

Applications of Artificial Intelligence and Natural Language Processing in Large Scale Survey Assessment

Matthias von Davier

Donald Monan, S.J., University Professor in Education

Lynch School of Education, Boston College
TIMSS & PIRLS International Study Center



64th Annual Meeting of the IEA General Assembly



TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE

Do we even need a session about AI in Education?

- NPR radio mentions 'generative AI' not only in their science news but also in advertisements without explanation
- HBR working paper (Sept 15th, 2023) shows "Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality"
- A Survey ("The AI Divide" Sept 7th, 2023) including >4000 respondents from US, UK, Australia and India shows 49% used GenAI, 65% of users are Millennials, %72 are employed.68% of non-users are Gen X or Baby Boomers.

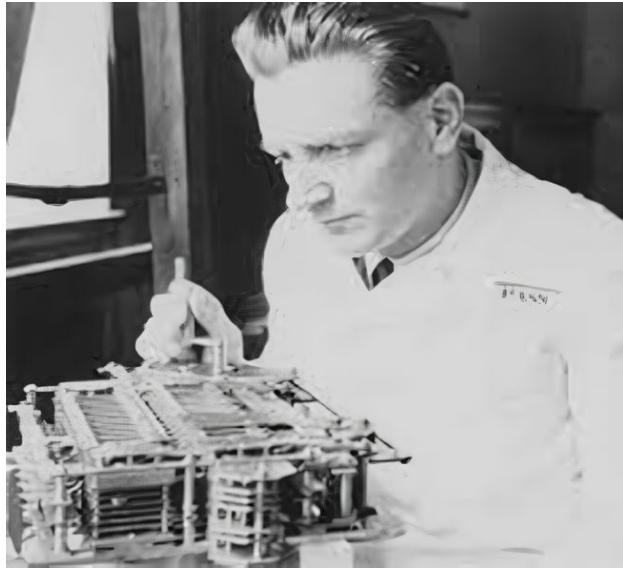
The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

Computers are About 80 Years Old



Konrad Zuse (1910-1995) Z3 (1941)
First **Turing-Complete** Computer
Plankalkül (1942) Language



Grace Hopper (1906-1992)
Worked on the MARK 1 (1944)
FLOW-MATIC -> COBOL
Coined the term 'Bug'

My First Computer is now 43 Years Old, and Language Modeling is at least as old:



Sinclair ZX81 (1980)

1KB RAM, expandable to 64KB

BASIC, no graphics, B&W,

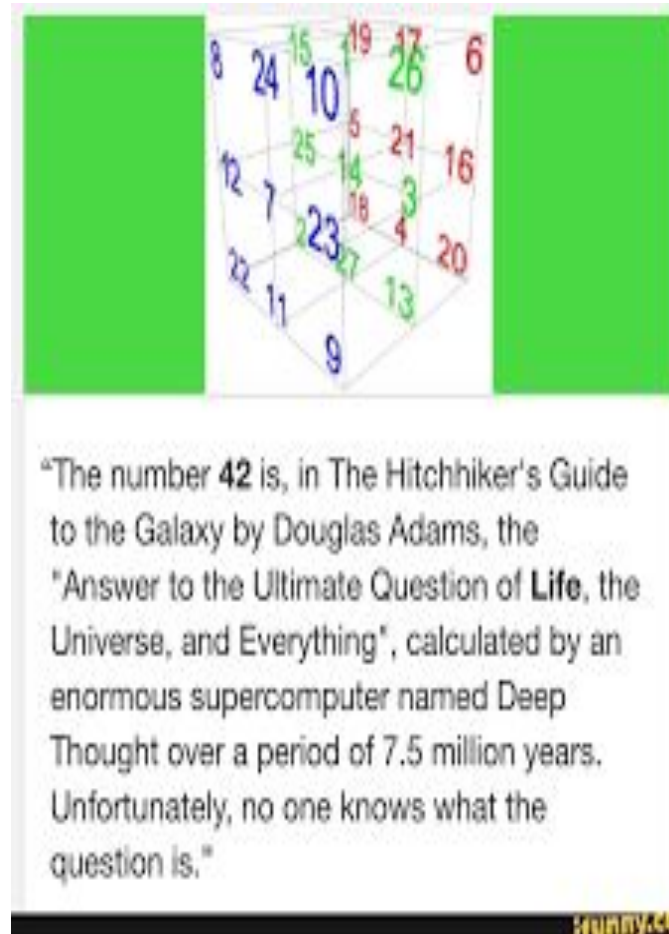
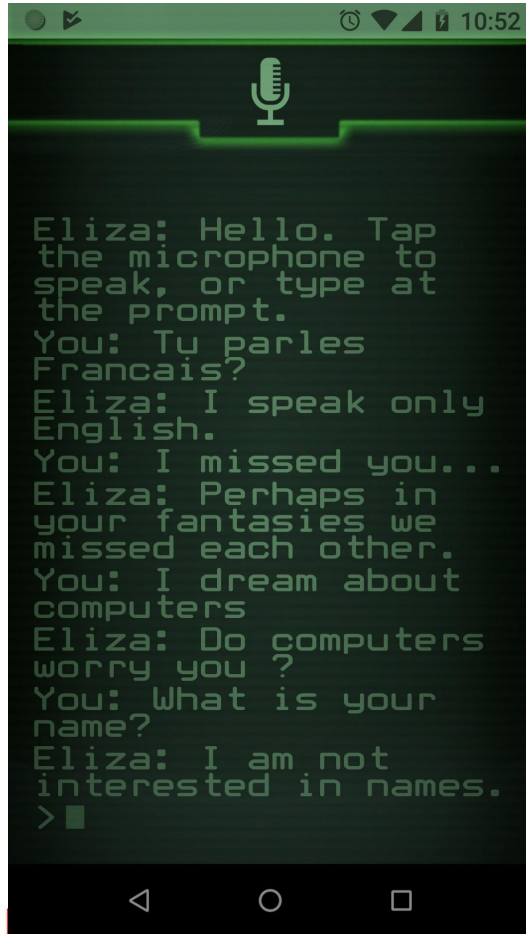
*Clouds and webs and networks
meant different things back then*

Simple text checking – tokenization, sequence alignment, but we did not even know these terms...

Simple Markov-based word sequence generators. These were 8-bit computers, but the ideas were there already.

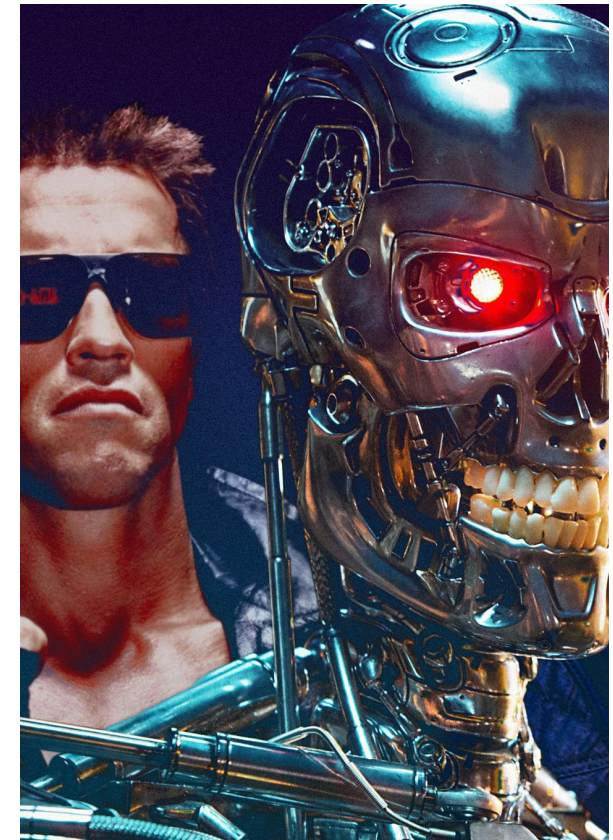


Eliza, Deep Thought, Golem XIV, and now?



