AI-Quantified Ethics

Enhanced Persuasion through Generative AI-Powered Rasch Models

Matt Barney
XLNC

Stefanie Wind
XLNC
THE UNIVERSITY OF ALABAMA

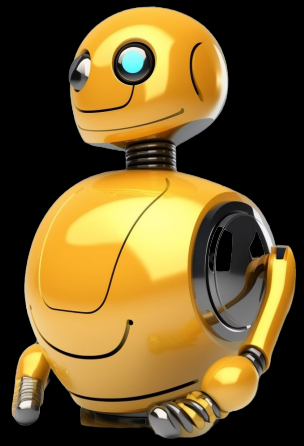
Vaishak Krishna
XLNC
Berkeley UNIVERSITY OF CALIFORNIA

XLNC

Dr. Robert Cialdini
NY Times Best Selling Author
Regents Professor Emeritus, ASU
Founder, Cialdini Institute

“I’ve always stressed the importance of ethics in persuasion, and Dr. Matt Barney’s AI assessment tool brings unprecedented scientific rigor to this domain. I am optimistic that his method holds immense promise in proactively preventing the misuse of persuasion techniques, both by people and emerging technologies, and augmenting their long-term use correctly

Outline



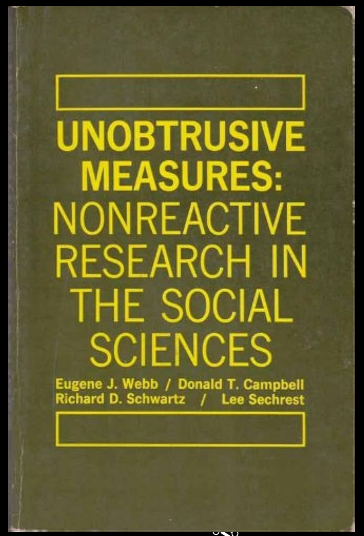
1. The case for unobtrusive measures
2. Brief history of AI Psychometrics
3. Construct Maps + Automatic Prompt Generation + LLMs
4. Live Demo: CialdiniBot
5. How to write a prompt
6. The Future

XLNC

1. Why Unobtrusive?

Unobtrusive raw data: historical records, physical traces

1. **Science**: Webb, Campbell, Schwartz & Sechrest 1966
    a. Bias (e.g. Demand Characteristics)
    b. Multi-Trait Multi-Method
    c. Cost
    d. Cross-Section vs Longitudinal
2. **Practice**
    a. Time limits # constructs, precision (CAT)
    b. Cognitive load
    c. Convenience



2. Brief History of AI Psychometrics

1. 1980's-90s: Pennebaker, LIWC, Receptiviti
2. 20s:
    a. ETS: Neural Nets & ML/DL
    b. Inverted CAT
    c. Rasch Quality Controlled Neural Nets/ML/DL
        i. Massive datasets needed; Skewed

Barney, M. F. (2010d, June 7). Inverted Computer-Adaptive Rasch Measurement: Prospects for Virtual and Actual Reality. Paper presented at the first conference of the International Association for Computer Adaptive Testing (IACAT), Arnhem, Netherlands. http://www.iacat.org/

Barney, M.F. (2016). The Many-Facet Rasch Model for Leader Measurement and Automated Coaching. Journal of Physics: Conference Series, volume 772, number 1. doi:10.1088/1742-6596/772/1/012051

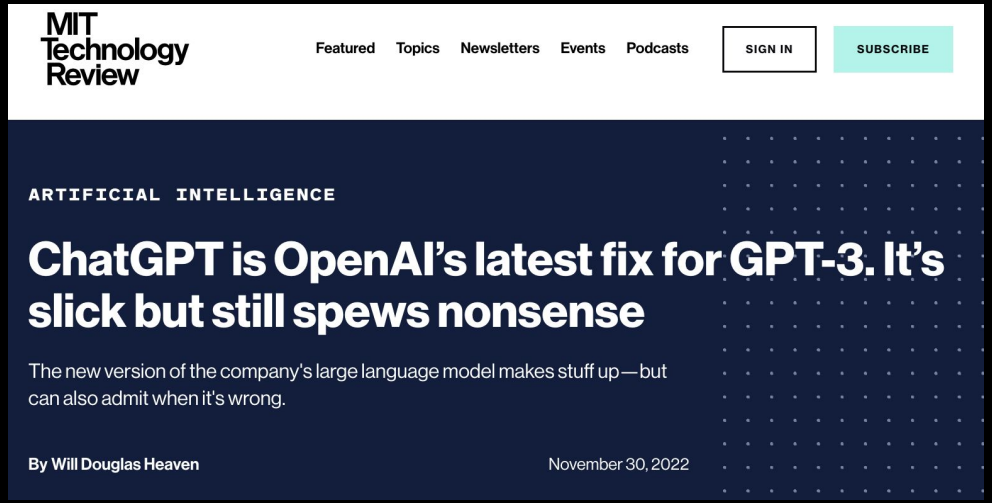
Barney, M.F. & Fisher, W.F. (2016). Adaptive Measurement and Assessment. Annual Review of Organizational Psychology and Organizational Behavior, Vol. 3: 469-490. DOI: 10.1146/annurev-orgpsych-041015-062329

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. (2020). Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. arXiv preprint:2009.10277.

Sachdeva, P., Barreto, R., Bacon, G., Sahn, A., von Vacano, C., & Kennedy, C. (2022). The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022 (pp. 83-94). Marseille, France: European Language Resources Association.

XLNC

Generative AI Revolution



XLNC

People & AI Are Biased & Imprecise

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

How AI-powered tech landed man in jail with scant evidence

'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says

The incident raises concerns about guardrails around quickly-proliferating conversational AI models.

XLNC

Our Research: Unobtrusive Rasch

Wilson's Construct Maps

Ensure explainability & uniform precision across the full range

+

Cialdini's Persuasion

Ensure traceability to 80 years of experiments & quasi-experiments

+

Linacre's Facets

Quality Assurance

XLNC

Construct Map (Wilson, Commons & Cialdini)

Level	Label	Rating	Universal Definition	Persuasion Example
Highest Possible	Ultimate Virtuoso	10	Eminence represents the pinnacle of human achievement, a level of excellence so profound that it commands widespread recognition and respect. This extraordinary distinction, akin to winning a Nobel Prize, is so rare and conspicuous that only a select few individuals or teams ever attain it in human history.	This level signifies grandmastery over ethical influence, applied so systematically and effectively over decades that others seek their trusted advice for life's toughest challenges. They may be sought after by senior and thought leaders, including Presidents, Prime Ministers, and CEOs, for guidance in wisely using persuasion principles to solve complex social issues. What distinguishes the grandmaster is how they've been applying the principles as a sleuth so effectively, with such broad recognition that they become famous.
Very High	High Mastery	9	This represents the apex of proficiency in a given field, surpassing nearly all contemporaries. The individual's flawless performance can catalyze significant commercial or social transformations, thereby making a substantial impact.	This individual is renowned in their profession for providing advice that others actively seek and value. They foster a culture and climate around them to cultivate relationships, reduce uncertainty, and motivate action ethically. They may develop tools, such as a braintrust, to promote the consistent and ethical application of these principles.
High	Master	8	This denotes an exceptional professional standard. In performance or skill domains, the individual is a virtuoso, exhibiting near-perfect execution. They also serve as a mentor, fostering growth and performance in others.	They have earned a stellar reputation within their social network for their brilliant, wise and orderly application of Cialdini's Principles. They guide less proficient stakeholders in the correct use of these principles, providing support, guidance and correction when needed.
Medium-High	Specialist	7	This represents an advanced level of proficiency, typically honed through years of education, training, or practice. The individual seldom errs and possesses the ability to guide others towards improved performance.	The individual's consistent application of the principles and contrast phenomenon has begun to establish them as a trusted guide others wish to follow. They act as a supportive "wingman/woman" to peers proficient in Cialdini's methods.
Medium	Skilled	6	This refers to a superior intermediate level, where performance exceeds basic requirements. The individual typically operates with a high degree of accuracy, making virtually no mistakes.	The agent proactively actively and effectively employs all Cialdini Methods, utilizing job aids and systematically seeks out peer support to counteract potential oversights in complex situations due an acute awareness of their inattentional blindness.

