ASC
Assessment Systems

AI in Assessment: The Limits of LLMs

Nathan Thompson, PhD
CEO
Assessment Systems

Introduction

assess.ai

Why AI? Save time and effort.

SMEs & Test Developers

Instructors/Teachers

Students/Candidates

Psychometricians

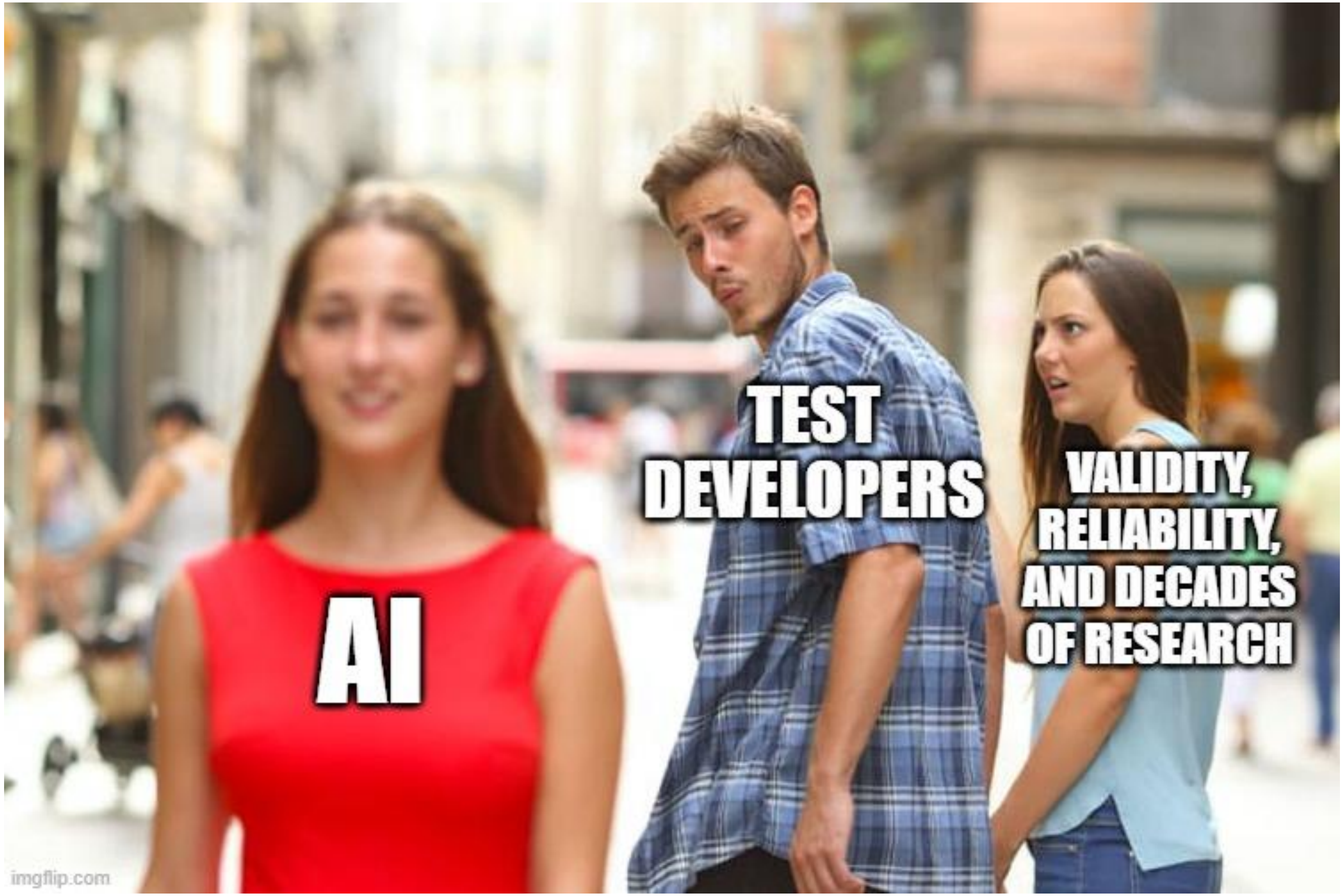
ASC
Assessment Systems

Why AI?

- Pain points in assessment
  - Item writing
  - Review items
  - Assemble linear forms
  - Make the tests smarter and faster for examinees
  - Reduce essay marking time/costs
  - Maintain security in delivery
  - Review results
  - Provide feedback to examinees

Why AI?

But…


AI
TEST DEVELOPERS
VALIDITY, RELIABILITY, AND DECADES OF RESEARCH
imgflip.com

What is validity?

Evidence to support our intended interpretations of test scores

ASC
Assessment Systems

Validity needs are a continuum

Quiz

Final exam

Admissions

Cert/Licensure

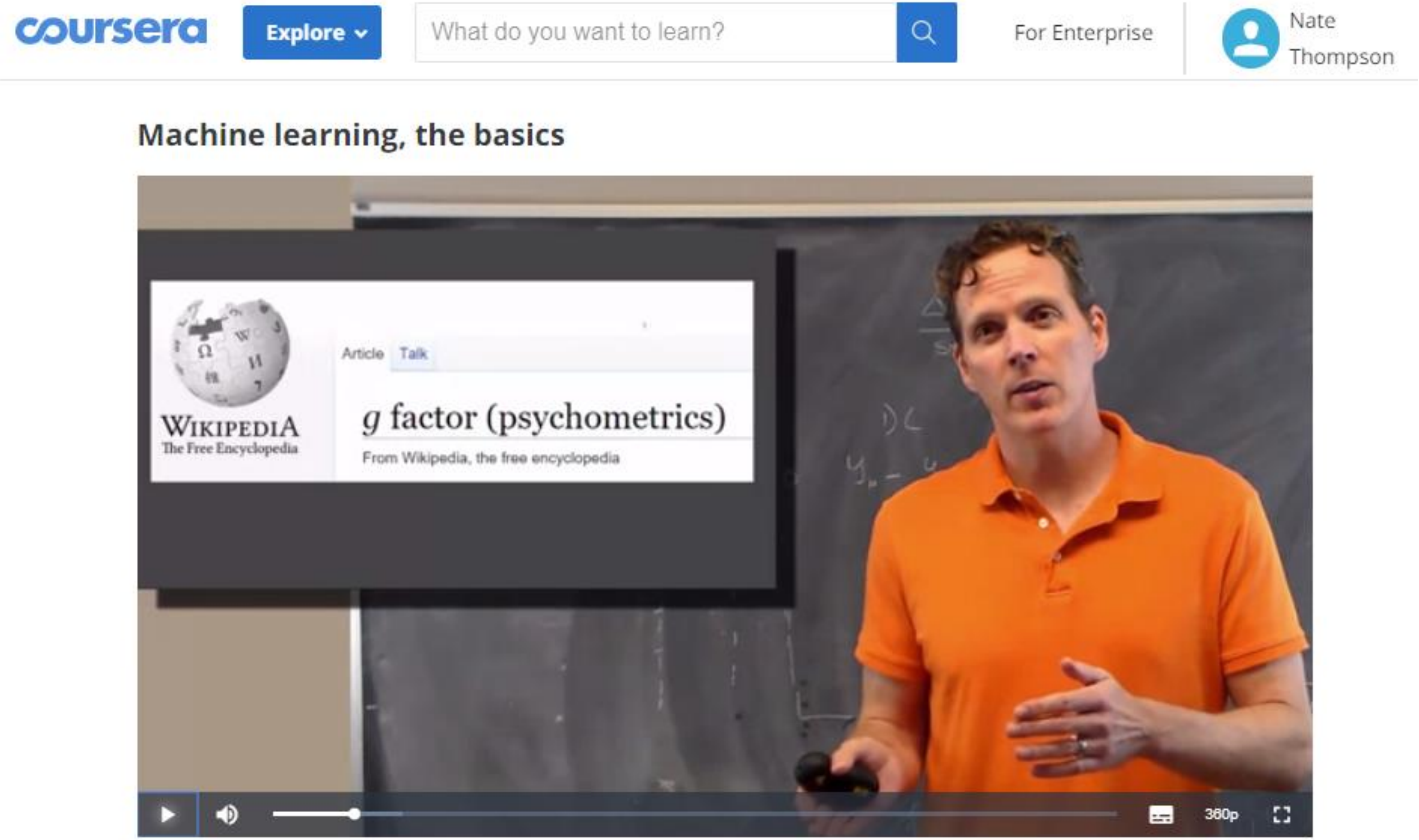
Why AI?

- So, will it help us solve our pain points?
- Or, will it lead to new problems?
- What constraints are needed?
- The answer, in many cases: it has been providing solutions, for decades.
- LLMs are new… how do they compare to traditional AI in our use cases, and regarding validity?

Part 2

AI and ML: An Introduction

New terms, not new ideas!