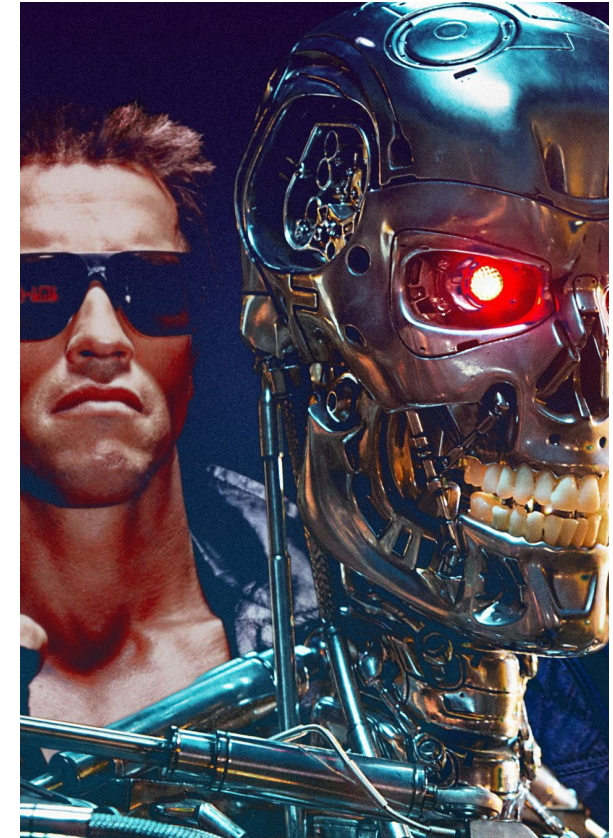
AI in Social Science, Literature, Media

- Weizenbaum, Joseph (January 1966). ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*. 9: 36–35.
- Adams, D. (2007). *The Hitchhiker's Guide to the Galaxy*. Random House.
- Lem, Stanisław (1981). *Golem XIV*. Wydawnictwo Literackie, Krakow.



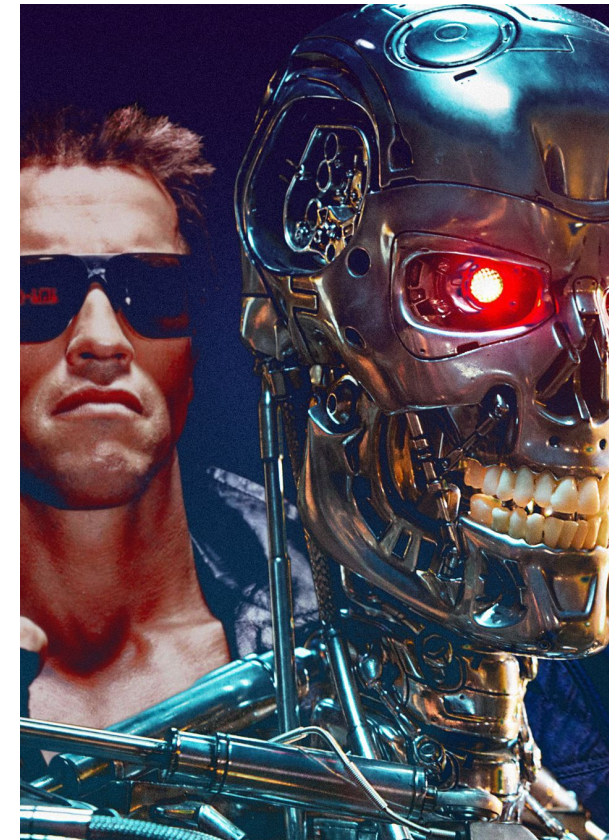
AI in Social Science, Literature, Media

- But AI did not really exist before 2023, right?
- This is all brand new and extremely unknown territory and hence very dangerous to apply!
- Education has not have touchpoints with AI, so we should rather wait until the dust settles, right?

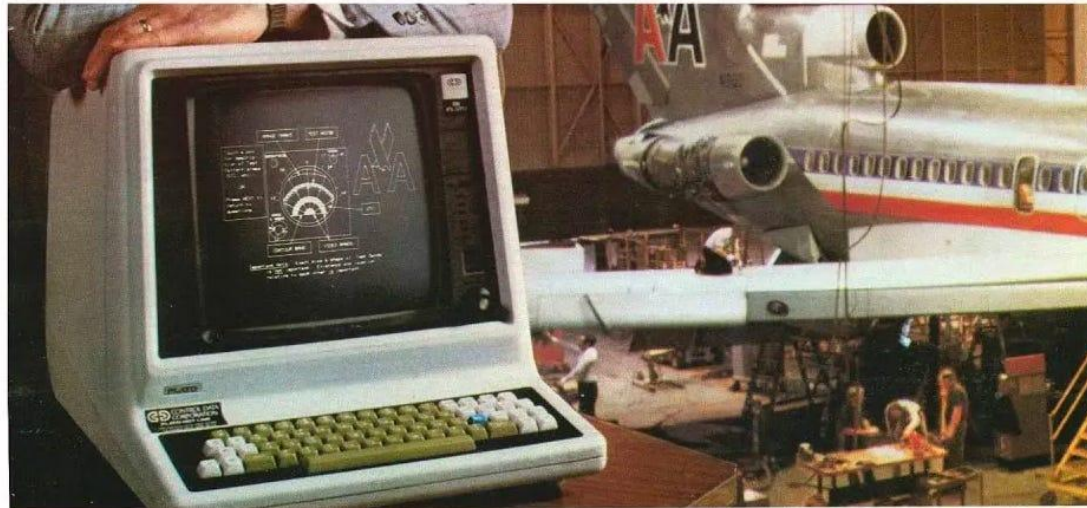


AI in Social Science, Literature, Media

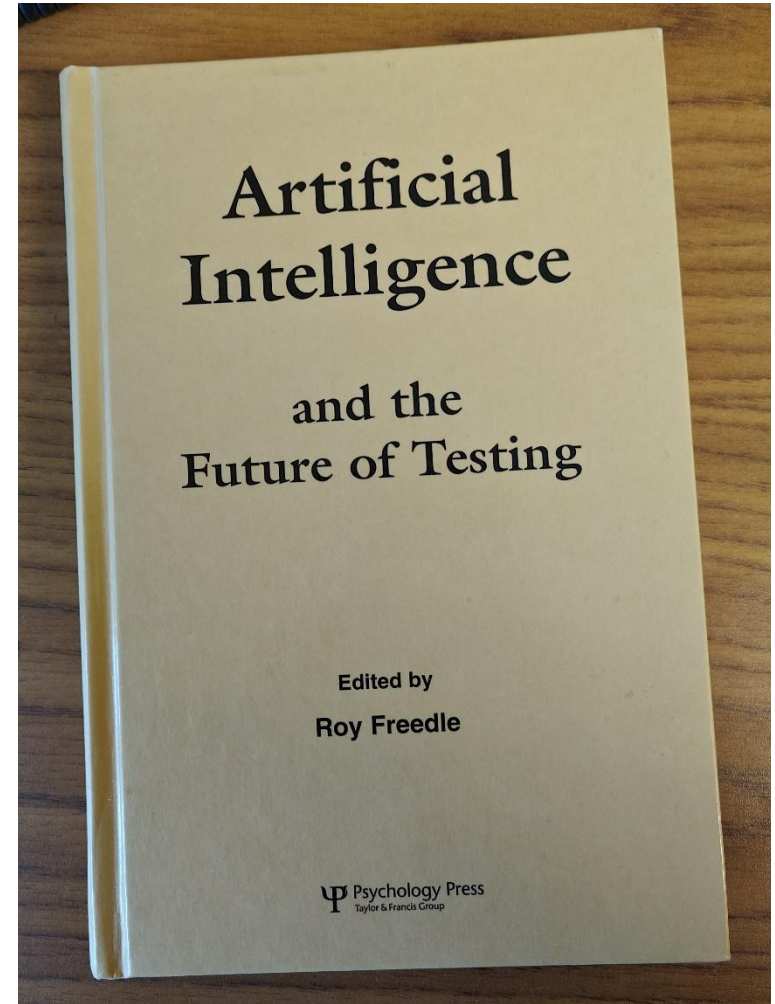
- But AI did not really exist before 2023, right?
- This is all brand new and extremely unknown territory and hence very dangerous to apply!
- Education has not have touchpoints with AI, so we should rather wait until the dust settles, right?
- **Not quite right! AI has been around 60+ years and AI in education is decades old!!!**



AI in Education? It must be new! (NO!)