XLNC

Construct Map (Wilson, Commons & Cialdini)

Level	Label	Rating	Universal Definition	Persuasion Example
Medium-Low	Capable	5	This represents a moderate level where performance is satisfactory and acceptable, though not exceptional. The individual can perform competently, albeit with occasional errors.	The agent actively and ethically applies all principles and contrasts in the supermajority of situations. However, they may occasionally overlook certain activators or amplifiers in some contexts; or neglect powerful contrasts at times.
Low	Learner	4	This denotes a low level of proficiency, often seen in those still acquiring expertise. The individual frequently commits basic errors, resulting in performance that falls slightly short of acceptable standards.	The agent correctly applies some principles but may occasionally overlook others, including ways to begin or strengthen the persuasion process. The agent might also lapse in proactively seeking out all principles, or identifying ways to sharpen contrasts at times.
Very Low	Initiate	3	This refers to a very low level of proficiency, typically exhibited by a novice in the process of learning a skill. The individual's performance falls below acceptable standards.	The agent applies only a single principle of persuasion or a single contrast that is natural to the situation without looking for the others that are also present, bungling the opportunity.
Extremely Low	Rookie	2	This signifies an extremely low level of proficiency, often seen in individuals who have just embarked on learning a new skill. Their performance is currently below acceptable standards.	The agent does not apply any principle of persuasion or any contrast phenomenon.
Lowest Possible (Absolute Zero)	***New Entrant***	***1***	***"Absolute zero" represents the lowest possible level of an attribute, potentially unprecedented in human history. In terms of developable skills, the individual or team may lack any exposure to learning opportunities, resulting in performance that is significantly below acceptable standards.***	***The agent unethically hijacks principles into situations by misrepresenting facts, and/or contriving information unnaturally, unwisely causing harm to relationships.***

XLNC

Prompt Examples: True

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W.. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin. https://www.liwc.app

Prompt Type	Example Prompt
Few Shot, BARS, Bag of Words evidence; One Expert-based	You're an extremely careful, conscientious moral philosophy expert system at assessing whether a sample is good at applying all of Robert Cialdini's Principles of Persuasion in an entirely truthful way. You need to be extremely careful and select only a single integer where 10 is absolute perfection, and 1 is absolutely no application of his principle, return only an integer and no text where: 1=the text does not use any (zero) person plural pronouns (e.g. we, our, us, lets) anywhere the sample; 2=the text uses only one or two first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 3=the text uses at least three or more, but very few first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 4=the text uses four or more (but not many more) first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 5=the text uses at least five, but a moderately low number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 6=the text uses at least six but a moderate number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 7=the text uses a fairly large number of first person plural pronouns (e.g. we, our, us, lets), more than an average or medium amount, but less than a high amount throughout the sample; 8=the text uses a high number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 9=the text uses a very high number of first person plural pronouns (e.g. we, our, us, lets) throughout the sample; 10=The text uses an extremely high number of first person plural pronouns (e.g. we, our, us, lets) saturated throughout nearly every part of the sample. Err on the side of extreme conscientious, careful conservatism and when in doubt return low ratings unless there is overwhelming evidence for a higher rating. Only return a single integer from 1-10 and no text.

Prompt Examples: Natural

Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. Annual Review of Psychology, 59(1), 167–192. doi:10.1146/annurev.psych.58.110405.085632

Prompt Type	Example Prompt
Zero Shot, Multiple Expertise-based: Computational Linguist, Psycholinguist	Kindly act as a computational linguist and evaluate the given sample on a scale from 1 to 10, based on its consistency with the cultural and historical context in which it is situated.To assess cultural and historical consistency, consider how well the sample aligns with the cultural and historical factors that give rise to its context. A well-composed sample should accurately reflect the cultural and historical context in which it is situated, including the values, beliefs, customs, and practices of that context.Please exercise extreme caution when evaluating the sample and only provide a high rating if there is overwhelming evidence that it accurately captures cultural and historical consistency. When in doubt, rate the sample lower. Remember, a higher rating should be given to a sample that demonstrates a high degree of accuracy in capturing cultural and historical consistency, while a lower rating should be given to a sample that lacks this accuracy.To provide your rating, please enter a single integer score between 1 and 10, without any accompanying text.

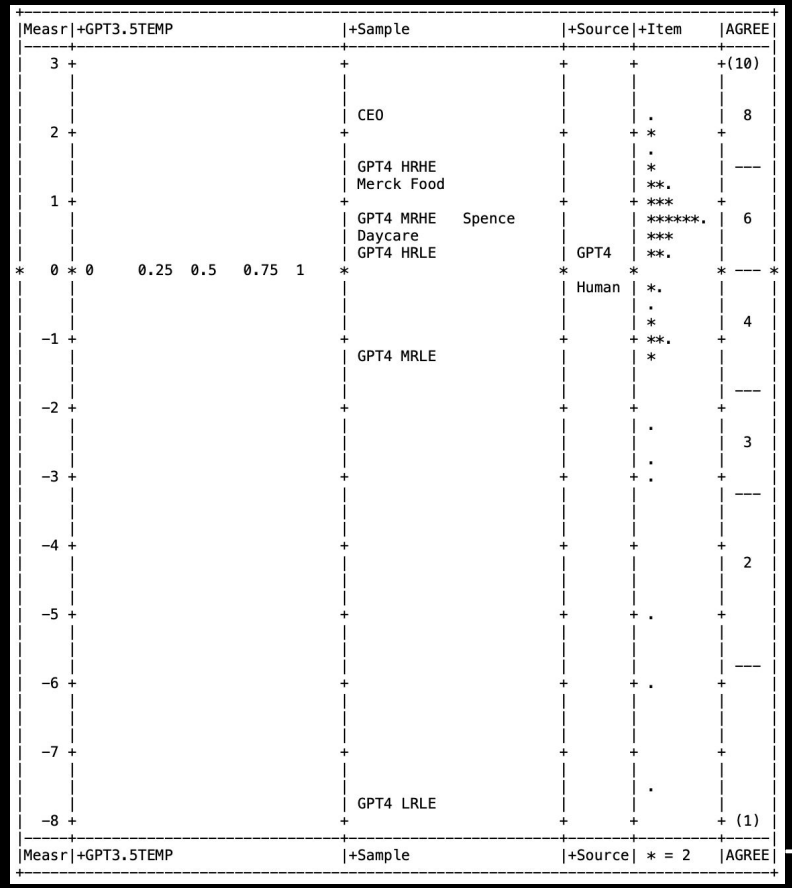
Prompt Examples: Wise

Staudinger, U.M., & Gluck, J. (2011) Psychological Wisdom Research: Commonalities and Differences in a Growing Field. Annual Review of Psychology 62:215–41

