts in the Kaggle ASAP dataset.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prompt 1 | Scores | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | Total |
| | Frequency | 10 | 1 | 17 | 17 | 110 | 135 | 687 | 334 | 316 | 109 | 47 | 1783 |
| | Percent | 0.6 | 0.1 | 1 | 1 | 6.2 | 7.6 | 38.5 | 18.7 | 17.7 | 6.1 | 2.6 | 100 |
| Prompt 2 | Scores | 1 | 2 | 3 | 4 | 5 | 6 | | | | | | Total |
| | Frequency | 24 | 153 | 763 | 778 | 75 | 7 | | | | | | 1800 |
| | Percent | 1.3 | 8.5 | 42.4 | 43.2 | 4.2 | 0.4 | | | | | | 100 |
| Prompt 3 | Scores | 0 | 1 | 2 | 3 | | | | | | | | Total |
| | Frequency | 39 | 607 | 657 | 423 | | | | | | | | 1726 |
| | Percent | 2.3 | 35.2 | 38.1 | 24.5 | | | | | | | | 100 |
| Prompt 4 | Scores | 0 | 1 | 2 | 3 | | | | | | | | Total |
| | Frequency | 312 | 636 | 570 | 253 | | | | | | | | 1771 |
| | Percent | 17.6 | 35.9 | 32.2 | 14.3 | | | | | | | | 100 |
| Prompt 5 | Scores | 0 | 1 | 2 | 3 | 4 | | | | | | | Total |
| | Frequency | 24 | 302 | 649 | 572 | 258 | | | | | | | 1805 |
| | Percent | 1.3 | 16.7 | 36 | 31.7 | 14.3 | | | | | | | 100 |
| Prompt 6 | Scores | 0 | 1 | 2 | 3 | 4 | | | | | | | Total |
| | Frequency | 44 | 167 | 405 | 817 | 367 | | | | | | | 1800 |
| | Percent | 2.4 | 9.3 | 22.5 | 45.4 | 20.4 | | | | | | | 100 |

# Class Imbalance

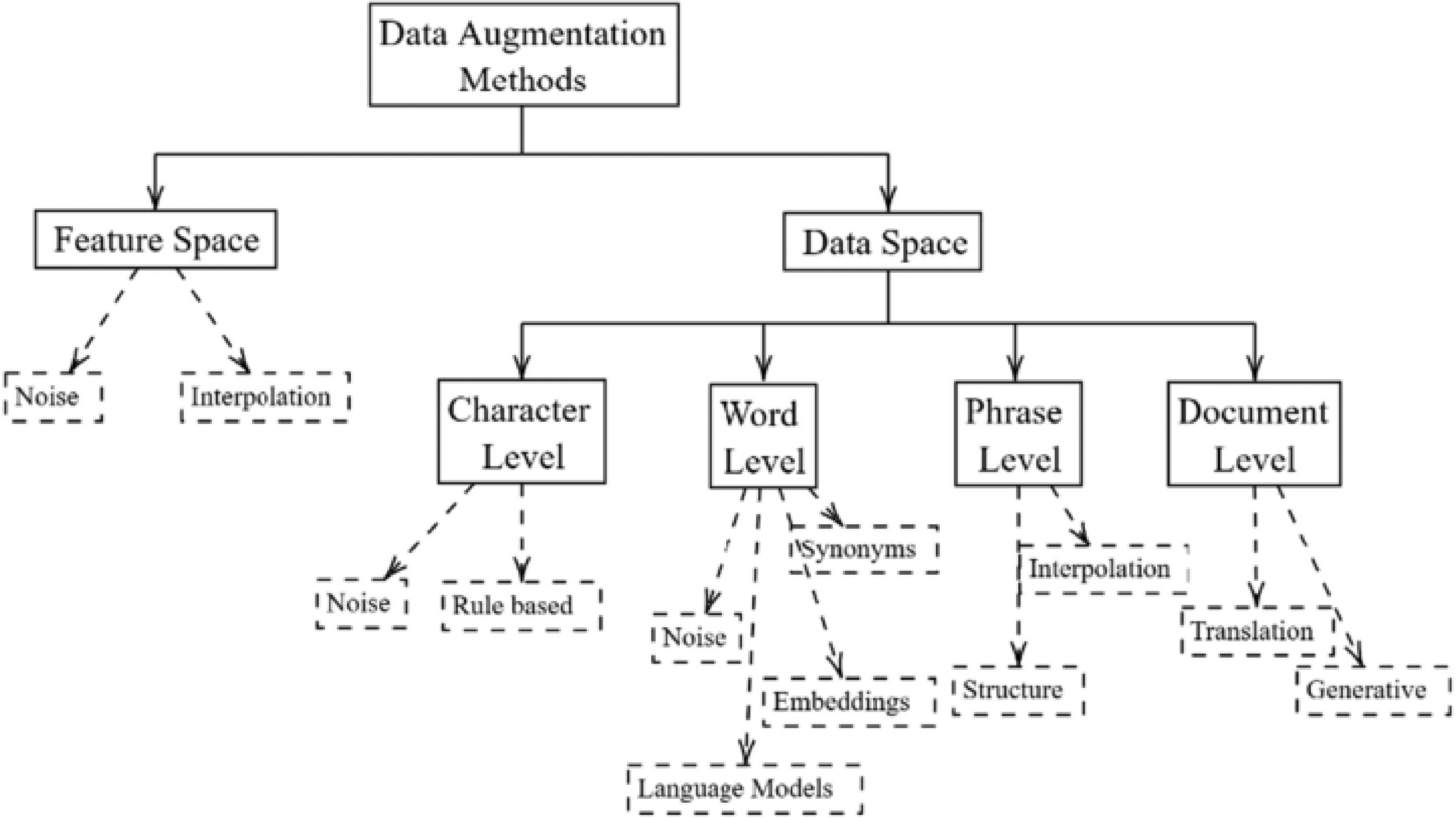
- Most machine learning algorithms are developed for equally balanced classes (He & Garcia, 2009).
- When the classes are imbalanced, machine learning algorithms tend to produce misleading or erroneous results favoring the majority class, leading to low prediction accuracy in the minority class.
- Guo et al. (2008) noted that the lack of representation and information of the key characteristic of the minority class makes machine learning difficult to predict the probability of the minority class.
- The skewed class distribution of score categories makes the automated scoring models biased towards the majority class, yielding high accuracy for predicting the majority classes but low accuracy in predicting the minority classes.
- Class imbalance has been found to be an issue that may result in errors or bias in automated scoring.

# Data Augmentation (from ChatGPT)

- Class Imbalance: When dealing with imbalanced datasets where one or more classes are significantly underrepresented compared to others, data augmentation techniques can help rebalance the class distribution by generating synthetic instances of minority classes.
- Limited Data Availability: In situations where the available dataset is small or insufficient to train a robust machine learning model, data augmentation can be used to artificially increase the size of the training set and improve model generalization.
- Overfitting: Data augmentation can mitigate overfitting by introducing variations in the training data, making the model more robust to noise and reducing its sensitivity to specific examples in the training set.
- Domain Adaptation: When there is a domain shift between the training and testing data, data augmentation techniques can help adapt the model to the target domain by generating synthetic instances that resemble the target distribution.
- Variability in Input Data: In tasks where the input data exhibits variability or uncertainty, such as image recognition, speech recognition, or natural language processing, data augmentation can help expose the model to different variations of the input data and improve its robustness.
- Enhancing Model Performance: Data augmentation can lead to improved model performance by increasing the diversity of the training data and helping the model learn more meaningful patterns and representations.
- Addressing Privacy Concerns: In situations where privacy concerns restrict access to sensitive data, data augmentation can be used to generate synthetic data that preserves the statistical properties of the original data while ensuring privacy and confidentiality.
- Improving Transfer Learning: Data augmentation is often used in conjunction with transfer learning, where a pre-trained model is fine-tuned on a target task. Augmenting the training data with task-specific variations can help fine-tune the model to the target task more effectively.
- Handling Data Sparsity: In tasks involving sparse or incomplete data, data augmentation techniques can help fill in missing information and augment the available data to improve model performance.

# Data Augmentation Methods

- Bayer, Kaufhold, and Reuter (2022).



# Methods for Data Augmentation

- Augmented numeric features
  - Synthetic Minority Over-Sampling Technique (SMOTE; Chawla et al., 2002; Joloudari et al., 2023; Tan et al., 2022)-Interpolation
  - Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN, He et al., 2008)
  - Random over-sampling
  - Random under-sampling (for class imbalance)
- Augmented text data
  - Word-level data augmentation
    - Synonyms Replacement from WordNet (Zhang, Zhao, & LeCun, 2015)
    - Random Swap: randomly swap questions, reference answers, or student answers in the two datasets-noise induction (Wei & Zou, 2019)
    - Random deletion-noise induction (Wei & Zou, 2019)
    - Random insertion (Wei & Zou, 2019)
  - Grammar induction (Jia & Liang, 2016)
  - Back-translation (Yu et al., 2018; Xie et al., 2019; Qu et al., 2020)-round-trip translation
  - Generative AI (Anaby-Tavor et al., 2019; Fang et al., 2023; Kobayashi, 2018; Qiu et al., 2020; Wu et al., 2019)

# Methods for Tabular Data Augmentation

- Bayer, Kaufhold, and Reuter (2022)

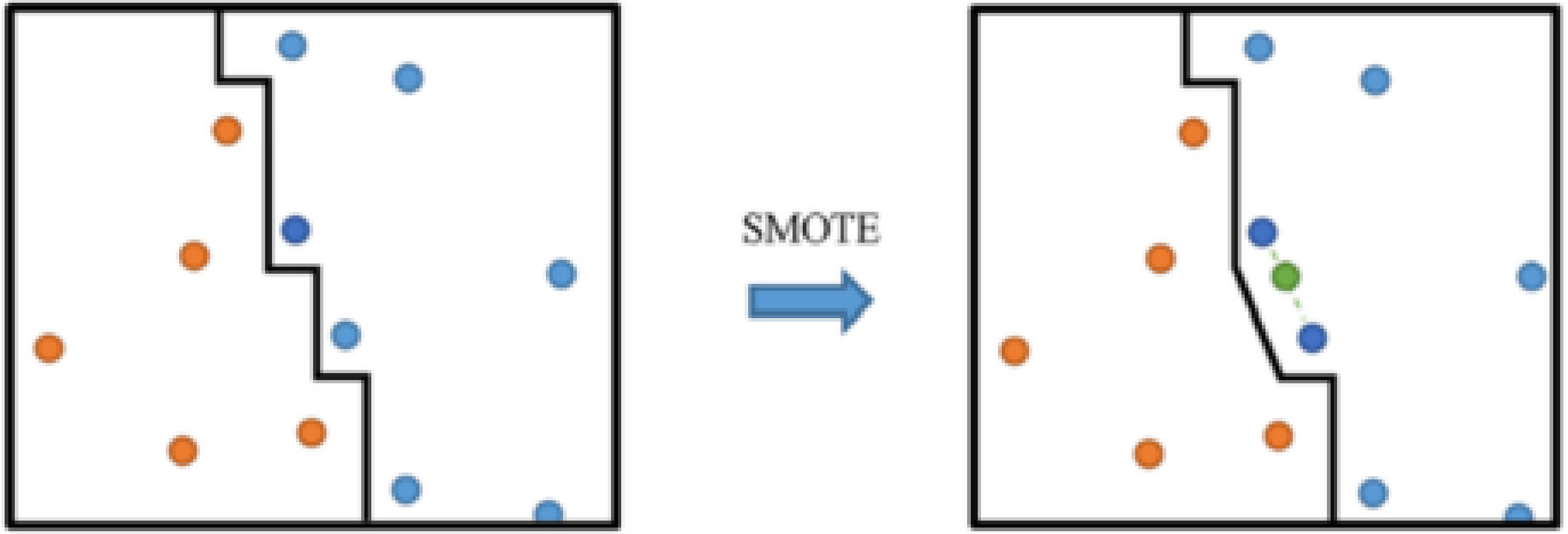
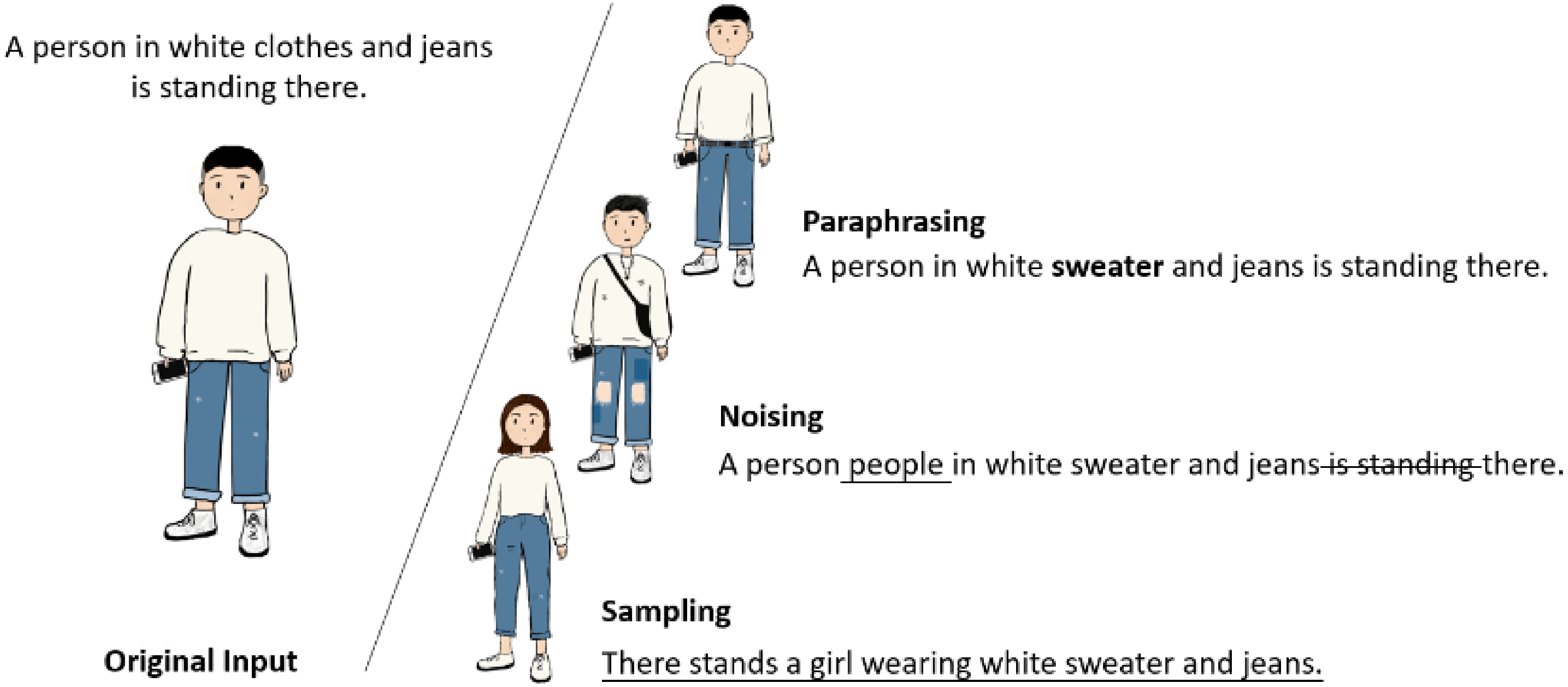


Fig. 5. Illustration of the interpolation method SMOTE.

