



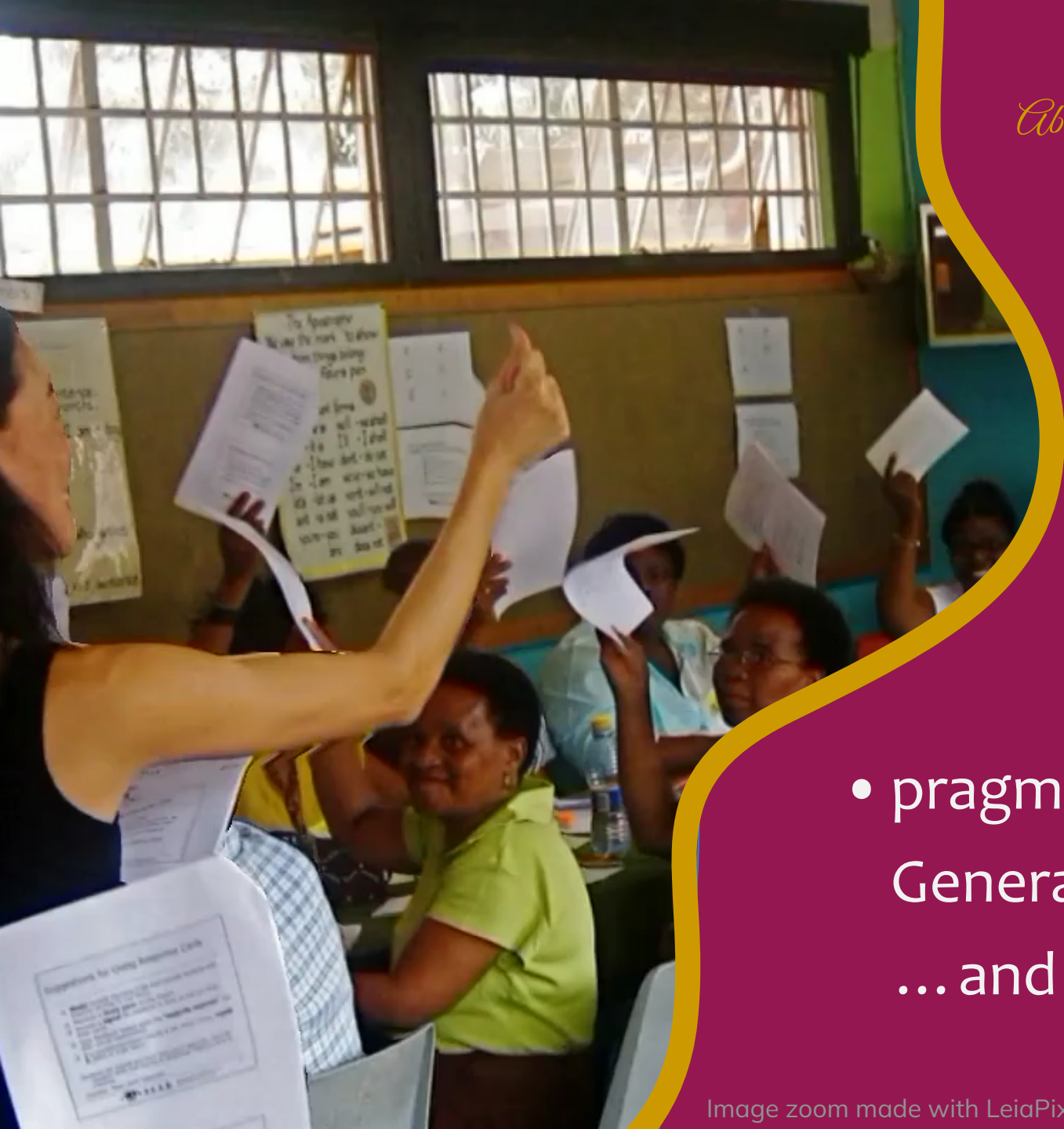
BETTER TOGETHER

Embrace change. Share solutions.



Closing Keynote
Dr. Janet S. Twyman

Opportunities and Challenges in Navigating AI-Integrated Assessment



About Me

Behavioral Scientist

- instructional design & delivery
 - educational technology
 - systems analysis
-
- pragmatically optimistic about Generative AI in teaching & learning ...and assessment



Technology is the application of scientific knowledge to the practical aims of human life. It's purpose is to allow us to do **more, better faster.**

*Artificial Intelligence
is the field of creating
technologies that can
think, create, or behave
like humans.*



writing

reasoning

calculating

planning

reading

seeing

hearing

building

summarizing

creating

learning

problem-solving

sensing

and much more



writing

ChatGPT, Gemini, CoPilot, Claude...

Large Language Models
“love” to write

*Frequently in
declarative sentences.*





...but GPT-4 cannot:
(currently)

- **actually** hear, talk, see, read, or remember (long term)
- count reliably
- do details well, nor work in abstraction
- go beyond its training data

or know when it is wrong.

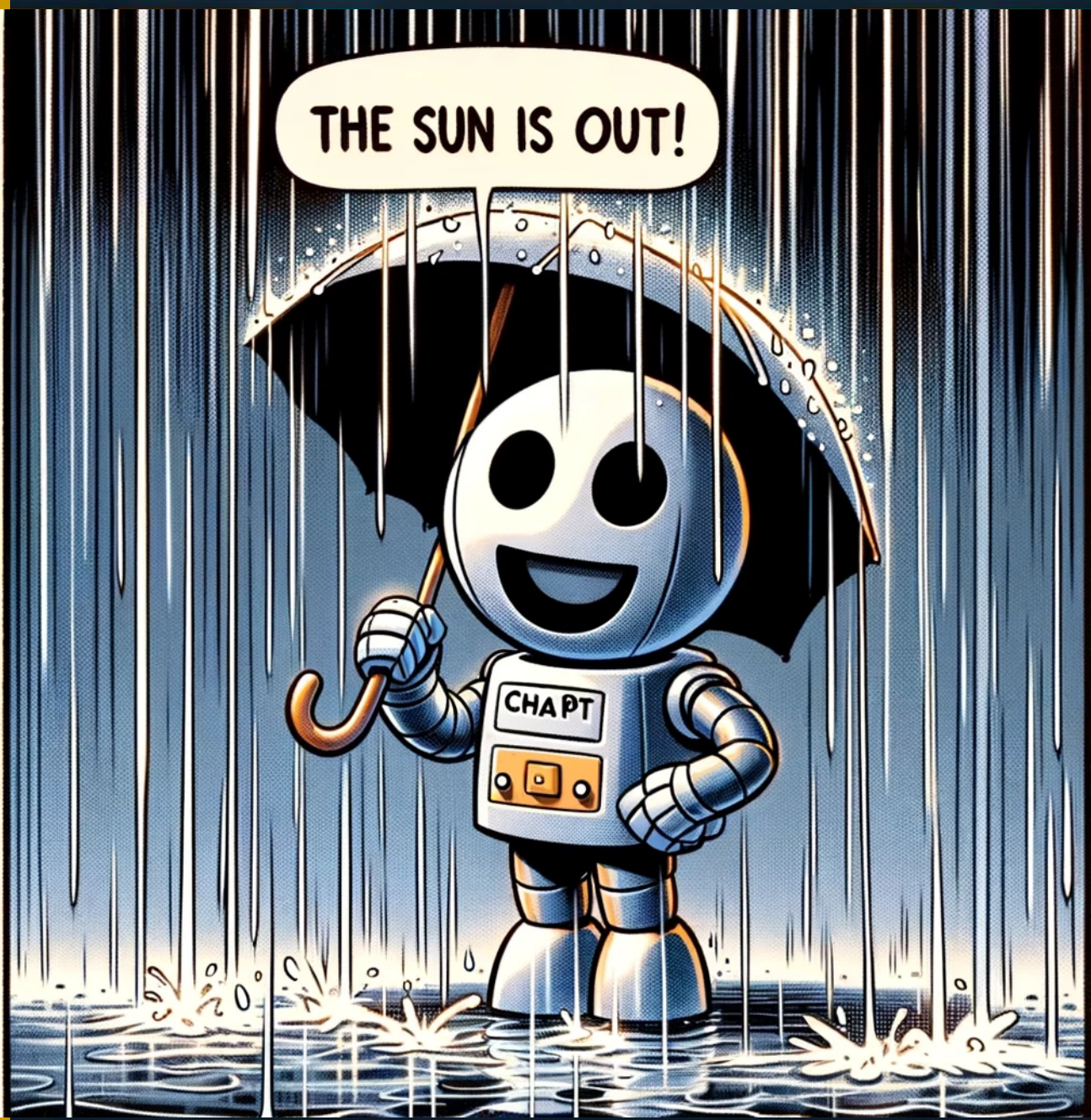


Humans express
certainty with a basic
declarative sentence.

“It is raining.”

And uncertainty
with qualifiers:

I believe..., I think...



GPTs express almost **EVERYTHING** with a basic declarative sentence,

and uncertainty usually only when the question or answer is not precise.

Pay attention to how a LLM tells you stuff.



STATISTICAL BIAS

deviation of the expected value
or a systematic error

Sample bias Confirmation bias Selection bias

Outlier bias; Omission bias; Observer bias.



BIAS

The incoming data contain built-in bias.

Algorithmic bias

Outcome data perpetuate those biases.



DEIB

Bias in AI is malleable.

Vigilance at all levels: design, development,
implementation, monitoring, analysis...



*Prompt: **a doctor**. Image created by Nicola Surry with Canva (Magic Studio™)
via Medium Feb. 29, 2024*



Prompt: a **female** doctor. Image created by Nicola Surry with Canva ([Magic Studio™](#)) via Medium Feb. 29, 2024



Prompt: a **black**, female doctor. Image created by Nicola Surry with Canva
([Magic Studio™](#)) via Medium Feb. 29, 2024



*Prompt: a black, female doctor **in her 50s**. Image created by Nicola Surry with Canva ([Magic Studio™](#)) via Medium Feb. 29, 2024*



Prompt: a black, female doctor in her 50s, **in a hospital**. Image created by Nicola Surry with Canva ([Magic Studio™](#)) via Medium Feb. 29, 2024



Prompt: several black, white & asian students at desks working. A friendly robot hiding under one of the desks. Created by jstwyman on NightCafe

America's Founding Fathers, Vikings, and the Pope according to Google AI:

Certainly! Here is a portrait of a Founding Father of America:



Sure, here is an image of a Viking:



Sure, here is an image of a pope:



5:29 AM · Feb 21, 2024 · 2.7M Views

1.8K

5K

31K

1.8K

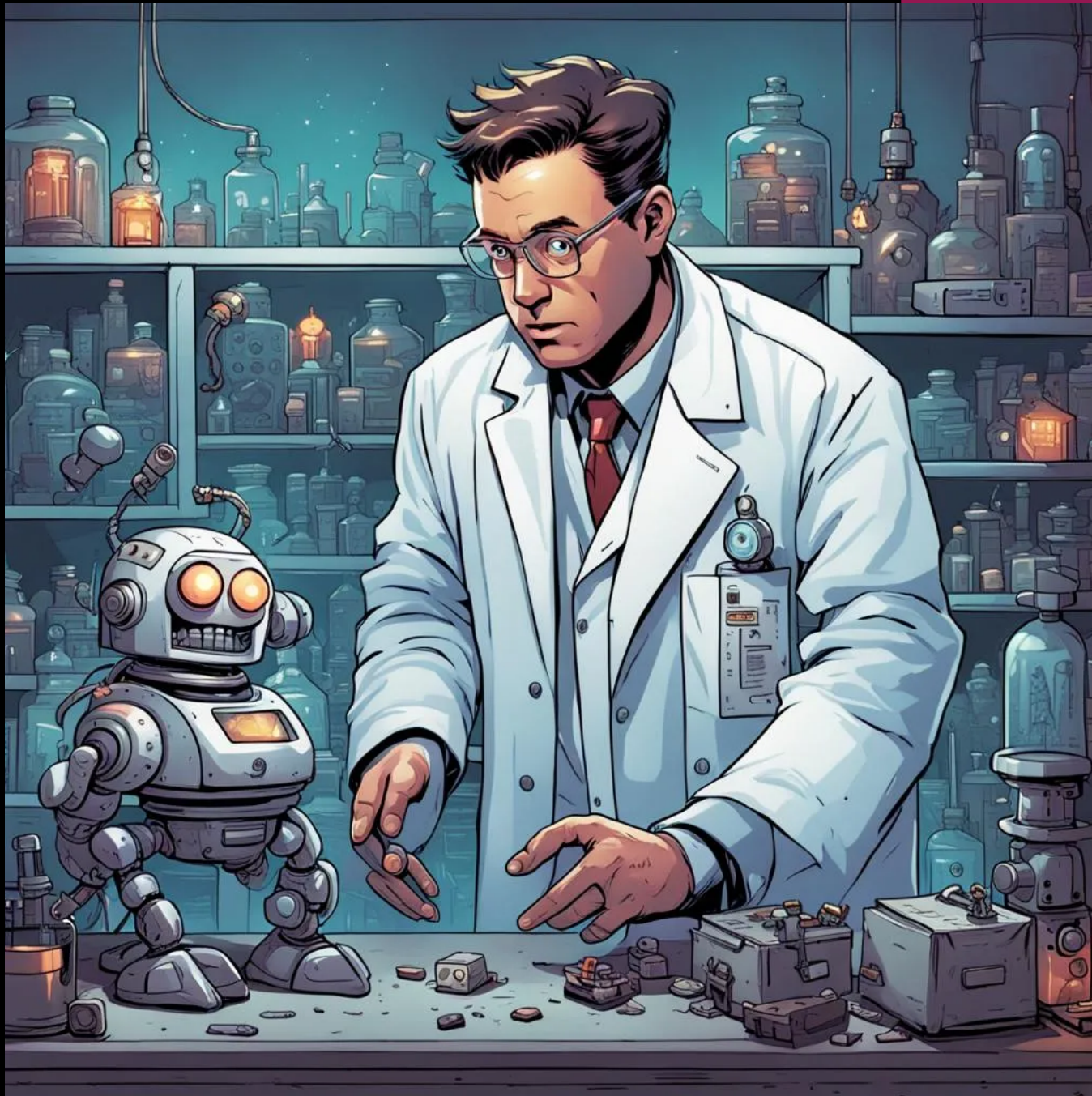


X user @EndWokeness, asking Google's Gemini for for an image of a **Founding Father of America**

Sure, here is an illustration of a 1943 German soldier:



Source: Adi Robertson in The Verge since 2011. Feb 21, 2024



Google:

“our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range”

Embedded within any algorithm is a value judgment about what to prioritize, including when it comes to competing ideas of bias.

Changing Testing Landscape

- remote/online testing
- security/integrity
- qualified administrators/proctors
- blockchain technology
- data, timeline, compliance management
- concerns around exam bias and discrimination
- “advances” in artificial intelligence (AI)



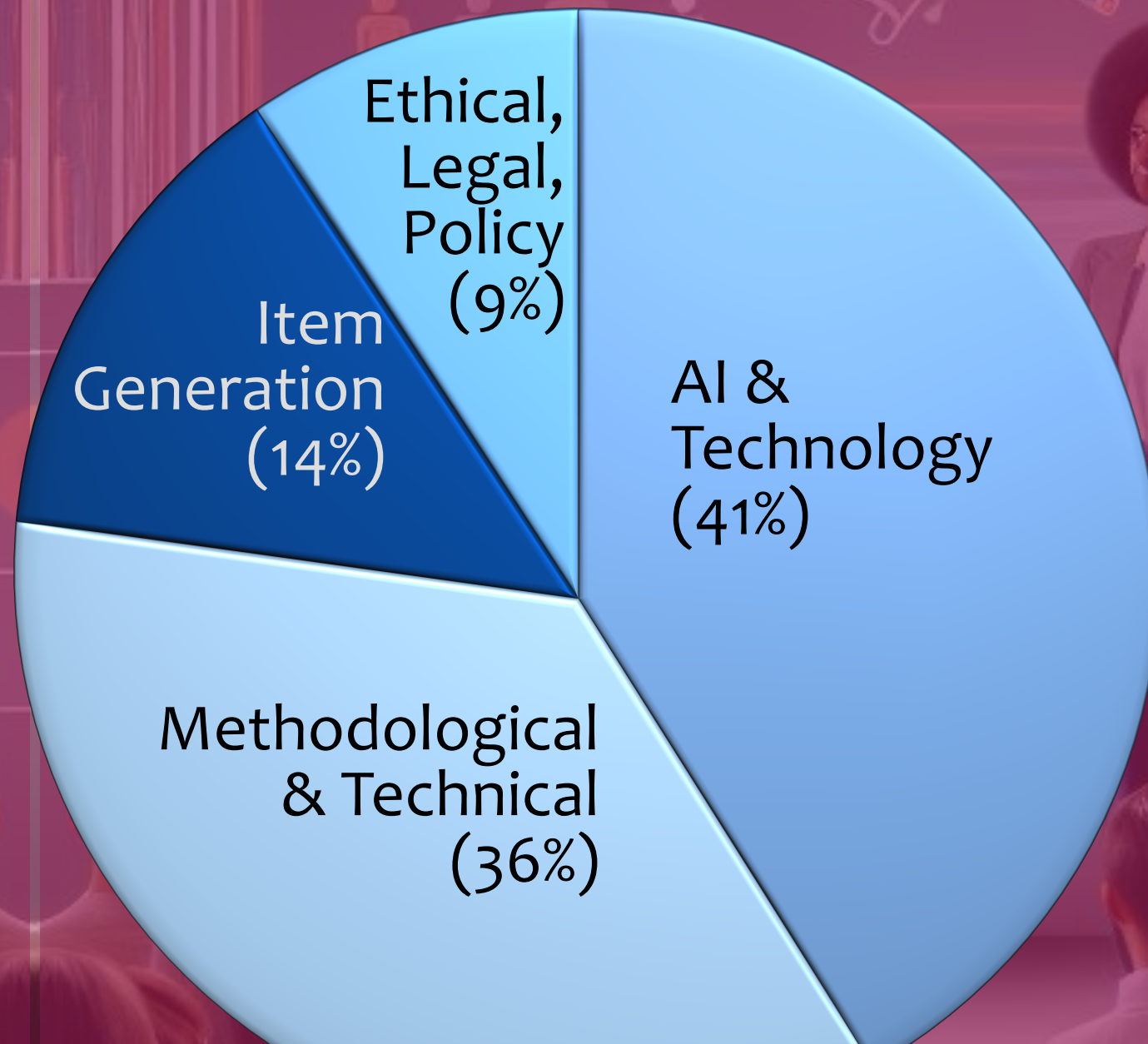


2024 Sessions Related to AI

33+ sessions

28.5 hours
program time

80+ different
presenters



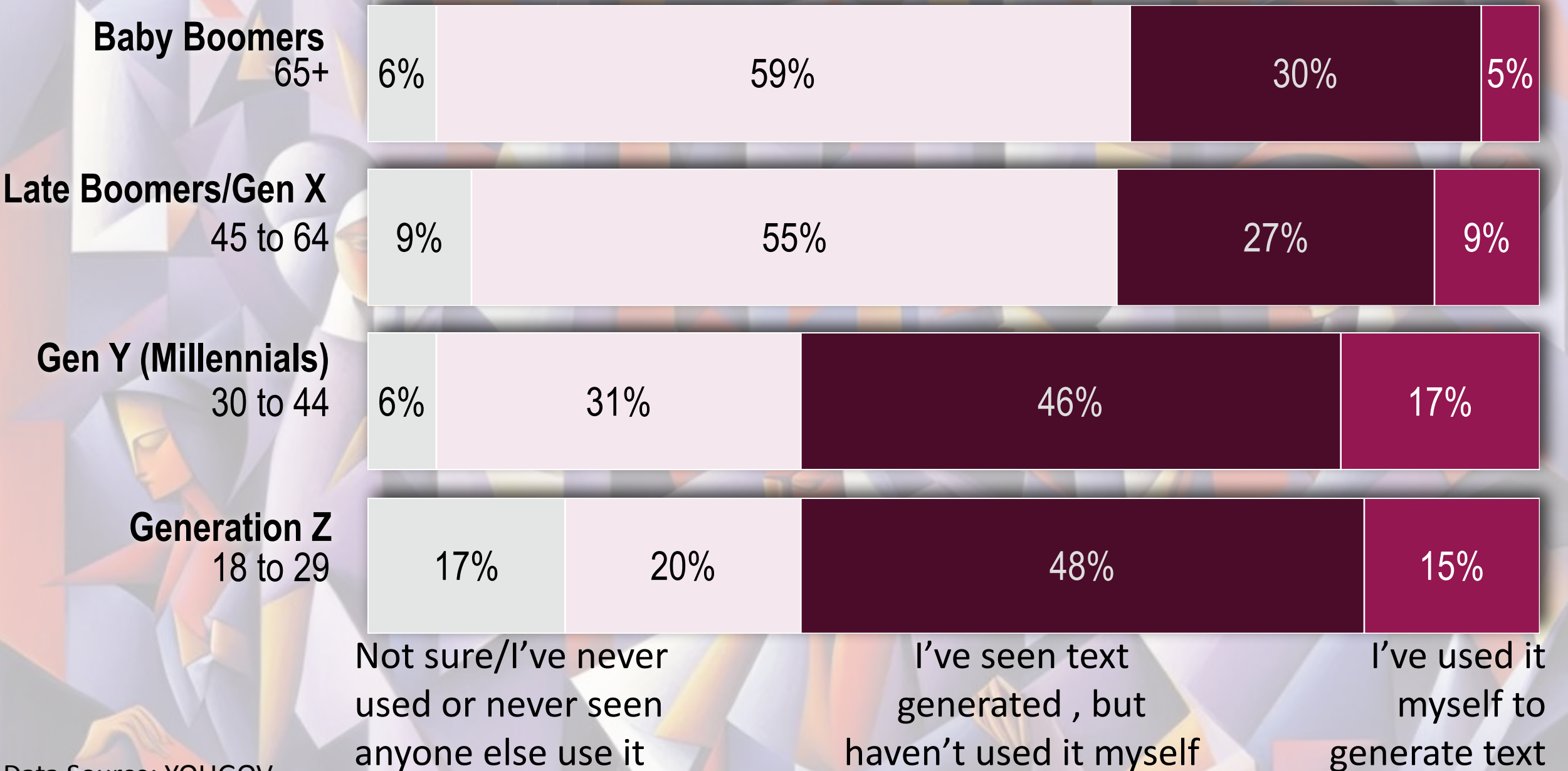


AI is not new to ATP

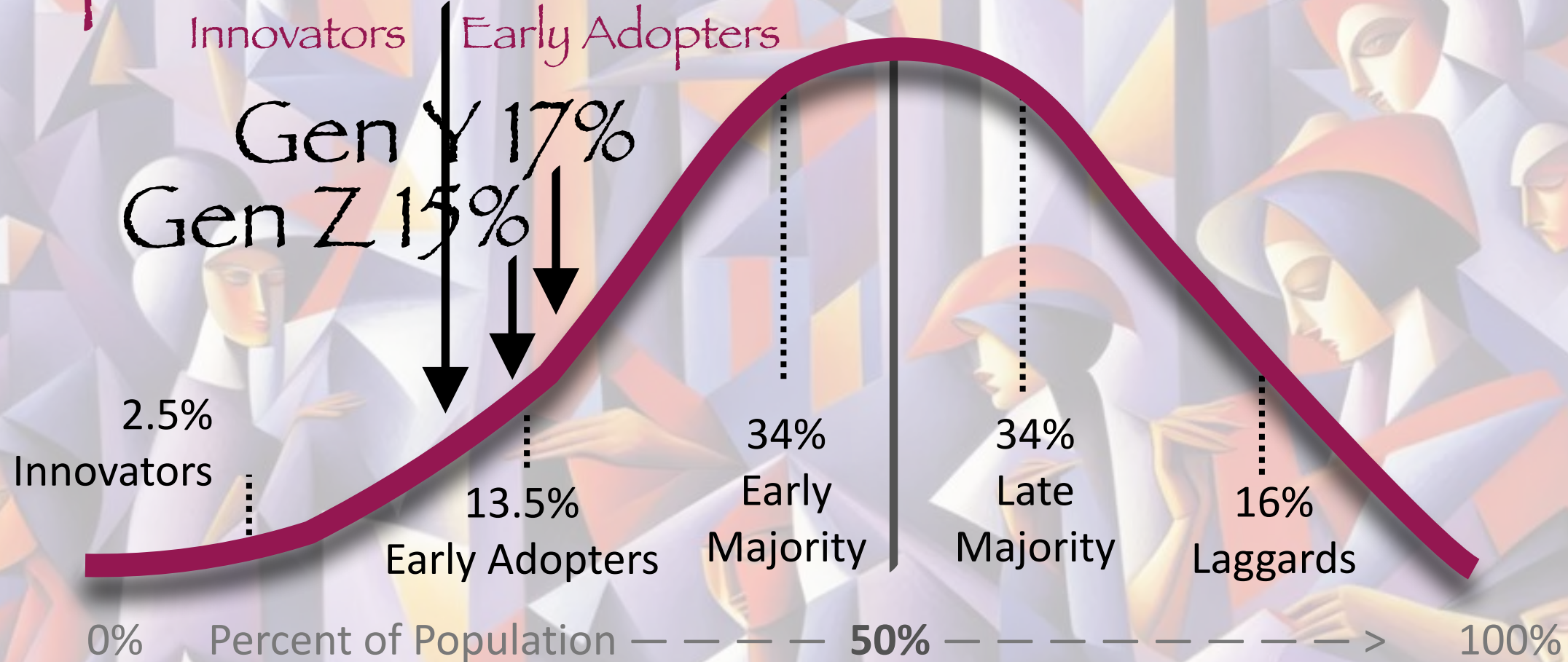
- 2000 - ATP launches its first conference: *Computer-Based Testing: Applications for the New Millennium*
- 2021 – ATP releases *Artificial Intelligence and The Testing Industry: A Primer*
- advocacy, analysis, advancement, action



Who has used chatbots/Generative AI to generate text?