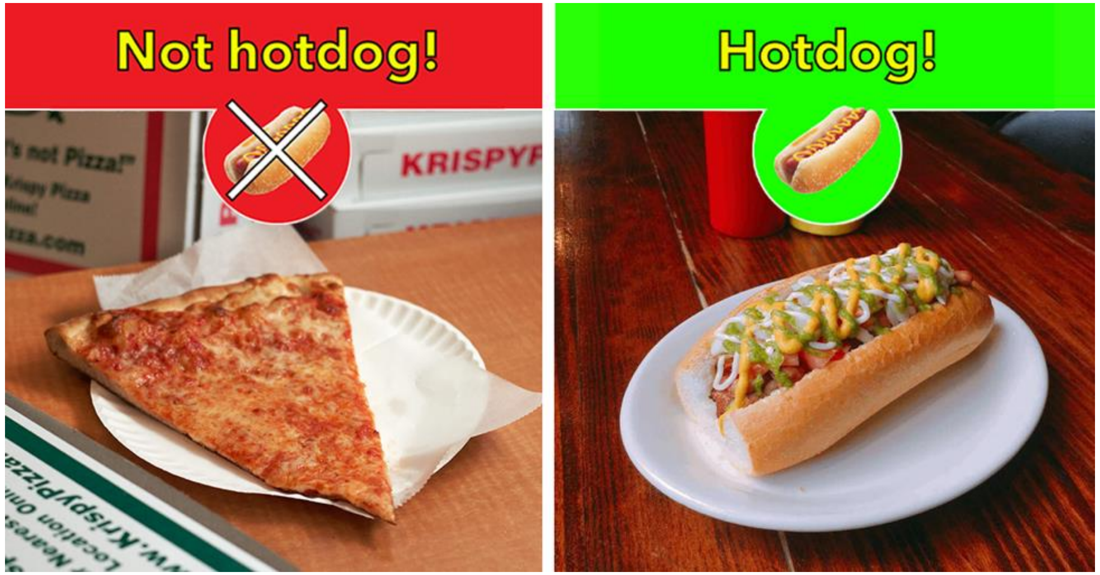


Definitions

- Machine Learning
  - The use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and **statistical models to analyze and draw inferences from patterns in data**. (*Oxford)*
- Artificial Intelligence
  - The theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, **decision-making**, and translation between languages. (*Oxford)*

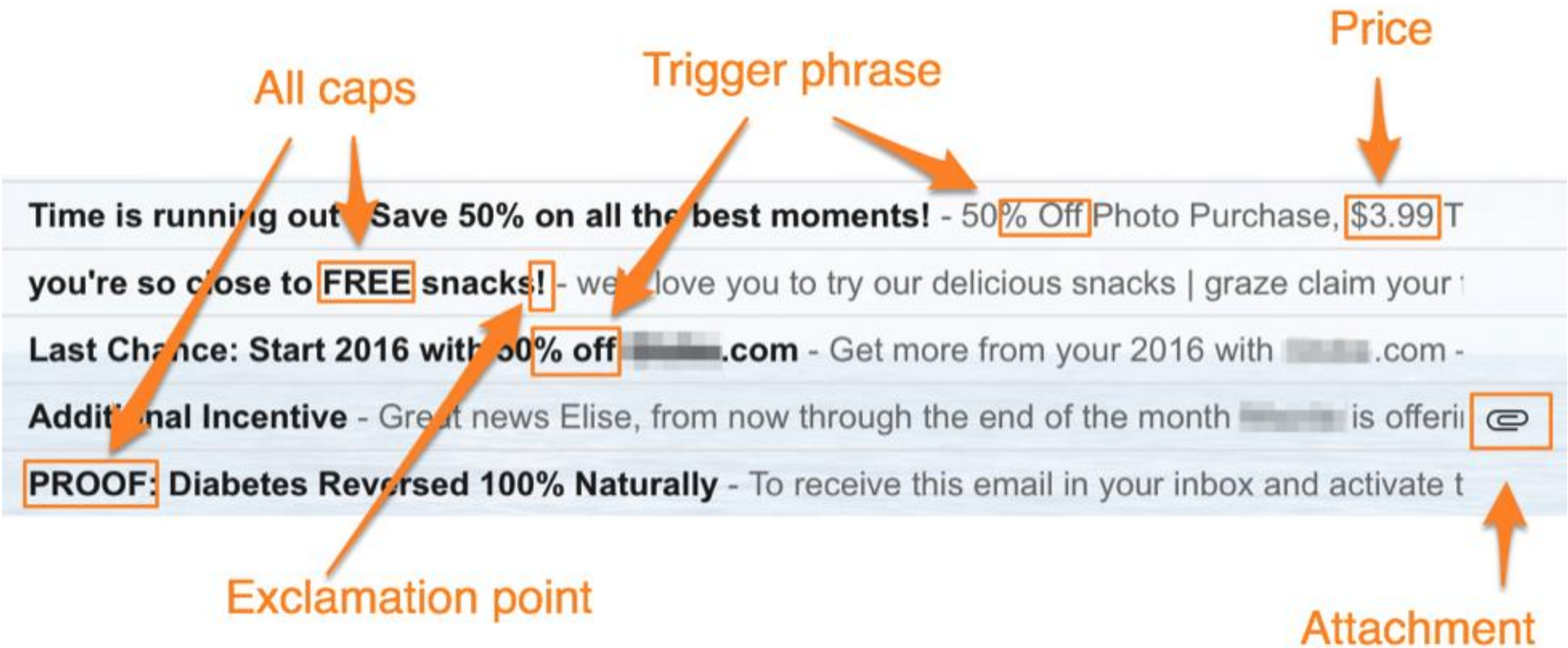
Definitions

- Machine Learning: **Unsupervised**
  - Looking for patterns with no particular goal
- Machine Learning: **Supervised**
  - Some sort of goal in mind (label/criterion/prediction)



Machine learning & AI

- Common example is logistic model to predict binary outcome...
- “Natural language processing”



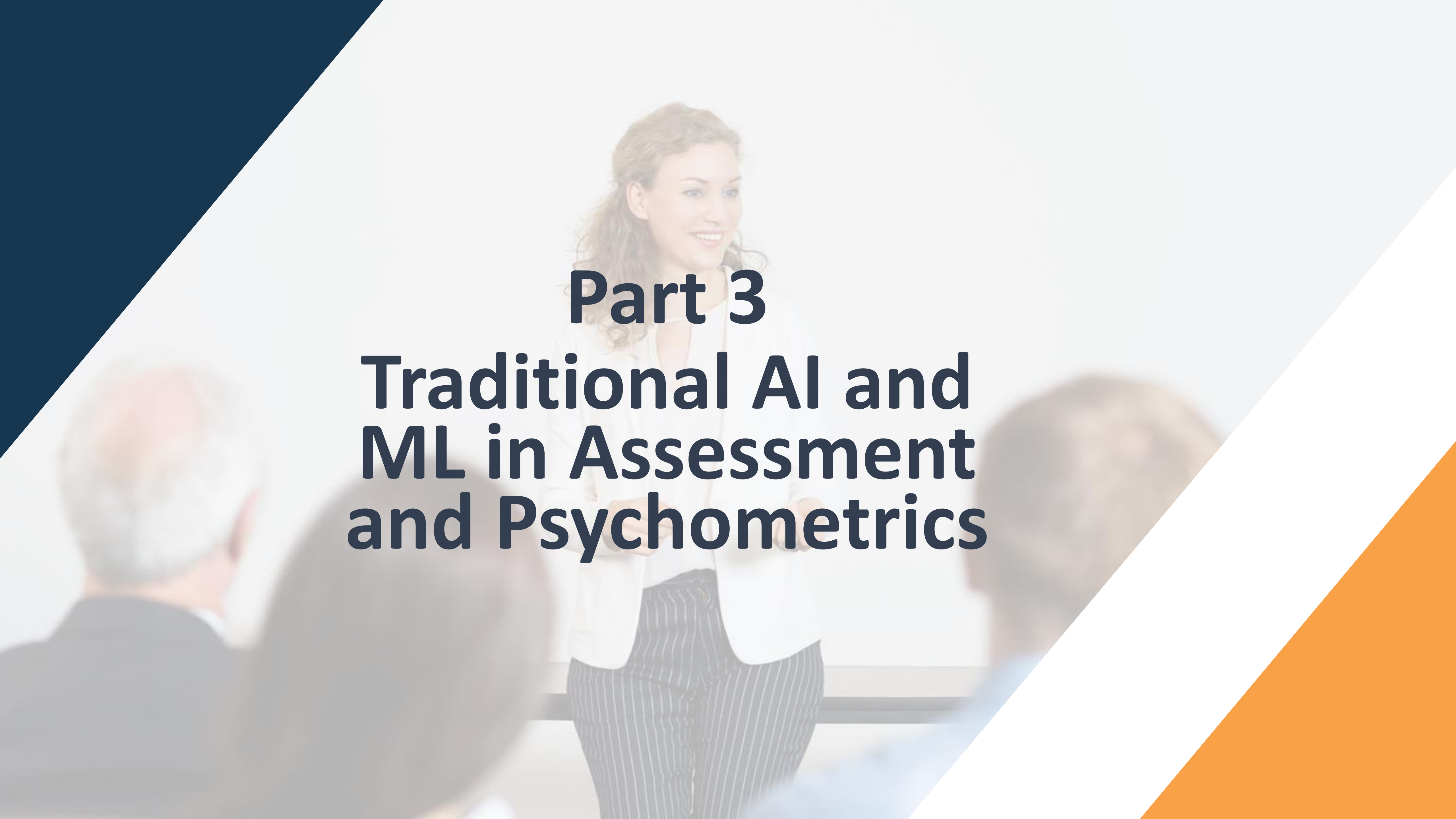
Time is running out: Save 50% on all the best moments! - 50% Off Photo Purchase, $3.99 T

you're so close to FREE snacks! - we'd love you to try our delicious snacks | graze claim your

Last Chance: Start 2016 with 50% off [redacted].com - Get more from your 2016 with [redacted].com -

Additional Incentive - Great news Elise, from now through the end of the month [redacted] is offeri