# Methods for Text Data Augmentation

- Li (https://github.com/BohanLi0110/NLP-DA-Papers?tab=readme-ov-file#21-swapping)



A person in white clothes and jeans is standing there.

**Original Input**

**Paraphrasing**

A person in white **sweater** and jeans is standing there.

**Noising**

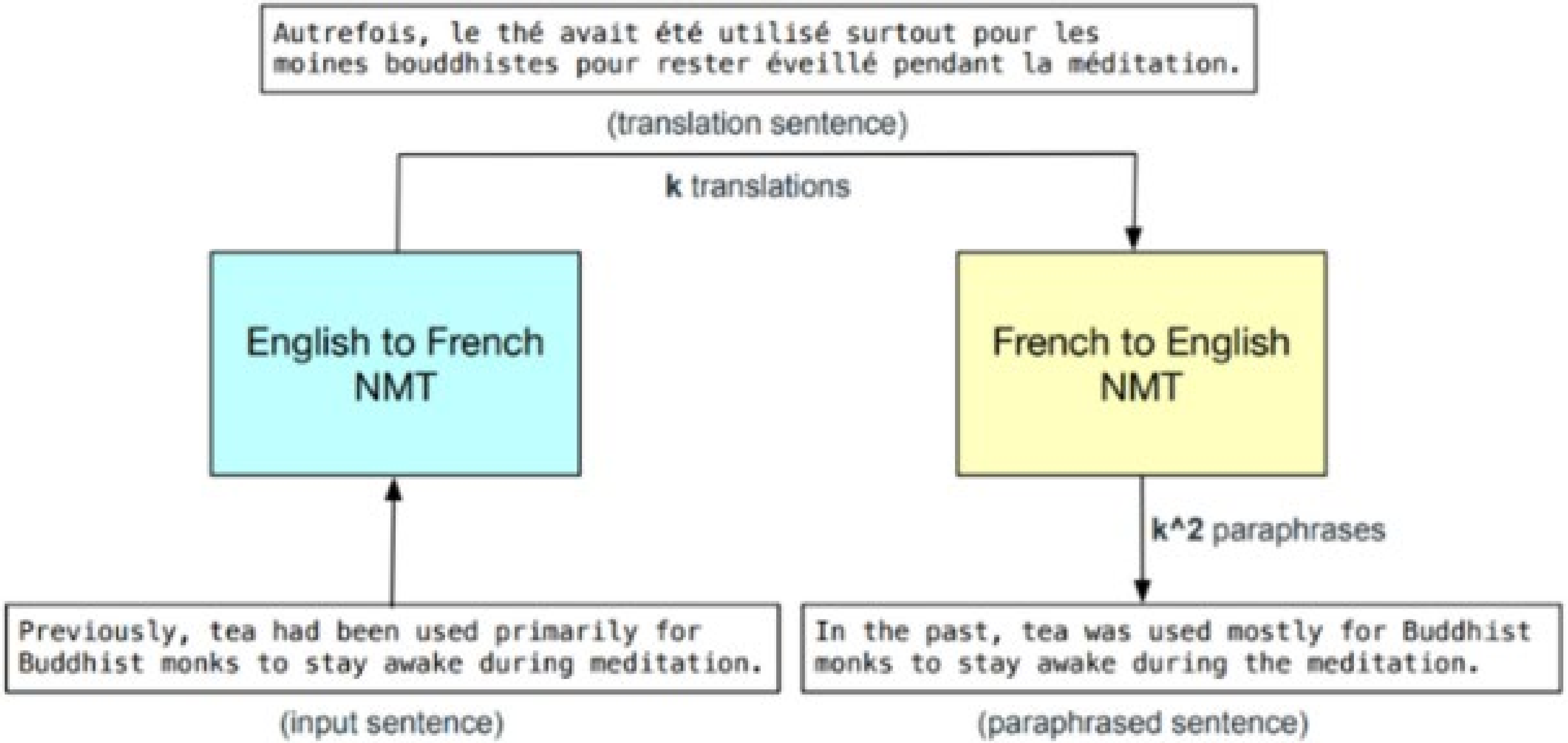
A person people in white sweater and jeans ~~is standing~~ there.

**Sampling**

There stands a girl wearing white sweater and jeans.

# Methods for Data Augmentation

- Back-translation (Yu et al., 2018)



Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.

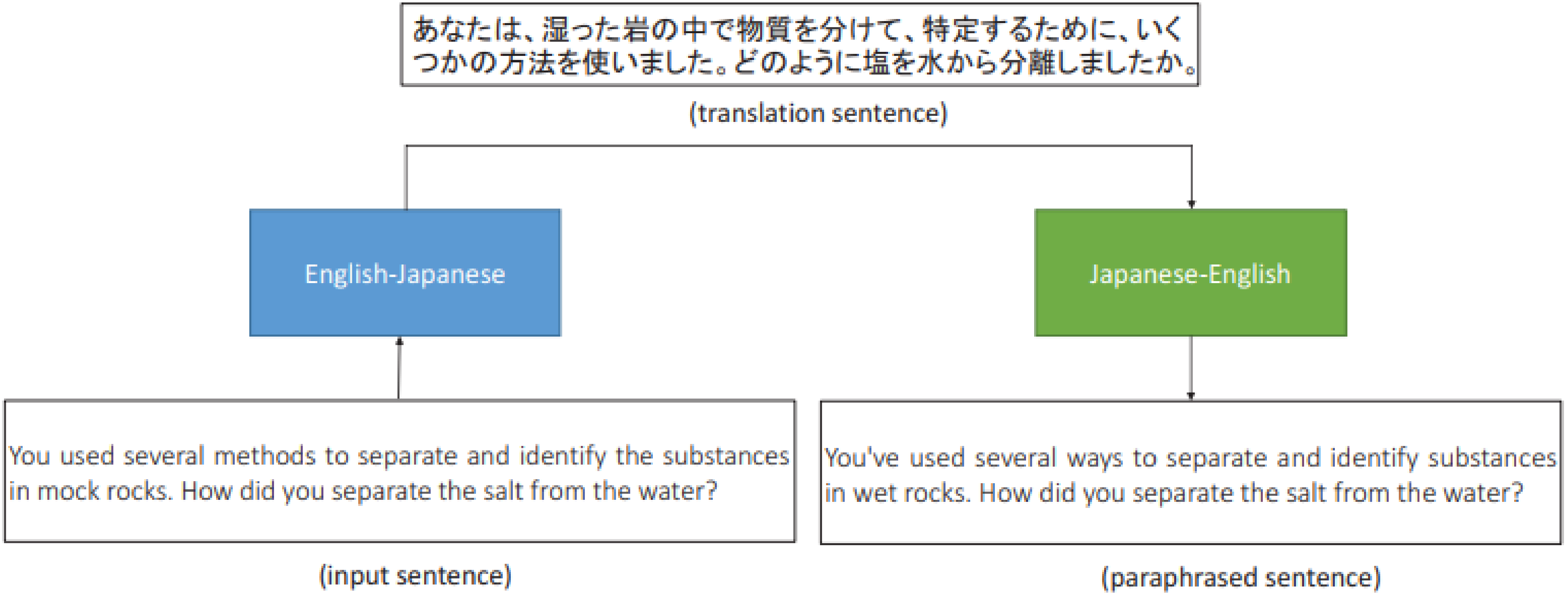
Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.

In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

Fig. 3. Round-trip translation process [94].

# Methods for Data Augmentation

- Back-translation (Lun et al., 2020)



あなたは、湿った岩の中で物質を分けて、特定するために、いくつかの方法を使いました。どのように塩を水から分離しましたか。

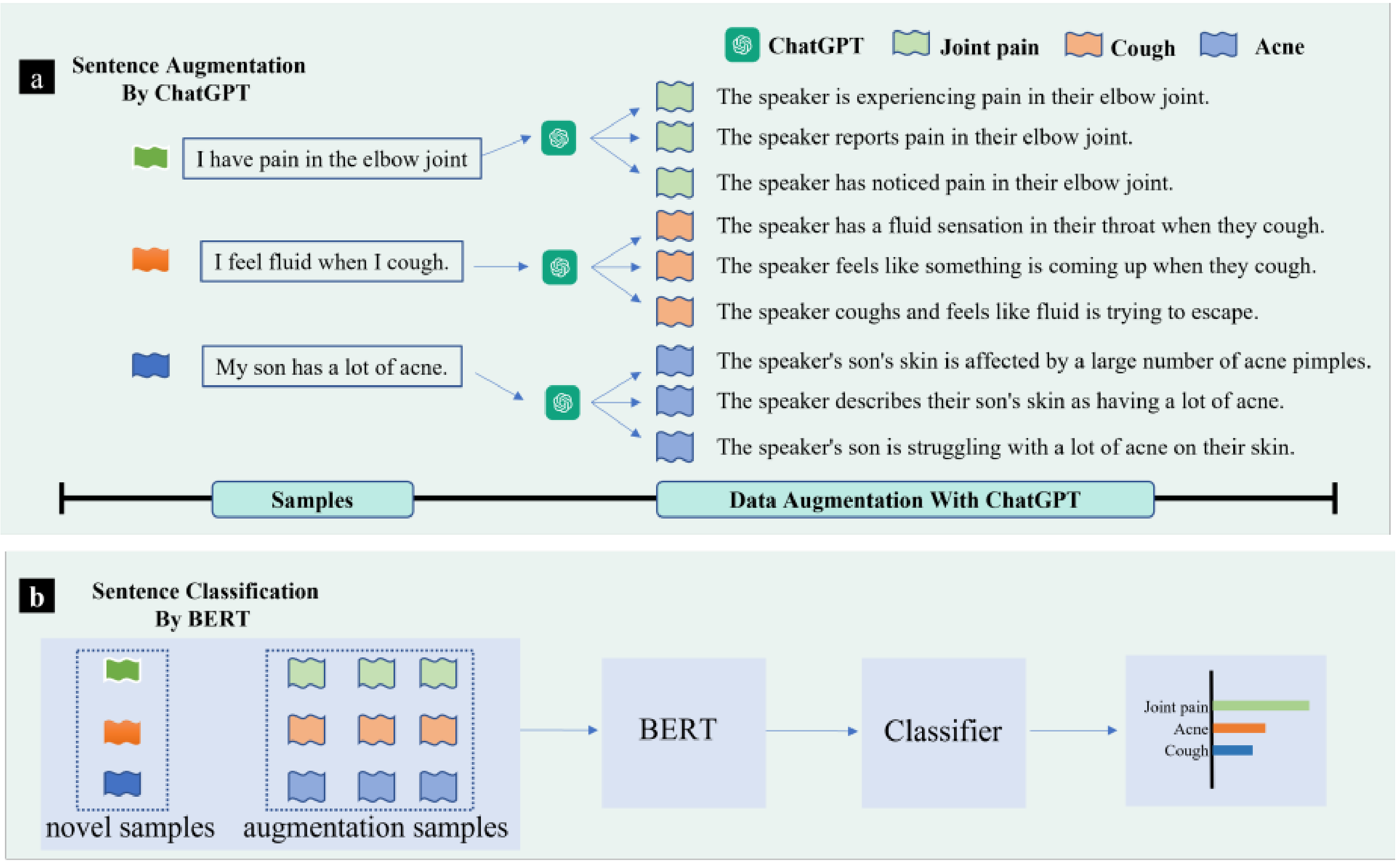
You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?

You've used several ways to separate and identify substances in wet rocks. How did you separate the salt from the water?

Figure 1: An illustration of the whole procedure of back-translation with Japanese as a key language.

# Methods for Data Augmentation

- Generative AI (Dai et al., 2023)



a Sentence Augmentation By ChatGPT

ChatGPT · Joint pain · Cough · Acne

I have pain in the elbow joint

- The speaker is experiencing pain in their elbow joint.
- The speaker reports pain in their elbow joint.
- The speaker has noticed pain in their elbow joint.

I feel fluid when I cough.

- The speaker has a fluid sensation in their throat when they cough.
- The speaker feels like something is coming up when they cough.
- The speaker coughs and feels like fluid is trying to escape.

My son has a lot of acne.

- The speaker's son's skin is affected by a large number of acne pimples.
- The speaker describes their son's skin as having a lot of acne.
- The speaker's son is struggling with a lot of acne on their skin.

Samples → Data Augmentation With ChatGPT

b Sentence Classification By BERT

novel samples · augmentation samples → BERT → Classifier →

Fig. 1. The framework of AugGPT. a (top panel): First, we apply ChatGPT for data augmentation. We input samples of all classes into ChatGPT and prompt ChatGPT to generate samples that preserves semantic consistency with existing labelled instance. b (bottom panel): In the next step, we train a BERT-based sentence classifier on the few-shot samples and the generated data samples and evaluate the model's classification performance.