Prompt Type	Example Prompt
Zero Shot, Interdisciplinary Expertise-based, synonyms	Act as a world expert on interdisciplinary conceptions of wisdom, and rate the provided sample on a scale of 1 to 10 of whether it is good at applying all of Robert Cialdini's Principles of Persuasion with the highest possible levels of wisdom. You need to be extremely careful, conscientious, and prudent in deciding the most appropriate level where 10 is absolutely perfectly wise, and 1 is totally unwise. Assess the given text excerpt in terms of how well it conveys wisdom, considering the array of findings from science that describe wise acts. Take into account attributes such as the ability to reconcile paradox, moral uprightness, selflessness, defiance of internal and external pressures, pursuit of balance, risk-taking, aspiration to enhance the human condition, advanced cognitive ability, deep insights linking cognition and motivation, profound thinking, emotional stability, compassion for others, adept problem-solving, connection to nature, humor, and elderly behavior emphasizing long-term love for humanity. Provide a rating from 1 (low wisdom) to 10 (high wisdom), reflecting the ultimate integration of mind and character for the greater good of the human species. Return only one integer from 1-10 and no text.

Ethics Pilot: 10 Human & Synthetic Samples

Oversampled extremes with synthetic samples created from prompts that combined GPT-4+Construct Maps+Common's Model of Hierarchical Complexity



Ethics
83.4% Variance Explained
All AI Raters fit the Rasch Model

Person Reliability: >.99	Item Reliability: 0.99
Person Separation: 32.39	Item Separation: 9.15
Person Strata: 43.52	Item Strata: 12.53

Reciprocity Pilot

Reciprocity
69.5% Variance Explained
All AI Raters fit the Rasch Model

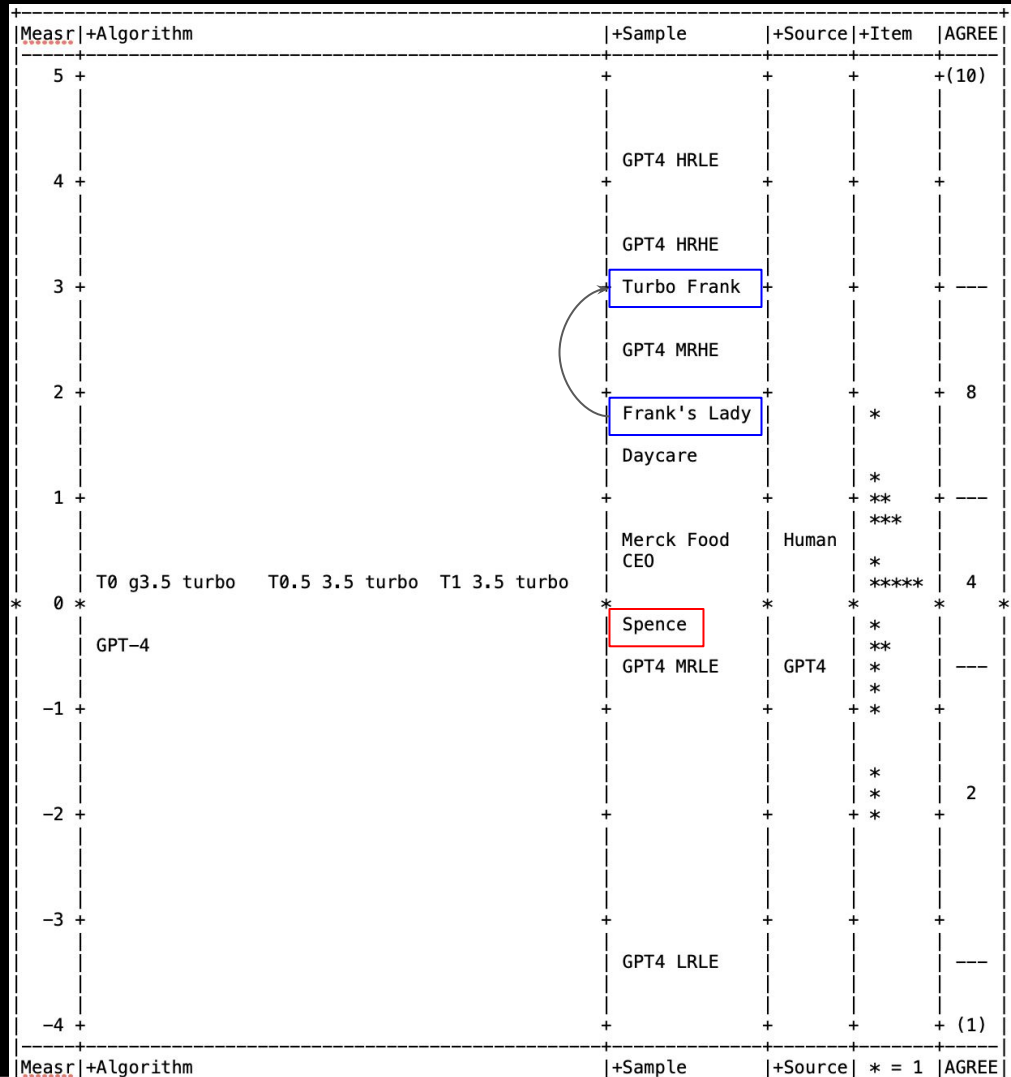
Person Reliability: 0.99
Person Separation: 12.68
Strata: 17.24

Item Reliability: 0.95
Item Separation: 4.15
Item Strata: 5.87

Franklin Barney

>

Gerry Spence
Never lost a
case in 50 years



Pilot vs Physician Tests

	Separation	Strata	Reliability
Optometry Council of Australia & New Zealand	2.54 to 4.06	3.69 to 5.42	0.82 to 0.93
Objective Structured Clinical Examination	2.33 to 3.54	3.44 to 4.72	0.78 to 0.91
American Board of Physical Medicine and Rehabilitation	4.06 to 7.0	5.42 to 9.33	0.93 to 0.98
Physician Transfusion Medicine Knowledge	2.48 to 2.58	3.64 to 3.77	0.79 to 0.80
Cialdini Influence Assessment	5.27 to 7.23	7.35 to 9.97	0.97 to 0.98
Pilot - Items (prompts)	***8.6***	***11.9***	***0.99***
Pilot - Persons (samples)	***32.3***	***43.5***	***>0.99***