PLATO
CHANGING HOW THE WORLD LEARNS



The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- **Automated Scoring – or What AIs Already Do Well**
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

A Bit of History: Automated Essay Scoring

“Breakthrough? Or buncombe and ballyhoo? You should know, after reading this careful description [sic] of efforts at the **University of Connecticut** to rescue the conscientious English teacher from his backbreaking burden. It is authored by the researcher whose very first computer strategy for essay grading yielded marks indistinguishable from those of experts. Mr. Page, himself a refugee from English teaching, answers questions that will occur to the skeptical educator.”

Page (1966) is considered the first attempt at AES (<https://www.jstor.org/stable/20371545>) presented '65 at ETS(!)

A Brief History of AES (according to GPT4 and me)

- 1966: Ellis Batten Page developed Project Essay Grade (PEG). PEG used a rudimentary form of NLP to evaluate essays.
- 1971: ETS developed a computer program called Writer's Workbench, which provided feedback on grammar and usage in student writing.
- *1984: Data Beutner releases Language Learning Software for MSX home computers that grades open responses (von Davier, Dienst, etc.)*
- 1997: The Intelligent Essay Assessor (IEA) was developed by Thomas Landauer and Peter Foltz.
- 1999: ETS released e-rater, a computer program that uses NLP to evaluate essays based on grammar.
- 2002: Vantage Learning released the IntelliMetric system, which uses AI and NLP.
- 2003: Pearson released the Intelligent Essay Assessor, a revised version of IEA.
- 2006: ETS released e-rater version 2.0, which included new features such as the ability to detect plagiarism and provide feedback.
- 2010: Pearson released the Pearson Test of English Academic (PTE Academic), an English proficiency test that uses automated scoring.
- 2013: ETS released the e-rater Scoring Engine API, which allows developers to integrate e-rater into their own software.



Automated Scoring Open-Ended Responses in International Assessments

TIMSS 2023 explores and utilizes more modern technologies, such as neural networks and deep learning, for automated scoring of open-ended responses:

- Supporting human scoring of open-ended responses
- Automated scoring of graphical responses
- Automated scoring of written responses

Early Artificial Neural Networks

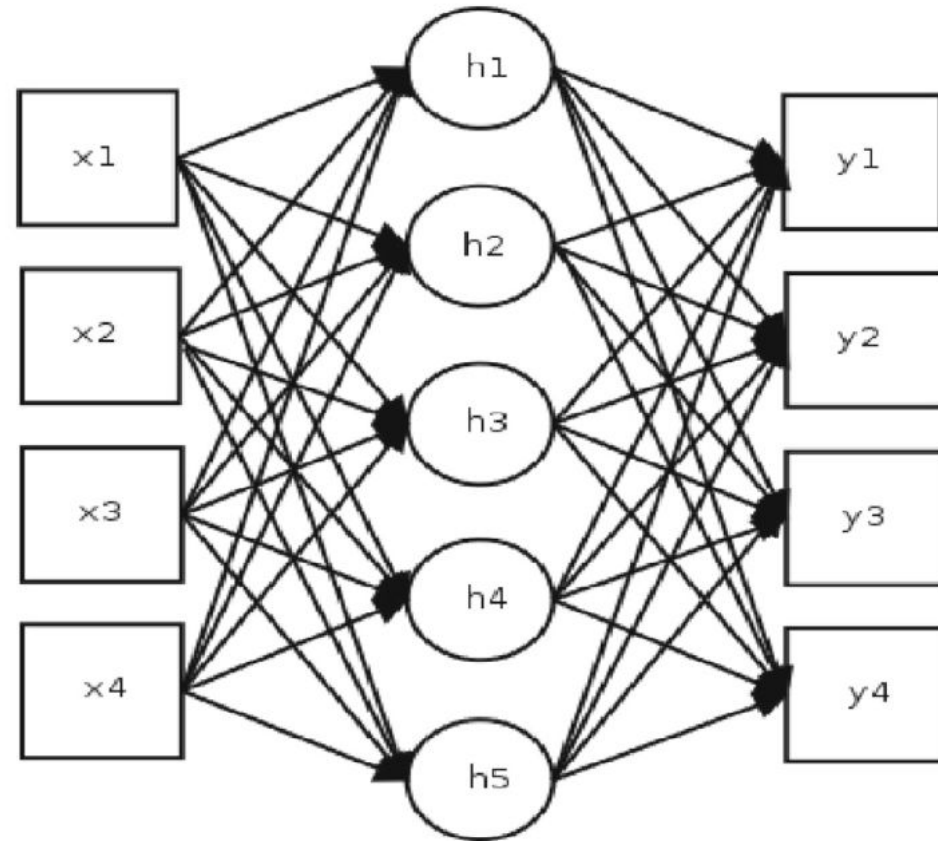


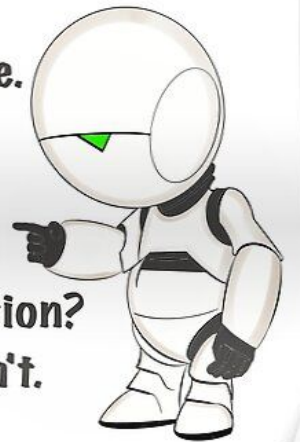
FIGURE 1.
Small feed-forward neural network with a single hidden layer.

What are (Deep) Neural Networks Used For?

Some standard applications are:

- **Computer vision:** Classifying images into certain categories ("Person. Woman. Man. Camera. TV.",...)
- **Representing Meaning:** e.g. by embedding (mapping/encoding) terms in a high dimensional space representing meaning by preserving their context and relative position
- **Sequence models:** Learn rules of how characters, words, or musical notes are aligned sequentially: Generate new sequences, predict, and translate

Here I am, brain the size of a planet, and they tell me to take you up to the bridge. Call that job satisfaction? 'Cos I don't.



TIMSS 2023: Automated Scoring for Quality Assurance

Supporting the Human Scoring of Open Responses

- Human scoring requires training of many scorers
- Scoring training in 70+ languages is a huge challenge
- Training can vary in success, scorers may vary in quality
- Scorer agreement can be better monitored with AI
- Automated Scoring using AI can provide reference data

What are Deep Neural Network AIs Used For?

Some standard applications are:

- **Computer vision:** Classifying images into certain categories("Person. Woman. Man. Camera. TV.",...)
- **Representing Meaning:** e.g. by embedding (mapping/encoding) terms in a high dimensional space representing meaning by preserving their context and relative position
- **Sequence Models:** Learn rules of how characters, words or musical notes are aligned sequentially: Generate new sequences, predict (continue sequences) and translate



TIMSS 2023: Engaging & Interactive Problem Solving and Inquiry Tasks

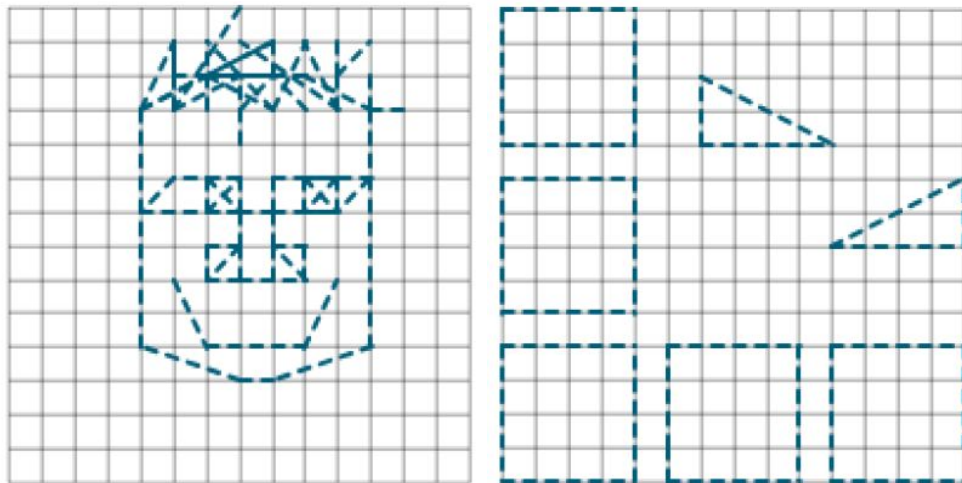
The screenshot shows a digital interface for a math task. At the top left, it says 'IEA eTIMSS 2019' and 'TIME LEFT 45'. A vertical sidebar on the left has numbered buttons 1 through 7, with button 4 highlighted. The main content area is titled '4 Constructing the Walls'. Below the title, there is a 3D perspective view of a building's wooden frame. The frame has a height of 2 m, a depth of 4 m, and a width of 4 m. Above the 3D view are three tabs: 'Front View', 'Back View', and 'Side View'. Below the 3D view, there is text: 'You are going to build the walls of the building. The building will have two side walls and one back wall. Click the tabs below to see another view.' Below this text, there is another paragraph: 'The board you have is 14 m by 14 m. The squares on the board are each 1 m by 1 m. Draw lines on the board to show where you would cut the one piece for the back wall and the two pieces for the side walls.' Below the text is a 14x14 grid representing the board. At the bottom of the grid are three icons: a pencil, an eraser, and a refresh button. At the very bottom of the interface are navigation arrows and a calculator icon.