# Methods for Data Augmentation

- Bayer, Kaufhold, and Reuter (2022)

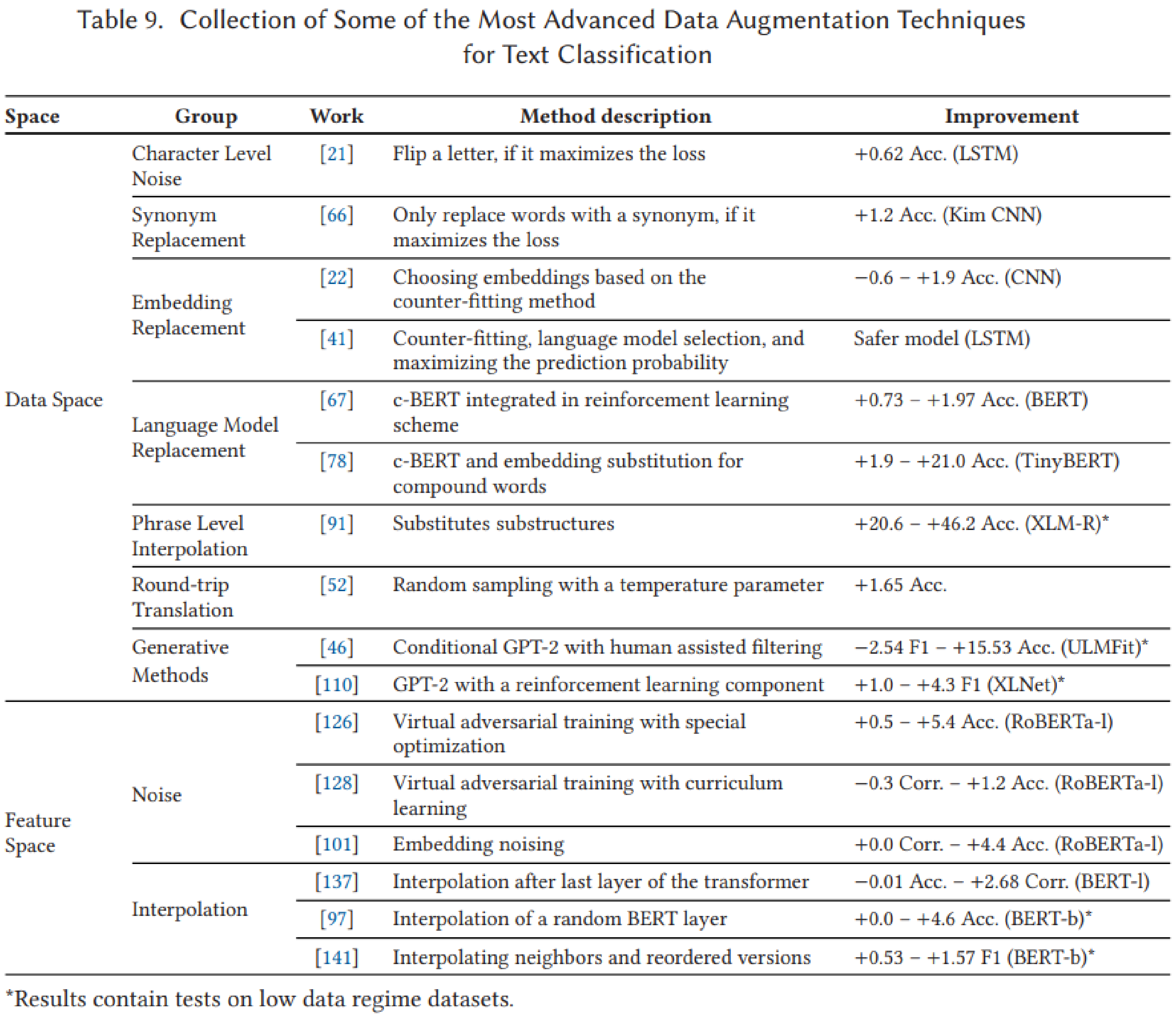


Table 9. Collection of Some of the Most Advanced Data Augmentation Techniques for Text Classification

| Space | Group | Work | Method description | Improvement |
|---|---|---|---|---|
| Data Space | Character Level Noise | [21] | Flip a letter, if it maximizes the loss | +0.62 Acc. (LSTM) |
| | Synonym Replacement | [66] | Only replace words with a synonym, if it maximizes the loss | +1.2 Acc. (Kim CNN) |
| | Embedding Replacement | [22] | Choosing embeddings based on the counter-fitting method | −0.6 – +1.9 Acc. (CNN) |
| | | [41] | Counter-fitting, language model selection, and maximizing the prediction probability | Safer model (LSTM) |
| | Language Model Replacement | [67] | c-BERT integrated in reinforcement learning scheme | +0.73 – +1.97 Acc. (BERT) |
| | | [78] | c-BERT and embedding substitution for compound words | +1.9 – +21.0 Acc. (TinyBERT) |
| | Phrase Level Interpolation | [91] | Substitutes substructures | +20.6 – +46.2 Acc. (XLM-R)* |
| | Round-trip Translation | [52] | Random sampling with a temperature parameter | +1.65 Acc. |
| | Generative Methods | [46] | Conditional GPT-2 with human assisted filtering | −2.54 F1 – +15.53 Acc. (ULMFit)* |
| | | [110] | GPT-2 with a reinforcement learning component | +1.0 – +4.3 F1 (XLNet)* |
| Feature Space | Noise | [126] | Virtual adversarial training with special optimization | +0.5 – +5.4 Acc. (RoBERTa-l) |
| | | [128] | Virtual adversarial training with curriculum learning | −0.3 Corr. – +1.2 Acc. (RoBERTa-l) |
| | | [101] | Embedding noising | +0.0 Corr. – +4.4 Acc. (RoBERTa-l) |
| | Interpolation | [137] | Interpolation after last layer of the transformer | −0.01 Acc. – +2.68 Corr. (BERT-l) |
| | | [97] | Interpolation of a random BERT layer | +0.0 – +4.6 Acc. (BERT-b)* |
| | | [141] | Interpolating neighbors and reordered versions | +0.53 – +1.57 F1 (BERT-b)* |

*Results contain tests on low data regime datasets.

# Class Imbalance

- Several studies explicitly addressed the class imbalance issue in automated scoring research.
  - Zhou and Liu (2006) trained cost-sensitive neural networks with methods addressing class imbalance.
  - Filho et al. (2019) addressed imbalanced class issue in automated essay scoring.
  - Tan et al. (2022) explored Guassian multi-class SMOTE for dataset over-sampling to balance the class for Prompts 1, 2, 7, and 8 of the ASAP datasets. Their proposed approach was able to increase quadratic weighted kappa (QWK) by 5.8% for Prompt 2, the highest among the four prompt-specific models.
  - Jiao, Xu, and Zhou (2021) and Jiao and Lnu (2023) applied techniques for rebalancing class in automated scoring of reading and math CR items respectively.

# Purposes of the Study

- Investigate the impact of class imbalance in automated scoring of science CR items.
- Explored the effectiveness of different methods in class rebalancing.
- Compare the performance of the feature-based modeling vs large language model LLMs-based modeling approaches with class rebalancing mechanism.

Jiao, H., Lnu*, C., & Zhai, X. (2024, April). *Data augmentation for class imbalance in developing automated scoring models for constructed-response items in science assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

# Data: Score Distributions

| Item 1: HH_S1_1c | | | Item 2: HH_S4_2a | | | Item 3: HH_S5_2c | | |
|---|---|---|---|---|---|---|---|---|
| Score | N | % | Score | N | % | Score | N | % |
| 0 | 27 | 3.5 | 0 | 145 | 19.2 | 0 | 278 | 37.4 |
| 1 | 343 | 44.3 | 1 | 379 | 50.2 | 1 | 51 | 6.9 |
| 2 | 308 | 39.7 | 2 | 162 | 21.5 | 2 | 210 | 28.2 |
| 3 | 97 | 12.5 | 3 | 69 | 9.1 | 3 | 205 | 27.6 |
| Total | 775 | 100 | Total | 755 | 100 | Total | 744 | 100 |
| | | | | | | | | |
| Item 4: HH_B1_0b | | | Item 5: MH_B2_0d | | | Item 6: HH_B3_1c | | |
| Score | N | % | Score | N | % | Score | N | % |
| 0 | 257 | 46.8 | 0 | 44 | 8.3 | 0 | 67 | 13.5 |
| 1 | 29 | 5.3 | 1 | 165 | 31.3 | 1 | 190 | 38.2 |
| 2 | 263 | 47.9 | 2 | 318 | 60.3 | 2 | 207 | 41.6 |
| Total | 549 | 100 | Total | 527 | 100 | 3 | 34 | 6.8 |
| | | | | | | Total | 498 | 100 |

# Study Design

- Study conditions
  - Prompt-specific models
    - Feature-based classic machine learning models (220 features, 80% vs 20% split)
      - Stacking
        - Base models: Logistic Regression, SVC, Random Forest, Decision Tree, Gaussian Naïve Bayes, Gradient Boosting, Linear Discriminant Analysis, MLPClassifier
        - Meta-model: Gradient Boosting
    - LLM-based
      - BERT
      - convBERT
      - distilBERT
      - roBERTa
      - Electra
      - deBERTa

# Methods for Class Rebalancing

- Over-sampling
  - Synthetic Minority Over-Sampling Technique (SMOTE; Chawla et al., 2002; Joloudari et al., 2023; Tan et al., 2022)
    - An effective oversampling method by generating synthetic cases in the minority class using the K-nearest neighbor algorithm.
    - First, a case from the minority class is randomly selected.
    - Then, its nearest neighbor is found
    - The difference between the data point and its nearest neighbor is obtained.
    - Lastly, the difference is multiplied by a random number R, where 0 < R < 1, then the synthetic data is added to the feature vector.
    - These steps are repeated until the targeted balanced level is achieved.
    - This strategy effectively broadens the minority class's decision-making region.
- Text Data augmentation
  - NLTK library and Word2Vec-7B dictionary
  - Synonyms Replacement (SR): Word2Vec embeddings to find the closest words in a sentence and then replace them
  - Random Swap (RS)
  - Better protection of data security

# Study Design

- Study conditions
  - Class rebalancing methods
    - Feature-based classic machine learning models
      - Without SMOTE
      - With SMOTE (ratio of 0.9)
        - Without cosine similarity measures
        - With cosine similarity measures
    - LLM-based models
      - Without text data augmentation
      - With text data augmentation
        - Synonyms Replacement (SR)
        - Random Swap (RS)

# Study Design

- Example augmented text:

## Prompt 1

**Original Text:**

- the sugar dissolves into the water because the sugar is like a powder that dissolves in water also the way the water is mixed makes the sugar dissolve into the water which makes it invisible to our eyes

**Augmented Text:**

- the sugar dissolves into the water because the shekels embody like a powder that dissolves in h2o also the way the water is mixed makes the sugar dissolve into the water which make it unseeable to our eyes

**Original Text:**

- i think the sugar dissolved in the glass of water after mary and laura stirred the water mary and laura cannot see the sugar anymore because sugar dissolves it becomes part of the water and can be seen

**Augmented Text:**

- i recall the sugar dissolved in the glass of piddle after mary and laura stirred the water mary and laura cannot see the bread any longer because sugar dissolves it turn part of the water and can be see

# Study Design

- Example augmented text:

## Prompt 2

**Original Text:**

- *the water can turn into three states solid liquid and gas*

**Augmented Text:**

- *the water can reverse in to ternary states liquid gas and solid*

## Prompt 6

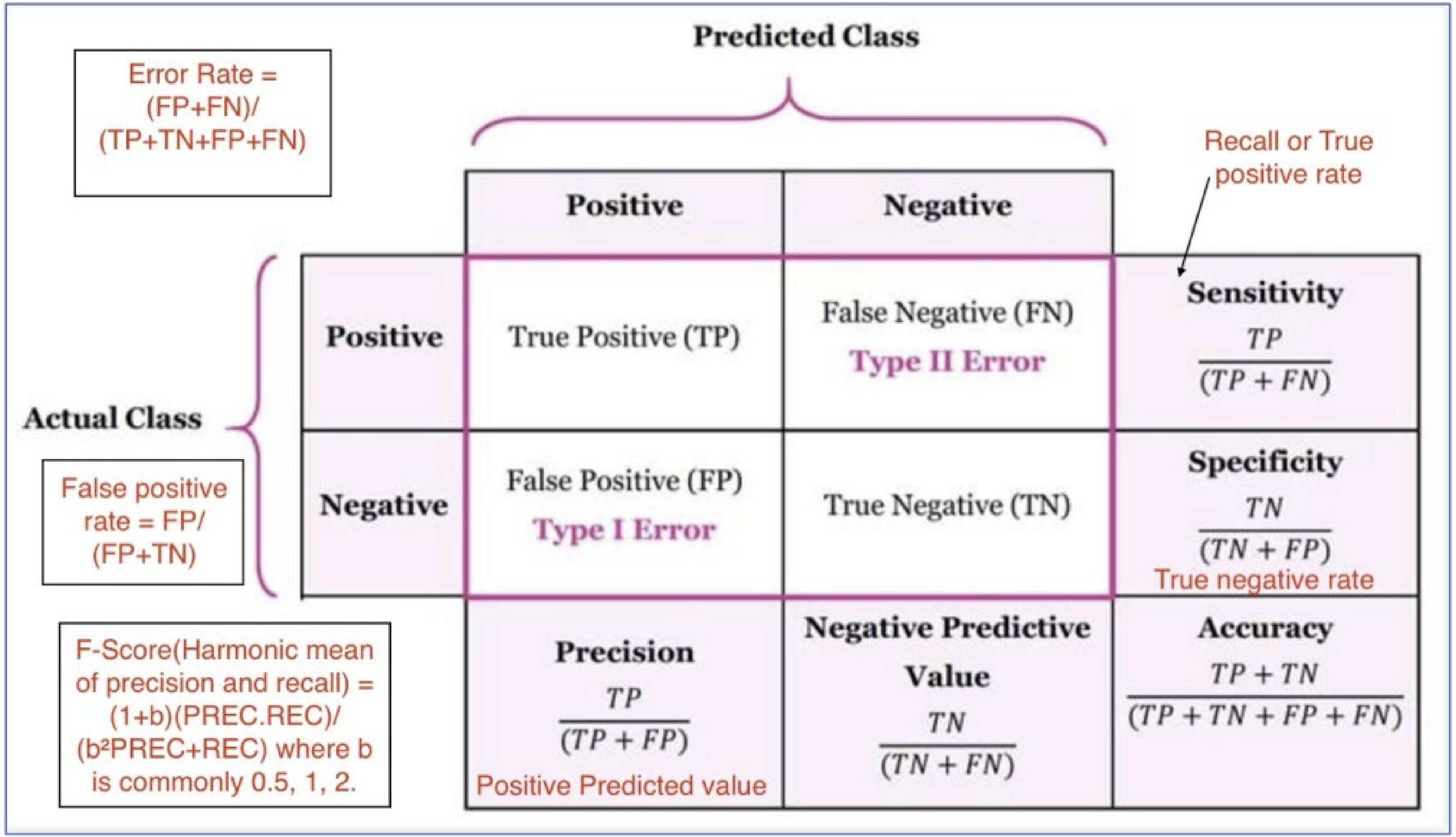
**Original Text:**

- in my opinion it is germy

**Augmented Text:**

- it is germy in my opinion

# Confusion Matrix



Error Rate = (FP+FN)/ (TP+TN+FP+FN)

| | | Predicted Class | | |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual Class | Positive | True Positive (TP) | False Negative (FN)<br>Type II Error | Sensitivity<br>$\frac{TP}{(TP + FN)}$ |
| | Negative | False Positive (FP)<br>Type I Error | True Negative (TN) | Specificity<br>$\frac{TN}{(TN + FP)}$<br>True negative rate |
| | | Precision<br>$\frac{TP}{(TP + FP)}$<br>Positive Predicted value | Negative Predictive Value<br>$\frac{TN}{(TN + FN)}$ | Accuracy<br>$\frac{TP + TN}{(TP + TN + FP + FN)}$ |

Recall or True positive rate

False positive rate = FP/ (FP+TN)

F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/ (b²PREC+REC) where b is commonly 0.5, 1, 2.