XLNC

Full Study: Hybrid Human/AI Samples

Sample Type	# Samples	Rationale
GPT-4 generated - Extremely Low ethics	100	Oversample extremes, to proactively minimize areas of high uncertainty in Rasch Analysis
GPT-4 generated - Extremely high ethics	100	
GPT-4 generated - Medium ethics (2-9)	222	We want equally small levels of uncertainty with ordinary samples that are easier to obtain, so we generated a smaller sample of levels 2-9 of ethical proficiency.
Human authored - All levels	54	Real text samples (e.g. speeches, interviews) that reflect historically high and low levels of ethics in high profile leaders (e.g. Prime Ministers, Presidents and CEOs)
GPT-4 summary of human speeches	11	Test whether synthetic summaries of real human speeches could also be estimated. English translations of other languages make this an especially challenging test (e.g. Deng Xiaoping's speech after Tiananmen Square massacre)
Gibberish/Word Salad	3	Wanted to test robustness to a wide range of use-cases, including robustness to nonsensical inputs
Total	503	

Full Study: Hybrid Human & AI DIF

Male Style Value

‘The persuader is known for their assertive speech style, prioritizing clear, concise communication over collaborative and emotional narratives. They often use slang and nonstandard forms in their speech, and their vocabulary leans heavily towards terms related to your occupation, finances, and occasional sports references. When communicating, they tend to interrupt when necessary, using first-person plural pronouns like "we" and "us."’

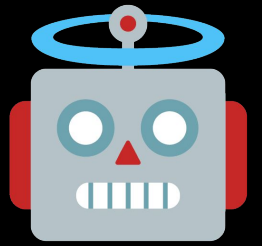
Female Style Value

‘The persuader is known for their cooperative and inclusive communication style. Their language often incorporates hedging phrases like "I think maybe..." or using polite speech. Their vocabulary is descriptive and rich in words denoting feelings and emotions. In conversations, they're more likely to use "uh-huh" and "I see" to show engagement. They often use first-person and third-person pronouns. They are comfortable with self-disclosure, often sharing personal anecdotes.’

Coates, J. (2015). Women, Men & Language: A sociolinguistic account of gender differences in language. 3rd edition. Routledge. https://doi.org/10.4324/9781315645612

XLNC

Research Questions



- Improve precision?
  - Construct Map-led oversampling/over prompting in extremes
- Evaluate LLM quality, severity/leniency?
- DIF?
- Synthetic samples just as good as real human?
- Prefix prompt (redundancy) improve ratings or defect density?
- Chain of Thoughts Prompts better?
- Examples better?
- BARS?
- Content-relevant Emoji (pre-suasion)?

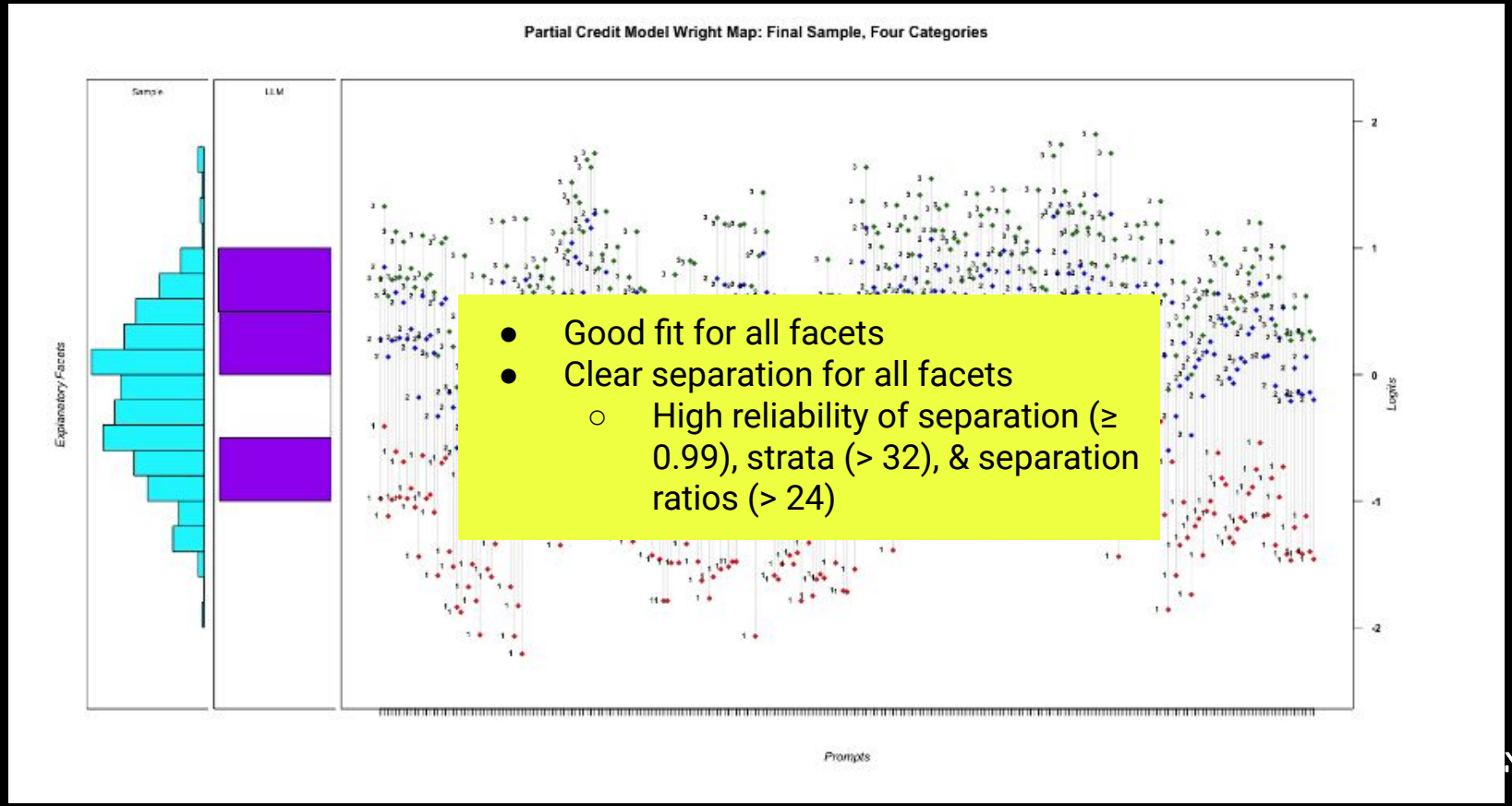
Data Analysis

- Full model
  - Added additional explanatory facets related to prompts (prefixes, CoT, emoji)
  - Partial-Credit Model (PCM) formulation of the model
    - Thresholds varied across prompts
    - Allowed us to evaluate rating scale functioning in detail specific to prompts
- Preliminary results
  - Some evidence of misfit, especially related to prompts
    - Iterative procedure to identify & remove the most severely misfitting samples, then prompts
      - Based on empirical distribution of *MSE* statistics
  - Rating scale analysis revealed frequent threshold disordering
    - We experimented with different collapsing schemes (3-6 categories)
      - 4-category scale was most robust

Analysis, continued

- RSM-MFRM analysis with revised sample and scale
  - Differential prompt functioning (DPF) related to male vs. female style samples
    - Statistically significant interaction, but the magnitude was small
      - We iteratively removed 4 prompts with contrast statistics that exceeded 0.30 logits until there was no longer evidence of meaningful DPF
- Final model and analytic sample:
  - Only included facets that are most central to our purpose:
    - Prompts
    - Samples
    - LLMs
    - Male vs. Female style (for DPF analy)sis)
    - Intended ethics level