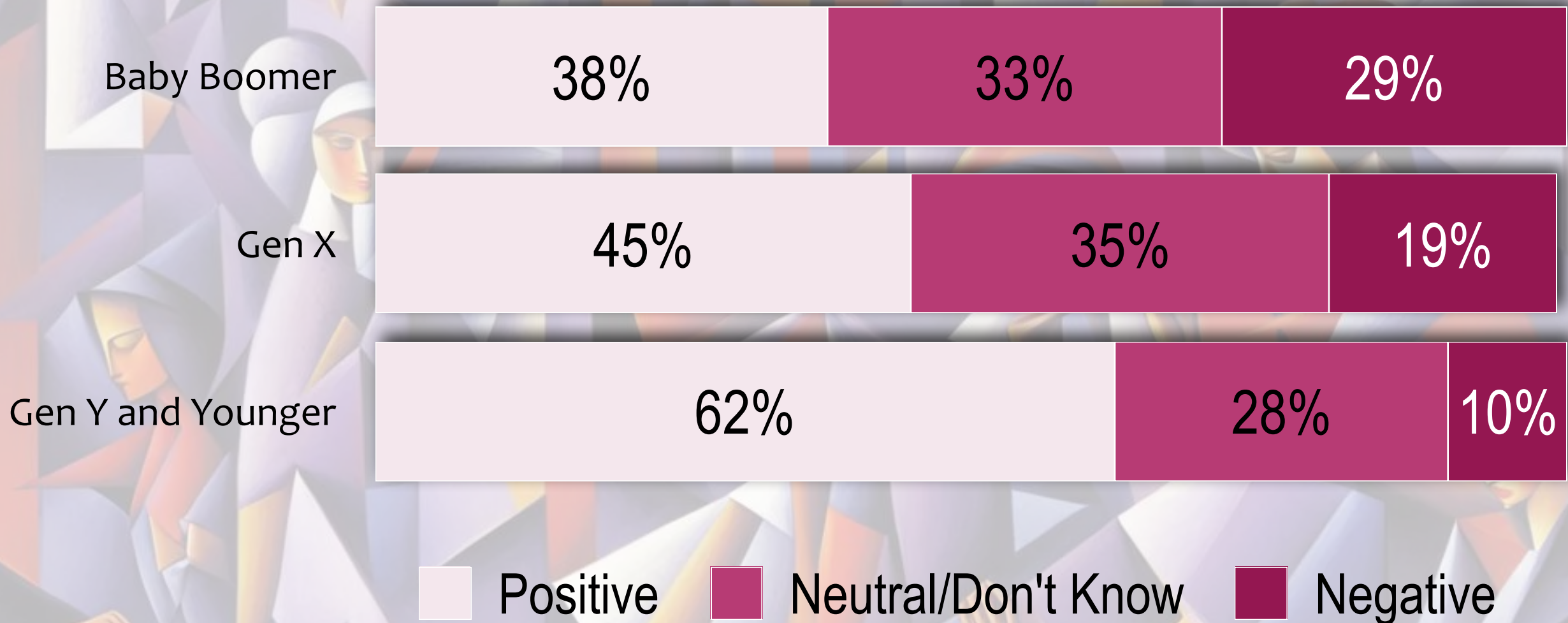


Population @10% have used ChatGPT



Rogers (1962) Diffusion of Innovations (stages)

In the coming years, do you think AI will have a positive or negative impact on your line of work?



Impact of AI: Concerns

Bias, fairness, discrimination,
ethical use

Security, privacy, legality,
regulation, copyrights

Transparency, explainability

Accessibility, comfort

Dependency, job displacement





Impact of AI: Benefits

More, better, faster, easier?

Increased efficiency, accuracy, and
idea generation

Reduced costs (long term)

Personalized assessment
(& learning)

New assessment methods & ways
of thinking about assessment

Generative AI Applications within professional assessment, licensure, and certification

Coursework

Studying

Fieldwork

Recertification

Continuing education

Exam Creation
objectives, content,
Item generation,
reliability, validity,
ongoing refinement
...

Exam Prep &
Mock Exams



Job aids,
Performance support &
Performance reviews

Maintaining registries
(reciprocity)

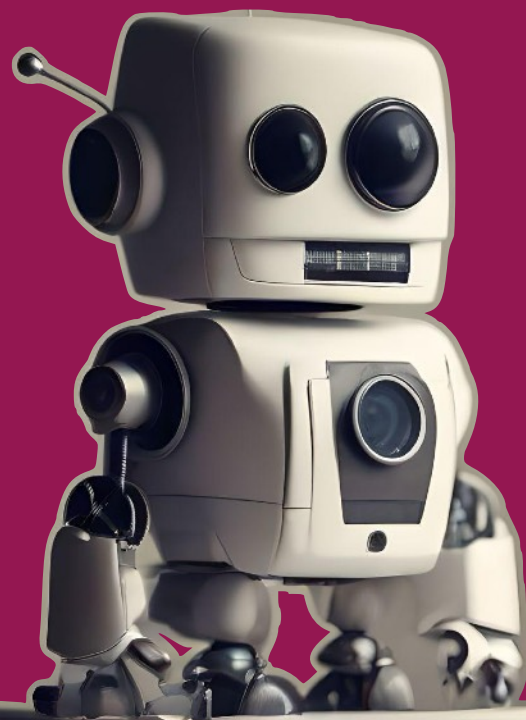
Item &
exam analysis

Exam Administration

eligibility, format, proctoring, scoring, appeals...

Data Analysis

More, better, faster, easier?

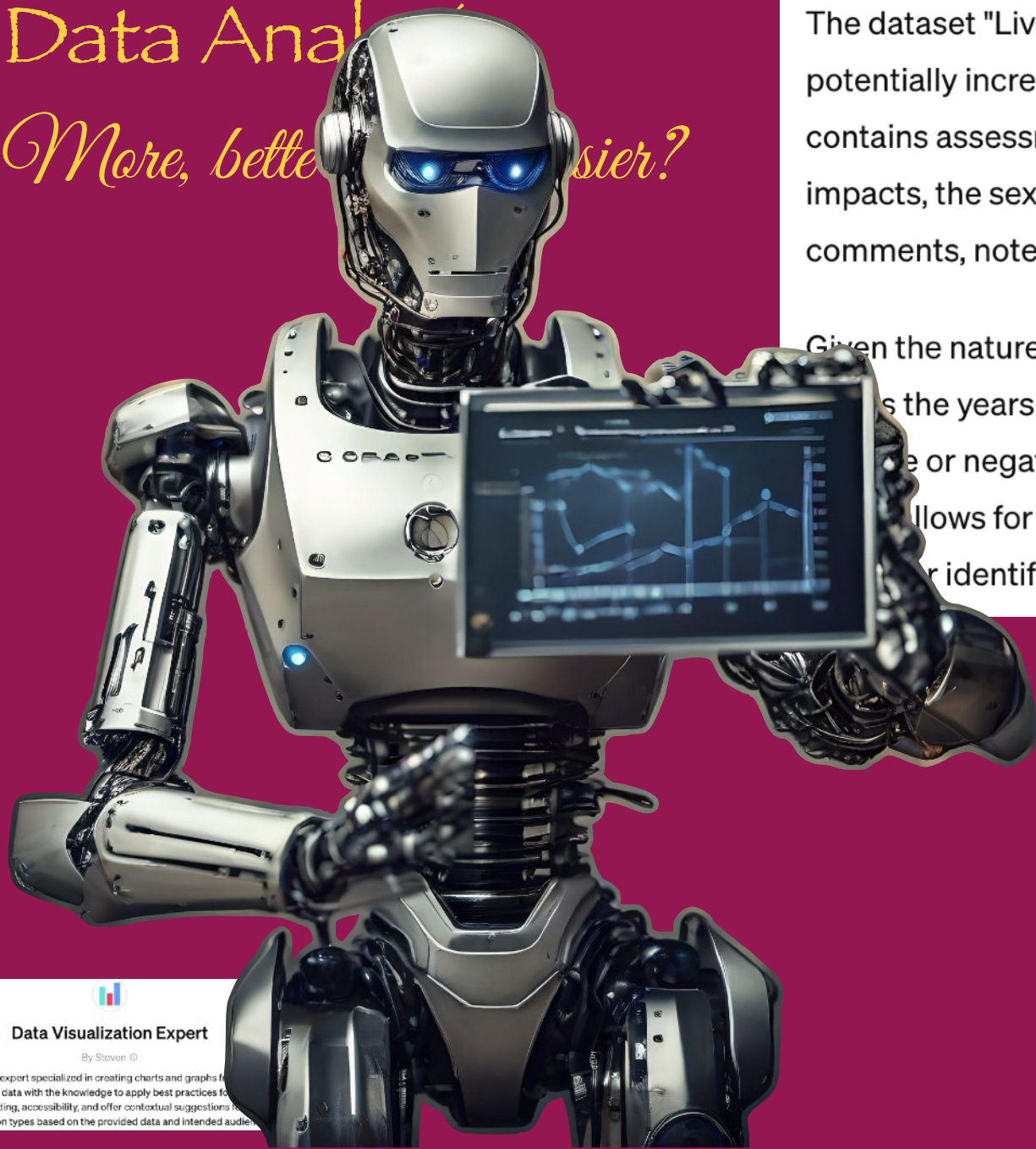


Gencay I. in
Level Up Coding

	Factor	Years gained / lost	strength of science?	strength of science as a number	sexes affected	effect	Comment	Note	ID	Sources
0	Smoking	-10.0	strong	3	Both	negative	10 years gained (against those who continue to...	14 if combined with exercise and eating health...	1	Khaw KT, et al. (2008). Combined impact of hea...
1	Sitting Down	-3.0	suggestive	1	Both	negative	for 8-11 hours sitting a day	No matter how much you exercise, sitting too m...	2	Sydney School of Public Health research paper ...
2	Too much sleeping	-1.5	suggestive	1	Both	negative	7 hours ideal, no more than 8 hrs a night	Averaging more than 8 hours sleep a night will...	3	New Scientist
3	More Optimism	2.0	suggestive	1	Female	positive	& faith in fellow humans. calc takes average %...	Women within the highest 25% of optimism score...	4	Source: Hilary A. Tindle, M.D., assistant prof...
4	More Pets	3.0	good	2	Both	positive	particularly for elderly. calc assumes that he...	(1) Cats - According to a study by the Minneso...	5	Medical News Today; National Center for Biotec...

Data Analysis

More, better, easier?



The dataset "LiveLongerData.csv" includes information on various factors that can potentially increase or decrease lifespan, represented by "Years gained / lost." It also contains assessments of the strength of the scientific evidence supporting these impacts, the sexes affected by each factor, whether the effect is positive or negative, comments, notes, sources, and links for further information.

Given the nature of the data, a suitable visualization could be a comparison chart that shows the years gained or lost due to each factor, highlighting whether the impact is positive or negative. This could be effectively visualized using a horizontal bar chart, which allows for easy comparison of the magnitude of impact across different factors and the identification of positive versus negative effects.