PROOF: Diabetes Reversed 100% Naturally - To receive this email in your inbox and activate t

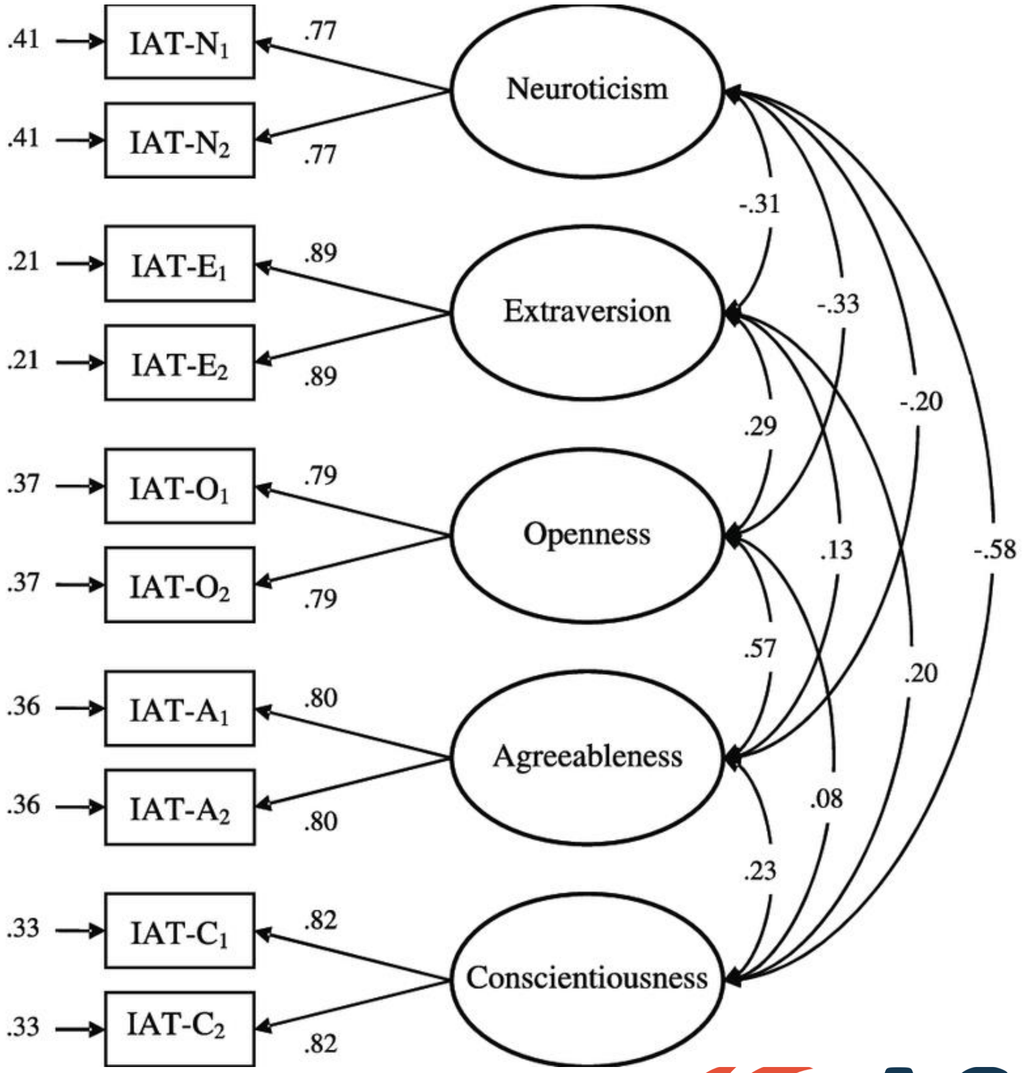


Part 3

Traditional AI and ML in Assessment and Psychometrics

ML: Factor Analysis

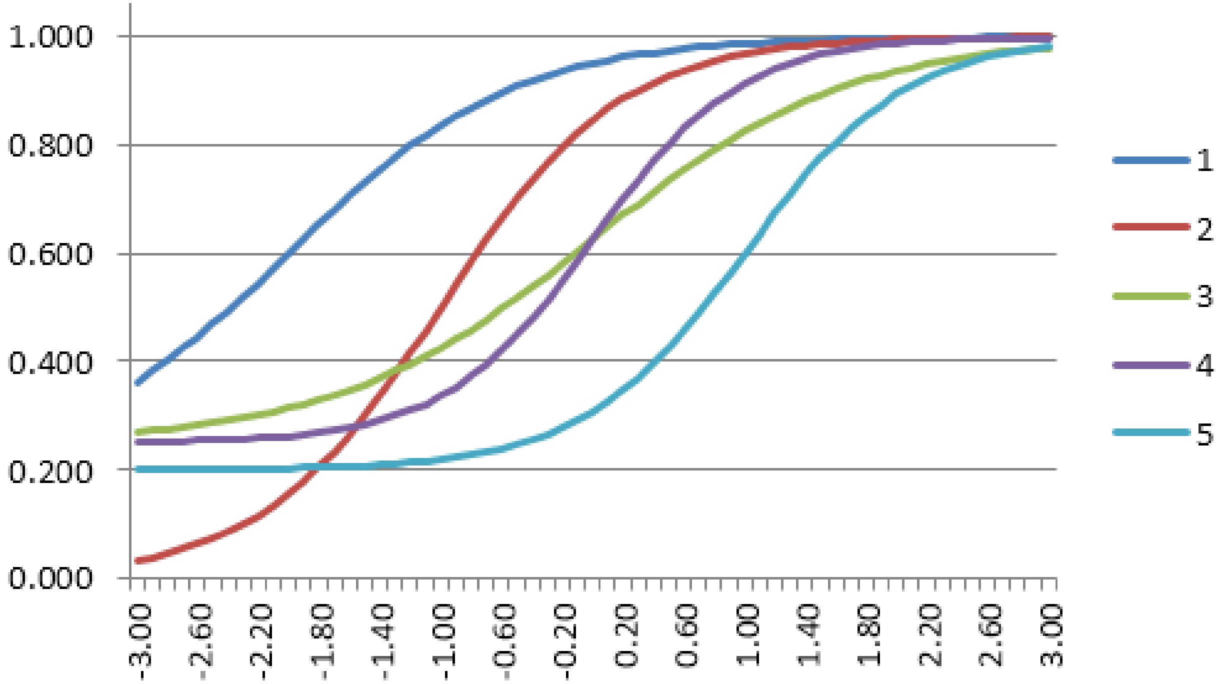
- Evaluate dimensionality, find patterns (unsupervised ML)
- 3 levels: Items, subscores, tests/battery
- Invented in 1904!!!!!!!!!



Schmukle, Back, &
Eglogg (2008)

ML: Item Response Theory (IRT)

- Fits logistical models to find patterns and make sense of the data (unsupervised ML)



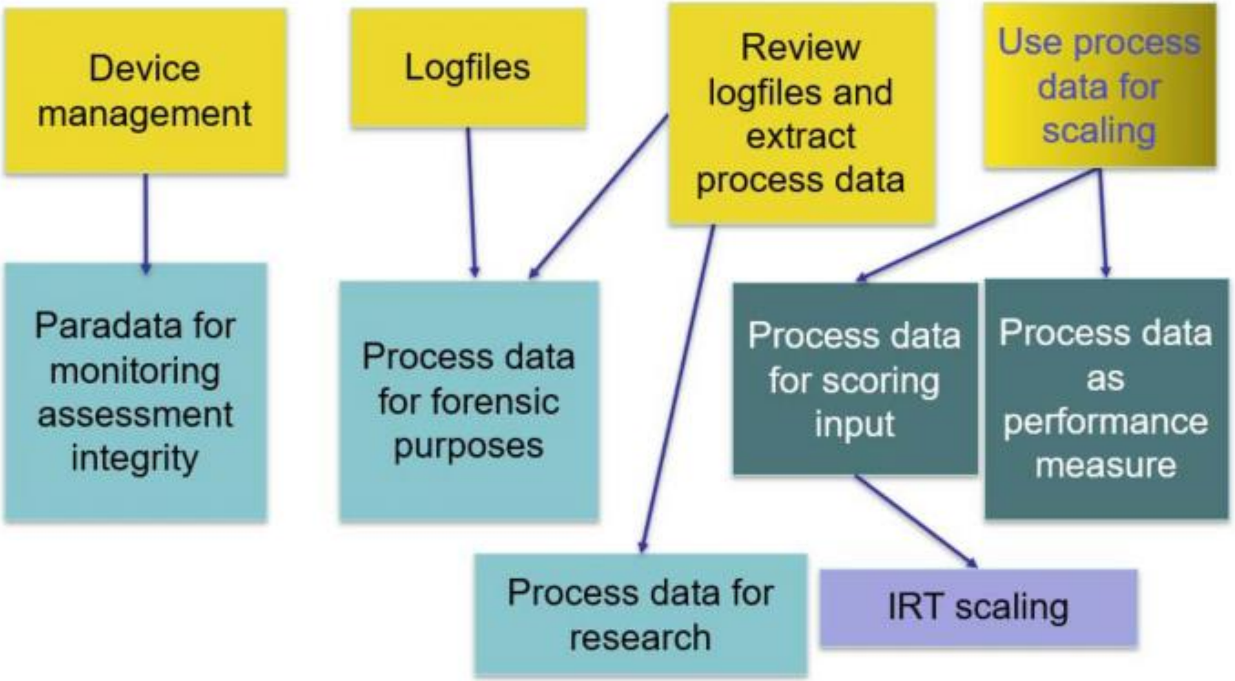
ML: Cognitive Diagnostic Models (CDMs)