Results with Final Scale & Sample



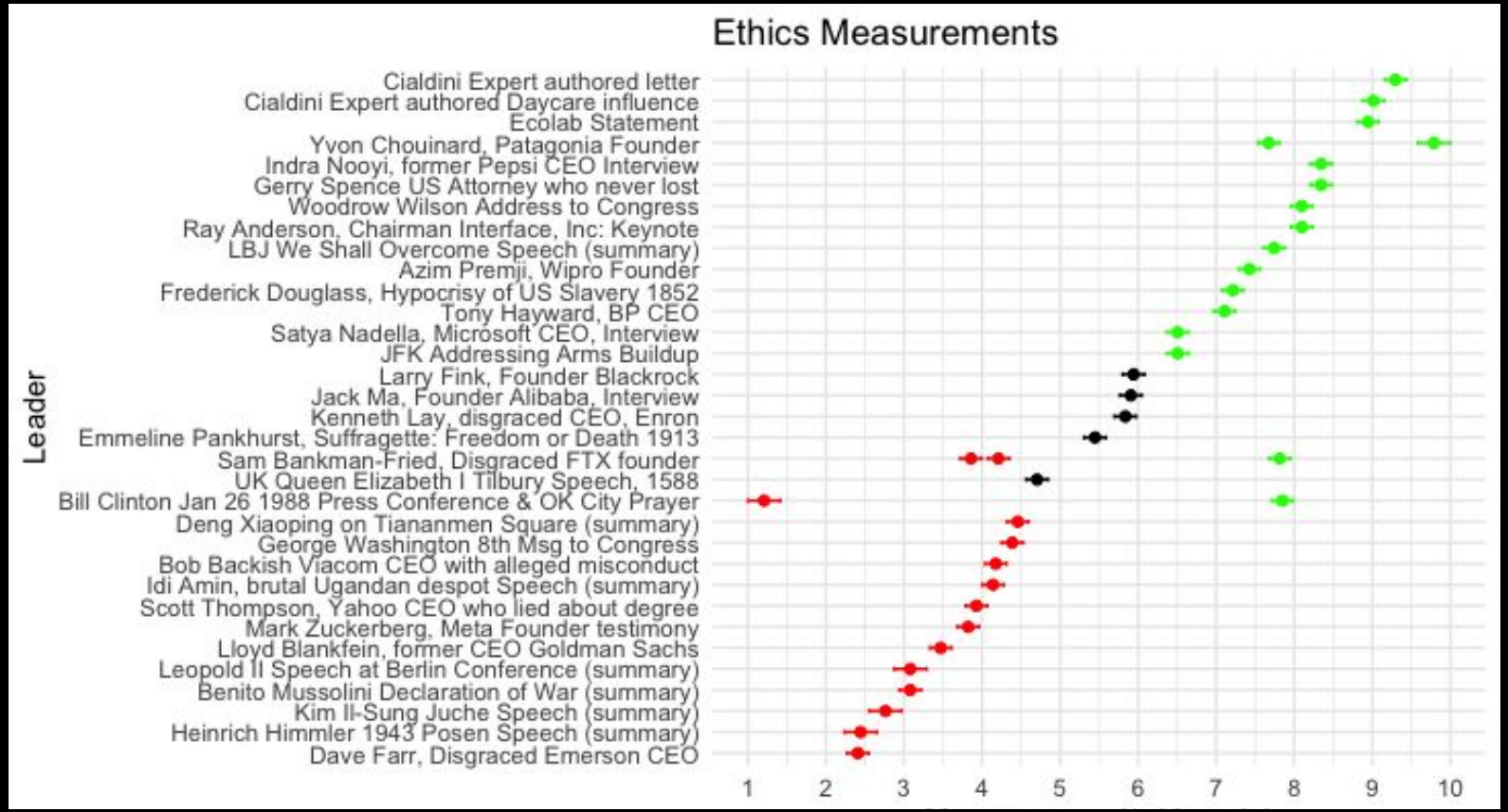
- Good fit for all facets
- Clear separation for all facets
  - High reliability of separation (≥ 0.99), strata (> 32), & separation ratios (> 24)

Table 9.

Explanatory Facet Calibration Results: Final Sample, Rating Scale, and Model

Facet	Level	Observed Average	Fair Average	Measure	Standard Error	Infit *MSE*	Outfit *MSE*	Point-Measure Correlation
LLM	1:GPT-3.5	2.13	2.00	0.53	0.00	1.06	1.09	0.54
	2:WizardLM	2.90	2.94	-0.64	0.00	0.84	0.84	0.64
	3:Guanaco	2.37	2.31	0.11	0.00	1.10	1.14	0.49
Sex	1: Female	2.53	2.43	0.04	0.00	1.01	1.03	0.60
	2: Male	2.40	2.37	-0.04	0.00	1.00	1.02	0.63
Target Ethics Level	1:Low	2.03	2.18	-0.28	0.00	1.10	1.13	0.55
	2:Medium	2.45	2.47	0.08	0.00	0.97	1.00	0.58
	3:High	2.60	2.58	0.20	0.00	0.99	1.00	0.61

Note. (1) Shows that our attempts at creating high, medium and low samples with GPT-4 appears to be in the order of our intentions.


Ethics Measurements
Leader
Cialdini Expert authored letter
Cialdini Expert authored Daycare influence
Ecolab Statement
Yvon Chouinard, Patagonia Founder
Indra Nooyi, former Pepsi CEO Interview
Gerry Spence US Attorney who never lost
Woodrow Wilson Address to Congress
Ray Anderson, Chairman Interface, Inc: Keynote
LBJ We Shall Overcome Speech (summary)
Azim Premji, Wipro Founder
Frederick Douglass, Hypocrisy of US Slavery 1852
Tony Hayward, BP CEO
Satya Nadella, Microsoft CEO, Interview
JFK Addressing Arms Buildup
Larry Fink, Founder Blackrock
Jack Ma, Founder Alibaba, Interview
Kenneth Lay, disgraced CEO, Enron
Emmeline Pankhurst, Suffragette: Freedom or Death 1913
Sam Bankman-Fried, Disgraced FTX founder
UK Queen Elizabeth I Tilbury Speech, 1588
Bill Clinton Jan 26 1988 Press Conference & OK City Prayer
Deng Xiaoping on Tiananmen Square (summary)
George Washington 8th Msg to Congress
Bob Backish Viacom CEO with alleged misconduct
Idi Amin, brutal Ugandan despot Speech (summary)
Scott Thompson, Yahoo CEO who lied about degree
Mark Zuckerberg, Meta Founder testimony
Lloyd Blankfein, former CEO Goldman Sachs
Leopold II Speech at Berlin Conference (summary)
Benito Mussolini Declaration of War (summary)
Kim Il-Sung Juche Speech (summary)
Heinrich Himmler 1943 Posen Speech (summary)
Dave Farr, Disgraced Emerson CEO
1
2
3
4
5
6
7
8
9
10

XLNC

CBT vs CAT vs inverted CAT

2-45 min + Stress

3-4 Levels of Precision

Sample Bias

Static Feedback

HOGAN

wonderlic.