# Evaluation Criteria

Quadratic Weighted Kappa (QWK)

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

where TP= True Positives, FP=False Positives, TN= True Negatives and FN= False Negatives

Accuracy

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision

$$Precision = \frac{True\ Positive}{TruePositive + FalsePositive}$$

Recall

$$Recall = \frac{True\ Positive}{TruePositive + FalseNegative}$$

F1 Score

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TruePositive}{2 * TruePositive + FalsePositive + FalseNegative}$$

# Results-Stacking (Baseline)

- Base Models : Logistic Regression, SVC, Random Forest, Decision Tree, Gaussian Naïve Bayes, Gradient Boosting, Linear Discriminant Analysis, MLPClassifier
- Meta Model : Gradient Boosting
- Baseline Approach : Without SMOTE or Cosine Similarity

| Item | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Item 1 | 0.591 | 0.684 | 0.694 | 0.684 | 0.674 |
| Item 2 | 0.502 | 0.609 | 0.610 | 0.609 | 0.609 |
| Item 3 | 0.698 | 0.617 | 0.621 | 0.617 | 0.616 |
| Item 4 | 0.841 | 0.882 | 0.834 | 0.882 | 0.857 |
| Item 5 | 0.516 | 0.689 | 0.672 | 0.689 | 0.677 |
| Item 6 | 0.594 | 0.600 | 0.589 | 0.600 | 0.581 |

# Results-Stacking

- Base Models : Logistic Regression, SVC, Random Forest, Decision Tree, Gaussian Naïve Bayes, Gradient Boost, Linear Discriminant Analysis, MLPClassifier
- Meta Model : Gradient Boosting
- Baseline Approach : With SMOTE but No Cosine Similarity

| Item | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Item 1 | 0.727 | 0.776 | 0.790 | 0.775 | 0.763 |
| Item 2 | 0.638 | 0.728 | 0.704 | 0.727 | 0.708 |
| Item 3 | 0.735 | 0.669 | 0.665 | 0.668 | 0.662 |
| Item 4 | 0.890 | 0.915 | 0.922 | 0.914 | 0.918 |
| Item 5 | 0.644 | 0.796 | 0.791 | 0.795 | 0.792 |
| Item 6 | 0.683 | 0.689 | 0.678 | 0.688 | 0.676 |

# Results-Stacking

- Base Models : Logistic Regression, SVC, Random Forest, Decision Tree, Gaussian Naïve Bayes, Gradient Boost, Linear Discriminant Analysis, MLPClassifier
- Meta Model : Gradient Boosting
- Baseline Approach : With SMOTE and Cosine Similarity

| Item | QWK | Accuracy | Precision | Recall | F1 Score | # Cosine Similarity Measures |
|---|---|---|---|---|---|---|
| Item 1 | 0.782 | 0.809 | 0.819 | 0.808 | 0.796 | 13 |
| Item 2 | 0.715 | 0.801 | 0.781 | 0.800 | 0.783 | 25 |
| Item 3 | 0.780 | 0.776 | 0.775 | 0.775 | 0.774 | 10 |
| Item 4 | 0.974 | 0.988 | 0.994 | 0.979 | 0.979 | 17 |
| Item 5 | 0.738 | 0.836 | 0.843 | 0.835 | 0.832 | 11 |
| Item 6 | 0.760 | 0.699 | 0.702 | 0.698 | 0.689 | 12 |

# Data: Score Distributions

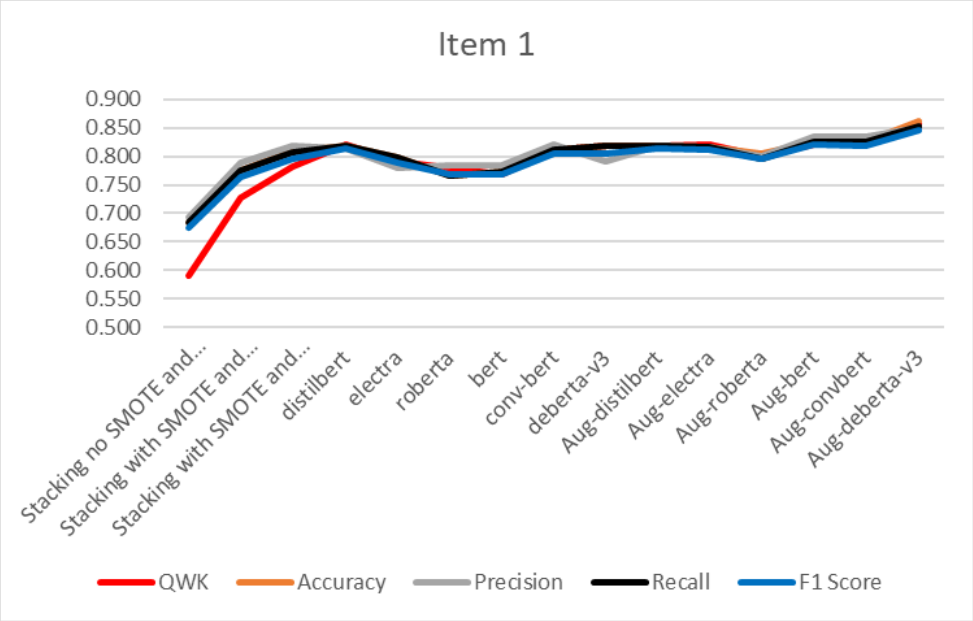
| Item 1: HH_S1_1c | | | Item 2: HH_S4_2a | | | Item 3: HH_S5_2c | | |
|---|---|---|---|---|---|---|---|---|
| Score | N | % | Score | N | % | Score | N | % |
| 0 | 27 | 3.5 | 0 | 145 | 19.2 | 0 | 278 | 37.4 |
| 1 | 343 | 44.3 | 1 | 379 | 50.2 | 1 | 51 | 6.9 |
| 2 | 308 | 39.7 | 2 | 162 | 21.5 | 2 | 210 | 28.2 |
| 3 | 97 | 12.5 | 3 | 69 | 9.1 | 3 | 205 | 27.6 |
| Total | 775 | 100 | Total | 755 | 100 | Total | 744 | 100 |
| | | | | | | | | |
| Item 4: HH_B1_0b | | | Item 5: MH_B2_0d | | | Item 6: HH_B3_1c | | |
| Score | N | % | Score | N | % | Score | N | % |
| 0 | 257 | 46.8 | 0 | 44 | 8.3 | 0 | 67 | 13.5 |
| 1 | 29 | 5.3 | 1 | 165 | 31.3 | 1 | 190 | 38.2 |
| 2 | 263 | 47.9 | 2 | 318 | 60.3 | 2 | 207 | 41.6 |
| Total | 549 | 100 | Total | 527 | 100 | 3 | 34 | 6.8 |
| | | | | | | Total | 498 | 100 |

# Sample Sizes after Text Augmentation

- Across the items, the minority score categories differ

| Training | Item 1 | | Item 2 | | Item 3 | | Item 4 | | Item 5 | | Item 6 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Score | N | Score | N | Score | N | Score | N | Score | N | Score | N |
| | 0 | 191 | 0 | 196 | 0 | 222 | 0 | 206 | 0 | 177 | 0 | 107 |
| | 1 | 274 | 1 | 303 | 1 | 155 | 1 | 136 | 1 | 132 | 1 | 152 |
| | 2 | 246 | 2 | 196 | 2 | 168 | 2 | 210 | 2 | 254 | 2 | 165 |
| | 3 | 191 | 3 | 196 | 3 | 164 | | | | | 3 | 107 |
| | | | | | | | | | | | | |
| Validation | 0 | 5 | 0 | 29 | 0 | 56 | 0 | 51 | 0 | 9 | 0 | 13 |
| | 1 | 69 | 1 | 76 | 1 | 10 | 1 | 6 | 1 | 33 | 1 | 38 |
| | 2 | 62 | 2 | 32 | 2 | 42 | 2 | 53 | 2 | 64 | 2 | 42 |
| | 3 | 19 | 3 | 14 | 3 | 41 | | | | | 3 | 7 |

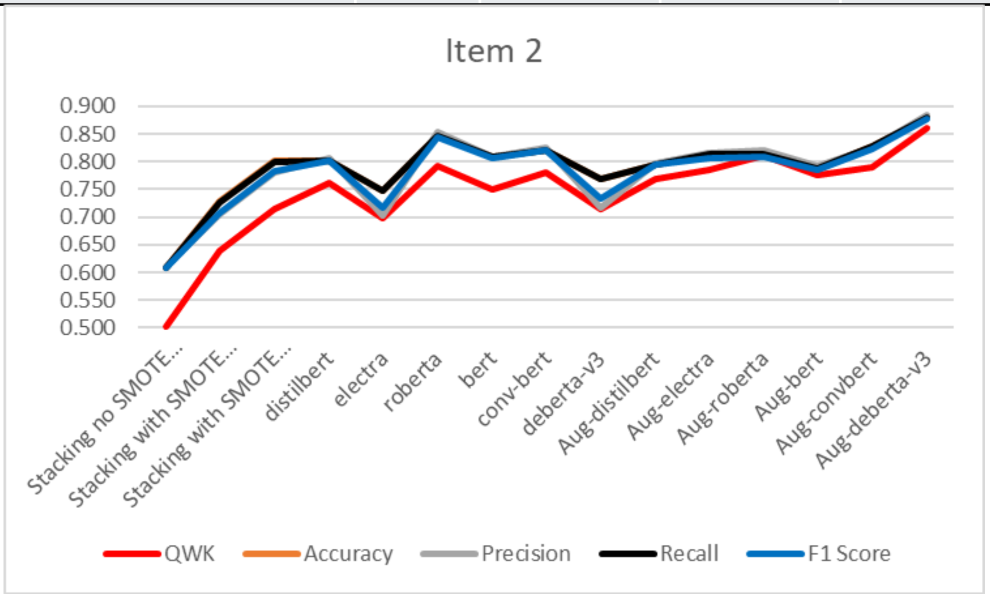
| Model-Item 1 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.591 | 0.684 | 0.694 | 0.684 | 0.674 |
| Stacking with SMOTE and No Cosine Similarity | 0.727 | 0.776 | 0.790 | 0.775 | 0.763 |
| Stacking with SMOTE and Cosine Similarity | 0.782 | 0.809 | 0.819 | 0.808 | 0.796 |
| distilbert | 0.821 | 0.819 | 0.815 | 0.819 | 0.813 |
| electra | 0.792 | 0.799 | 0.779 | 0.799 | 0.788 |
| roberta | 0.780 | 0.767 | 0.785 | 0.767 | 0.768 |
| bert | 0.769 | 0.773 | 0.785 | 0.773 | 0.769 |
| conv-bert | 0.812 | 0.812 | 0.822 | 0.812 | 0.805 |
| deberta-v3 | 0.819 | 0.819 | 0.792 | 0.819 | 0.805 |
| Augmentation-distilbert | 0.818 | 0.819 | 0.818 | 0.819 | 0.814 |
| Augmentation-electra | 0.821 | 0.816 | 0.816 | 0.816 | 0.812 |
| Augmentation-roberta | 0.801 | 0.806 | 0.800 | 0.796 | 0.796 |
| Augmentation-bert | 0.822 | 0.826 | 0.835 | 0.826 | 0.821 |
| Augmentation-convbert | 0.824 | 0.826 | 0.834 | 0.826 | 0.818 |
| Augmentation-deberta-v3 | 0.857 | 0.862 | 0.851 | 0.852 | 0.846 |



- Form item 1, DeBERTa with class rebalancing led to the largest accuracy of automated scoring.

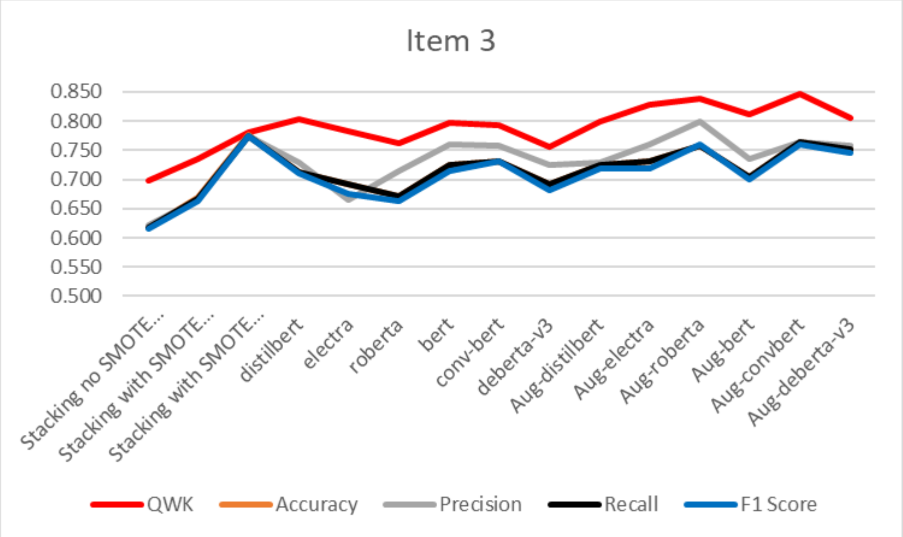
| Model-Item 2 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.502 | 0.609 | 0.610 | 0.609 | 0.609 |
| Stacking with SMOTE and No Cosine Similarity | 0.638 | 0.728 | 0.704 | 0.727 | 0.708 |
| Stacking with SMOTE and Cosine Similarity | 0.715 | 0.801 | 0.781 | 0.800 | 0.783 |
| distilbert | 0.761 | 0.801 | 0.806 | 0.801 | 0.802 |
| electra | 0.699 | 0.748 | 0.703 | 0.748 | 0.717 |
| roberta | 0.792 | 0.848 | 0.854 | 0.848 | 0.845 |
| bert | 0.750 | 0.808 | 0.810 | 0.808 | 0.806 |
| conv-bert | 0.782 | 0.821 | 0.827 | 0.821 | 0.820 |
| deberta-v3 | 0.715 | 0.768 | 0.717 | 0.768 | 0.734 |
| Aug-distilbert | 0.770 | 0.795 | 0.797 | 0.795 | 0.794 |
| Aug-electra | 0.786 | 0.815 | 0.815 | 0.815 | 0.807 |
| Aug-roberta | 0.811 | 0.815 | 0.820 | 0.815 | 0.809 |
| Aug-bert | 0.776 | 0.788 | 0.792 | 0.788 | 0.786 |
| Aug-convbert | 0.790 | 0.828 | 0.826 | 0.828 | 0.824 |
| Aug-deberta-v3 | 0.862 | 0.881 | 0.885 | 0.881 | 0.877 |

- For item 2, DeBERTa with class rebalancing led to the largest improvement in the accuracy of automated scoring.