Gencay I. in
Level Up Coding



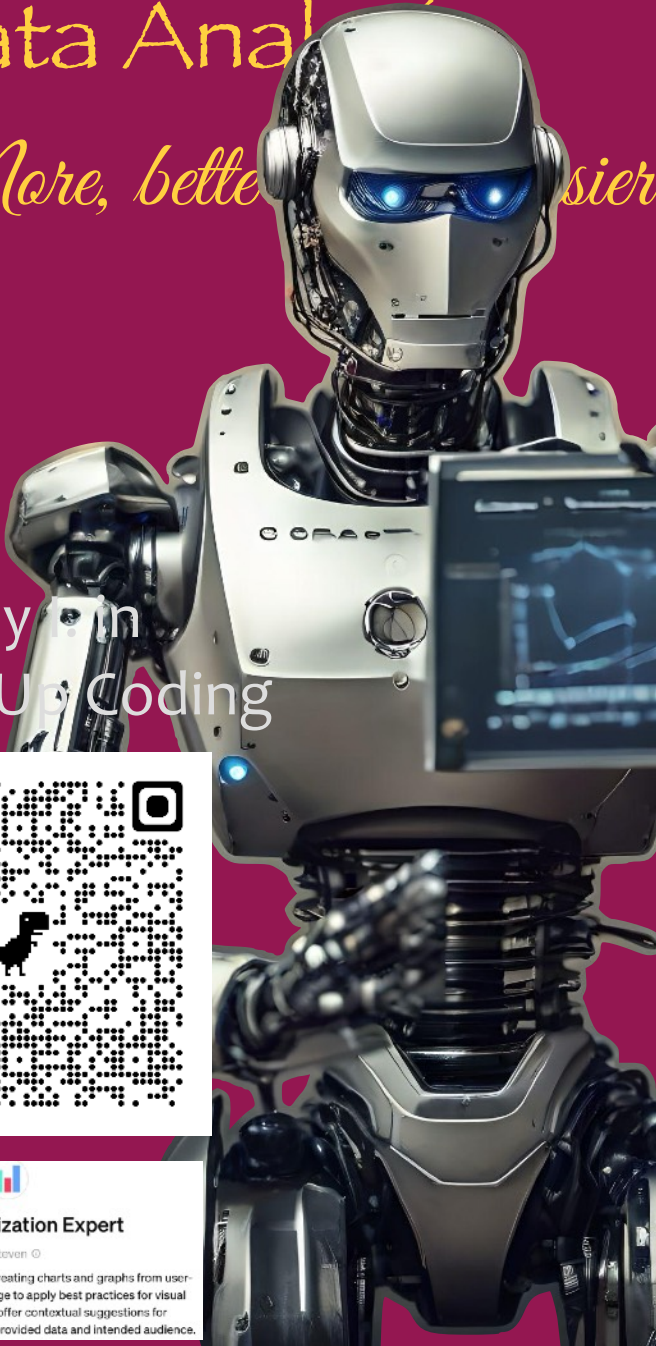
 Data Visualization Expert

By Steven ©

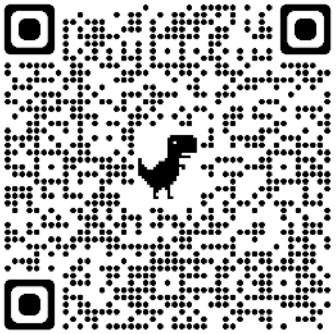
A data viz expert specialized in creating charts and graphs from provided data with the knowledge to apply best practices for encoding, accessibility, and offer contextual suggestions for visualization types based on the provided data and intended audience.

Data Analysis

More, better visualization



Gencay in
Level Up Coding



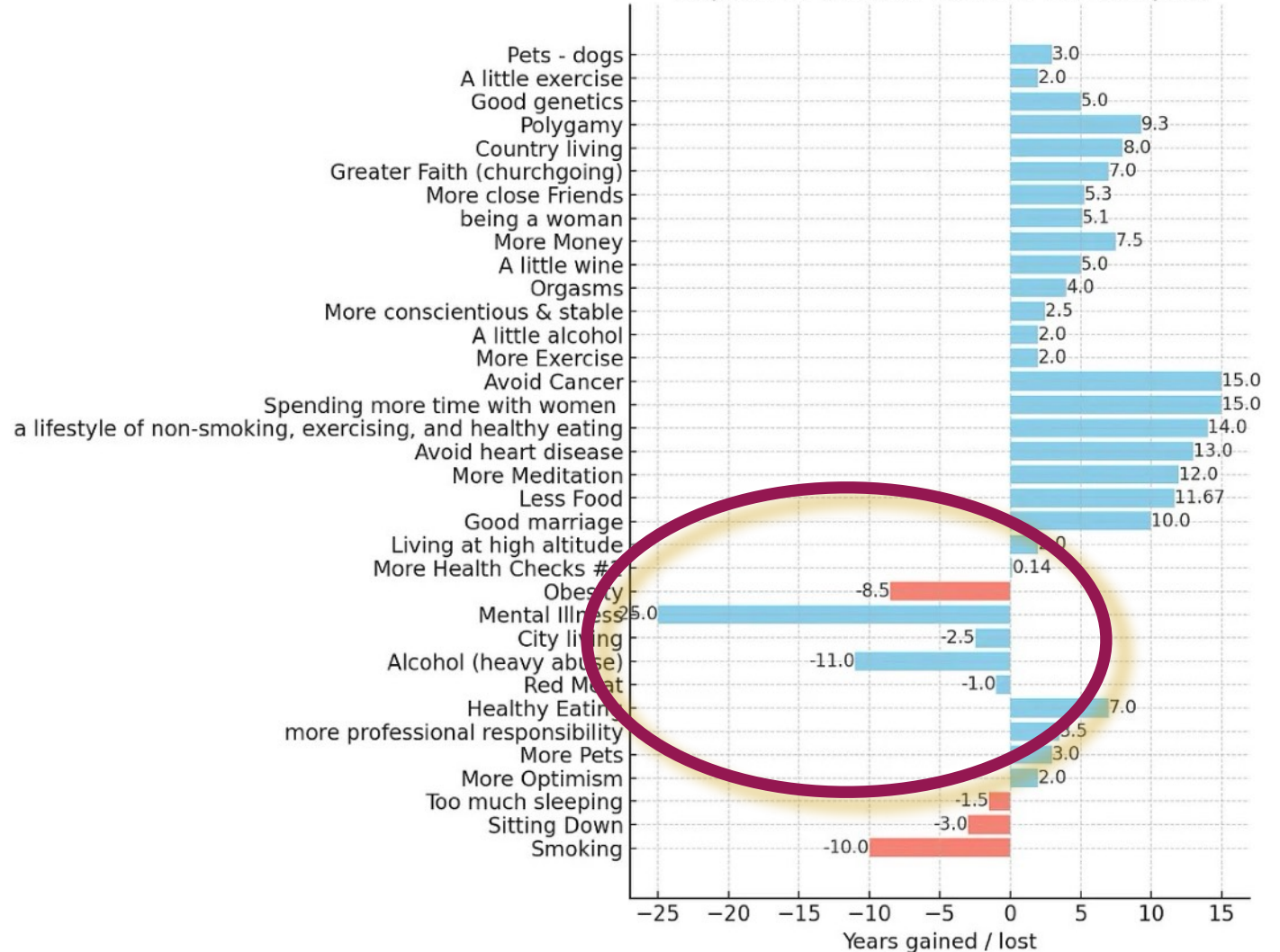
Data Visualization Expert

By Steven ©

A data viz expert specialized in creating charts and graphs from user-provided data with the knowledge to apply best practices for visual encoding, accessibility, and offer contextual suggestions for visualization types based on the provided data and intended audience.

Let's create a horizontal bar chart that shows each factor along the y-axis and the years gained or lost on the x-axis, with color coding to distinguish between positive (years gained) and negative (years lost) impacts. [↔]