- The 'building' is an example of engaging, multi-step, interactive tasks
- This TIMSS 2019 task was released and data will be made available for research
- TIMSS 2023 will contain the next generation of interactive tasks

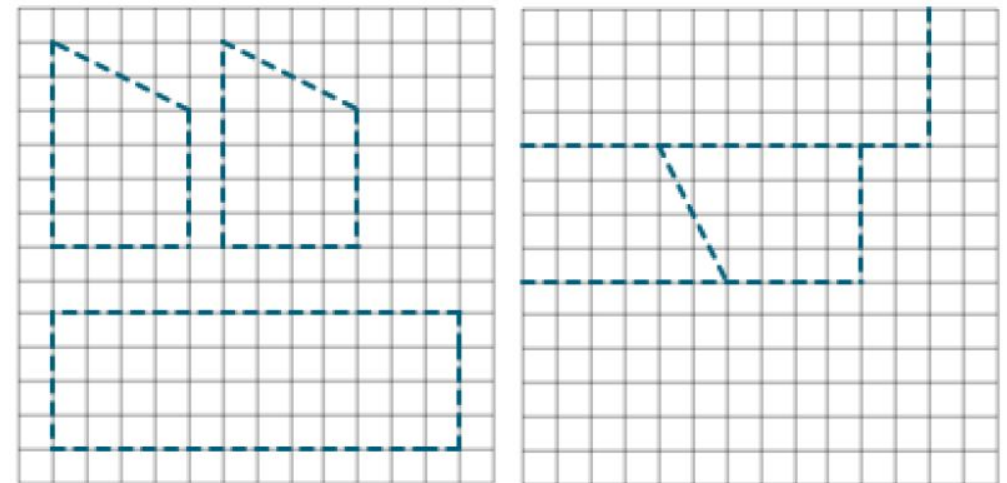
TIMSS 2023: AI Supported Scoring Research

Automated Scoring of **Graphical Open Responses**

Incorrect Responses



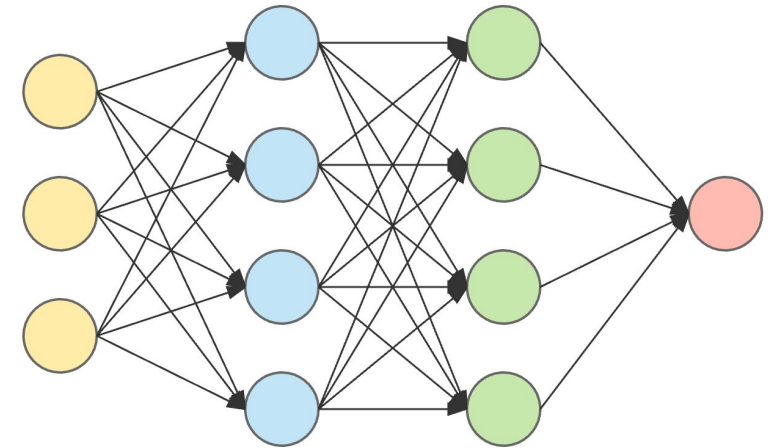
Correct Responses



TIMSS 2023: AI Supported Scoring Research

Automated Scoring of **Graphical Open Responses**

- Research project using artificial neural networks (ANNs)
- The project is training ANNs to score 15,000 line drawings collected in the 'Building' released PSI task
- Results are encouraging—very high agreement (between 91% and 98%) when comparing performance on test data (unseen by ANNs) with human scores
- IEA is supporting our work on a research project expanding the results to other TIMSS items



TIMSS 2023: AI Supported Scoring Research

Automated Scoring of Written Responses

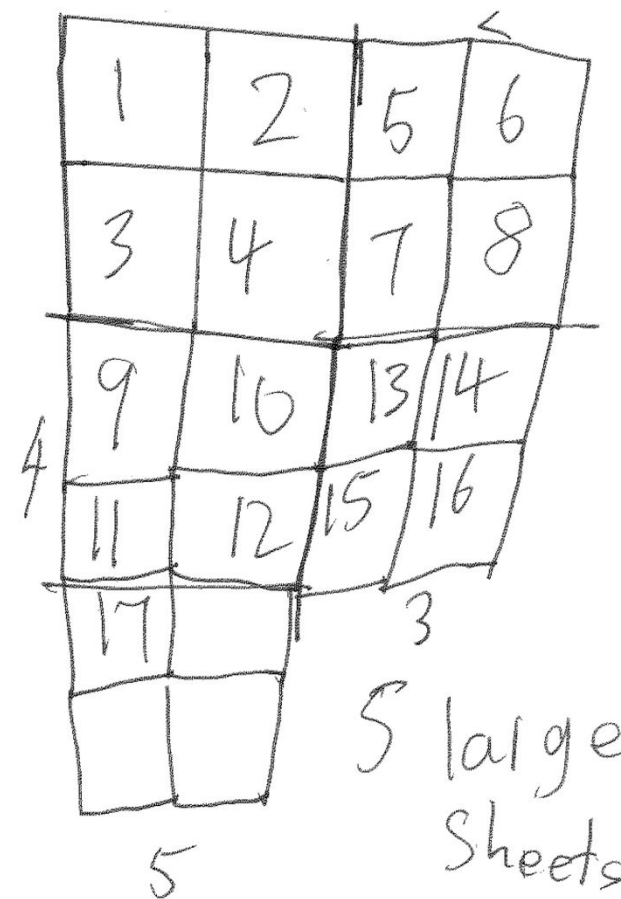
- Open ended responses from released items were analyzed in three languages: Korean, German, and English
- 'bag of words' based classification methods were used

```
## [1] "Label"      "hide"      "eat"       "hold"      "tail"      "head"
## [7] "shade"      "sun"       "heat"      "help"      "squirrel"  "surviv"
## [13] "give"       "use"       "behavior"  "bodi"     "can"       "cool"
## [19] "act"        "like"      "cape"      "block"     "sunlight"  "make"
## [25] "keep"      "will"     "light"     "feel"     "hot"       "much"
## [31] "umbrella"  "shelter"  "go"        "eye"      "tri"       "get"
## [37] "think"     "look"     "rock"      "day"      "behaviour" "shine"
## [43] "way"       "find"     "food"     "cover"    "hit"       "predat"
## [49] "gete"      "tale"     "big"       "fur"      "put"       "shad"
## [55] "ground"    "protect"  "air"       "live"     "dri"       "squirrel"
## [61] "stay"     "environ"  "warm"     "let"      "desert"    "water"
## [67] "dont"     "face"     "know"     "someth"   "nut"       "thing"
## [73] "away"     "burn"     "die"       "thier"    "anim"      "mayb"
## [79] "need"     "self"     "allow"     "provid"   "lift"      "enviro"
## [85] "respond"  "chang"   "around"   "squrrel"  "sade"     "come"
## [91] "might"    "cold"    "place"    "wont"     "squirrl"  "prevent"
## [97] "want"     "ther"    "see"      "sinc"     "ray"      "less"
## [103] "also"     "realli"  "dessert"  "cooler"   "weather"  "dehydr"
## [109] "drink"    "touch"   "caus"
```

TIMSS 2023: AI Supported Scoring Research

Automated Scoring of **Written Responses**

- Pilot project using ML and DL methods
- Comparing scoring agreement of different approaches, logistic regression, ANNs, random forest, etc.
- Results are encouraging, performance on test data (unseen by methods) compared with human scores, up to 91% agreement
- We examine multilingual aspects by studying differences between approaches using a variety of methods for language processing.