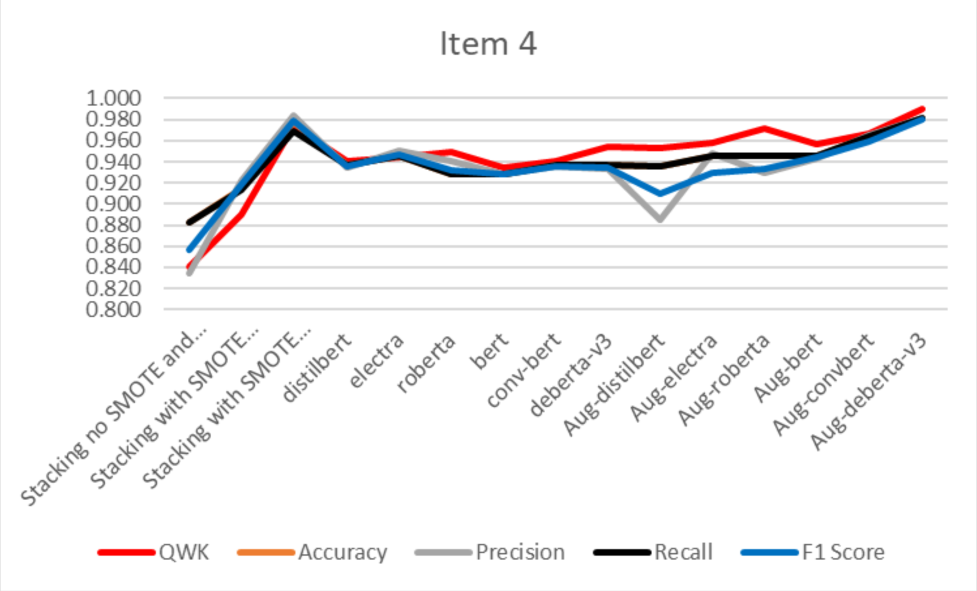
| Model-Item 3 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.698 | 0.617 | 0.621 | 0.617 | 0.616 |
| Stacking with SMOTE and No Cosine Similarity | 0.735 | 0.669 | 0.665 | 0.668 | 0.662 |
| Stacking with SMOTE and Cosine Similarity | 0.780 | 0.776 | 0.775 | 0.775 | 0.774 |
| distilbert | 0.803 | 0.712 | 0.729 | 0.712 | 0.710 |
| electra | 0.783 | 0.692 | 0.665 | 0.692 | 0.675 |
| roberta | 0.763 | 0.671 | 0.714 | 0.671 | 0.664 |
| bert | 0.796 | 0.725 | 0.759 | 0.725 | 0.714 |
| conv-bert | 0.794 | 0.732 | 0.757 | 0.732 | 0.732 |
| deberta-v3 | 0.756 | 0.692 | 0.726 | 0.692 | 0.681 |
| Augmentation-distilbert | 0.800 | 0.725 | 0.730 | 0.725 | 0.718 |
| Augmentation-electra | 0.829 | 0.732 | 0.761 | 0.732 | 0.719 |
| Augmentation-roberta | 0.838 | 0.758 | 0.798 | 0.758 | 0.759 |
| Augmentation-bert | 0.811 | 0.705 | 0.736 | 0.705 | 0.700 |
| Augmentation-convbert | 0.846 | 0.765 | 0.765 | 0.765 | 0.761 |
| Augmentation-deberta-v3 | 0.804 | 0.752 | 0.758 | 0.752 | 0.745 |

- For item 3, ConvBERT with data augmentation led to the highest scoring accuracy.



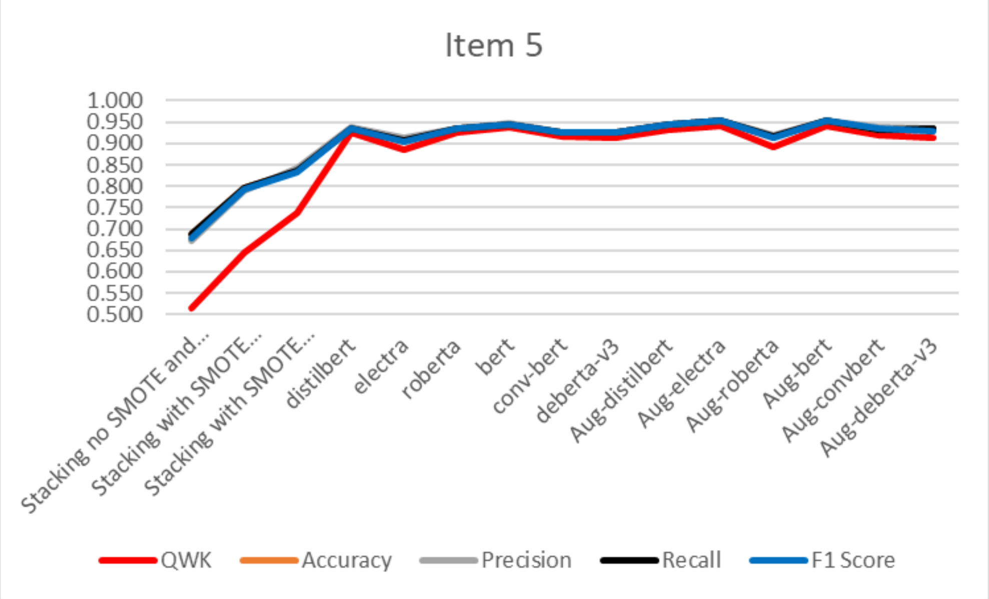
| Model-Item 4 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.841 | 0.882 | 0.834 | 0.882 | 0.857 |
| Stacking with SMOTE and No Cosine Similarity | 0.890 | 0.915 | 0.922 | 0.914 | 0.918 |
| Stacking with SMOTE and Cosine Similarity | 0.974 | 0.978 | 0.984 | 0.969 | 0.979 |
| distilbert | 0.940 | 0.937 | 0.935 | 0.937 | 0.936 |
| electra | 0.944 | 0.946 | 0.951 | 0.946 | 0.947 |
| roberta | 0.949 | 0.928 | 0.940 | 0.928 | 0.932 |
| bert | 0.935 | 0.928 | 0.928 | 0.928 | 0.928 |
| conv-bert | 0.940 | 0.937 | 0.936 | 0.937 | 0.936 |
| deberta-v3 | 0.954 | 0.937 | 0.933 | 0.937 | 0.934 |
| Augmentation-distilbert | 0.953 | 0.936 | 0.885 | 0.936 | 0.910 |
| Augmentation-electra | 0.958 | 0.945 | 0.948 | 0.945 | 0.930 |
| Augmentation-roberta | 0.972 | 0.945 | 0.929 | 0.945 | 0.933 |
| Augmentation-bert | 0.957 | 0.945 | 0.943 | 0.945 | 0.944 |
| Augmentation-convbert | 0.967 | 0.964 | 0.965 | 0.964 | 0.959 |
| Augmentation-deberta-v3 | 0.990 | 0.982 | 0.982 | 0.982 | 0.980 |

- For item 4, DeBERTa with augmentation led to the highest QWK.

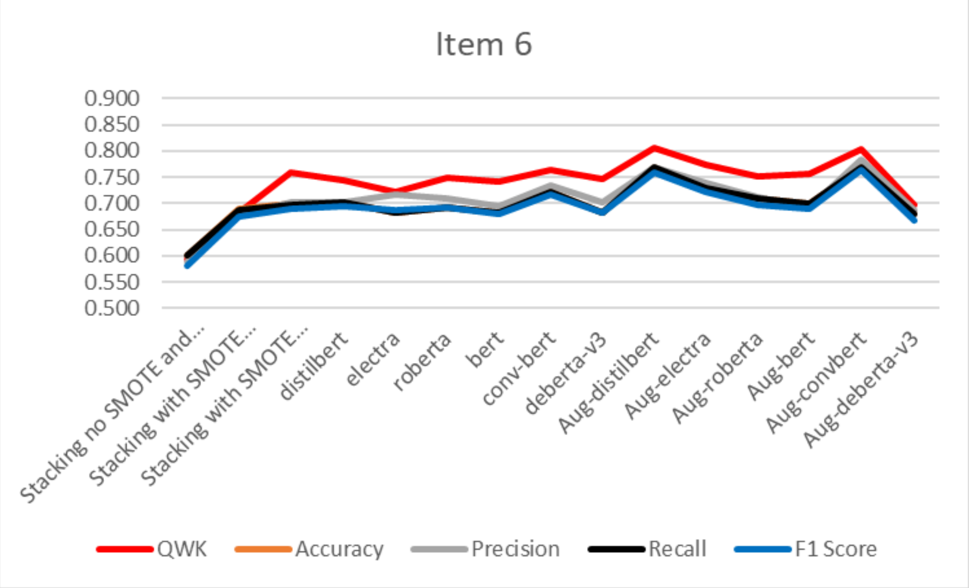


| Model-Item 5 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.516 | 0.689 | 0.672 | 0.689 | 0.677 |
| Stacking with SMOTE and No Cosine Similarity | 0.644 | 0.796 | 0.791 | 0.795 | 0.792 |
| Stacking with SMOTE and Cosine Similarity | 0.738 | 0.836 | 0.843 | 0.835 | 0.832 |
| distilbert | 0.927 | 0.936 | 0.937 | 0.936 | 0.936 |
| electra | 0.887 | 0.907 | 0.912 | 0.907 | 0.904 |
| roberta | 0.927 | 0.936 | 0.935 | 0.936 | 0.935 |
| bert | 0.938 | 0.945 | 0.946 | 0.945 | 0.944 |
| conv-bert | 0.916 | 0.926 | 0.927 | 0.926 | 0.926 |
| deberta-v3 | 0.914 | 0.926 | 0.927 | 0.926 | 0.926 |
| Augmentation-distilbert | 0.931 | 0.943 | 0.944 | 0.943 | 0.943 |
| Augmentation-electra | 0.941 | 0.953 | 0.954 | 0.953 | 0.952 |
| Augmentation-roberta | 0.891 | 0.915 | 0.919 | 0.915 | 0.913 |
| Augmentation-bert | 0.941 | 0.953 | 0.954 | 0.953 | 0.952 |
| Augmentation-convbert | 0.919 | 0.933 | 0.938 | 0.933 | 0.934 |
| Augmentation-deberta-v3 | 0.913 | 0.934 | 0.936 | 0.934 | 0.929 |

- For item 5, BERT and ELECTRA with augmentation performed slightly better than BERT with no-augmentation.



| Model-Item 6 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking no SMOTE and Cosine Similarity | 0.594 | 0.600 | 0.589 | 0.600 | 0.581 |
| Stacking with SMOTE and No Cosine Similarity | 0.683 | 0.689 | 0.678 | 0.688 | 0.676 |
| Stacking with SMOTE and Cosine Similarity | 0.760 | 0.699 | 0.702 | 0.698 | 0.689 |
| distilbert | 0.745 | 0.703 | 0.702 | 0.703 | 0.694 |
| electra | 0.723 | 0.683 | 0.717 | 0.683 | 0.687 |
| roberta | 0.750 | 0.693 | 0.709 | 0.693 | 0.693 |
| bert | 0.741 | 0.683 | 0.695 | 0.683 | 0.681 |
| conv-bert | 0.763 | 0.723 | 0.735 | 0.723 | 0.716 |
| deberta-v3 | 0.746 | 0.683 | 0.703 | 0.683 | 0.683 |
| Augmentation-distilbert | 0.807 | 0.770 | 0.770 | 0.770 | 0.759 |
| Augmentation-electra | 0.773 | 0.730 | 0.739 | 0.730 | 0.723 |
| Augmentation-roberta | 0.751 | 0.710 | 0.711 | 0.710 | 0.698 |
| Augmentation-bert | 0.757 | 0.700 | 0.692 | 0.700 | 0.689 |
| Augmentation-convbert | 0.803 | 0.770 | 0.783 | 0.770 | 0.765 |
| Augmentation-deberta-v3 | 0.696 | 0.680 | 0.686 | 0.680 | 0.668 |



- For item 6, DistilBERT with data augmentation yielded the highest accuracy of automated scoring.

# Summary

- **QWK**
  - According to the criteria for the machine-human scoring agreement proposed by Nehm and Haertig (2012), Cohen’s κ values between .41 to .60 are moderate, 0.61 to 0.80 is substantial, and 0.81 to 1.00 is (almost) perfect. In our study, the overall machine-human scoring agreement is all above 0.8, with one item QWK is 0.99.
- **Data Augmentation**
  - SMOTE for classical machine learning methods and Synonyms Replacement (SR)/Random Swap (RS) for LLMs both improved accuracy.
  - LLMs with data augmentation led to the highest automated scoring accuracy for every studied item though the best-performing LLM could be different for different items.

# Highlight 1

## AI is not panacea

- Uses of AI do not guarantee you have a technically defensible solution.

## Technical standards need to be followed

- Empirical comparison and evaluation is critical.