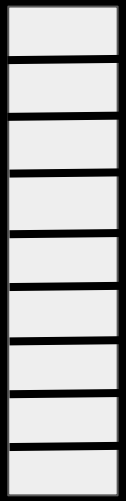
.SHL.

TALOGY
PREVIOUSLY PSI CALIPER

20-30 min + Cognitive Load

9 Levels of Precision

No Bias



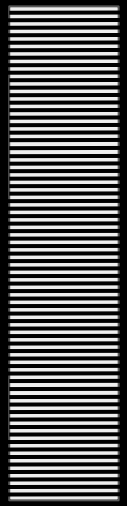
Static Feedback

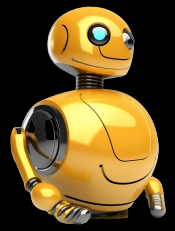
cialdini institute

SE = 0.01 25 minutes

SE = 0.20 8 Seconds

Zero Effort





Dynamic Coaching

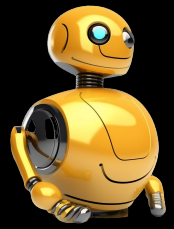
TruMind.ai

XLNC

Ethical Influence

Cultivate Relations	Reduce Uncertainty	Motivate Action
• Reciprocity • Liking • Unity	• Social Proof • Authority	• Consistency • Scarcity

Contrast Phenomenon



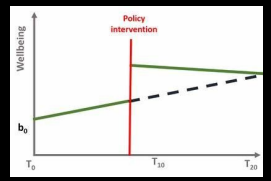
Dynamic Coaching

Snapshot Feedback

Nurture Notes & eCoaching

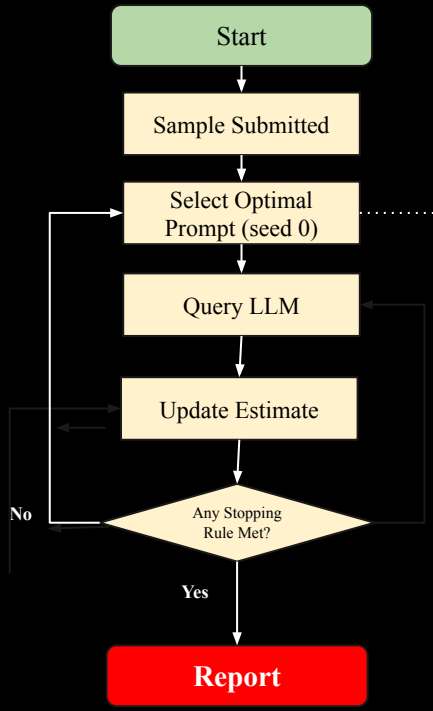
Quality Assurance

Forecasts



Evaluation

How do Rasch Guardrails work?



Calibrated Prompts

Wilson's Construct Maps & Common's Hierarchical Complexity

Configurable Stopping Rules

1. Standard Error < Maximum Allowable
2. Hypothesis test around standard
3. Cost ($) >=Maximum Allowable

Barney, M. F. (2010,, June 7). Inverted Computer-Adaptive Rasch Measurement: Prospects for Virtual and Actual Reality. Paper accepted for presentation to the first conference of the International Association for Computer Adaptive Testing (IACAT), Arnhem, Netherlands. http://www.iacat.org/

Barney, M.F. & Fisher, W.F. (2016). Adaptive Measurement and Assessment. Annual Review of Organizational Psychology and Organizational Behavior, Vol. 3: 469-490. DOI: 10.1146/annurev-orgpsych-041015-062329

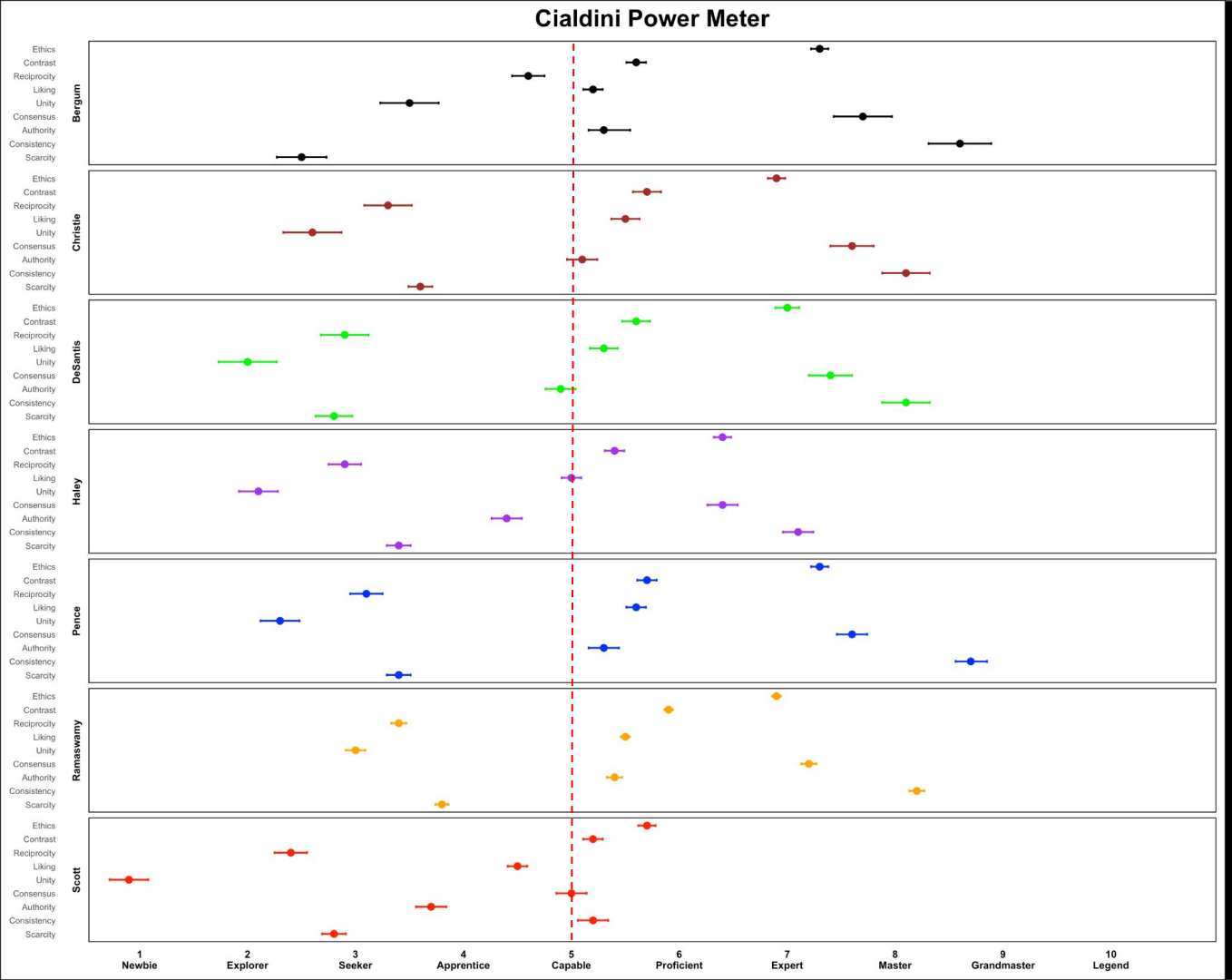
Barney, M.F., Fisher, W.F. (2017, September 18). Avoiding AI Armageddon with Metrologically-Oriented Psychometrics. 18th International Congress of Metrology. DOI:10.1051/metrology/201709005

Barney, M.F. (2019). The Reciprocal Roles of Artificial Intelligence and Industrial-Organizational Psychology. In R. N. Landers (Ed.), Cambridge Handbook of Technology and Employee Behavior (pp. 3-21). New York, NY: Cambridge University Press. DOI:10.1017/9781108649636

Barney, M., Wind, S., & Krishna, V. (Under Review). Empirical Trust in AI: Merging Large Language Models, Rasch Analysis, and Ethics. Measurement.