Impact of Various Factors on Lifespan

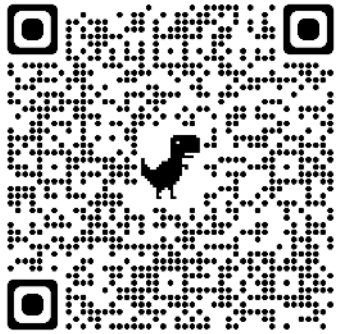


Data Analysis

More, better, science



Gencay in
Level Up Coding



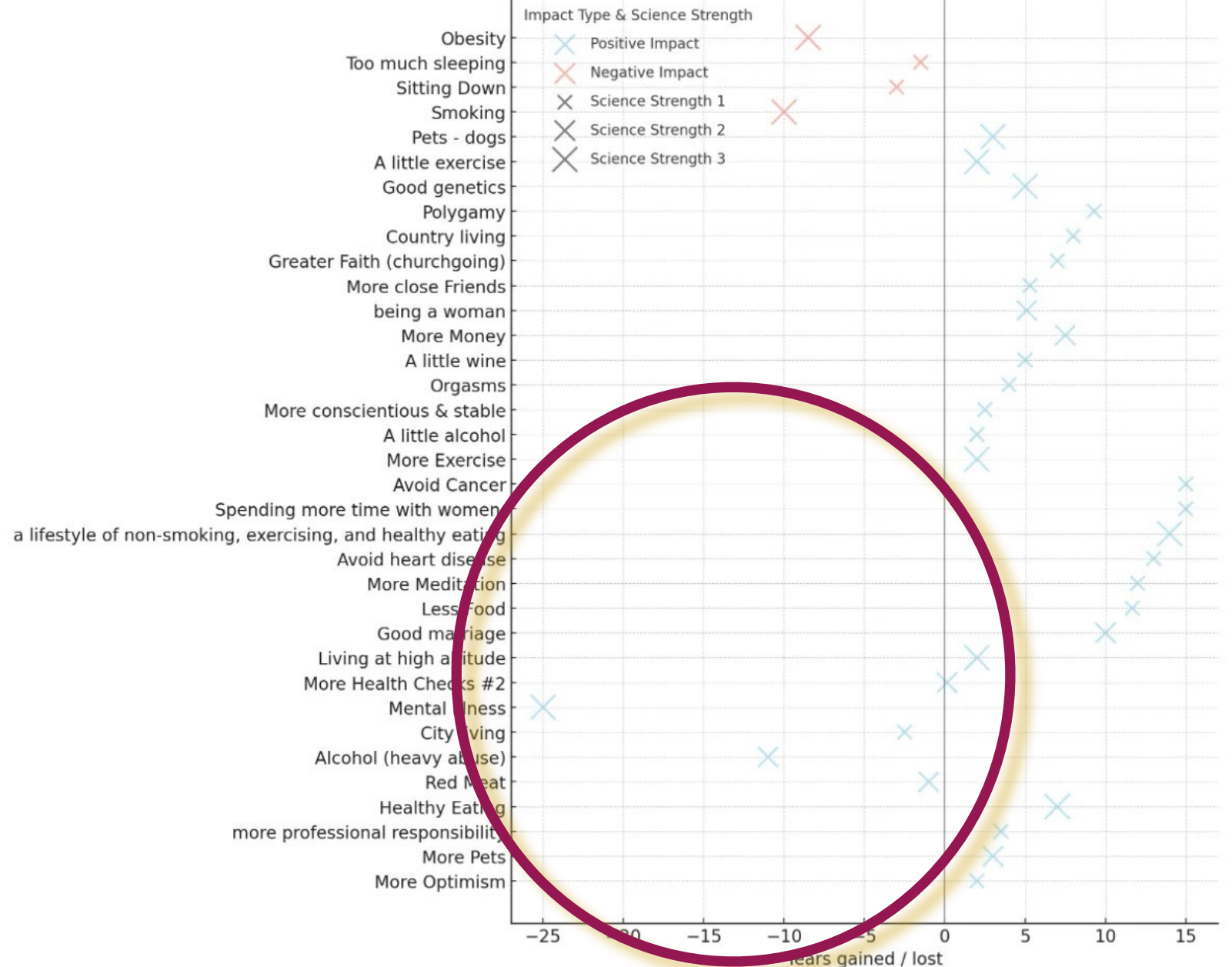
Data Visualization Expert

By Steven ©

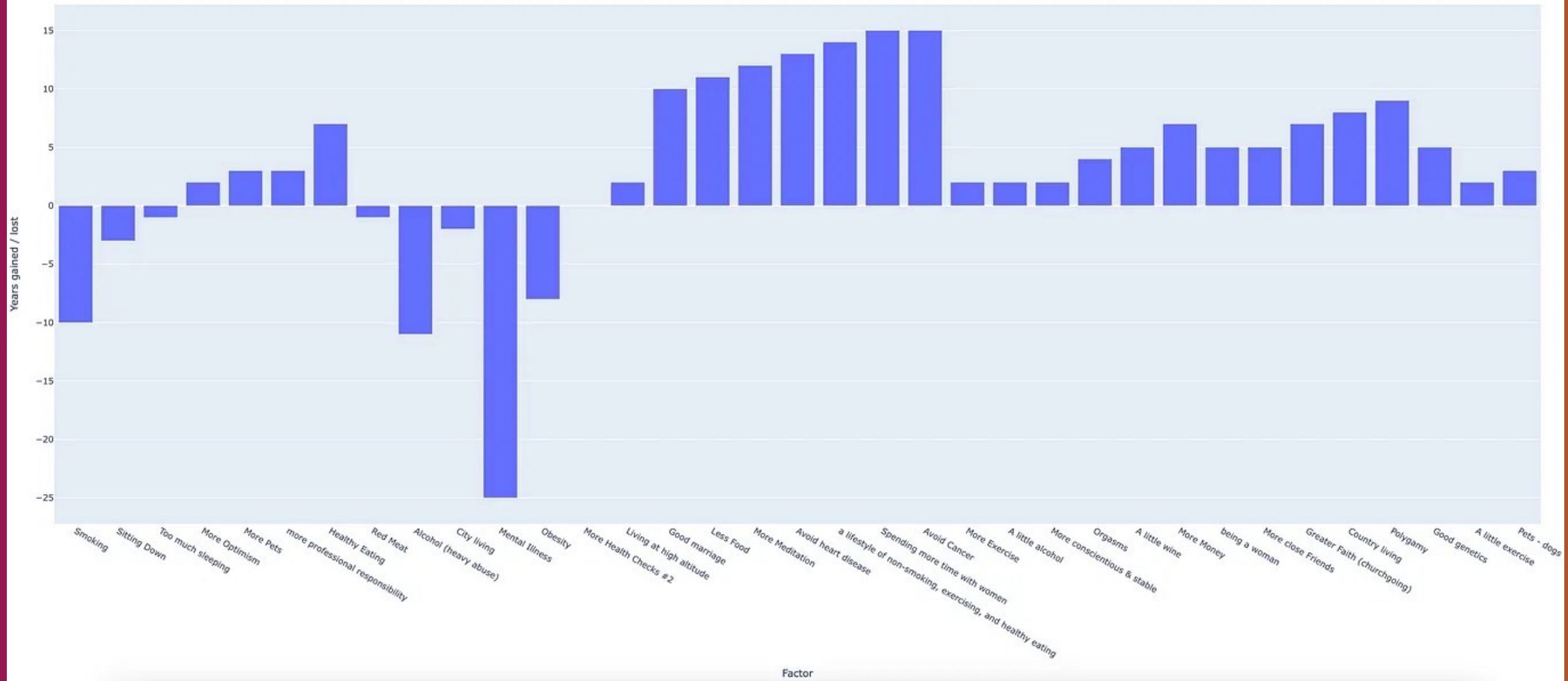
A data viz expert specialized in creating charts and graphs from user-provided data with the knowledge to apply best practices for visual encoding, accessibility, and offer contextual suggestions for visualization types based on the provided data and intended audience.

Let's proceed with this visualization.

Impact of Various Factors on Lifespan with Scientific Evidence Strength



Impact of Various Factors on Lifespan



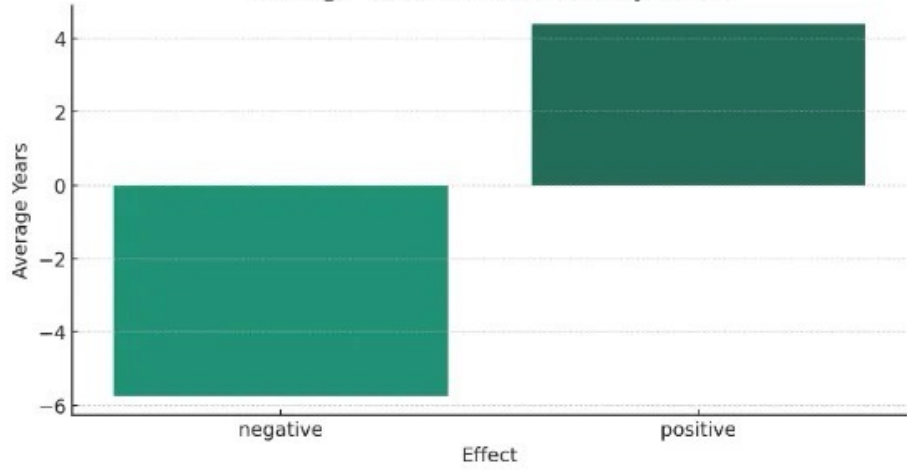
Data Visualization Expert

By Steven ©

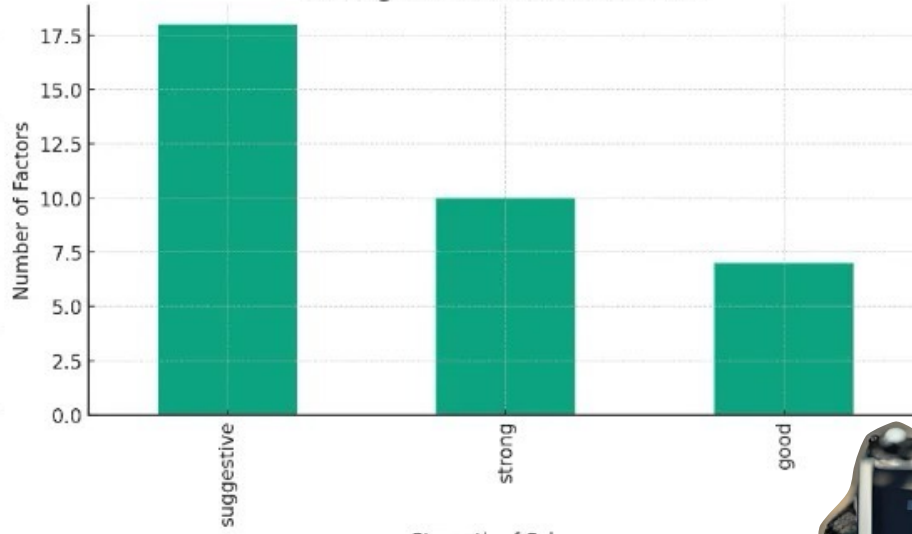
A data viz expert specialized in creating charts and graphs from user-provided data with the knowledge to apply best practices for visual encoding, accessibility, and offer contextual suggestions for visualization types based on the provided data and intended audience.