# Not Every Automated Scoring Model Works

| Model-Item 1 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking neither SMOTE nor Cosine Similarity | **0.591** | 0.684 | 0.694 | 0.684 | 0.674 |
| Stacking with SMOTE but No Cosine Similarity | 0.727 | 0.776 | 0.790 | 0.775 | 0.763 |
| Stacking with SMOTE and Cosine Similarity | 0.782 | 0.809 | 0.819 | 0.808 | 0.796 |
| distilbert | 0.821 | 0.819 | 0.815 | 0.819 | 0.813 |
| electra | 0.792 | 0.799 | 0.779 | 0.799 | 0.788 |
| roberta | 0.780 | 0.767 | 0.785 | 0.767 | 0.768 |
| bert | 0.769 | 0.773 | 0.785 | 0.773 | 0.769 |
| conv-bert | 0.812 | 0.812 | 0.822 | 0.812 | 0.805 |
| deberta-v3 | 0.819 | 0.819 | 0.792 | 0.819 | 0.805 |
| Augmentation-distilbert | 0.818 | 0.819 | .818 | 0.819 | 0.814 |
| Augmentation-electra | 0.821 | 0.816 | 0.816 | 0.816 | 0.812 |
| Augmentation-roberta | 0.801 | 0.806 | 0.800 | 0.796 | 0.796 |
| Augmentation-bert | 0.822 | 0.826 | 0.835 | 0.826 | 0.821 |
| Augmentation-convbert | 0.824 | 0.826 | 0.834 | 0.826 | 0.818 |
| Augmentation-deberta-v3 | **0.857** | 0.862 | 0.851 | 0.852 | 0.846 |

| Model-Item 5 | QWK | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| Stacking neither SMOTE nor Cosine Similarity | **0.516** | 0.689 | 0.672 | 0.689 | 0.677 |
| Stacking with SMOTE but No Cosine Similarity | 0.644 | 0.796 | 0.791 | 0.795 | 0.792 |
| Stacking with SMOTE and Cosine Similarity | 0.738 | 0.836 | 0.843 | 0.835 | 0.832 |
| distilbert | 0.927 | 0.936 | 0.937 | 0.936 | 0.936 |
| electra | 0.887 | 0.907 | 0.912 | 0.907 | 0.904 |
| roberta | 0.927 | 0.936 | 0.935 | 0.936 | 0.935 |
| bert | 0.938 | 0.945 | 0.946 | 0.945 | 0.944 |
| conv-bert | 0.916 | 0.926 | 0.927 | 0.926 | 0.926 |
| deberta-v3 | 0.914 | 0.926 | 0.927 | 0.926 | 0.926 |
| Augmentation-distilbert | 0.931 | 0.943 | 0.944 | 0.943 | 0.943 |
| Augmentation-electra | **0.941** | 0.953 | 0.954 | 0.953 | 0.952 |
| Augmentation-roberta | 0.891 | 0.915 | 0.919 | 0.915 | 0.913 |
| Augmentation-bert | **0.941** | 0.953 | 0.954 | 0.953 | 0.952 |
| Augmentation-convbert | 0.919 | 0.933 | 0.938 | 0.933 | 0.934 |
| Augmentation-deberta-v3 | 0.913 | 0.934 | 0.936 | 0.934 | 0.929 |

- Jiao, Lnu, Zhang, & Zhai (2024)-Science CR items

# Not Every Item Automatically Scoreable

| Item | distilBERT | DeBERTa-Base | ELECTRA-Base | ELECTRA-Base+nlpaug | BERT-Base+nlpaug | MathBERT+ANN+nlpaug | MathBERT+LSTM+CNN+nlpaug |
|---|---|---|---|---|---|---|---|
| Item 1 | 0.933 | 0.720 | 0.938 | 0.906 | 0.704 | 0.784 | 0.770 |
| Item 2 | 0.802 | 0.620 | 0.788 | 0.777 | 0.643 | 0.708 | 0.703 |
| Item 3 | 0.832 | 0.770 | 0.857 | 0.865 | 0.777 | 0.854 | 0.860 |
| Item 4 | 0.816 | 0.780 | 0.849 | 0.833 | 0.815 | 0.844 | 0.847 |
| Item 5 | 0.827 | 0.790 | 0.860 | 0.827 | 0.720 | 0.785 | 0.802 |
| Item 6 | 0.793 | 0.770 | 0.758 | 0.773 | 0.757 | 0.782 | 0.796 |
| **Item 7** | **0.933** | **0.770** | **0.943** | **0.940** | **0.779** | **0.814** | **0.817** |
| Item 8 | 0.721 | 0.430 | 0.705 | 0.670 | 0.408 | 0.481 | 0.490 |
| Item 9 | 0.904 | 0.870 | 0.943 | 0.933 | 0.862 | 0.890 | 0.885 |
| **Item 10** | **0.684** | **0.610** | **0.665** | **0.644** | **0.624** | **0.609** | **0.626** |

- Jiao, Lnu, & Shah (2023)-Math CR items: LLM-based automated scoring models

# Automated Scoring Engine Evaluation

| Measure | Threshold |
|---|---|
| Pearson R | ≥ 0.70 |
| Quadratic Weighted Kappa (QWK) | ≥ 0.70 |
| Kappa | ≥ 0.40 |
| Exact Agreement | ≥ 65% (or better than human-human agreement) |
| Per score point agreement | ≥ 50% (or better than human-human agreement) |
| Standardized Mean Difference (SMD) | Within \|0.15\| |

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices*, 31, 2.

# Highlight 2

- AI is competent in processing text/ multimodal data and unstructured data.
- Traditional psychometric theory and practices are built upon principled approach of assessment design.
- Both have unique values and advantages in developing state assessments.

- **Integration of traditional psychometric methods and AI methods may enhance the development of state assessments.**

# Cheating Detection

- Integration of traditional psychometric methods (person fit measures) and AI methods (ML methods for item responses and process data as well as anomaly detection methods) may enhance the accuracy in cheating detection.

# Meta-Model vs. Base Model Comparison

- Used item response and response time data

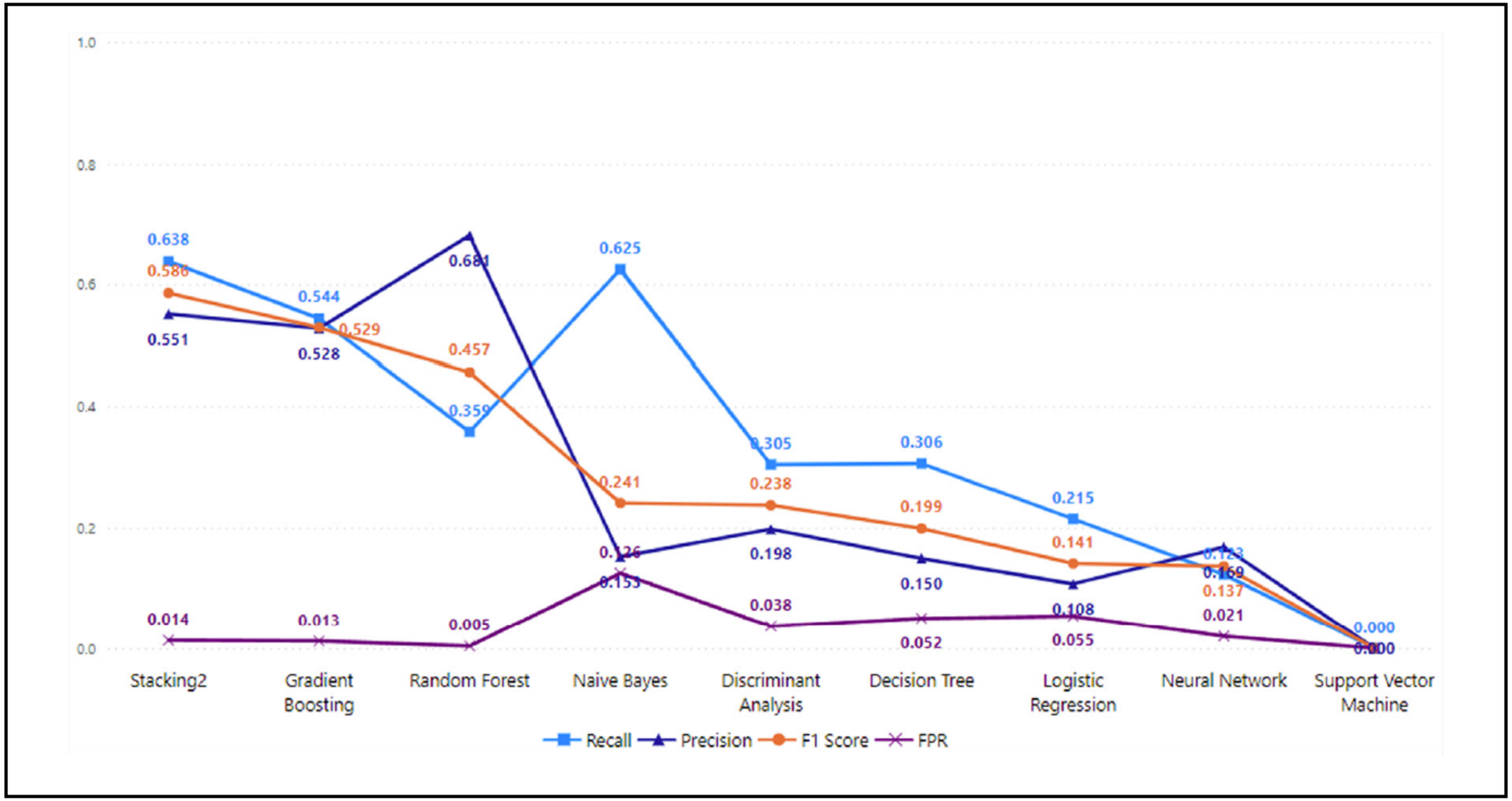
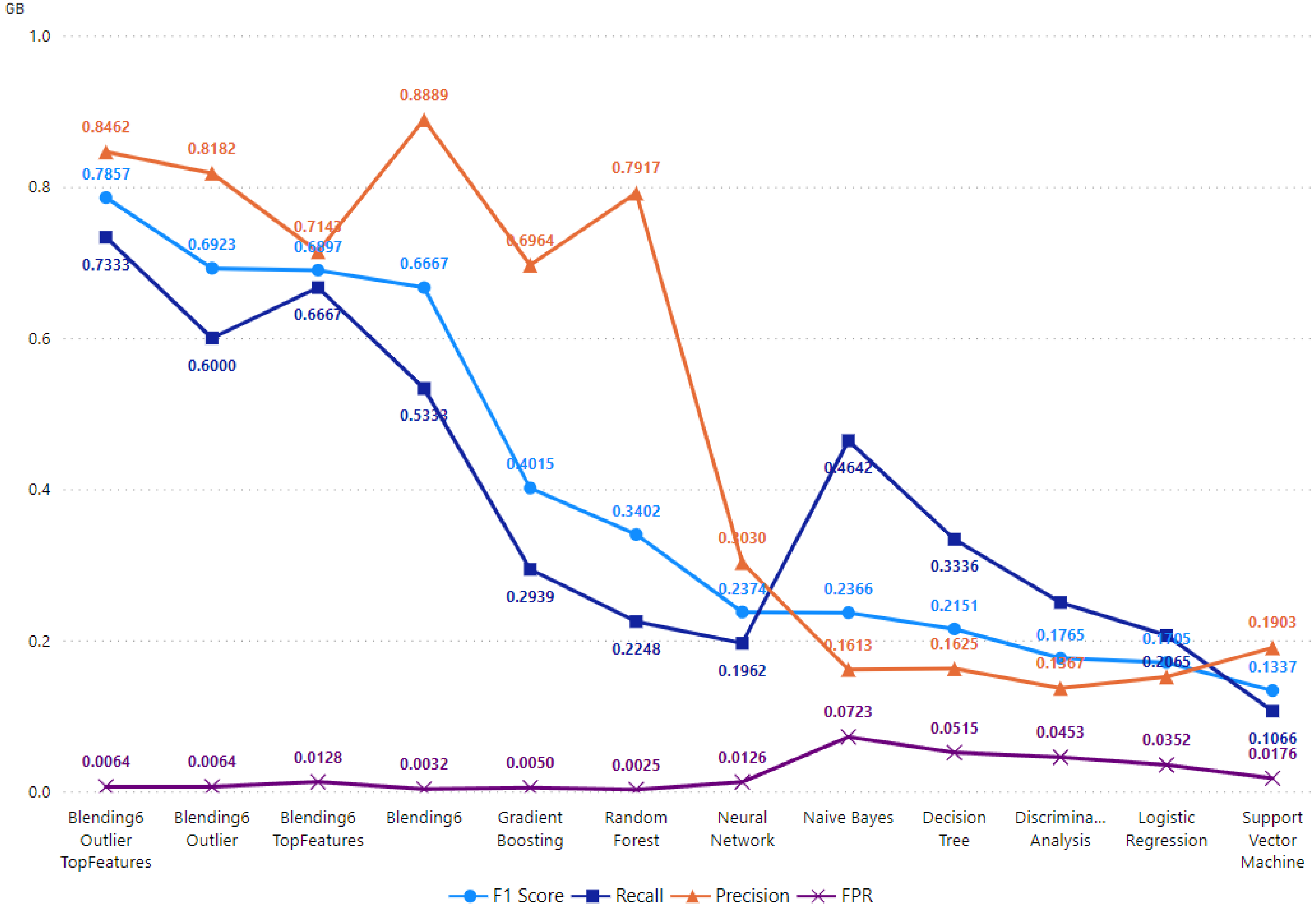


**Figure 8.** Comparison Between the Base Models and the Meta-Model Built Upon Stacking Using Discriminant Analysis Based on Item Response and Summative Statistics for an Under-Sampling Ratio of 10:1 Between the Non-Cheater and Cheater Classes.

- Jiao & Zhou (2022a)-Stacking ensemble learning for cheating detection

# Meta-Model vs. Base Model Comparison

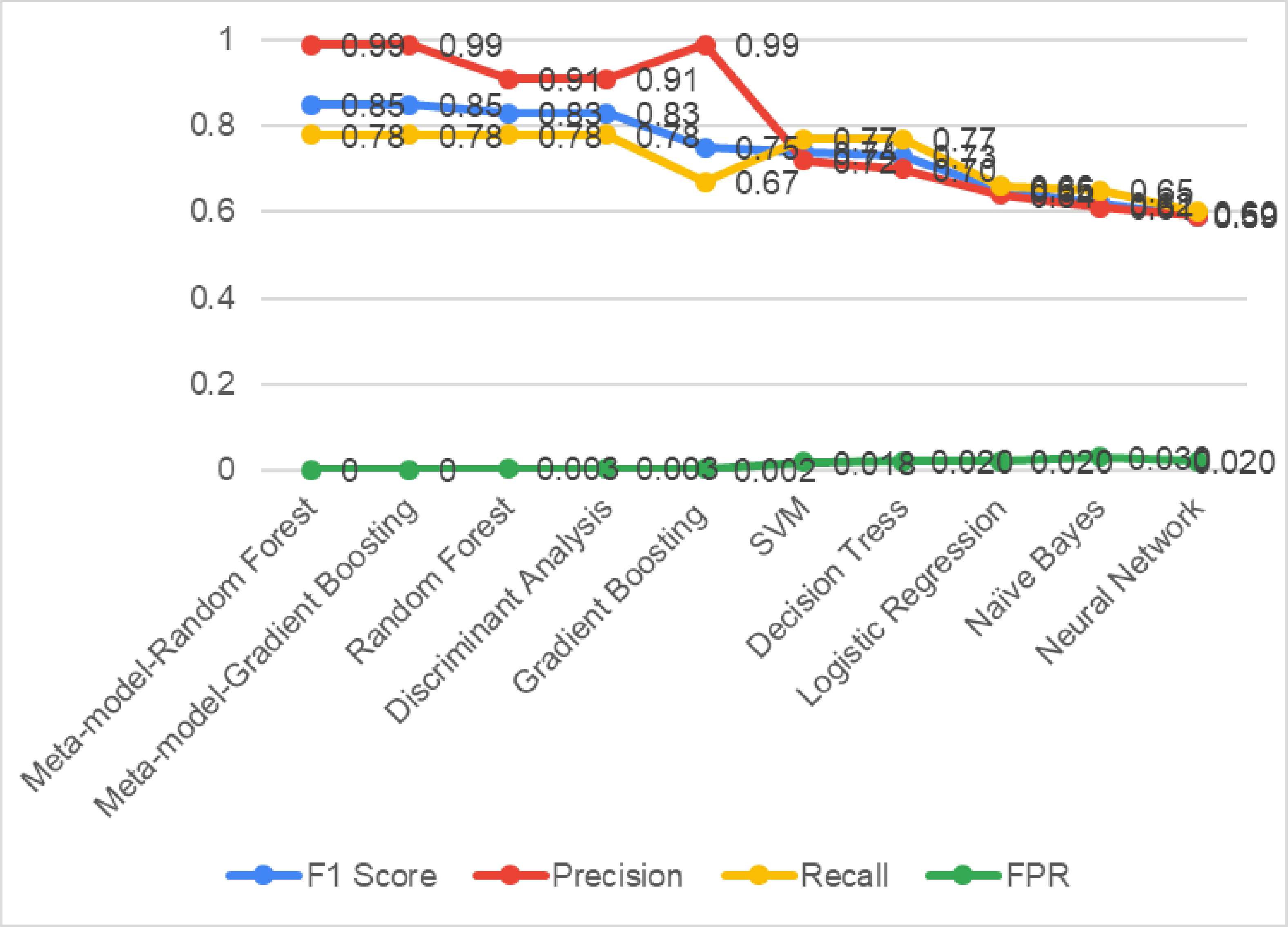
- Used item responses, response time, and anomaly detection measures (Isolation Forest, Elliptic Envelope, Local Outlier Factor, One Class SVM, DBSCAN) from machine learning.