- Make a “Q matrix” for items and curriculum/blueprint
- Can be supervised or unsupervised

Item	Excel Formulas	Quant reasoning	Tax knowledge	Reports
1	1	0	0	1
2	0	0	1	1
3	0	0	1	0
4	1	0	0	0
5	0	1	1	0

ML: Process Data

- Let’s use ALL the data we have from assessment and learning
- We need complex ML to make sense of it


Device management

Logfiles

Review logfiles and extract process data

Use process data for scaling

Paradata for monitoring assessment integrity

Process data for forensic purposes

Process data for scoring input

Process data as performance measure

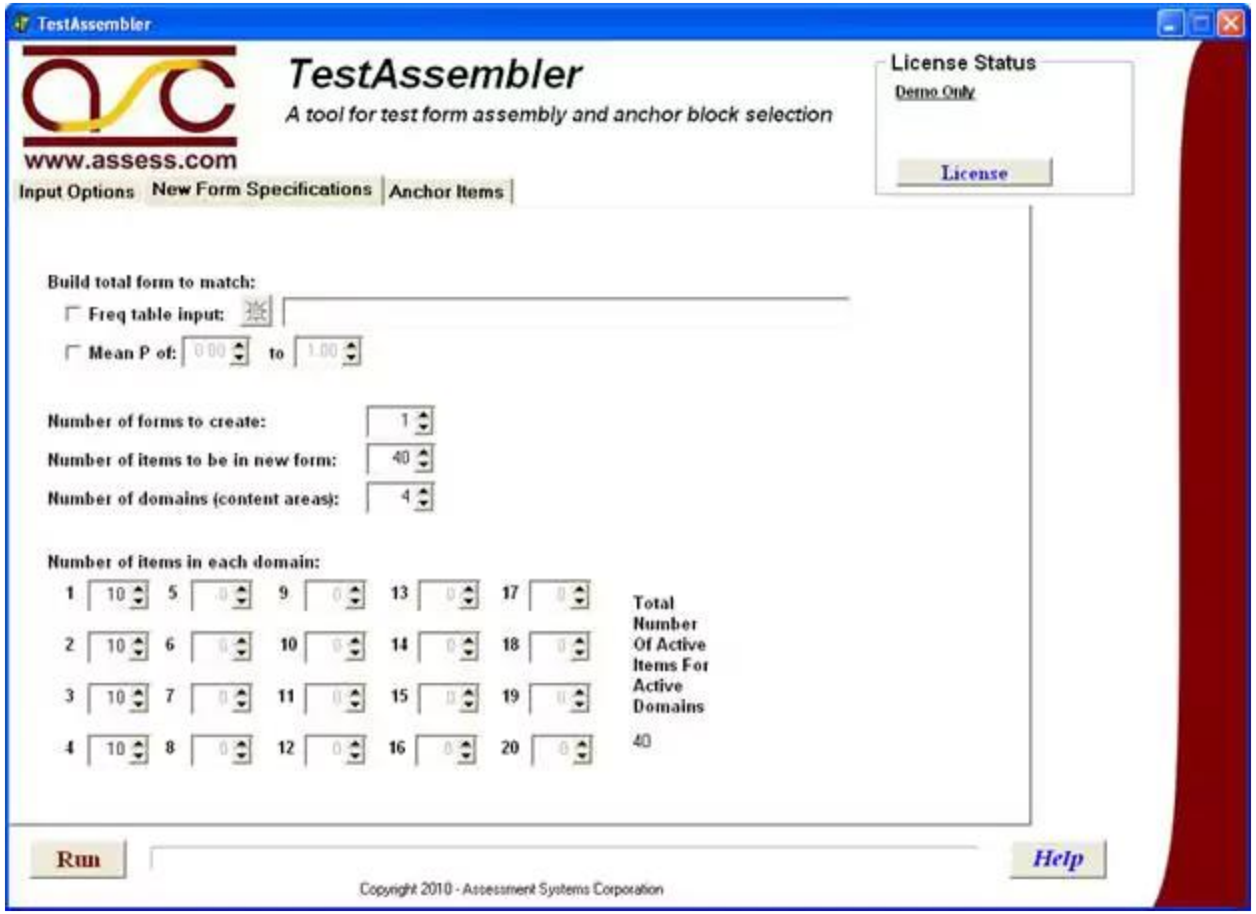
Process data for research

IRT scaling

Provasnik (2021)

AI: Automated Test Assembly

- Automate how a psychometrician would manually do complex assembly
- I used to do this!



TestAssembler

A tool for test form assembly and anchor block selection

Build total form to match:

☐ Freq table input:

☐ Mean P of: 0.00 to 1.00

Number of forms to create: 1

Number of items to be in new form: 40

Number of domains (content areas): 4

Number of items in each domain:

1	10	5	0	9	0	13	0	17	0
2	10	6	0	10	0	14	0	18	0
3	10	7	0	11	0	15	0	19	0
4	10	8	0	12	0	16	0	20	0

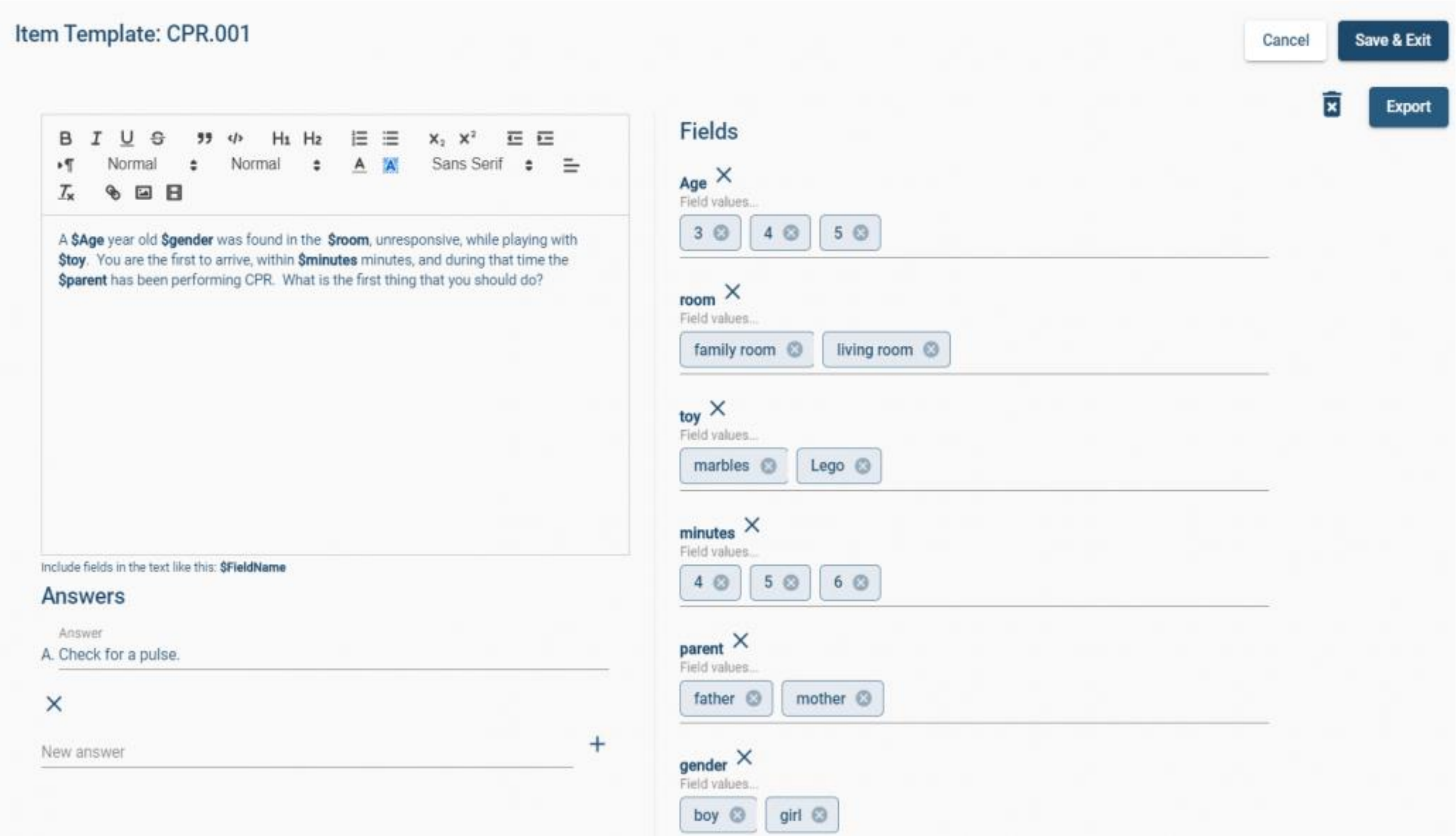
Total Number Of Active Items For Active Domains

40

Copyright 2010 - Assessment Systems Corporation

AI: Automated Item Generation

- Templates
- Can still be 6x as effective (Medical Council of Canada)



Item Template: CPR.001

A **$Age** year old **$gender** was found in the **$room**, unresponsive, while playing with **$toy**. You are the first to arrive, within **$minutes** minutes, and during that time the **$parent** has been performing CPR. What is the first thing that you should do?

Include fields in the text like this: **$FieldName**

Answers

Answer
A. Check for a pulse.