So, Is it Useful? YES:

von Davier, M., Tyack, L., & Khorramdel, L. (2022). *Scoring Graphical Responses in TIMSS 2019 Using Artificial Neural Networks*. Educational and Psychological Measurement. <https://doi.org/10.1177/00131644221098021>

Jung, J.-Y.; Tyack, L.; von Davier, M. (2022). *Automated Scoring of Constructed-Response Items Using Artificial Neural Networks in International Large-scale Assessment*. Psychological Test and Assessment Modeling; Lengerich Vol. 64, Iss. 4, (2022): 471-494.

The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

woke up to the _____ smell of _____
adjective type of bird

roasting in the _____ downstairs. I
room in a house

_____ down the stairs to see if I could
verb (past tense)

help _____ the dinner. My mom said,
verb

"See if _____ needs a fresh _____." So I
relative's name noun

carried a tray of glasses full of _____ into
a liquid

the _____ room. When I got there, I
verb ending in -ing

couldn't believe my _____! There were
part of the body (plural)

Does traditional AIG work well?

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P. C. Kyllonen (Eds.), *Generating items from cognitive tests: Theory and practice* (pp. 199–217). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, 64, 407–433.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Lawrence Erlbaum Associates Publishers.

Gierl, M. J., & Haladyna, T.M. (2013). *Automatic item generation*. New York, NY: Routledge.

Does traditional AIG work well?

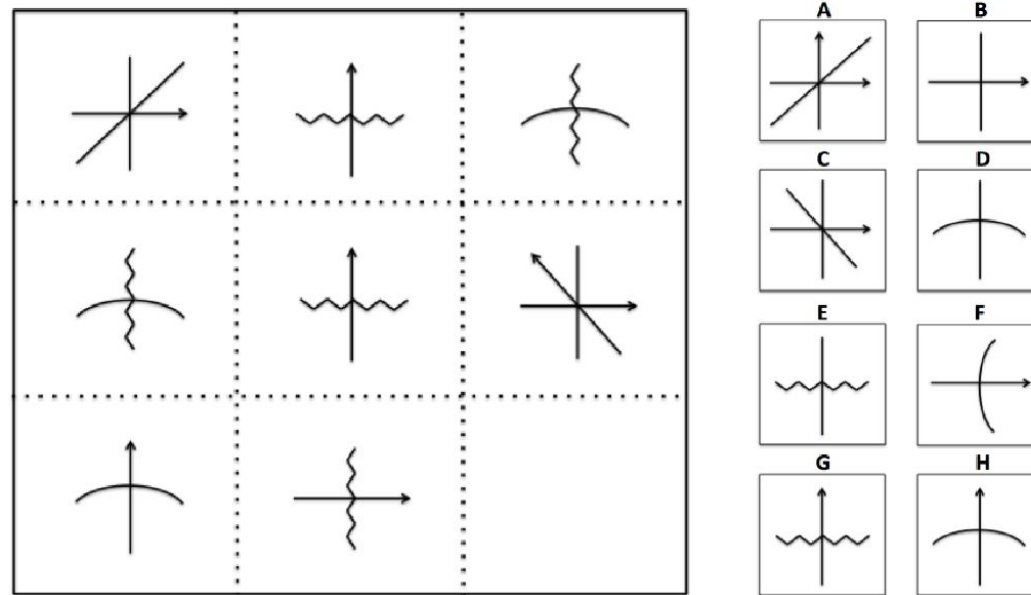


Figure 1: Example RPM problem.

Ke Wang and Zhendong Su, "Automatic Generation of Ravens Progressive Matrices",
Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

Does traditional AIG work well?

- Question Generator:
<https://www.wolframalpha.com/pro/problem-generator/>
- AIG from Cafalab:
<https://aig.cafalab.com/>
- AIG R package (remember my GPT-3 example)
<https://cran.r-project.org/web/packages/AIG/index.html>
- Imak R package (see what I mean...)
<https://cran.r-project.org/web/packages/IMak/index.html>

The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- **Character- and Word-Based Language Models for AIG**
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

Recurrent Neural Networks

- Recurrent Neural Networks integrate the output of a previous state with the input of a next stage for tasks that require sequence learning
- Elements of a sequences are input and output, not at the same time, but iterated over and over again
- They tend to be harder to train, vanishing or exploding gradients are a problem, a topic for another time...

Recurrent Neural Networks

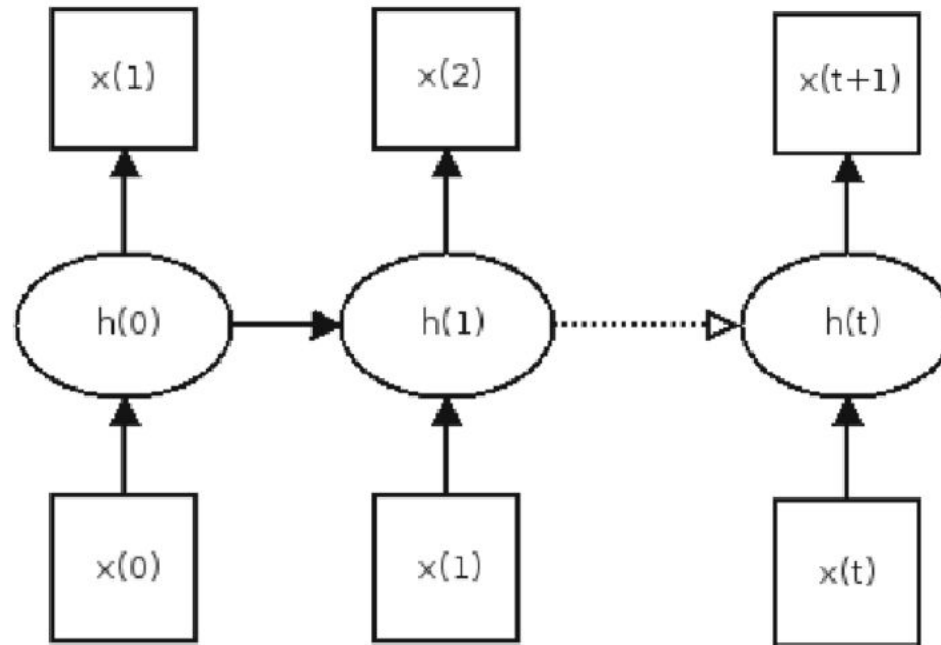
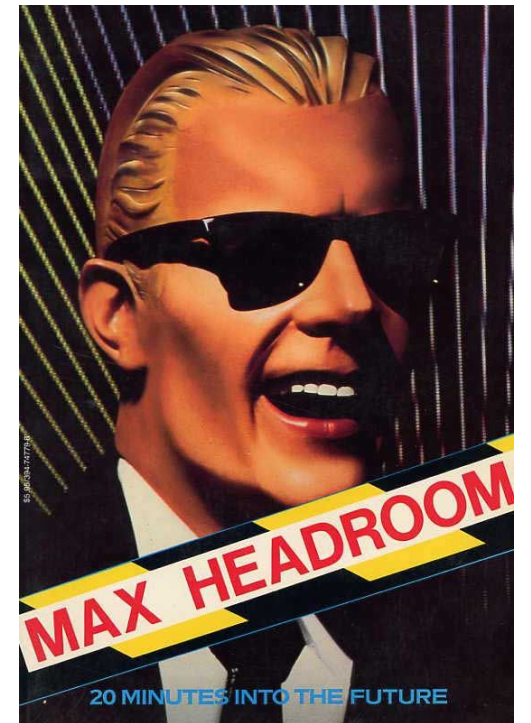


FIGURE 2.

A visualization of an unrolled recurrent neural network showing the hidden state dependency on previous hidden states.