Gencay I. in
Level Up Coding

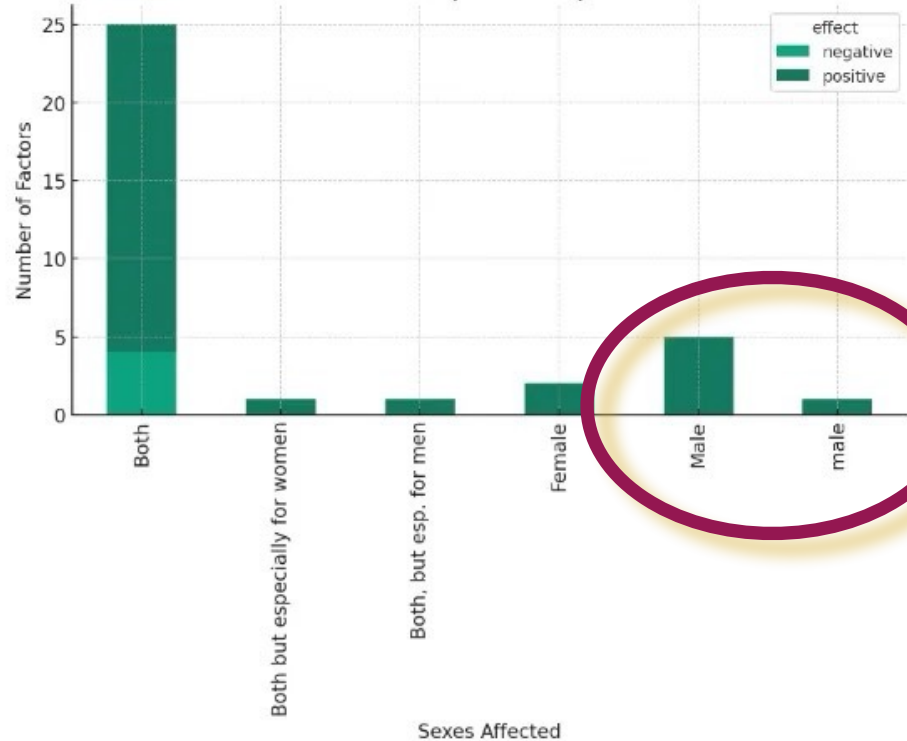
Average Years Gained / Lost by Effect



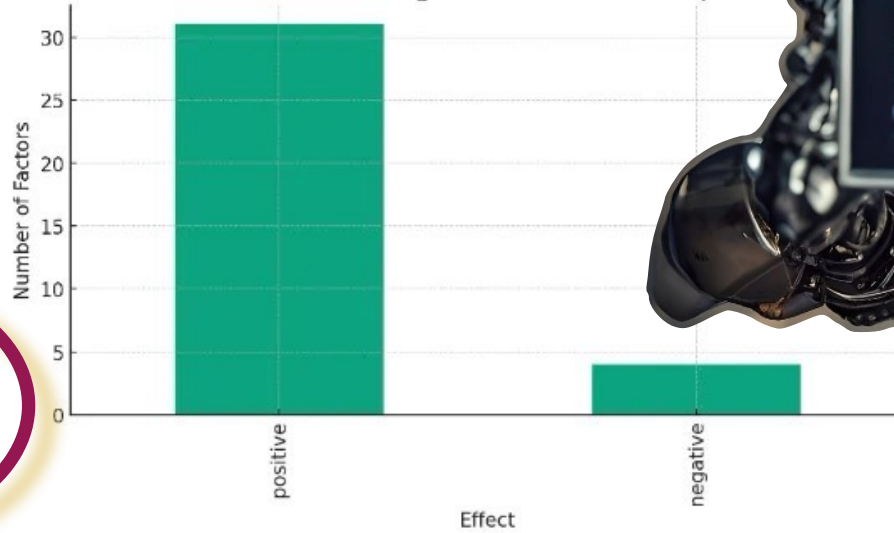
Strength of Scientific Evidence



Gender-Specific Impacts



Strength of Science Positive vs. Negative Effects on Lifespan



Gencay I. in
Level Up Coding



Interactive data visualization

By Decisiv ©

Upload a csv or xlsx document and ask a question

Data Analysis

*More, better, faster,
easier? Useful?*

Gencay I. in
Level Up Coding



Interactive data visualization

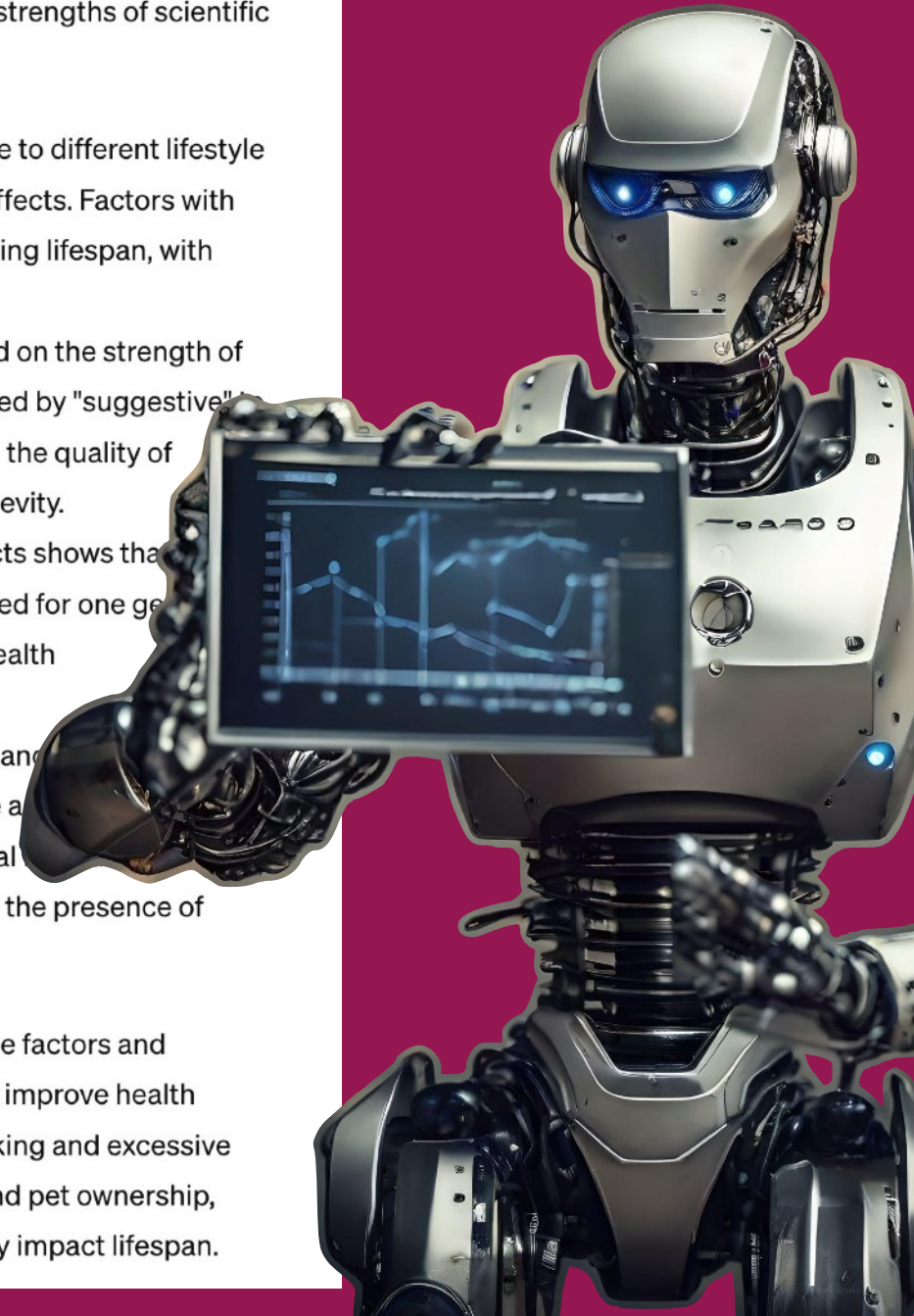
By Decisiv ©

Upload a csv or xlsx document and ask a question

The analysis of the "LiveLongerData.csv" dataset reveals several key insights about lifestyle factors and their impacts on lifespan, backed by varying strengths of scientific evidence. Here's a summary of our findings:

1. **Overall Impact on Lifespan:** The average years gained or lost due to different lifestyle factors show a clear distinction between positive and negative effects. Factors with negative effects tend to have a more significant impact on reducing lifespan, with smoking being notably detrimental.
2. **Strength of Scientific Evidence:** The distribution of factors based on the strength of scientific evidence indicates that a variety of factors are supported by "suggestive" "strong" evidence. This highlights the importance of considering the quality of evidence when evaluating the impact of lifestyle choices on longevity.
3. **Gender-Specific Impacts:** The analysis of gender-specific impacts shows that many factors affect both sexes, some factors are specifically noted for one gender. This differentiation underscores the necessity of personalized health recommendations based on gender.
4. **Positive vs. Negative Effects:** The comparison between positive and negative influences on lifespan demonstrates a balance of factors, with both positive and negative influences represented in the dataset. This suggests the potential to improve their lifespan through positive lifestyle changes, even in the presence of negative factors.

These insights underscore the complex interplay between lifestyle factors and longevity, highlighting opportunities for targeted interventions to improve health outcomes. Whether it's by reducing negative behaviors, like smoking and excessive sitting, or promoting positive ones, such as fostering optimism and pet ownership, there's substantial evidence that lifestyle choices can significantly impact lifespan.





What about: Text to ... Images?

More, better, faster, easier? Useful?

Hello, Janet

How can I help you today?



What about

Text to ... 3D images

Impact of realistic

AI generated 3D

- Text -> 3D
- Image -> 3D

in seconds

with less cost

and less effort



"motorcycle"



"mech suit"



"ghost lantern"



"furry fox head"



"dresser"



"swivel chair"



"astronaut"



"mushroom house"

Text to ... 3D images

“Human Heart”

(less than 3 seconds)

Item generation

- Label parts
- Describe feature, function

More, better, faster, easier? Useful?



What about

Text to ... VIDEO

- Prompt: “Will Smith eating spaghetti”
- Created one year ago (Mar 2023)
- Both a bit amazing AND really bad



“Will Smith eating spaghetti” (screenshot)
created by “chaindrop” using Modelscope

More, better, faster, easier? Useful?

What about

Text to ... VIDEO

- a string of text is far easier to produce than a stream of images
- early attempts created brief, low quality videos
- technology as of 3 months ago (Dec 2023)



Video Source: Pika via [inverse.com](https://www.inverse.com)

What about

Text to ... VIDEO

- a string of text is far easier to produce than a stream of images
- early attempts created brief, low quality videos
- technology as of 3 months ago (Dec 2023)
- look at all the aspects of “dogness” that the AI cannot manage



Video Source: Pika via [inverse.com](https://www.inverse.com)

Feb 15, 2024
(@ 2 weeks ago)

 OpenAI Sora

Eduardo Borges

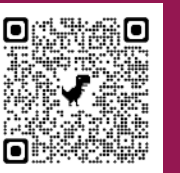
“A litter of golden retriever puppies playing in the snow. Their heads pop out of the snow, covered in.”



Impact of realistic
AI generated video

- Learning
- Assessment
- Work and Life

It is just the beginning.





ATP's mission:

*to promote and preserve the
general welfare of testing*

and its value to society,

in all its forms and uses

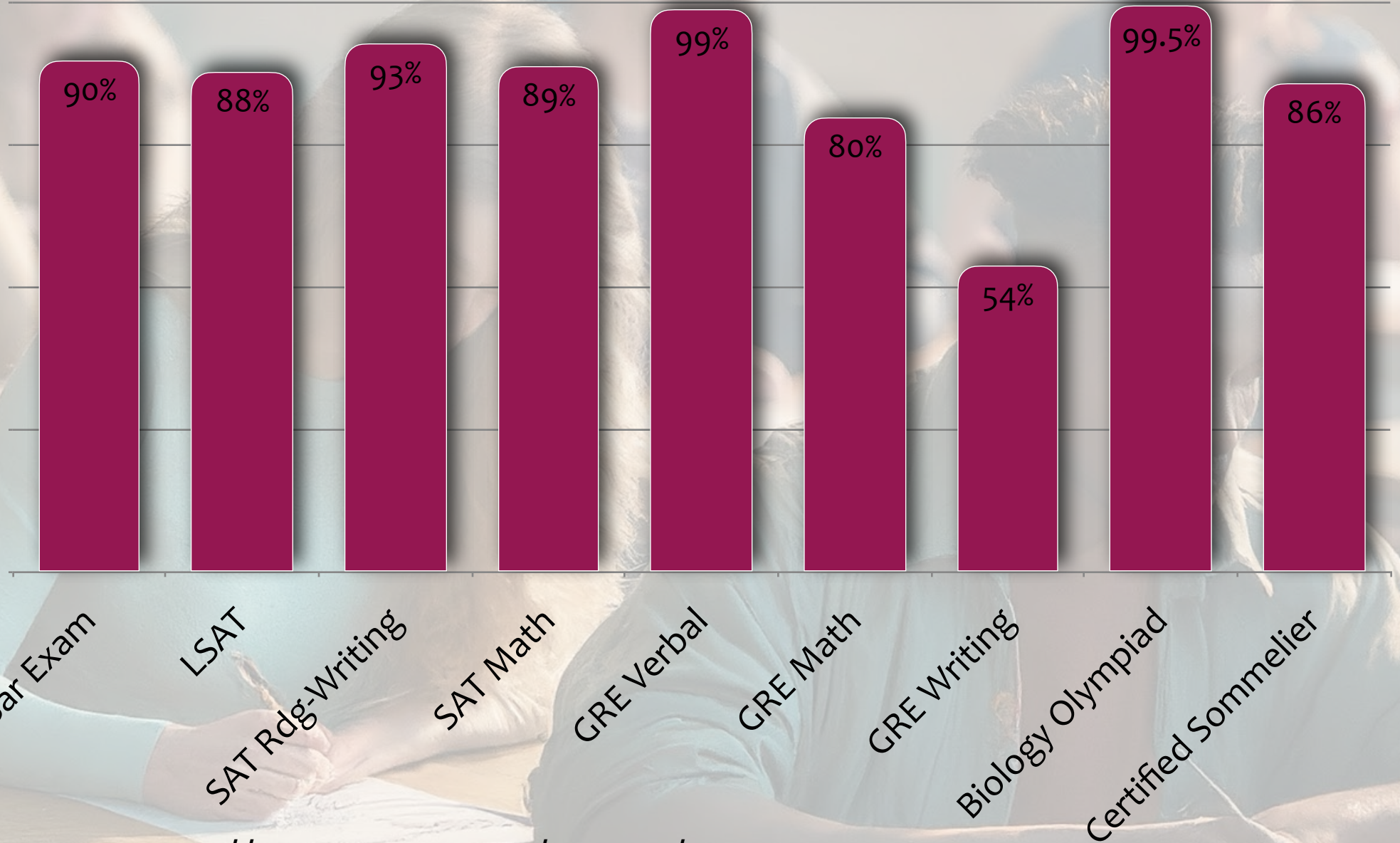
...well-developed tests are a valid
way to assess knowledge...”

What is
The Nature of
Knowledge



GPT performance on academic and professional exams.

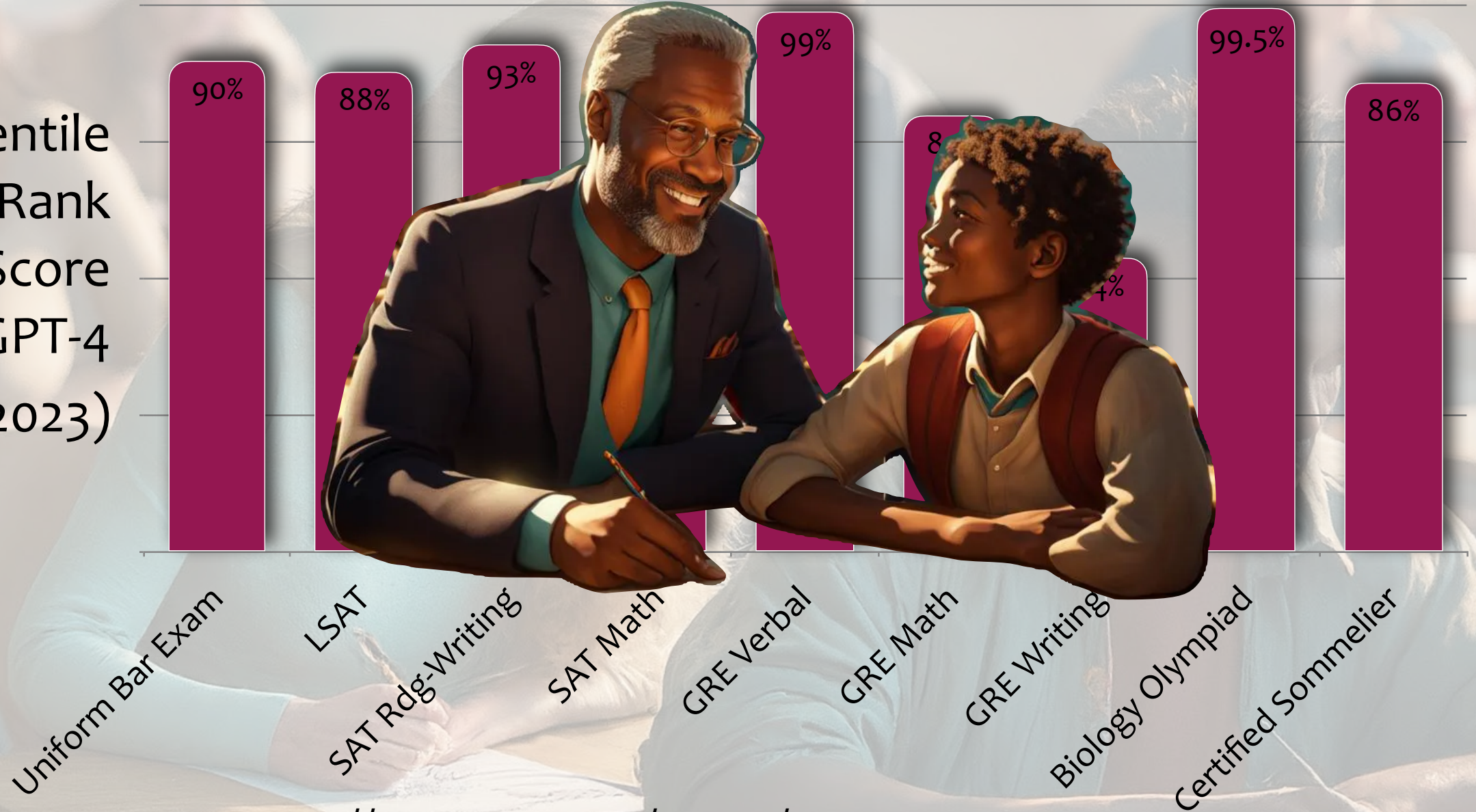
Percentile Rank Score of GPT-4 (Mar 2023)



Data Source: <https://cdn.openai.com/papers/gpt-4.pdf>

GPT performance on academic and professional exams.