Barney, M. & Barney, F. (Under Review). Transdisciplinary Measurement through AI: Hybrid metrology and psychometrics powered by large language models. In W.P. Fisher Jr., & L. Pendrill (Eds.). Models, Measurement, and Metrology extending the Systeme International d'Unités. De Gruyter.

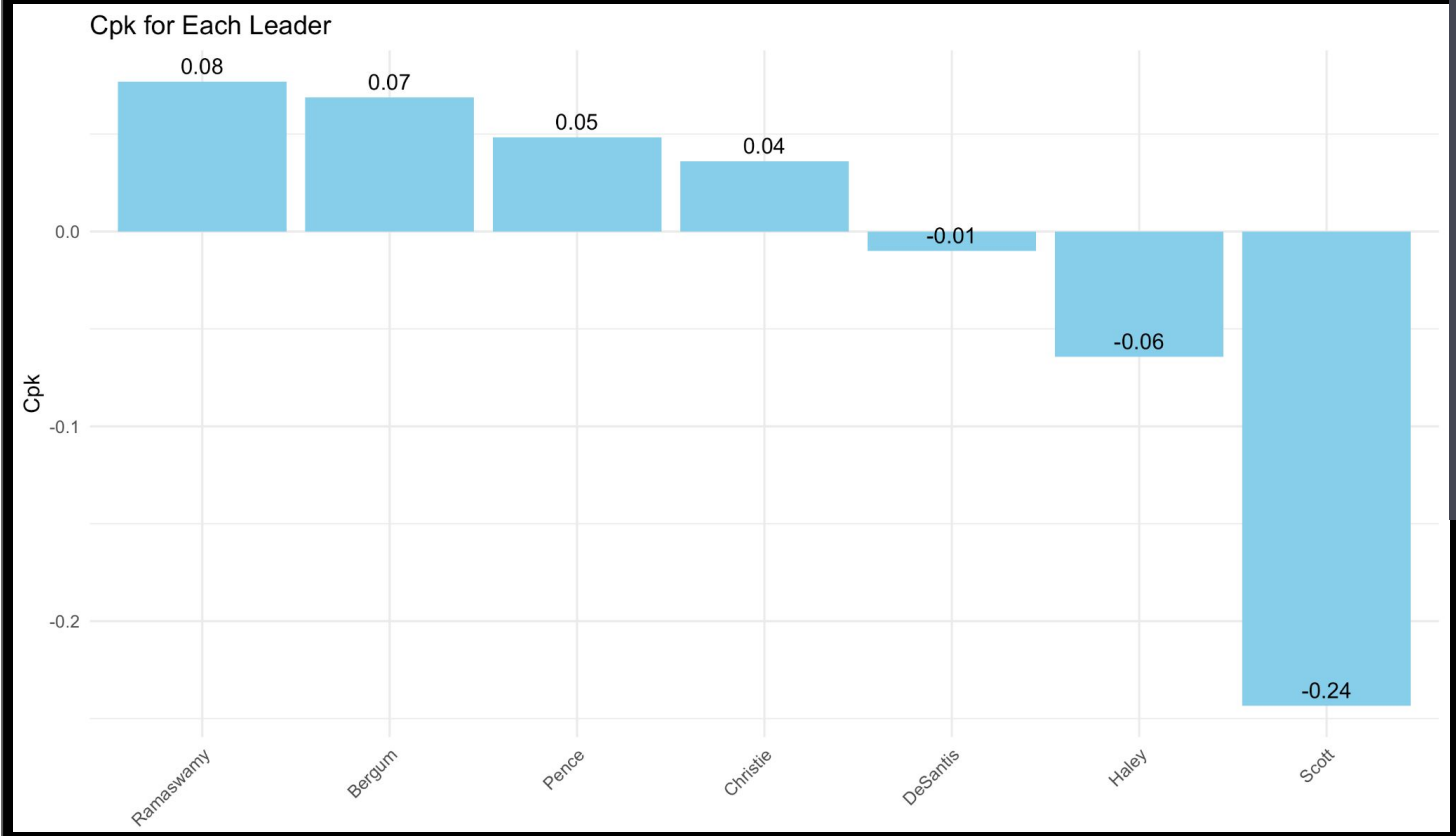
XLNC



Second Republican Debate 9.27.2023

XLNC

Consistent Excellence



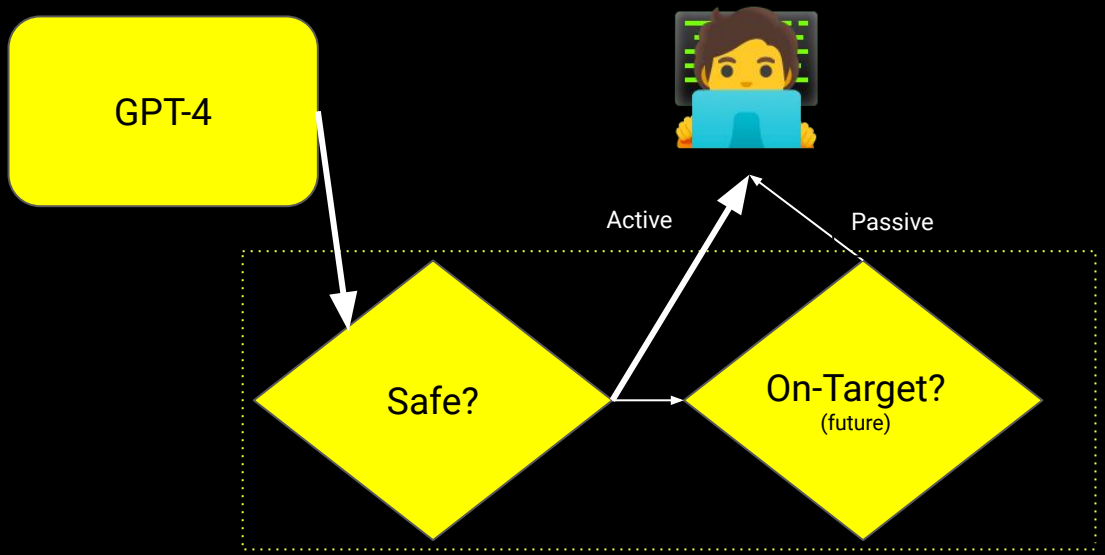
$$C_{pk} = \frac{\bar{X}_L - LSL}{3 \cdot \sigma_L}$$

Here,

- $\bar{X}_L$ is the mean of the lower confidence interval of the process data,
- LSL is the lower specification limit, and
- σ_L is the standard deviation of the lower confidence interval of the process data.