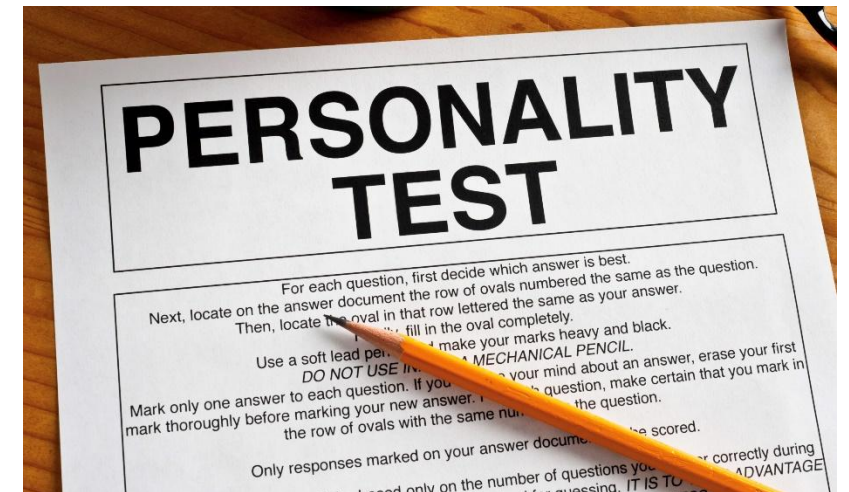
Generative Language Models

- Models that predict the next character (or word) of a text based on the sequence of previously observed characters (or words)
- Depending on how much text is there to learn, these models can produce surprisingly similar text sequences that 'sound' a lot like the original sequence



Generating BIG-5 Items with RNNs

- Personality items seemed like a low hanging fruit
- Almost a no-brainer, as learning rules for text generation is the bread and butter of these sequence models
- However, generated text sequences are often just 'surface similar', and generated text may be non-sensical
- May be of benefit in this use case (remember the MMPI?)



Generating Items with RNNs

- Generative language models need large-ish databases
- “More is better” – for once – holds here
- Open-source items used to generate new personality items
- There are 3320 personality items available at <https://ipip.ori.org/AlphabeticalItemList.htm>
- There are also many translations available upon request <https://ipip.ori.org/newItemTranslations.htm>
- Interesting to study NLG using translated items vs. translating

Generating Items with RNNs

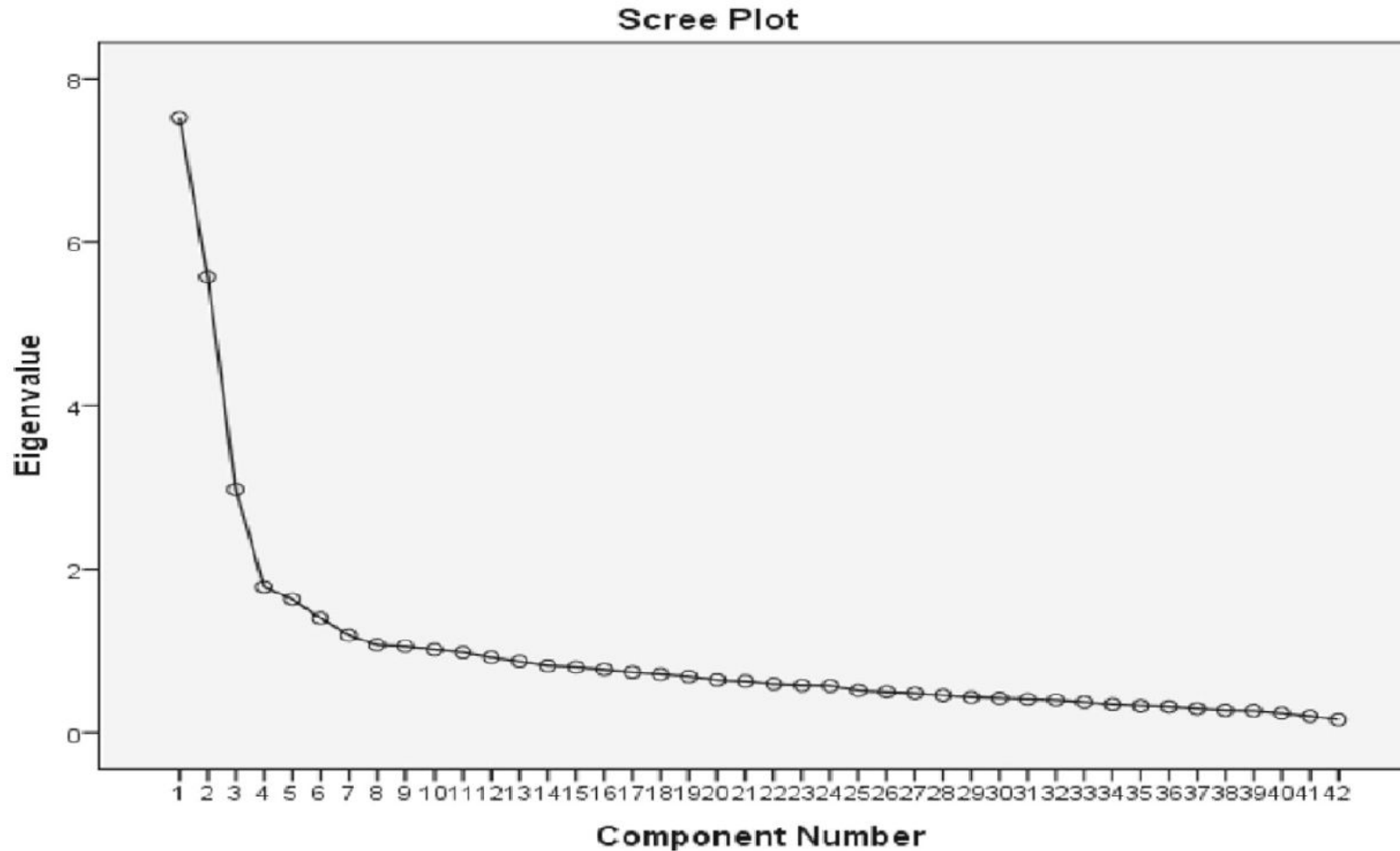
I check on things out of nothing	1
I plan my life based on other people	1
I know that I am troubled by praise	1
I usually control others	1
I work hard to get others to do things	1
I tend to be lazy	
I have few artistic interests	
I go out of my way to meet outside advice	1
I don't like to draw attention to myself	
I am a reserved person	
I talk to a lot of different people at parties	
I am outgoing and sociable	
I prefer to be the perfect leader	1

Evaluating The Generated Items

- Cost me \$87.50,...
- and 2:30 hours of my time
- using Amazon Turk, collected ~300 respondents
- ~43 items, 22 generated, rest existing/published items, from 5-factor inventories
- factor structure & loadings very promising



Evaluating The Generated Items



Summary from 2018

- 2018: AIG using RNNs seemed very promising
- More research: generating more complex items
- More complex networks to enable distractor generation and item stems would be useful
- Automated QC and selection seemed possible using *Generative Adversarial Networks (GANs)*

So, Is it Useful? YES:

von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, Volume 83, Issue 4, pp 847–857.
<https://doi.org/10.1007/s11336-018-9608-y>

The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- **Transformers, Small and Large(r) for NLG**
- Does GPT4 pass the Item-Writing Turing-Test?

What Are Transformers?

- Large Artificial neural networks
- Add positional encoding
- Attention units identify key topics of the sequence
- Integrates previous states of the sequence processing
- Used in
 - machine translation
 - natural language generation
 - summarization, etc.