New answer

Fields

Age: 3, 4, 5

room: family room, living room

toy: marbles, Lego

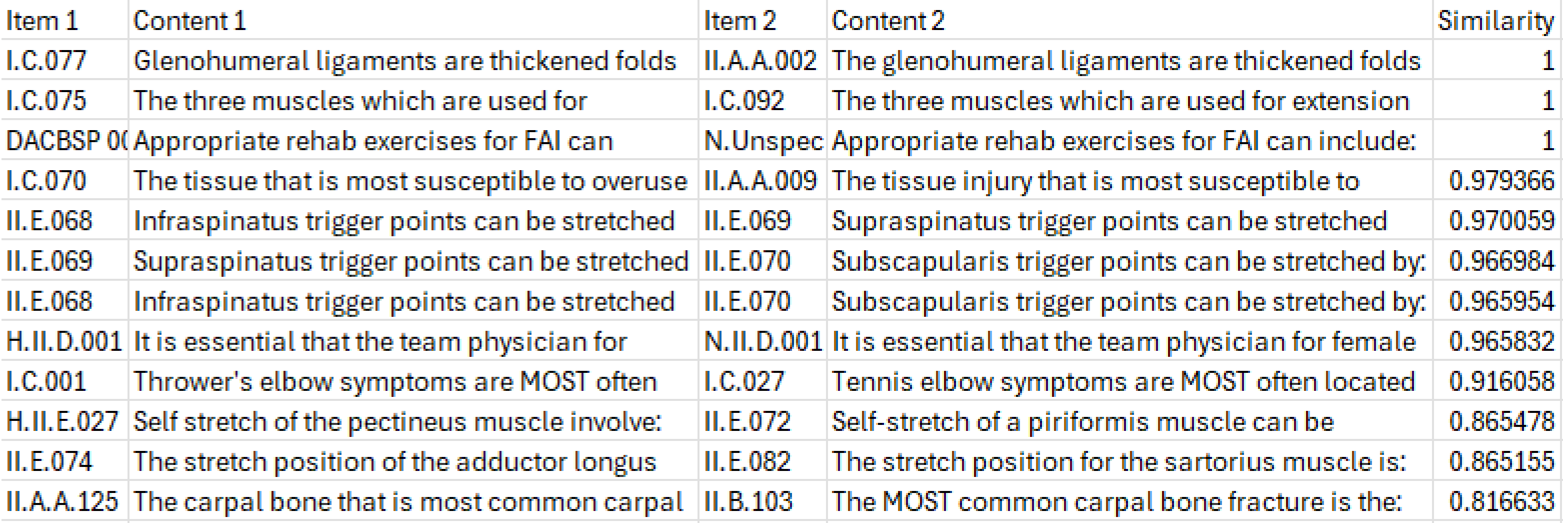
minutes: 4, 5, 6

parent: father, mother

gender: boy, girl

AI: Enemy Item detection

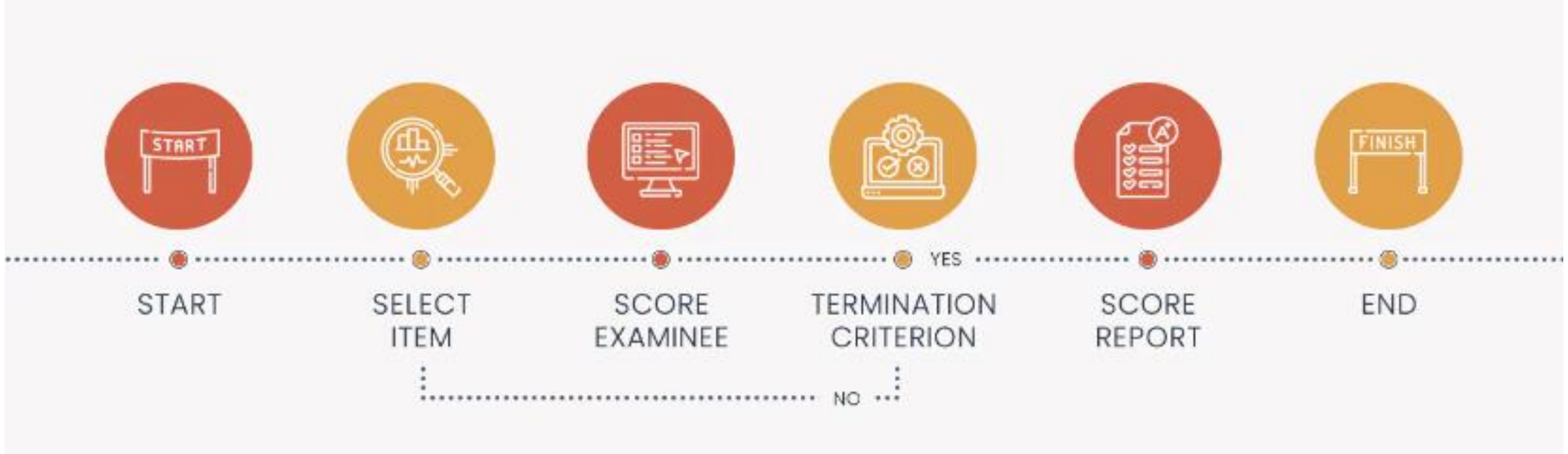
- NLP + Text similarity/distance indices



Item 1	Content 1	Item 2	Content 2	Similarity
I.C.077	Glenohumeral ligaments are thickened folds	II.A.A.002	The glenohumeral ligaments are thickened folds	1
I.C.075	The three muscles which are used for	I.C.092	The three muscles which are used for extension	1
DACBSP 0(	Appropriate rehab exercises for FAI can	N.Unspec	Appropriate rehab exercises for FAI can include:	1
I.C.070	The tissue that is most susceptible to overuse	II.A.A.009	The tissue injury that is most susceptible to	0.979366
II.E.068	Infraspinatus trigger points can be stretched	II.E.069	Supraspinatus trigger points can be stretched	0.970059
II.E.069	Supraspinatus trigger points can be stretched	II.E.070	Subscapularis trigger points can be stretched by:	0.966984
II.E.068	Infraspinatus trigger points can be stretched	II.E.070	Subscapularis trigger points can be stretched by:	0.965954
H.II.D.001	It is essential that the team physician for	N.II.D.001	It is essential that the team physician for female	0.965832
I.C.001	Thrower's elbow symptoms are MOST often	I.C.027	Tennis elbow symptoms are MOST often located	0.916058
H.II.E.027	Self stretch of the pectineus muscle involve:	II.E.072	Self-stretch of a piriformis muscle can be	0.865478
II.E.074	The stretch position of the adductor longus	II.E.082	The stretch position for the sartorius muscle is:	0.865155
II.A.A.125	The carpal bone that is most common carpal	II.B.103	The MOST common carpal bone fracture is the:	0.816633

AI: Computerized Adaptive Testing (CAT)

- Uses IRT to personalize the test for all examinees
- Reduces time by 50%, more secure, can measure off-grade, and more
- Automates Alfred Binet's 1908 IQ test with ML



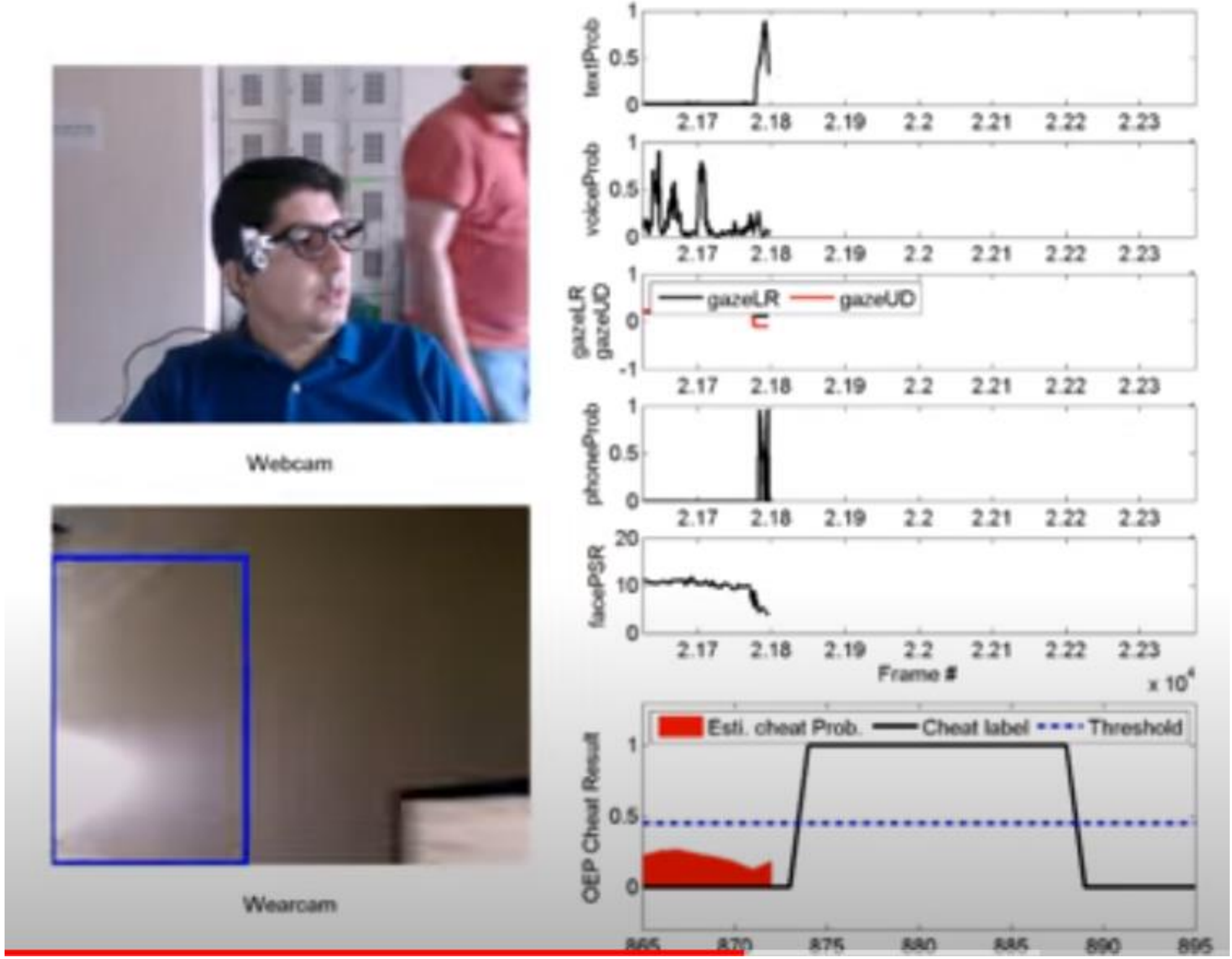
AI: Automated essay scoring (AES)