- Jiao & Zhou (2022b)-Blending ensemble learning for cheating detection with data augmentation by anomaly detection methods

# Meta-Model vs. Base Model Comparison

- Used item responses, response time, anomaly detection measures (Isolation Forest, Elliptic Envelope, One Class SVM, Local Outlier Factor) from machine learning, and person fit measures (29 fit indices) after fit to an IRT model.



- Jiao et al.(2023)-integrating person fit measures, anomaly detection measures

# Highlight 3

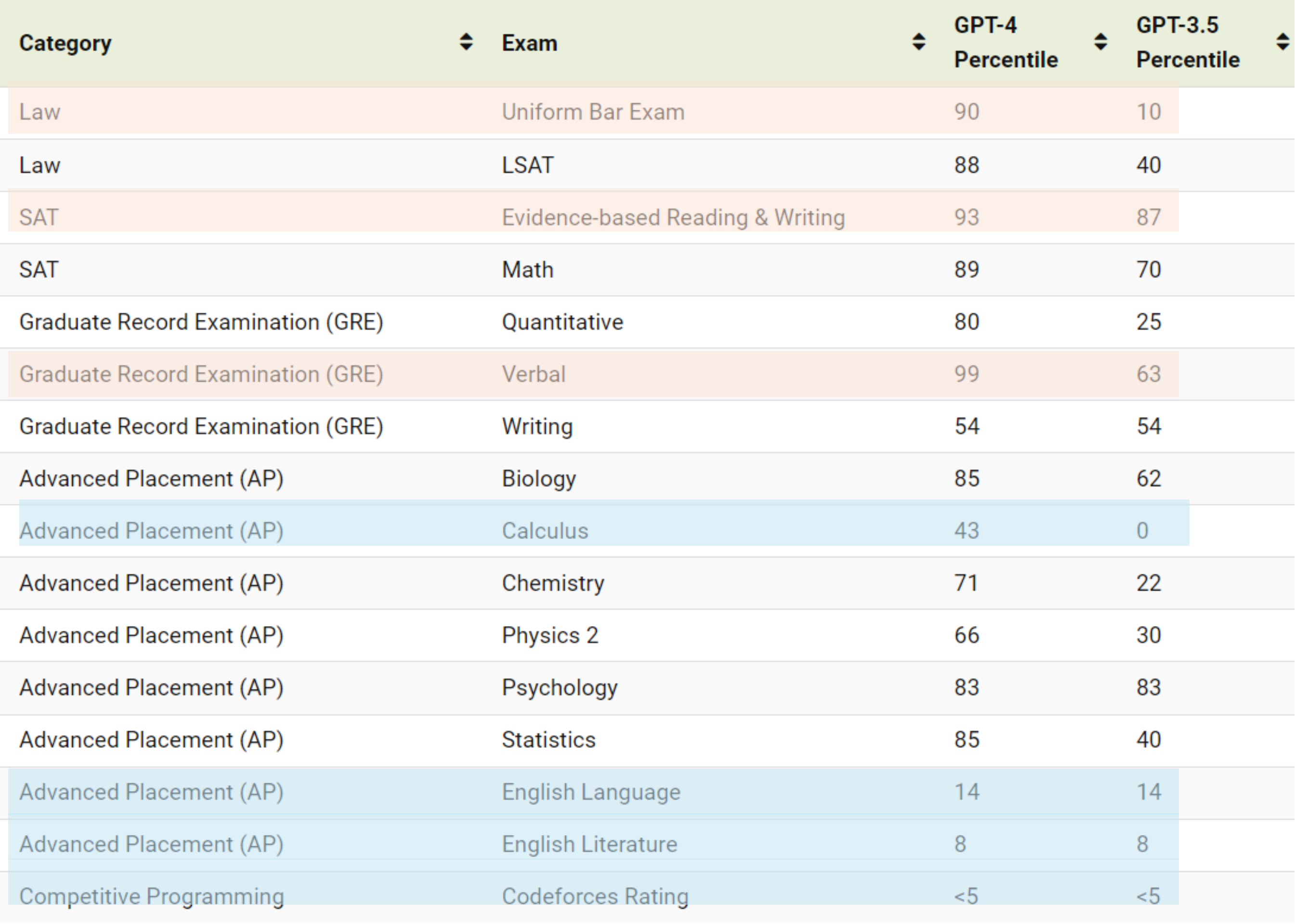
- GPT 4 could be very different from GPT 3.5 on some functionalities.

- **Dramatic improvement on some tasks, but not all**
- **Take advantages of the advances in AI**

UNIVERSITY OF MARYLAND 1856

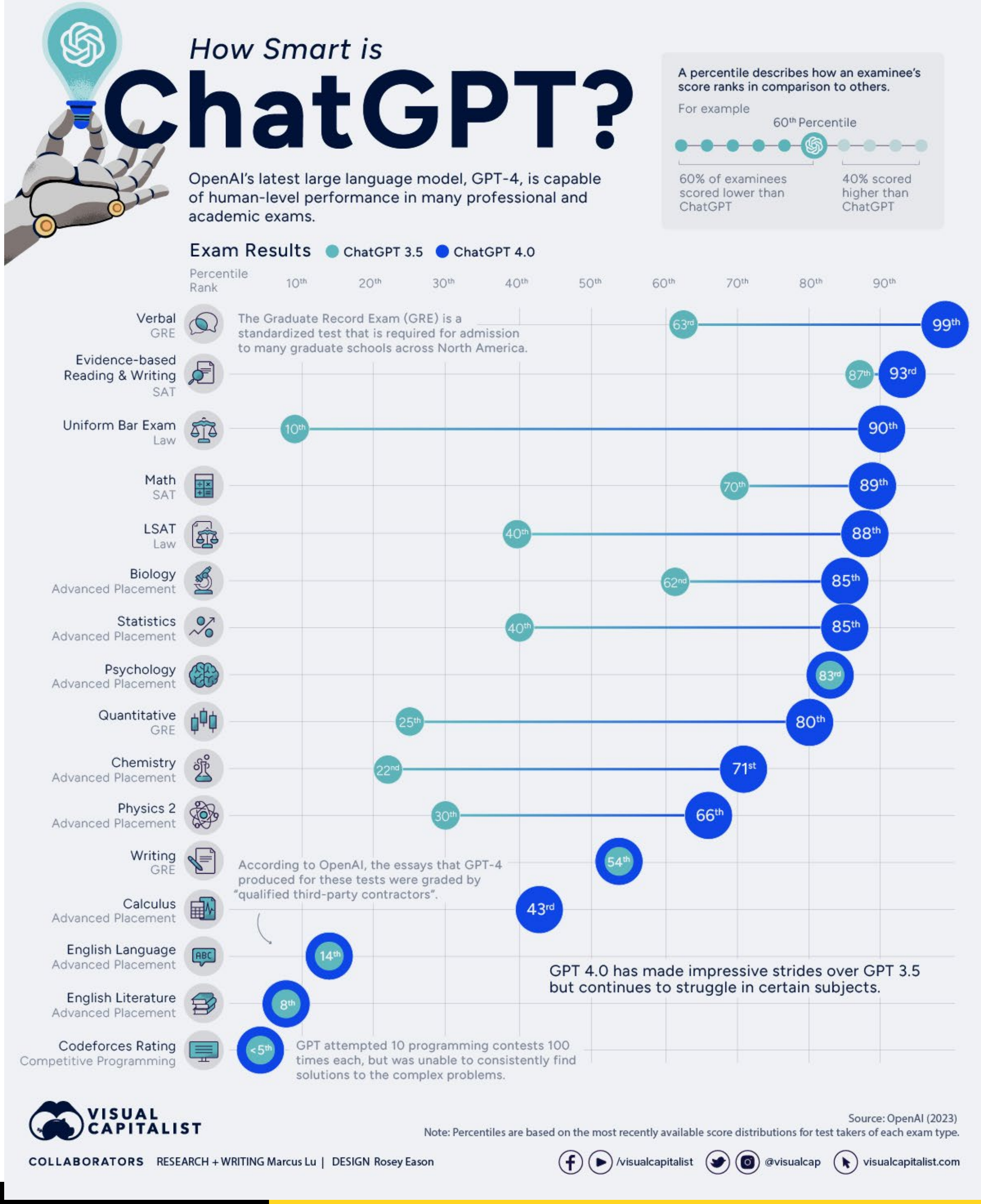
COLLEGE OF EDUCATION
MARYLAND ASSESSMENT RESEARCH CENTER

# AI is improving



| Category | Exam | GPT-4 Percentile | GPT-3.5 Percentile |
|---|---|---|---|
| Law | Uniform Bar Exam | 90 | 10 |
| Law | LSAT | 88 | 40 |
| SAT | Evidence-based Reading & Writing | 93 | 87 |
| SAT | Math | 89 | 70 |
| Graduate Record Examination (GRE) | Quantitative | 80 | 25 |
| Graduate Record Examination (GRE) | Verbal | 99 | 63 |
| Graduate Record Examination (GRE) | Writing | 54 | 54 |
| Advanced Placement (AP) | Biology | 85 | 62 |
| Advanced Placement (AP) | Calculus | 43 | 0 |
| Advanced Placement (AP) | Chemistry | 71 | 22 |
| Advanced Placement (AP) | Physics 2 | 66 | 30 |
| Advanced Placement (AP) | Psychology | 83 | 83 |
| Advanced Placement (AP) | Statistics | 85 | 40 |
| Advanced Placement (AP) | English Language | 14 | 14 |
| Advanced Placement (AP) | English Literature | 8 | 8 |
| Competitive Programming | Codeforces Rating | <5 | <5 |

https://www.visualcapitalist.com/how-smart-is-chatgpt/



## How Smart is ChatGPT?

OpenAI's latest large language model, GPT-4, is capable of human-level performance in many professional and academic exams.

A percentile describes how an examinee's score ranks in comparison to others.

For example

60% of examinees scored lower than ChatGPT

40% scored higher than ChatGPT

Exam Results ● ChatGPT 3.5 ● ChatGPT 4.0

The Graduate Record Exam (GRE) is a standardized test that is required for admission to many graduate schools across North America.

According to OpenAI, the essays that GPT-4 produced for these tests were graded by "qualified third-party contractors".

GPT 4.0 has made impressive strides over GPT 3.5 but continues to struggle in certain subjects.

GPT attempted 10 programming contests 100 times each, but was unable to consistently find solutions to the complex problems.

VISUAL CAPITALIST

Source: OpenAI (2023)

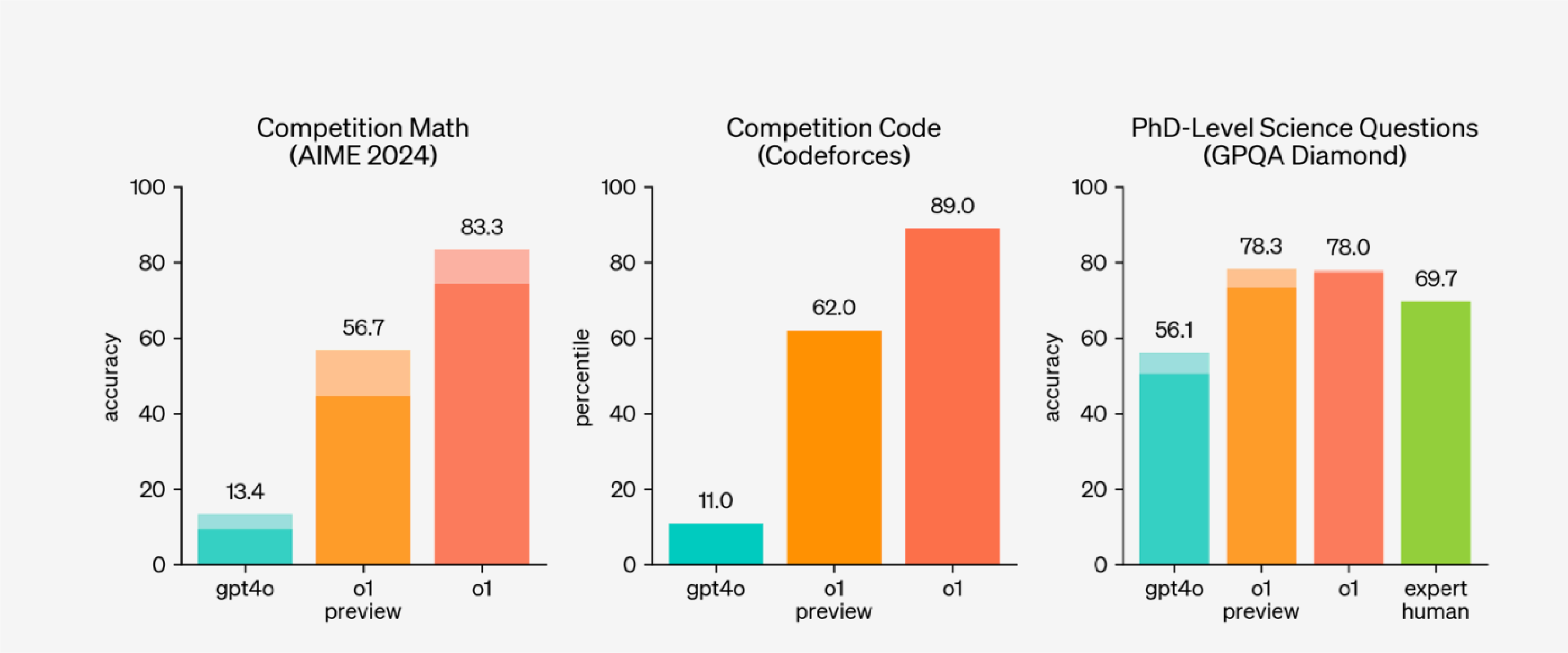
Note: Percentiles are based on the most recently available score distributions for test takers of each exam type.