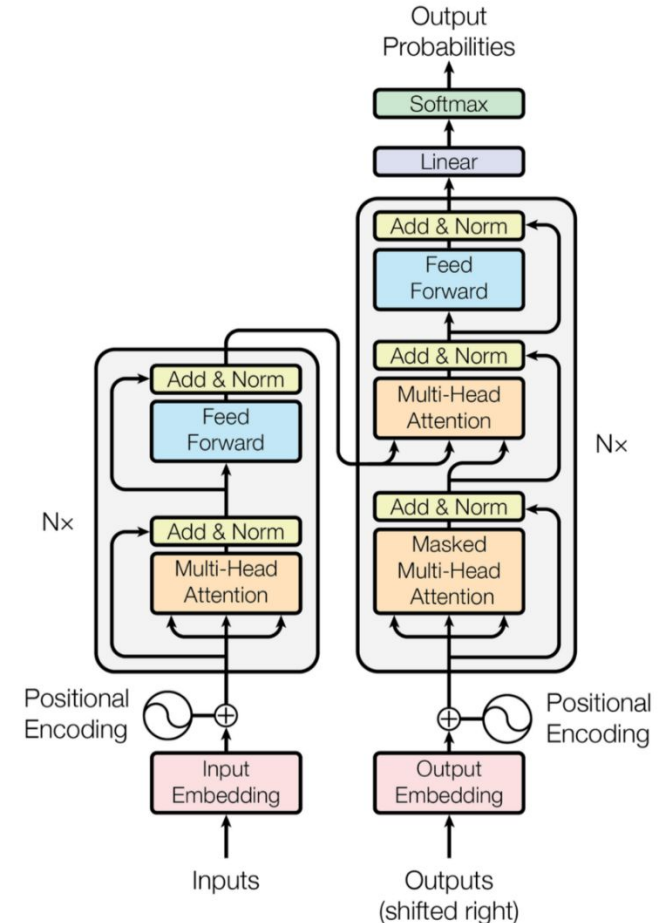
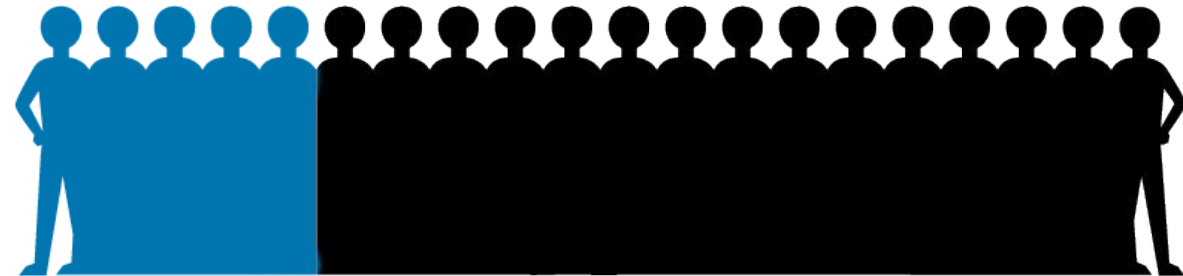


Figure 1: The Transformer - model architecture.

2019: Retraining Transformers

- GPT-2, BERT are based on ‘transformer architectures’
- These are very large language models based on deep neural networks, trained on very large amounts of data
- In the application to AIG for medical vignettes, I retrained the GPT-2 medium (345 million parameters) with **>4GB of text** taken from open access articles **from PubMed**
- Training from ground up would take weeks, retraining is like fine-tuning a model, adjusting the parameters to a population



A Year Later, Retraining Transformers

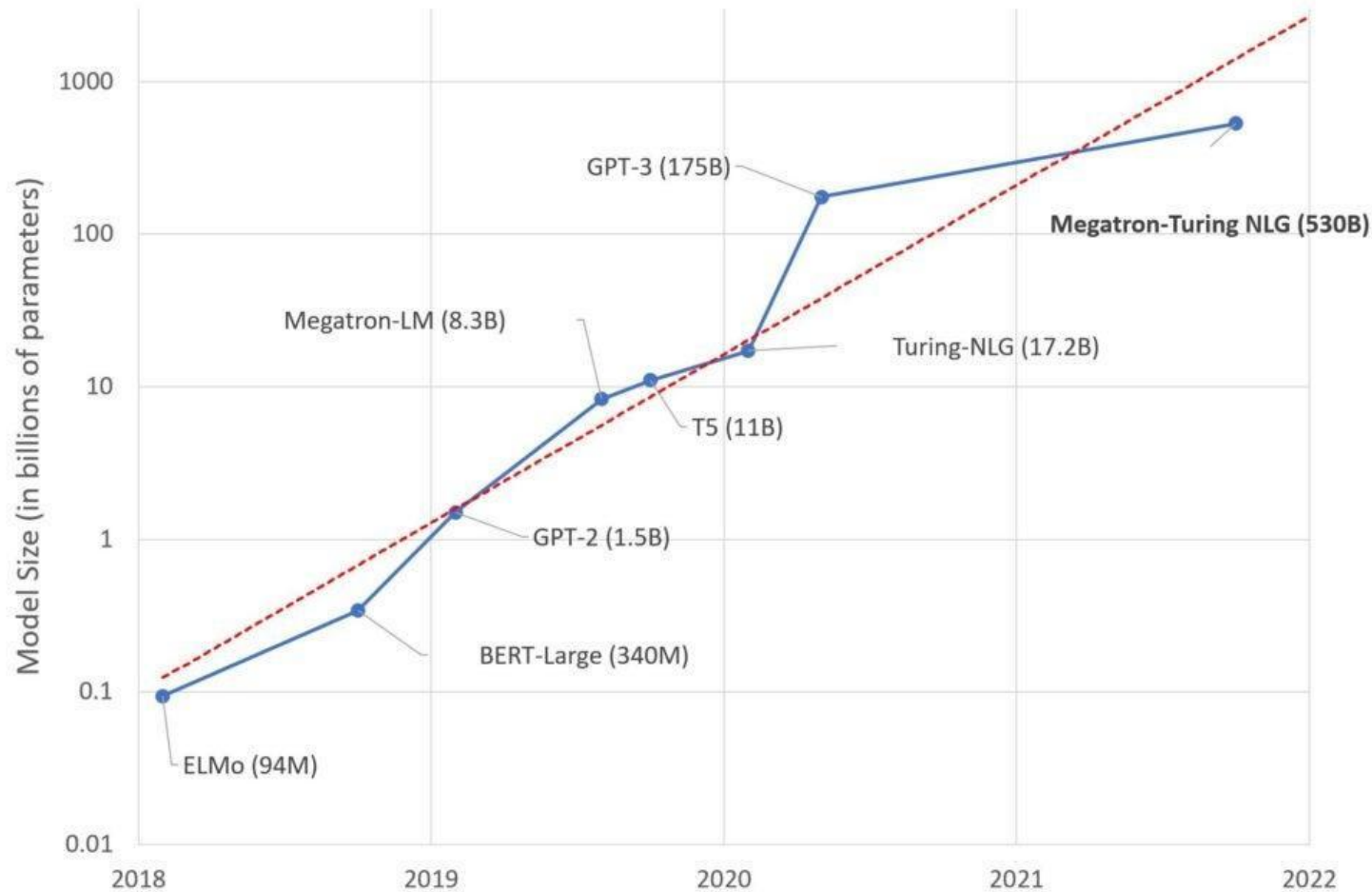
von Davier, M. (2019). Training Optimus Prime, M.D.:
Generating Medical Certification Items by Fine-Tuning
OpenAI's GPT-2 Transformer Model.

<https://arxiv.org/abs/1908.08594>

What Use Cases are there for Transformers?

- GPT = General Pretrained Transformer
 - 2019: GPT-2 (OpenAI) – 1.5 billion parameters (~neural connections)
 - 2020: GPT-3 (OpenAI) – 175 billion parameters
 - 2021: Megatron-Turing (NVIDIA/Microsoft) – 530 billion parameters
- Used for Natural Language Generation (NLG) in Tasks such as:
 - [Writing assistants](#) and [Story-Based Games](#)
 - [Chatbots](#)
 - [Summarization of text](#)
 - [Question answering \(also with numerical data\)](#)
 - Code Completion ([Programming support](#))
 - [Automated Item Generation](#) (also others: commercial [Finetune](#))

State of affairs as of 10/2021



- 530 billion parameter Megatron Turing announced by Microsoft & NVIDIA October 11th, '21
- an estimated 86 billion neurons in human brain, *but how many synapses?*
- in 3 years, will NVIDIA make a Megatron-Deep-Thought-Lovelace-Zuse-Golem XIV = 23 trillion parameter general AI?

(OOOOPS!)

What can 'vanilla' NLG models do in 2021?

Prompt: Lilly was a five-year-old very inquisitive spider. She wanted to know how things work. Her brothers and sisters were not like her. That is why Lilly chose to venture out into the world and explore. How do plants grow? Where does the water in the river come from? She decided that if nobody could tell her, she would find out herself.