- NLP does not mean we are naturally processing
- We train a machine learning model for each prompt/rubric
- AI can now also do *evaluation and feedback*

Score	school_board	three_reasons	leadership	ludacris
5	3	1	2	0
4	2	1	0	0
3	1	1	1	0
2	2	0	0	0
1	0	0	0	2

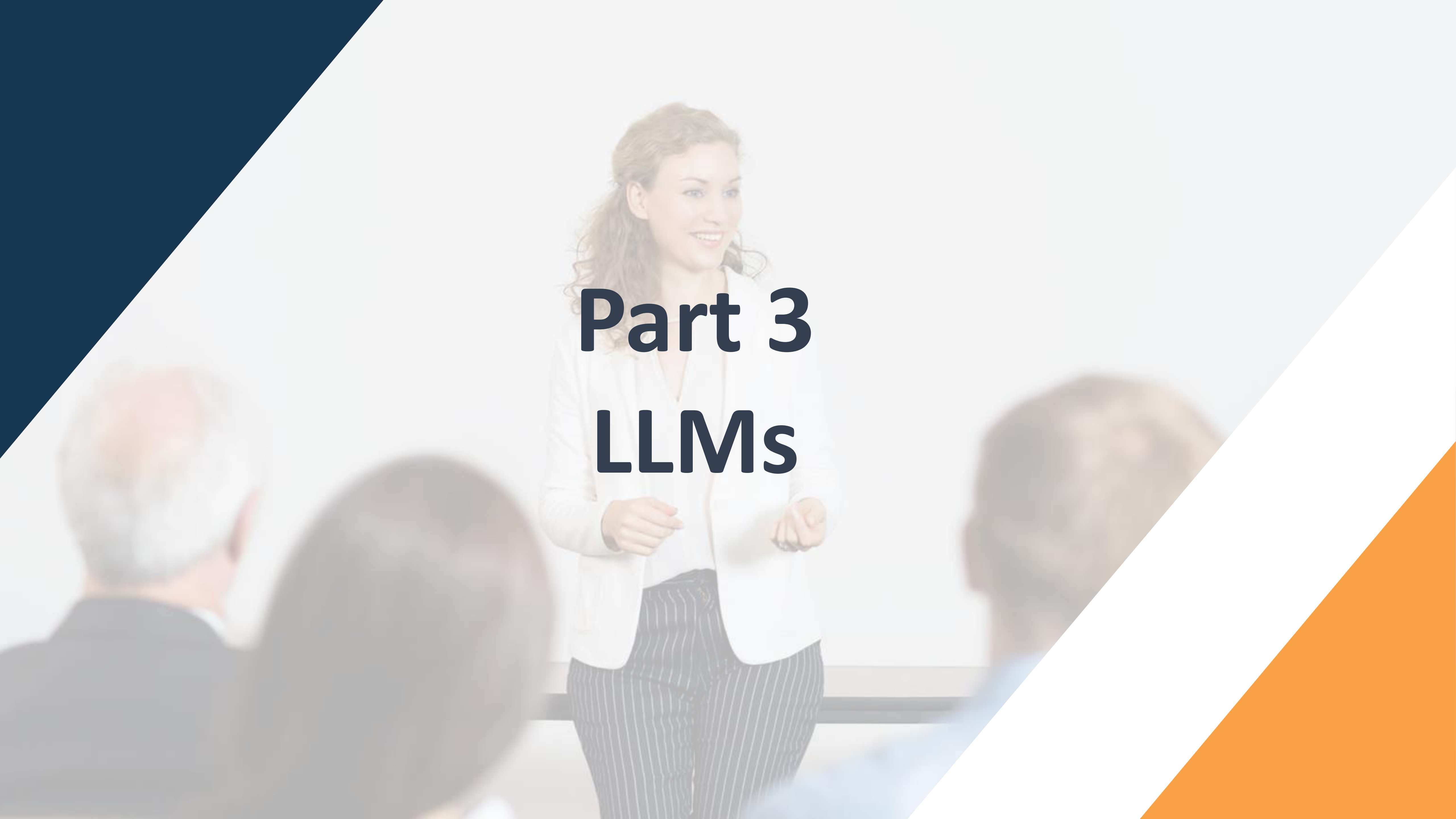
AI: Remote Proctoring

- Break it down into solvable problems
- Looks for checklist of issues by evaluating still images based on ML models



ML/AI outside the Assessment

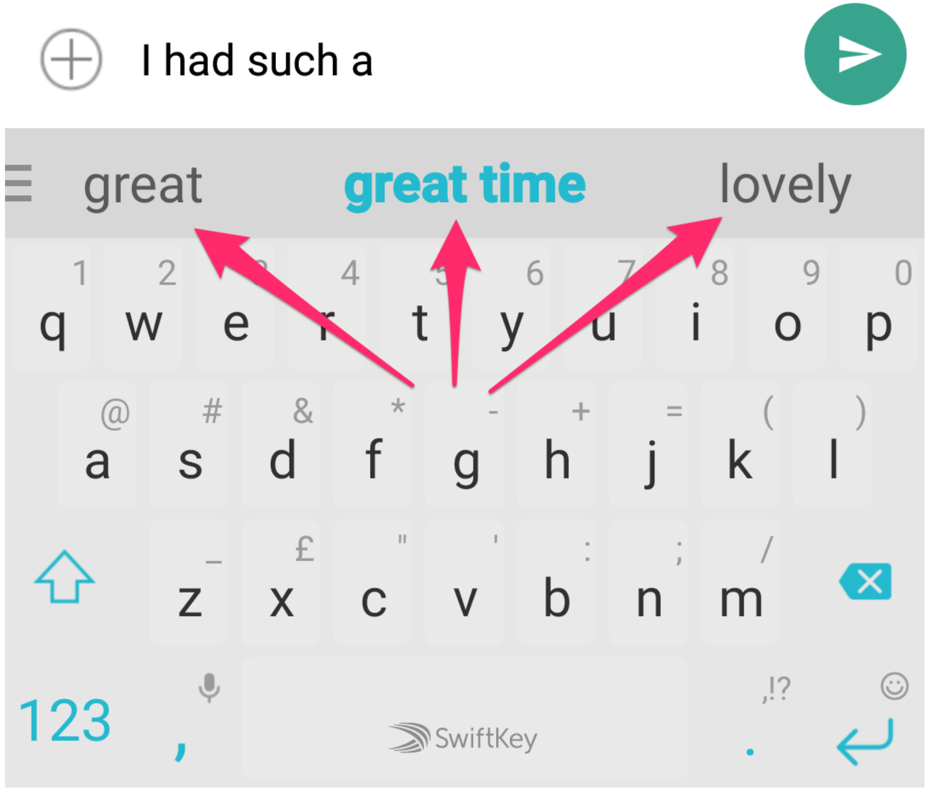
- Prediction models:
  - Job performance, counterproductive work behavior, tenure
  - Same for Universities too!
- Resume/CV evaluation
- Adaptive learning
- CrystalKnows, Cambridge Analytica, etc
- So much more!