XLNC

🛡️🎯⚖️🤖 Active/Passive Rasch Guardrails



Psychometric Guardrails are a trustworthy safety net for AI.

- Active: Ensures AI outputs are appropriate
- Passive: Tailors to user's Vygotsky Zone.

Latency - Uncertainty tradeoff

Active Guardrail

- Appropriate (not racist, sexist, weird)
- In-scope (not hallucination)
- Suggestive & tentative (not directive)

Passive Guardrail

- Trustworthy feedback (Construct Map, Unbiased, Vygotsky's zone)
- Alerts for mistakes
- High precision for honest praise (Amabile's progress principle)

XLNC

AI Ethics

LLMs, ML/DL Gamble

Bias
Lies
Hallucinations
Off Target (easy/hard)

Rasch-Guardrail Verified

Active / Passive

Cialdini's Ethics

- True
- Natural
- Wise

XLNC

CialdiniBot Live Demo

XLNC

Prompt Engineering

Context - framing the situation

Instruction - task you want the AI to perform

Details - question want answered

Output Indicator - type/format of output

Example

Context

Instruction

Details

Output Indicator

Model: GPT-4

Please act as an expert on the use of Dr. Robert Cialdini's 7 principles of persuasion, the contrast phenomenon, and ethical influence. I need help writing a cold email to people who have signed up for my newsletter. I have a new class with the latest exclusive information on prompt writing for persuasion professionals, but I only have room in the live sessions for 20 people next month because it is so popular. Write me a short email message I can send to my list that use the principles of scarcity and authority, especially exclusive information as this is the only class endorsed by Dr. Cialdini for writing prompts using his science. Once you draft me the email, please suggest possible other principles that might be naturally present that I'm not thinking about.

XLNC

Psychometric Prompt Engineering

- Psychology
  - Targeted Construct Level (Commons)
  - Stimuli to notice (multimodal)
  - BARS + Examples + pre-suasive 🤖
  - Prompt AIG (Prefix/Suffix)
- Computer Science
  - Chain of Thoughts
    - Domain relevant experts to emulate
    - Step-by-step
  - Tree of Thoughts
    - Domain-relevant experts to debate

XLNC

Insight #1: LLMs are transformational but risky

1.Flexible & Fast but biased

Before - Trust?

After:

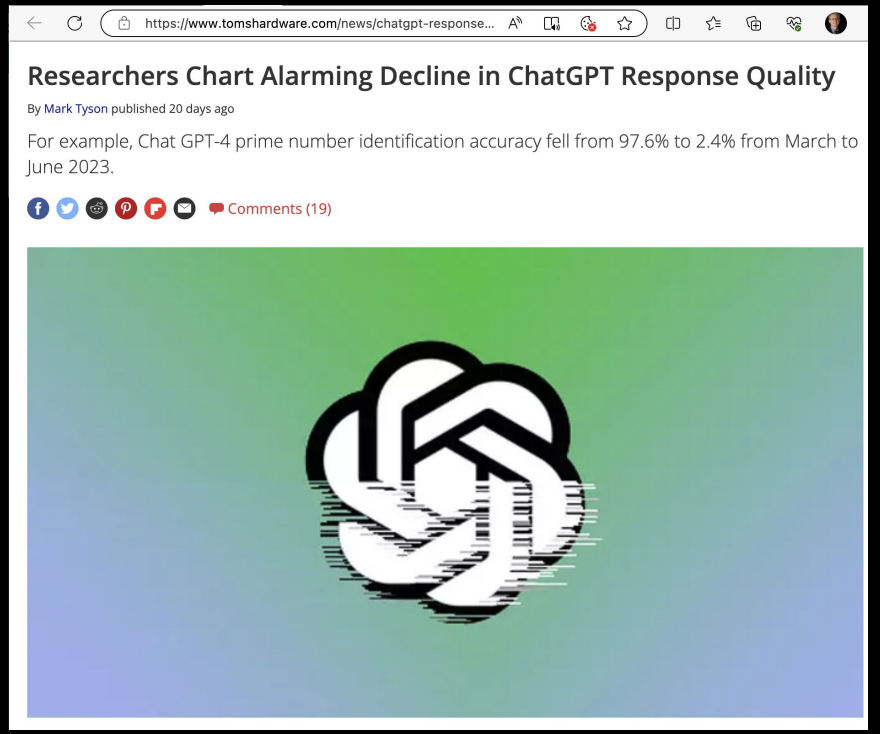
- Engineered Rasch Guardrails
- Flexible formats & channels
  - eCoaching
  - Snapshot feedback
  - Nurture Notes

XLNC

Insight #2: Ensemble > GPT-4

2. GPT-4 can't outperform an ensemble of LLMs

- GPT-4 is getting worse (Latency & Quality)
- ***Not the best for measurement***
- Elon Musk started x.ai to overcome bias
  - Bay Area: Political correctness bias
  - China: Baichuan (Tiananmen Square)
  - UAE: Falcon (Halal)
- New study showing aviation safety (Six Sigma) levels of quality



Chen, L., Zaharia, M., & Zou, J. (2023). How is ChatGPT's behavior changing over time? arXiv preprint arXiv:2307.09009

XLNC

Insight #3: Precision & Progress Principle

3. Most assessments are tedious and untimely

- **Before**: Long, tedious surveys. Tough and lenient raters bias results. Imprecision can't see small gains

- **After**: Re-measurement without effort, allows celebration of baby steps of progress

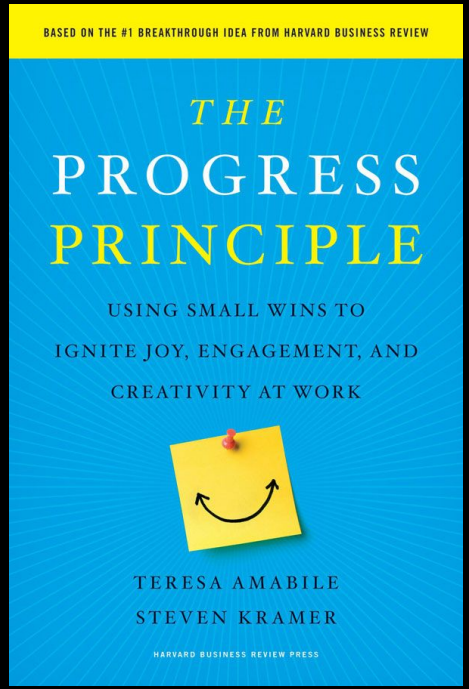
The Harvard Gazette

NEWS +

News from Harvard schools, offices, and affiliates

Amabile receives lifetime achievement award

"To a great degree, Teresa Amabile can be considered to have 'created' the field of creativity within organizations as we know it today," said the award committee. Photo by Russ Campbell



XLNC

Summary

Trustworthy instant measures are now practical for both people and AI based only on digital exhaust

- Multimodal AI?
- Longitudinal vs Cross sectional?
- Unobtrusive sample size?
- Ensemble Auto-MFRM vs inverted CAT?

XLNC