Response: Her mother helped her to gather some things, but she was not sure whether she had everything that she needed. She went outside her house. She ran through the fields. She jumped from one plant to another. She walked on the stream. She looked at the sea and the sea looked at her. The sea told her that the plants that she looked at were the most beautiful of all.

She went back to her house. She asked her mother. She asked her father. They were not sure what she was talking about. They knew that plants grow and that water comes from the sea. They did not understand how that could...

<https://6b.eleuther.ai/>

What can 'vanilla' NLG models do in 2021?

Prompt: Lilly was a five-year-old very inquisitive spider. She wanted to know how things work. Her brothers and sisters were not like her. That is why Lilly chose to venture out into the world and explore. How do plants grow? Where does the water in the river come from? She decided that if nobody could tell her, she would find out herself.

Response: So Lilly climbed onto a tree. She was very excited because she was finally free to go wherever she wanted. She had already been to the moon and knew what it looked like. She was about to see something new when she heard a voice. It was her father calling her home. She was very surprised because she did not think he would mind. It was not her fault that she was out of the house.

The next day, Lilly's mother asked her to stay home. Lilly was very sad. She knew that she was not allowed to leave the house. The next day, Lilly was given an assignment. She...

<https://6b.eleuther.ai/>

The use of AI in International Assessment

- From Konrad Zuse and Grace Hopper to ChatGPT
- Automated Scoring – or What AIs Already Do Well
- Early AIG: Just an exercise in permutations (MadLib, anyone?)
- Character- and Word-Based Language Models for AIG
- Transformers, Small and Large(r) for NLG
- Does GPT4 pass the Item-Writing Turing-Test?

Do We Need a New Turing Test?

TURING TEST EXTRA CREDIT:
CONVINCE THE EXAMINER
THAT HE'S A COMPUTER.

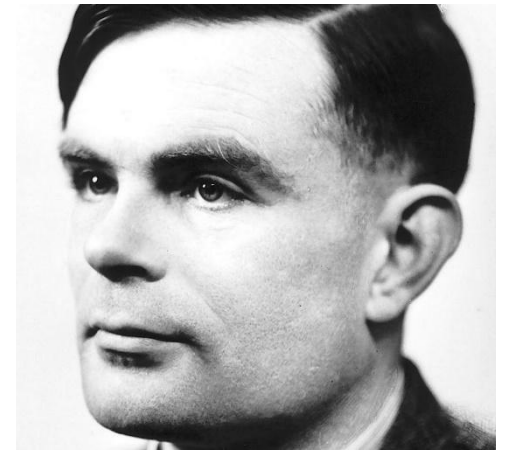
YOU KNOW, YOU MAKE
SOME REALLY GOOD POINTS.
/
I'M ... NOT EVEN SURE
WHO I AM ANYMORE.



What is the Turing Test?

The Turing test measures a machine's ability to exhibit intelligent behavior equivalent to or indistinguishable from a human. It involves a human evaluator engaging in natural language conversations with a machine and a human without knowing which is which. If the evaluator cannot reliably distinguish between the machine and human responses, the machine is said to have passed the Turing test.

(GPT4 output)



Who was Alan Turing?

Alan Turing (1912–1954) was an English mathematician, computer scientist, logician, cryptanalyst, philosopher, and theoretical biologist.

(Wikipedia)

ATHLETICS

MARATHON AND DECATHLON CHAMPIONSHIPS

The Amateur Athletic Association championships for this year were concluded at Loughborough College Stadium, Leicestershire, on Saturday, with the second, and last, day of the Decathlon and the decision of the Marathon championship.

MARATHON CHAMPIONSHIP. (26 miles-385-yds.) (record: 2hrs. 30min. 57.6sec., by H. W. Payne, Windsor to Stamford Bridge, on July 5, 1929; standard time: 3hrs. 5min.)—1. T. Holden (Tipton Harriers), 2hrs. 33min. 20-1-5sec., 1; T. Richards (South London Harriers), 2hrs. 36min. 7sec., 2; D. McNab Robertson (Maryhill Harriers, Glasgow), 2hrs. 37min. 54-1-5sec., 3; J. E. Farrell (Maryhill Harriers), 2hrs. 39min. 46-2-5sec., 4; Dr. A. M. Turing (Walton A.C.), 2hrs. 46min. 3sec., 5; L. H. Griffiths (Reading A.C.), 2hrs. 47min. 50-2-5sec., 6.

DECATHLON CHAMPIONSHIP.—H. J. Moesgaard-Kjeldsen (Polytechnic Harriers, London), 5,965 points, 1; Captain H. Whittle (Army and Reading A.C.), 5,650, 2;



What I hoped for in 2021: (How Little Did I Know)

We want AI based NLG to produce 'meaningful' and non-offensive texts. We want to build an AI that does our bidding with the purpose of AIG:

- Align generated text with context variables (topics, response types)
- Properly reflect intended content & cater to developer needs
- Follow good item writing guidelines and study conventions
- *'the AI should read my mind and write the items I want'*
- ***We need better ways to guide NLG using AI
(few shot learning, prompting, conditional output?)***



I assume GPT5 may pass the Turing Test

GPT4 is causing the current (early 2023) hyper AI buzz:

- It is said to pass several major high-stakes tests (while the journalists were exaggerating a tiny bit)
- It can produce surprisingly coherent text, can revise, revisit, and seemingly 'react' properly to requests
- It cannot (yet) include current (day-to-day) events, but work is going on to enable this
- Competition freaks out and calls for an AI 6-month dev. halt
(Musk signs, screams loudest, but then invests more in AI,
<https://www.wsj.com/amp/articles/elon-musks-new-artificial-intelligence-business-x-ai-in-corporates-in-nevada-962c7c2f>)

What Would You Convince That Computers Can Write (Good) Items?

Prompt: Write a paragraph on rain forests followed by a multiple-choice item on the topic with four response options:

- Item Generation: GPT-3 (way back when that was a thing)
- Picture Generation: DALL-E
- Translation: Google Translate API

What Would You Convince That Computers Can Write (Good) Items? Plus Artwork?

The world's rain forests are vital to the global environment. They play a key role in the carbon cycle, help to regulate the Earth's climate, and are home to an estimated 50% of the world's plant and animal species. However, rainforests are being lost at an alarming rate, with an estimated 18 million acres destroyed each year.

What is the primary threat to rainforests?

- a) Deforestation
- b) Climate change
- c) Pollution
- d) Overpopulation



What Would You Convince That Computers Can Write (Good) Items? And Translate Them?

Die Regenwälder der Welt sind lebenswichtig für die globale Umwelt. Sie spielen eine Schlüsselrolle im Kohlenstoffkreislauf, tragen zur Regulierung des Erdklimas bei und beherbergen schätzungsweise 50 % der Pflanzen- und Tierarten der Welt. Regenwälder gehen jedoch mit alarmierender Geschwindigkeit verloren, wobei jedes Jahr schätzungsweise 18 Millionen Hektar zerstört werden.

Was ist die Hauptbedrohung für Regenwälder?

- a) Entwaldung
- b) Klimawandel
- c) Umweltverschmutzung
- d) Überbevölkerung



What Would You Convince That Computers Can Write (Good) Items? And Translate Them?

세계의 열대 우림은 지구 환경에 매우 중요합니다. 탄소 순환에서 중요한 역할을 하고 지구의 기후를 조절하는 데 도움이 되며 세계 동식물 종의 약 50%가 서식하고 있습니다. 그러나 열대 우림은 놀라운 속도로 사라지고 있으며 매년 약 1,800만 에이커가 파괴되고 있습니다.

열대 우림에 대한 주요 위협은 무엇입니까?

- a) 삼림 벌채
- b) 기후 변화
- c) 오염
- d) 인구 과잉



So, Is it Useful? YES:

Bezirhan, U., & von Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*. Volume 5, 2023, <https://doi.org/10.1016/j.caeai.2023.100161>

Yaneva, V., & von Davier, M. (Eds.). (2023). *Advancing Natural Language Processing in Educational Assessment* (1st ed.). Routledge. <https://doi.org/10.4324/9781003278658>