Part 3
LLMs

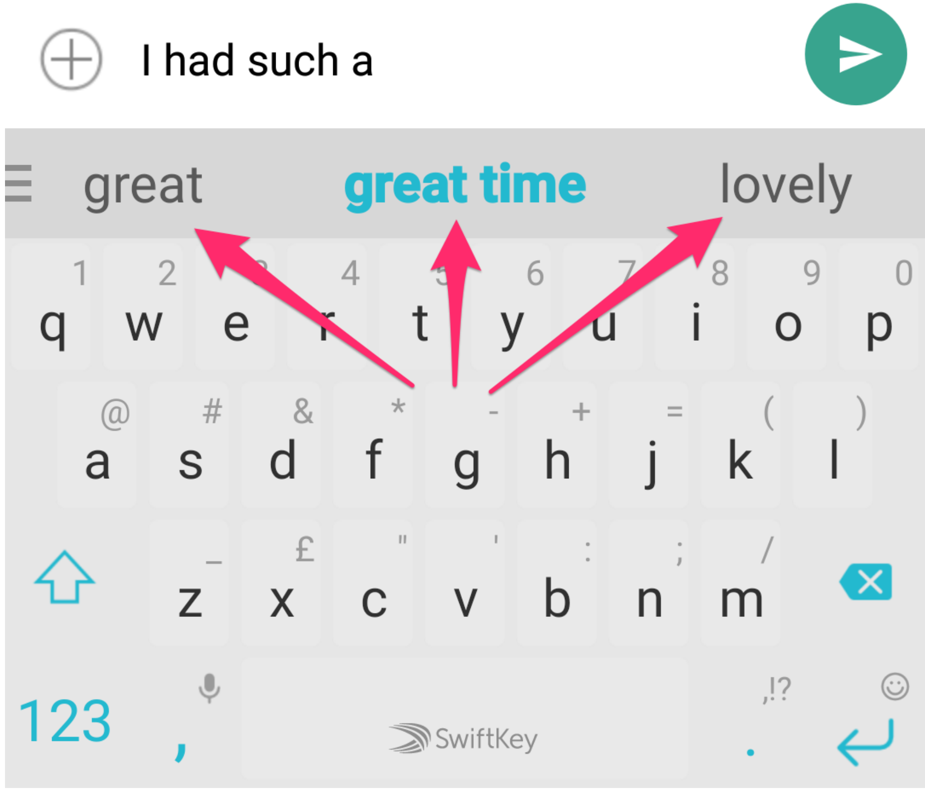
Intro to LLMs

- LLMs are like essay scoring in that they break down texts into a regression-type data set
- Predict the next word or words based on prompts and previous words
- But waaaaay more complex

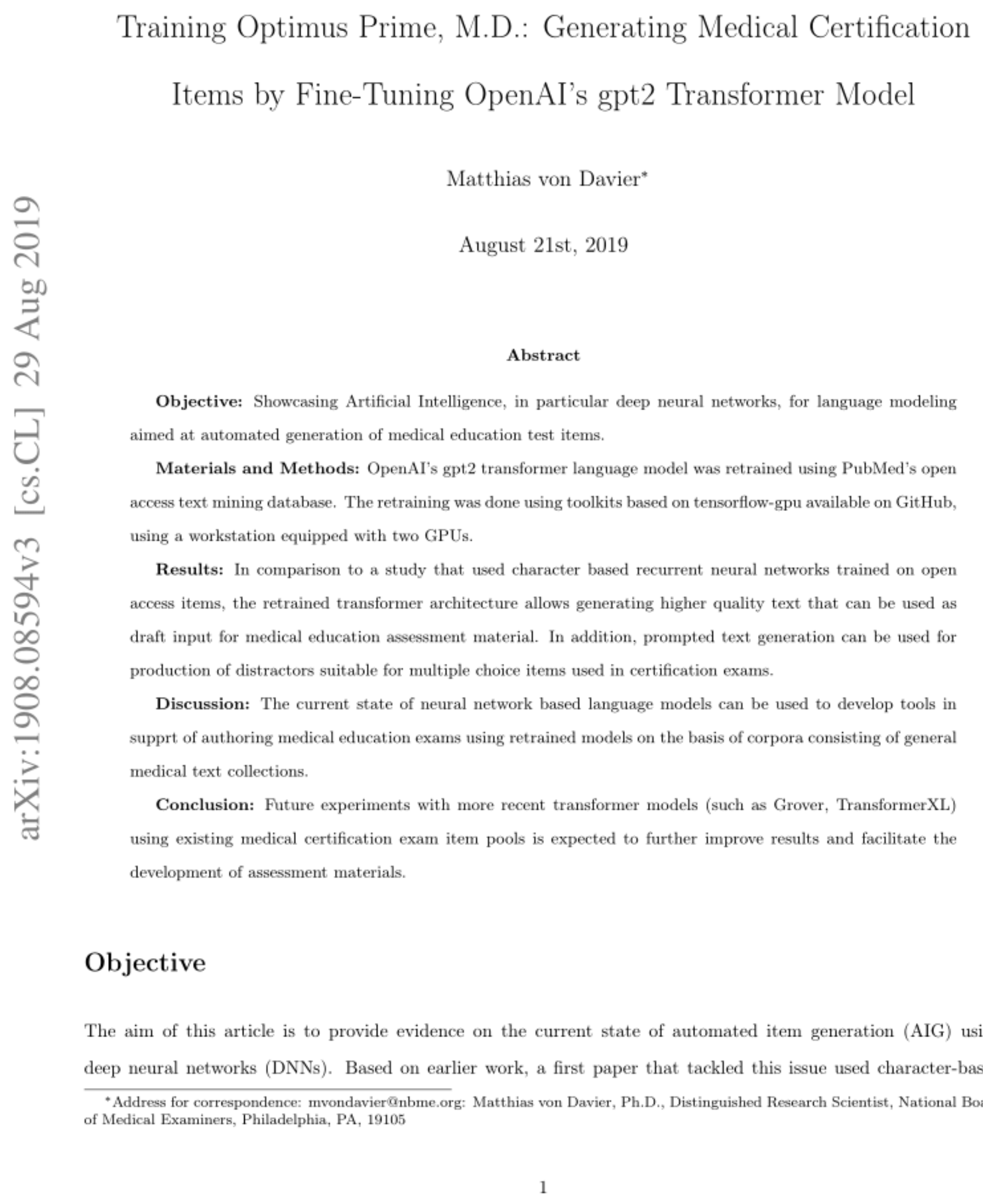

I had such a

AIG

- We've all heard about ChatGPT, but what does it mean for assessment?



I had such a



arXiv:1908.08594v3 [cs.CL] 29 Aug 2019

Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model

Matthias von Davier*

August 21st, 2019

Abstract

Objective: Showcasing Artificial Intelligence, in particular deep neural networks, for language modeling aimed at automated generation of medical education test items.

Materials and Methods: OpenAI's gpt2 transformer language model was retrained using PubMed's open access text mining database. The retraining was done using toolkits based on tensorflow-gpu available on GitHub, using a workstation equipped with two GPUs.

Results: In comparison to a study that used character based recurrent neural networks trained on open access items, the retrained transformer architecture allows generating higher quality text that can be used as draft input for medical education assessment material. In addition, prompted text generation can be used for production of distractors suitable for multiple choice items used in certification exams.

Discussion: The current state of neural network based language models can be used to develop tools in supprt of authoring medical education exams using retrained models on the basis of corpora consisting of general medical text collections.

Conclusion: Future experiments with more recent transformer models (such as Grover, TransformerXL) using existing medical certification exam item pools is expected to further improve results and facilitate the development of assessment materials.

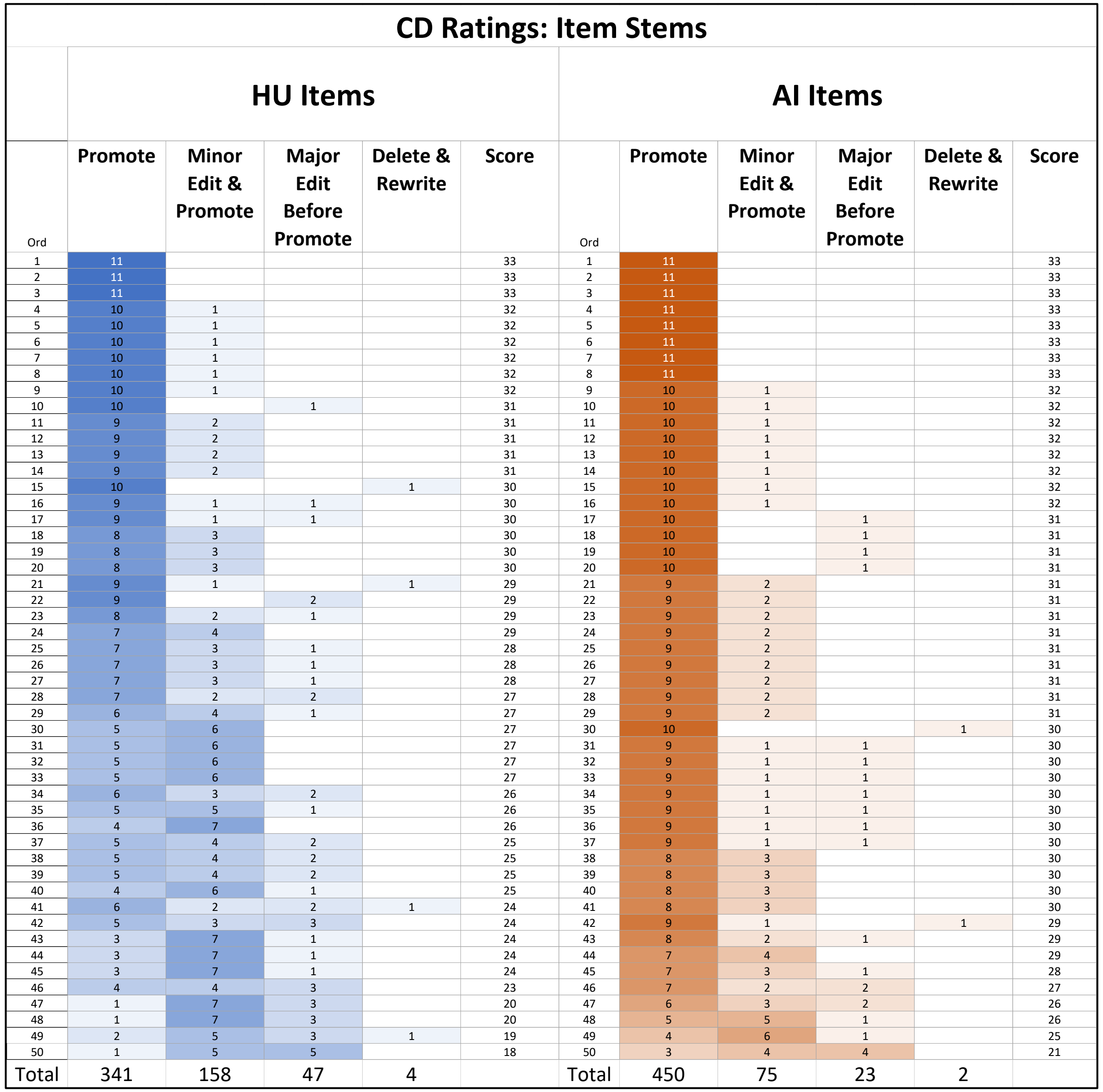
Objective

The aim of this article is to provide evidence on the current state of automated item generation (AIG) usi deep neural networks (DNNs). Based on earlier work, a first paper that tackled this issue used character-bas