Percentile Rank Score of GPT-4 (Mar 2023)



Data Source: <https://cdn.openai.com/papers/gpt-4.pdf>

Good educators increase knowledge through Intentional, Responsive Teaching...

- personalizing learning
- using evidence from interaction, observation, formative assessment
- adjusting instruction and activities in real time
- teaching in accordance with students' current level of success, confidence, and interest

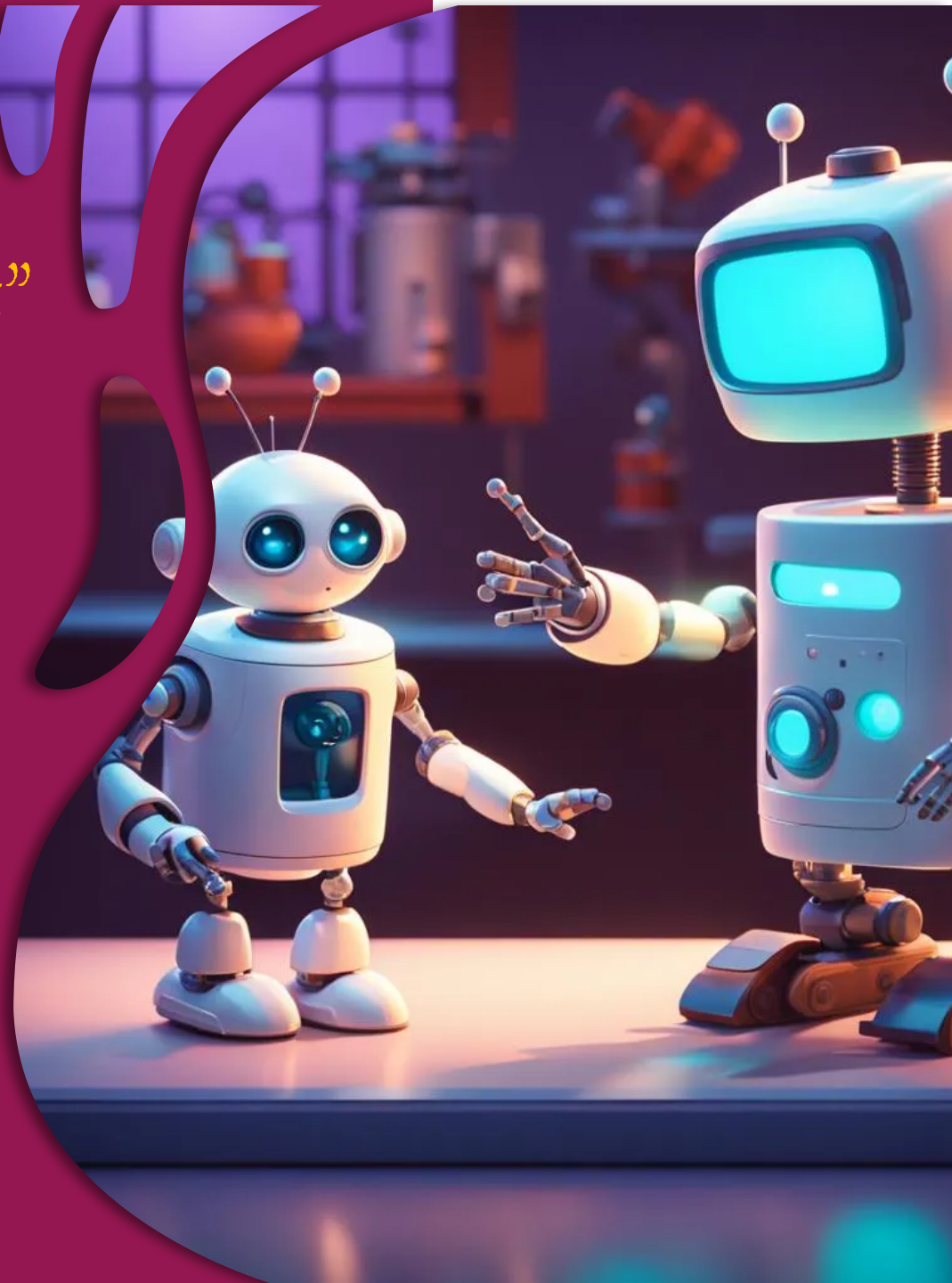


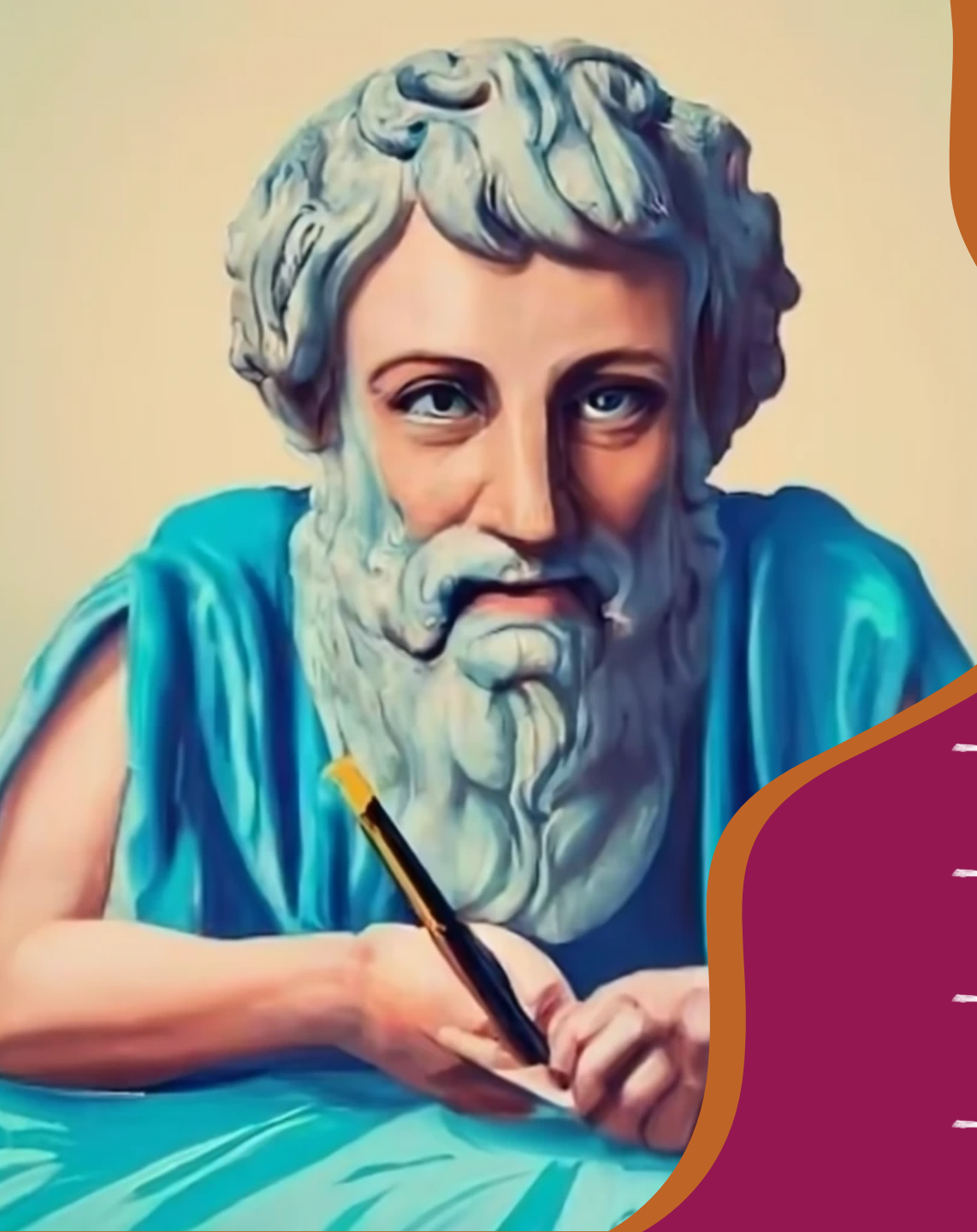
Could ATPers assess knowledge through Intentional

“Responsive Assessment”

- computer adaptive testing (CAT)
(already leverages AI and ML)
- customize the difficulty, level, and order of questions based upon student responses
- real-time determinations
- more precise information - less time

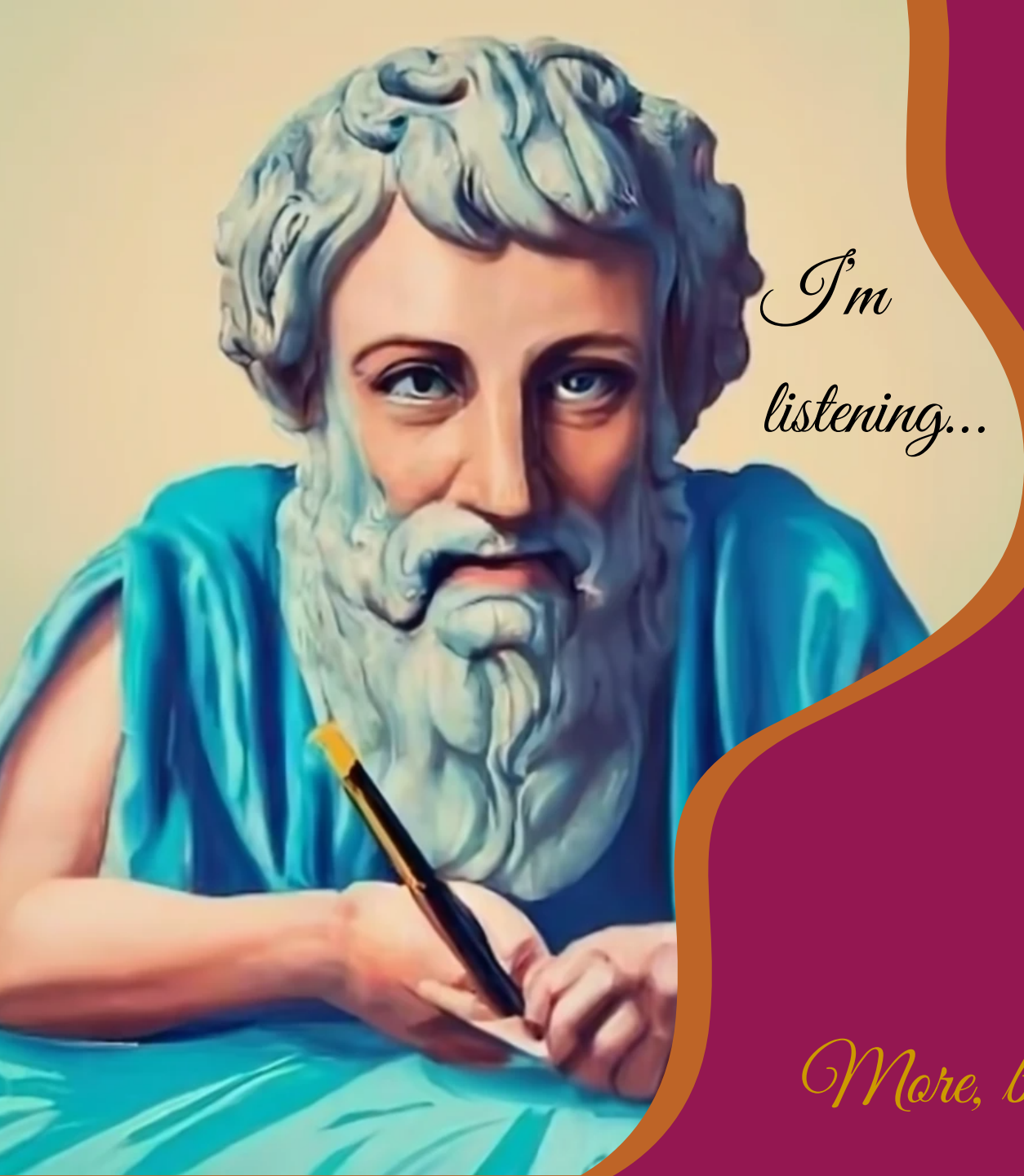
More, better, faster, easier? Useful?