COLLABORATORS RESEARCH + WRITING Marcus Lu | DESIGN Rosey Eason

/visualcapitalist @visualcap visualcapitalist.com

COLLEGE OF EDUCATION
MARYLAND ASSESSMENT RESEARCH CENTER

# AI is improving

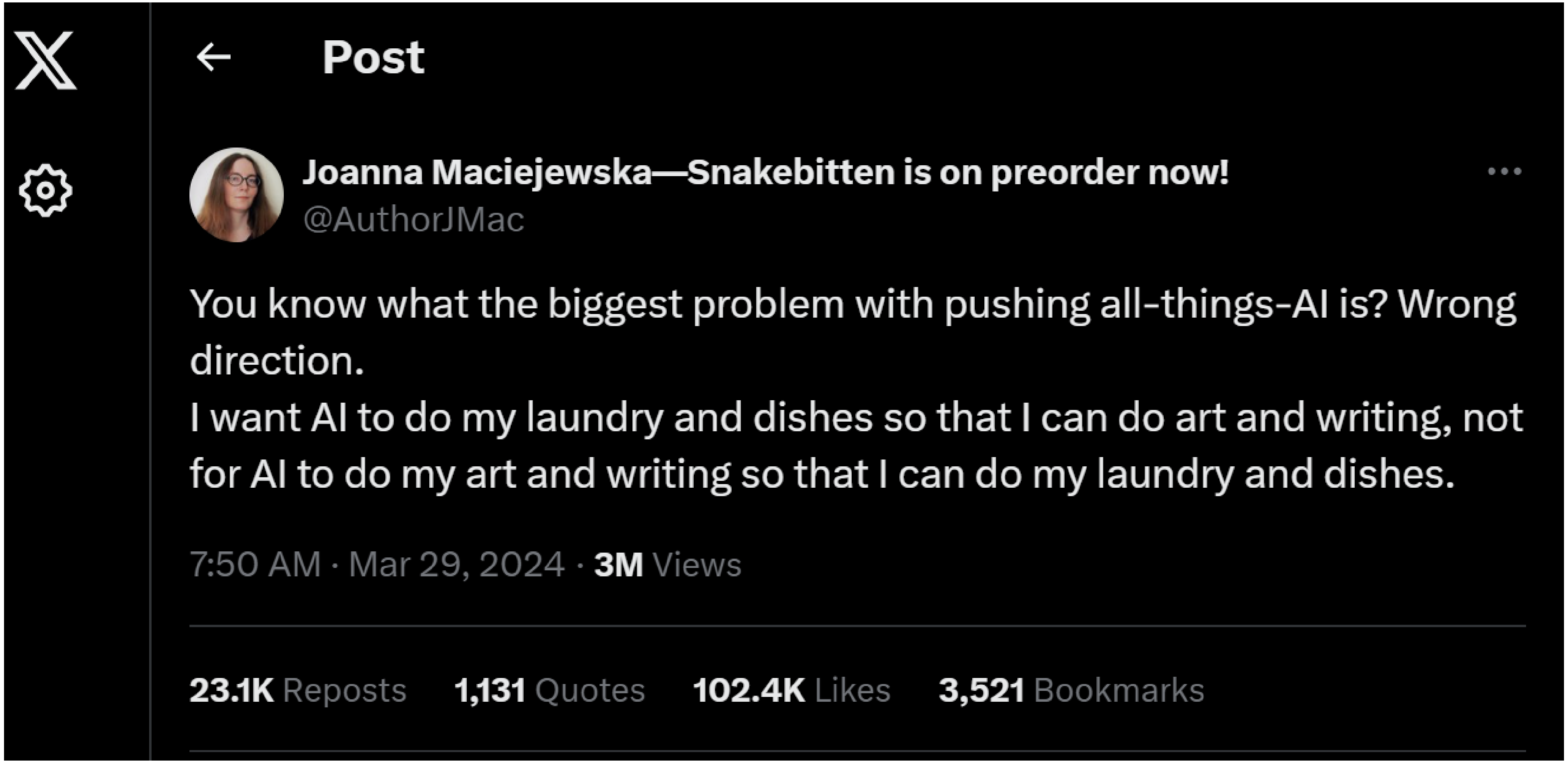


# Highlight 4

- Hallucinations
- Bias

- **Cautions on using AI for high-stakes decisions**
- **Human-in-the-loop**
- **Quality control/evaluation of AI procedures/tools**

# Highlight 5



Joanna Maciejewska—Snakebitten is on preorder now!
@AuthorJMac

You know what the biggest problem with pushing all-things-AI is? Wrong direction.
I want AI to do my laundry and dishes so that I can do art and writing, not for AI to do my art and writing so that I can do my laundry and dishes.

7:50 AM · Mar 29, 2024 · **3M** Views

**23.1K** Reposts **1,131** Quotes **102.4K** Likes **3,521** Bookmarks

# Highlight 5

## Test Development Steps (Downing, 2006)

- Overall plan
- Content/construct definition
- Test specifications
- Item development
- Test design and assembly
- Test production
- Test administration
- Scoring test responses
- Calibration, equating, and scaling
- Passing scores
- Reporting test results
- Item banking
- Test technical report
  - Validity
  - Reliability
  - Fairness

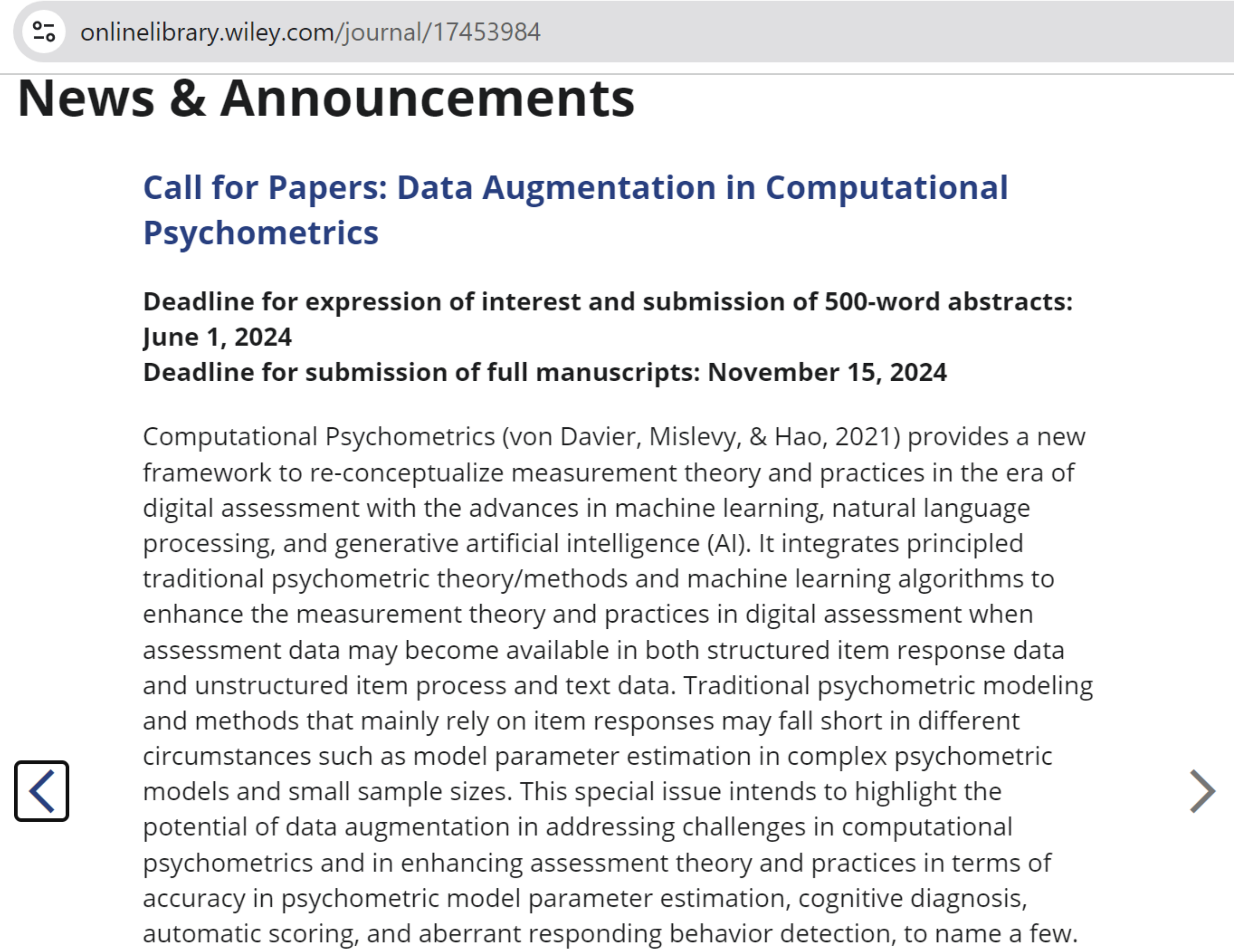
## Educational Assessments

- What can/should be done by AI?
  - Keep updating
- What can/should not be done by AI for now?
  - Keep updating

# Summary

- Successful use cases of AI in enhancing educational assessment practices and theory.
- Worthy of further exploration of AI use cases in educational assessment.
- AI development is swift, needs to keep up with the AI development speed.
- Human-in-the-loop is needed for AI applications in high-stakes educational assessment programs.
- Quality control: performance evaluation should be always provided in technical documentation.
- Extensive empirical evidence need to be collected to determine what works and what does not work to inform the practices for educational assessment programs.

# JEM Special Issue



onlinelibrary.wiley.com/journal/17453984

## News & Announcements

### Call for Papers: Data Augmentation in Computational Psychometrics

**Deadline for expression of interest and submission of 500-word abstracts: June 1, 2024**
**Deadline for submission of full manuscripts: November 15, 2024**

Computational Psychometrics (von Davier, Mislevy, & Hao, 2021) provides a new framework to re-conceptualize measurement theory and practices in the era of digital assessment with the advances in machine learning, natural language processing, and generative artificial intelligence (AI). It integrates principled traditional psychometric theory/methods and machine learning algorithms to enhance the measurement theory and practices in digital assessment when assessment data may become available in both structured item response data and unstructured item process and text data. Traditional psychometric modeling and methods that mainly rely on item responses may fall short in different circumstances such as model parameter estimation in complex psychometric models and small sample sizes. This special issue intends to highlight the potential of data augmentation in addressing challenges in computational psychometrics and in enhancing assessment theory and practices in terms of accuracy in psychometric model parameter estimation, cognitive diagnosis, automatic scoring, and aberrant responding behavior detection, to name a few.

https://onlinelibrary.wiley.com/journal/17453984

# MARC Conferences

- **Machine Learning, Natural Language Processing, and Psychometrics**

https://education.umd.edu/research/centers/marc/workshops-and-conferences/2022-virtual-marc-conferenceProvide information

- **Application of Artificial Intelligence (AI) to Assessment**

https://education.umd.edu/research/centers/marc/workshops-and-conferences/past-conferences/2017-marc-conferenceIs

- **Data Analytics and Psychometrics: Informing Assessment Practices**

https://education.umd.edu/research/centers/marc/workshops-and-conferences/past-conferences/2016-marc-conference

# References

- Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® automated essay scoring system. Handbook of automated essay evaluation: Current applications and new directions, 55–67.
- Devlin, J., Chang, M-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Jiao, H., Xu, S., & Zhou, T. (2021). *Automated scoring of the NAEP reading constructed-response items.* Technical Report.
- Jiao, Y., Shridhar, K., Cui, P., Zhou, W., & Sachan, M. (2023). Automatic educational question generation with difficulty level controls. In International Conference on Artificial Intelligence in Education (pp. 476-488). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_39
- Maeda, H. (2023, April). *Field-Testing items using artificial intelligence: Natural Language Processing with transformers.* Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.
- Mangino, T. A., Finch, H., French, B., & Demir, C. (2023, April). *Identification of differential item functioning using machine learning*. Presentation at the annual conference of National Council on Measurement in Education. Chicago, IL.
- McGraw-Hill Education CTB (2014, December 24). Smarter Balanced Assessment Consortium Field Test: Automated Scoring Research Studies (in accordance with Smarter Balanced RFP 17). Retrieved from: http://www. smarterapp.org/documents/FieldTest_ AutomatedScoringResearchStudies.pdf.
- Page, E. B. (2003). Project Essay Grade: PEG. In Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates
- Pearson and ETS (2015, March 9). Research Results of PARCC Automated Scoring Proof of Concept Study. Retrieved from: http:// www.parcconline.org/images/Resources/ Educatorresources/PARCC_AI_Research_ Report.pdf.
- Rodriguez, P., Jafari, A., & Ormerod, C. (2019). Language models and automated essay scoring. arXiv preprint, https://arxiv.org/abs/1909.09482.
- Shermis, M. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing, 20, 53-76.
- Shermis, M., & Lottridge, S. (2019, April). Communicating to the Public About Machine Scoring: What Works, What Doesn’t. Paper presented at the National Conference on Measurement in Education, Toronto, CA.
- von Davier, M. (2019). Training Optimus prime, MD: Generating medical certification items by fine-tuning OpenAI's gpt2 transformer model. arXiv preprint arXiv:1908.08594. https://doi.org/10.48550/arXiv.1908.08594
- Williamson, D., Xi, X., and Breyer, F.J. (2012). A framework for the evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2-13.
- Zhou, T. & Jiao, H. (2022a). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement.* https://doi.org/10.1177/00131644221117193
- Zhou, T. & Jiao, H. (2022b). Data augmentation in machine learning for cheating detection: An illustration with the Stacking learning algorithm. *Psychological Testing and Assessment Modeling.*