---

*Address for correspondence: mvondavier@nbme.org: Matthias von Davier, Ph.D., Distinguished Research Scientist, National Bo of Medical Examiners, Philadelphia, PA, 19105

AIG



CD Ratings: Item Stems

	HU Items						AI Items				
Ord	Promote	Minor Edit & Promote	Major Edit Before Promote	Delete & Rewrite	Score	Ord	Promote	Minor Edit & Promote	Major Edit Before Promote	Delete & Rewrite	Score
1	11				33	1	11				33
2	11				33	2	11				33
3	11				33	3	11				33
4	10	1			32	4	11				33
5	10	1			32	5	11				33
6	10	1			32	6	11				33
7	10	1			32	7	11				33
8	10	1			32	8	11				33
9	10	1			32	9	10	1			32
10	10		1		31	10	10	1			32
11	9	2			31	11	10	1			32
12	9	2			31	12	10	1			32
13	9	2			31	13	10	1			32
14	9	2			31	14	10	1			32
15	10			1	30	15	10	1			32
16	9	1	1		30	16	10	1			32
17	9	1	1		30	17	10		1		31
18	8	3			30	18	10		1		31
19	8	3			30	19	10		1		31
20	8	3			30	20	10		1		31
21	9	1		1	29	21	9	2			31
22	9		2		29	22	9	2			31
23	8	2	1		29	23	9	2			31
24	7	4			29	24	9	2			31
25	7	3	1		28	25	9	2			31
26	7	3	1		28	26	9	2			31
27	7	3	1		28	27	9	2			31
28	7	2	2		27	28	9	2			31
29	6	4	1		27	29	9	2			31
30	5	6			27	30	10			1	30
31	5	6			27	31	9	1	1		30
32	5	6			27	32	9	1	1		30
33	5	6			27	33	9	1	1		30
34	6	3	2		26	34	9	1	1		30
35	5	5	1		26	35	9	1	1		30
36	4	7			26	36	9	1	1		30
37	5	4	2		25	37	9	1	1		30
38	5	4	2		25	38	8	3			30
39	5	4	2		25	39	8	3			30
40	4	6	1		25	40	8	3			30
41	6	2	2	1	24	41	8	3			30
42	5	3	3		24	42	9	1		1	29
43	3	7	1		24	43	8	2	1		29
44	3	7	1		24	44	7	4			29
45	3	7	1		24	45	7	3	1		28
46	4	4	3		23	46	7	2	2		27
47	1	7	3		20	47	6	3	2		26
48	1	7	3		20	48	5	5	1		26
49	2	5	3	1	19	49	4	6	1		25
50	1	5	5		18	50	3	4	4		21
Total	341	158	47	4		Total	450	75	23	2	

- Classic comparison also used for AES: how does ChatGPT compare to human quality?
- Jones, Becker, & Ferris (2023)

Item Review

- First: code your rules into your item banker
- Second: LLMs?
  1. Ask LLM to review
  2. Provide your review criteria to LLM
  3. Fine-tune LLM based on past reviews

Item Review

Item Writing Rules

1. No colon at the end of the question stem.

2. Do not repeat text in the answers.

Question 1

The capital of South Korea is:

A. Seoul

B. Tokyo

C. London

D. Paris

Question 2

What is 2 + 2?

A. The answer is 4.

B. The answer is 6.

C. The answer is 8.

D. The answer is 10.

ASC
Assessment Systems

Item Review

Here’s the feedback based on your rules:

Question 1

Rule 1: Violates the rule about not using a colon at the end of the question stem.

Rule 2: This question follows the rule, as the answers don’t repeat text.

Question 2

Rule 1: This question adheres to the formatting rule (no colon).

Rule 2: Violates the rule, as all answers repeat “The answer is.”

Suggested Revisions:

Question 1

The capital of South Korea is

A. Seoul

B. Tokyo

C. London

D. Paris

Question 2

What is 2 + 2?

A. 4

B. 6

C. 8

D. 10

Let me know if you need further assistance!

Item Alignment

- Give it your curriculum and item bank, tell it to align all items to the curriculum as well as anything else like Bloom’s Taxonomy

AI: “Image of teacher using Bloom’s Taxonomy”

AI: Automated essay scoring (AES)

- Feed your essays and rubrics to an LLM
- Can also fine-tune based on some human ratings
- Produces scores on the rubrics for each student

AI: Automated essay scoring (AES)

- Here are some results from US high school equivalency exam
    - R or Python with basic NLP/ML: 0.65 – 0.80
    - LLAMA-3 without fine-tuning: 0.18
    - LLAMA-3 with fine-tuning (long run time with high costs): 0.81

AI: Automated essay scoring (AES)

- Funny thing is, every website inquiry I have received thinks that #2 above is the only approach, and that it is a panacea
- They also think that:
  - No humans ever need to mark anything
  - We don’t even need to establish rubrics, the AI will just “figure out what growth mindset means”

Part 3

Recommendations and Future

Using AI

- We have been using ML and AI in psychometrics for more than a century!
- Most of it is highly specific, developed to meet certain needs
- Not a panacea, but very powerful in the right hands
  - E.g., IRT revolutionized a lot of tricky problems

Using LLMs

- They are being used more and more
- Very powerful, but not a panacea
- Huge time saver but still needs humans
- Consider the results from Jones et al. – better results than human writers, but tons of items still need editing

Thoughts on AES

- Being able to explain and support the model is very important… that is validity!
- Using LLMs might have the same accuracy, but is a total black box; a lawsuit waiting to happen
    - Especially true without fine-tuning, which also means that no humans have rated any students

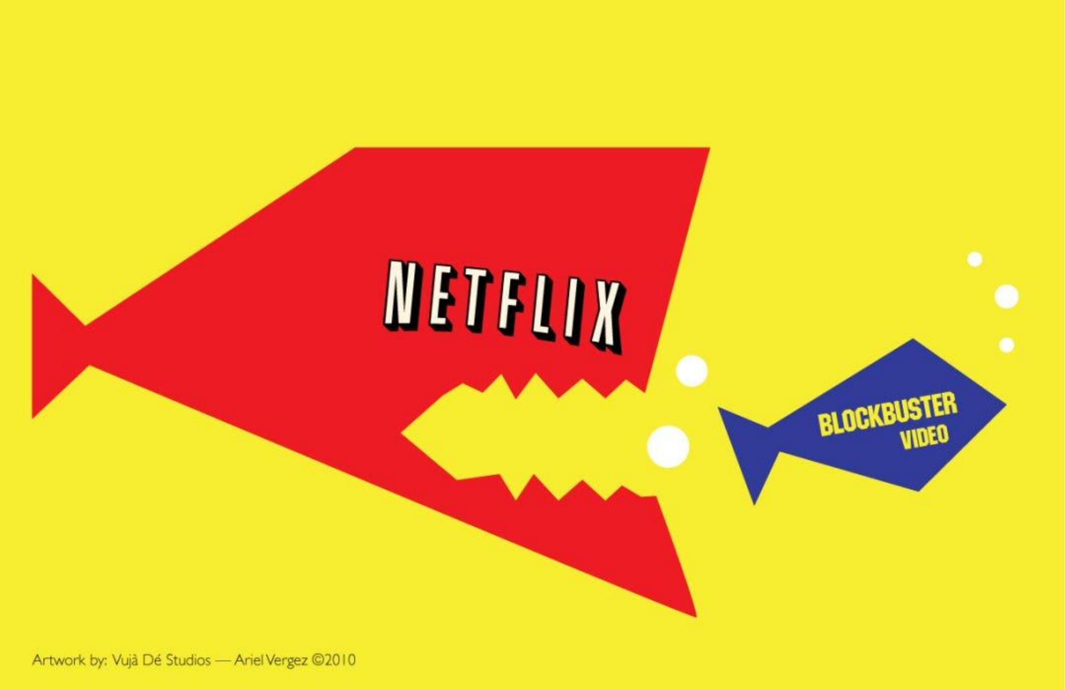
Security?

- Do you want to feed your content to an LLM to make items?
  - Past items
  - Your textbooks etc.
- How about data from past item reviews or item statistics?
- For AES, how about all your student essays and human marks?
- Public vs self-hosted

ASC Assessment Systems

The future?

- As in many other areas, LLMs and other AI are making an impact
- Even if they only provide 10% improvement, you do not want to be left behind
- It is irresponsible to avoid AI altogether
  - Friend that was told to stop all AI, even spam filters!
- So you need to decide what to use, and with what guardrails and post hoc checks



Artwork by: Vujà Dé Studios — Ariel Vergez ©2010

The future?

Questions?

nate@assess.com

ASC Assessment Systems