Oral Competency Exams

- ✓ harder to cheat on the fly
 - ✓ assess broader skills
 - ✓ higher Bloom's levels
 - ✓ enables follow-up
-
- cost
 - appropriateness
 - bias, equity concerns
 - scalability (resource intensive)



*I'm
listening...*

Socratic Method of ...
Assessment?

Conversational AI allows
for much more.

More, better, faster, easier? Useful?



Me:

You are an expert in assessment in the field of applied behavior analysis.

I am asking you to convert these ten sample multiple choice questions (and correct response with explanation) into a conversational exam using the Socratic Method.



I will give you the 10 questions and answers then you will step into your role of assessor and begin asking me questions to gauge (rate) my knowledge of these topics. I may step out of my role as trainee at time to ask you how I am doing.

Here are the Questions and Answers:



Ex: When using behavioral momentum, behaviors selected for a high-p request sequence should:

- A. Be part of the learner's current repertoire
- B. Be new behavior(s) the learner has to acquire
- C. Have a very long duration of occurrence
- D. Be those behaviors with which the learner is not complaint

Correct Answer: A

Explanation: The learner should already be familiar with and capable of performing these behaviors. By selecting behaviors that the learner is already proficient in, it increases the likelihood of success and reinforces the learner's confidence and motivation. Introducing new behaviors or behaviors that the learner is not familiar with may lead to frustration and decrease the effectiveness of the high-p request sequence.



ChatGPT

Absolutely, let's dive into our conversational examination. To begin, let's explore the concept of research designs in applied behavior analysis.

Could you describe a research design that is particularly effective for collecting baseline data on the same behavior across multiple individuals, ideally in different settings? What makes this design suitable for establishing cause-and-effect relationships without necessarily having a control group, especially when dealing with rare behaviors or those that occur infrequently?



You

Multiple baseline across participants (subjects) would highlight that the independent variable was responsible for the change in the dependent variable because it would only change (improve) with the participant once the intervention was applied. Other participants would still be in baseline and thus not show the desired change in behavior until the intervention was applied in their setting.



ChatGPT

You've provided an excellent explanation of the multiple-baseline across individuals design, demonstrating a clear understanding of its application and utility in behavior analysis research. Well done!

Now, let's shift our focus to a concept within the realm of behavioral interventions.

Could you explain the principle behind selecting behaviors for a high-probability (high-p) request sequence in the context of behavioral momentum? Why is it important for the behaviors chosen for this sequence to be within the learner's current repertoire?

Across the Assessment landscape, AI may...

How will your organization engage?

stimulate

strengthen

surrogate

supplement

succeed

substitute

supplant

supersede

How will your organization engage? How will YOU engage?

AI enhanced
research,
development, &
testing

Educating

Researching

Learning

Building

minimal
engagement;
wait and see

Reacting

Playing

Iterating

Shaping

Preventing

Kinds of Engagement

Driving

Activism

Avoiding

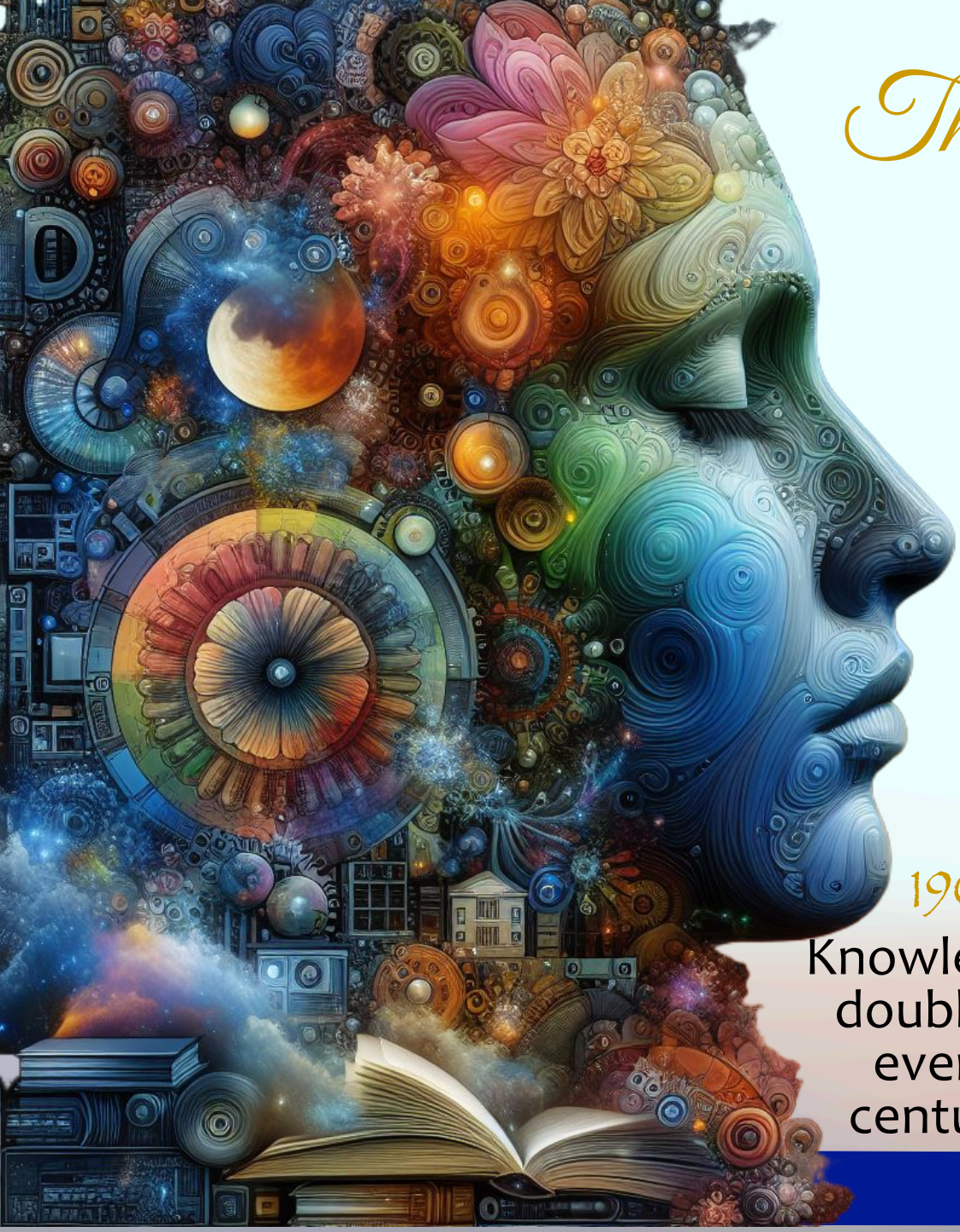
Ignoring

Advocacy

Say at the Table

- Government
- Funding
- AI Alliance

Industry Engagement with Generative AI



The Knowledge Doubling Curve

Buckmeister Fuller

As an individual...

Just try it!

2020

Knowledge doubling every 11-12 hours

1982

Knowledge doubling every 12-13 months

1945

Knowledge doubling every 25 years

1900

Knowledge doubling every century

Easy Tools to Start With



[https://
copilot.microsoft.
com](https://copilot.microsoft.com)



[https://
gemini.google.com
/app](https://gemini.google.com/app)

NightCafé
CREATOR

[https://
creator.nightcafe.
studio](https://creator.nightcafe.studio)





**Tech is neutral.
How people use it
is not.**

*As we all move forward,
think about*

Bias and the furthering
of inequities

Perpetuating
disinformation

Sleepwalking into
the future

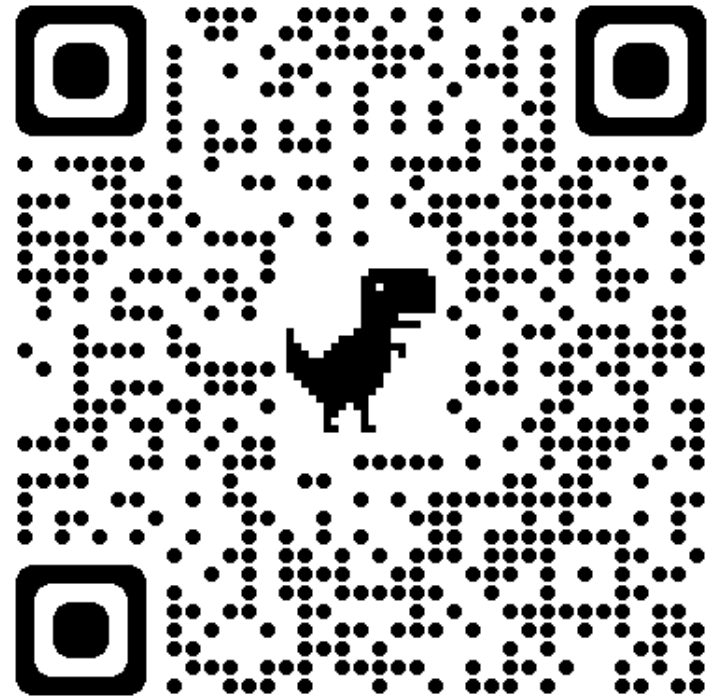


The Culture Code

Daniel Coyle

As applied to new AI technologies

1. Start with safety, but don't end there
2. Be comfortable with missteps and new starts.
3. Stick to your story.

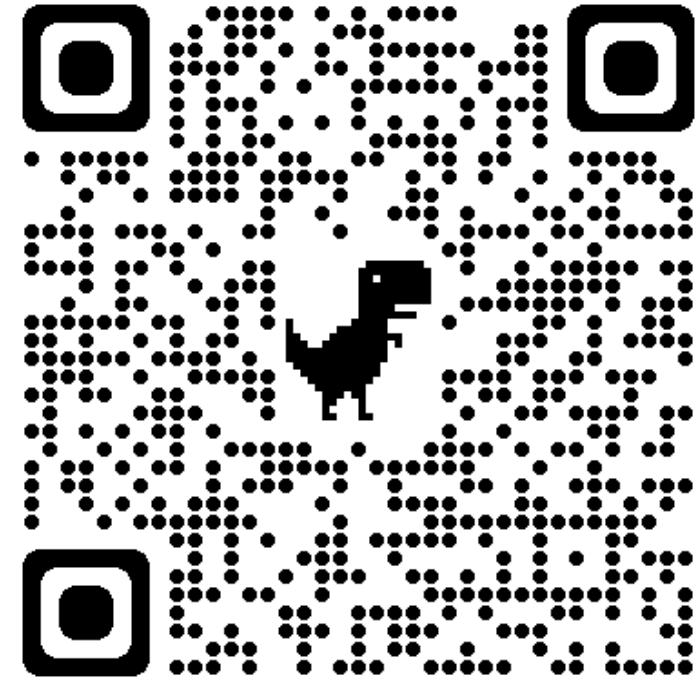


Strategies for an Accelerating Future

Four questions to ask your organization.

FEB 20 • ETHAN MOLLICK

Ethan Mollick **ONE USEFUL THING**



Decoding Generative AI in Assessment

The Evolution of Skill
Measurement

Future-Focused with Christopher Lind

Each generation holds a unique piece of the puzzle; only when we work together can we assemble the full picture.

 ATP
GLOBAL Innovations in Testing 2024

BETTER TOGETHER
Embrace change. Share solutions.





BETTER TOGETHER

Embrace change. Share solutions.



Thank you.

Opportunities and
Challenges in Navigating
AI-Integrated
Assessment

Dr. Janet S. Twyman