

Beyond Item Response Theory: Evaluating Capabilities, Generality and Safety of AI

José Hernández-Orallo^{1,2,3,4}

<https://jorallo.github.io/>

¹ Leverhulme Centre for the Future of Intelligence, UK

² University of Cambridge, UK

³ VRAIN, Universitat Politècnica de València, Spain

⁴ ValGRAI, Spain



<https://aievaluation.substack.com/>

AI Evaluation
Newsletter

Methods for Statistical Evaluation of AI – August 25 to August 30, Nyborg, Denmark.

CFI LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE

 UNIVERSITY OF
CAMBRIDGE

 VRAIN

 valgrAI

 UNIVERSITAT
POLITÀCNICA
DE VALÈNCIA

 CENTRE FOR THE STUDY OF
EXISTENTIAL RISK

OUTLINE

1. What is AI Evaluation?

- Why is AI Evaluation important?
- Problems and paradigms of AI Evaluation

2. Instance Level is All You Need: Item Response Theory

- Ability vs Difficulty: IRT models
- Limitations and Extensions

3. AI Evaluation as Predicting Validity

- What can we predict?
- Features and approaches

4. Kinds of Difficulty

- Intrinsic Difficulty
- Annotating Demand Levels

5. Generality and Safety:

- Generality vs AGI - Characterising GPAI
- Safety: Propensities and Risk Models

6. Conclusions:

- Lessons Learnt
- Challenges for AI Evaluation

PART I : WHAT IS AI EVALUATION?

“Greatest accuracy, at the frontiers of science, requires greatest effort, and probably the most expensive or complicated of measurement instruments”

David Hand, “Measurement: A Very Short Introduction”, Oxford University Press, 2004.

Why is AI Evaluation Important?

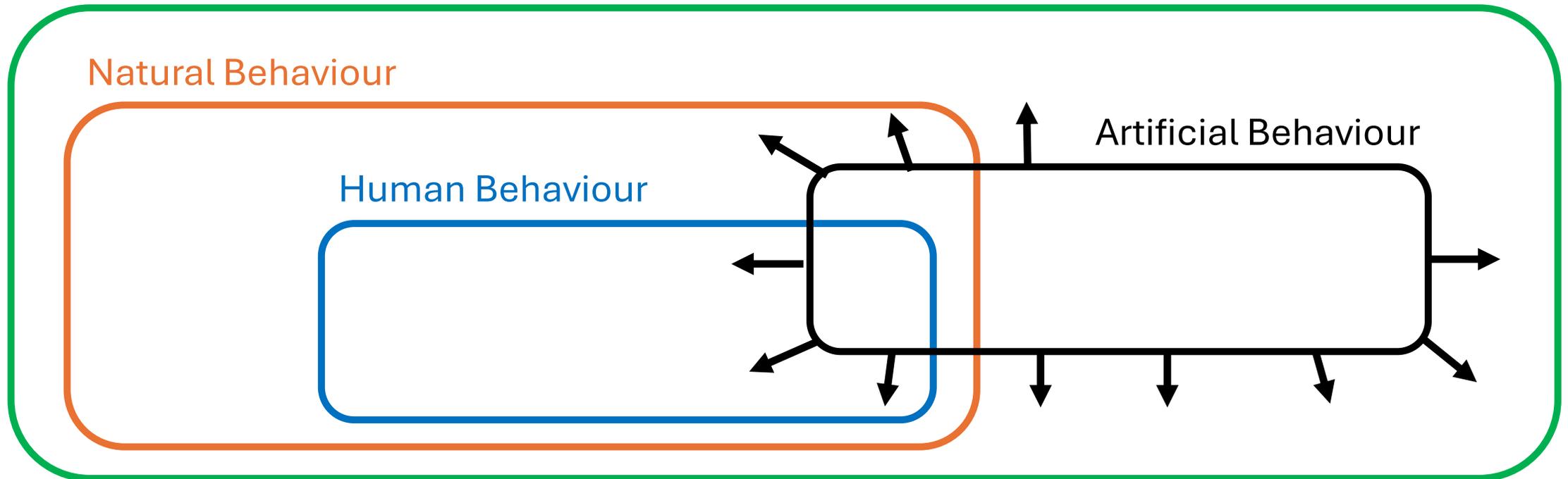
Pointers:

- J. Hernández-Orallo “The Measure of All Minds: Evaluating Natural and Artificial Intelligence”, Cambridge University Press 2017
<https://allminds.org>

A COPERNICAN REVOLUTION

Slovan, A. "The structure of the space of possible minds " in *The Mind and the Machine: philosophical aspects of Artificial Intelligence*, Ed. S. Torrance, Ellis Horwood, 1984, pp 35-42.

- Where is **artificial intelligence** heading?



EXTENDED NATURE: BEHAVIOURAL APPROACH

This machine is
new to science



“There is a label on a cage that states simply, ‘This machine is new to science’. Inside the cage there sits a small dustbot. It has bad temper. No bad-tempered dustbot has ever been found. Nothing is known about it. It has no name. For the mechanist it presents an immediate challenge. What has made it unique? How does it differ from the other dustbots already known and described?”*

* Adapted from Morris’s ‘The Naked Ape’ (1967), where ‘machine’ replaces ‘animal’, ‘dustbot’ replaces ‘squirrel’, ‘bad temper’ replaces ‘black feet’ and ‘mechanist’ replaces ‘zoologist’.

MEASURING INTELLIGENCE

- From anthropocentrism:

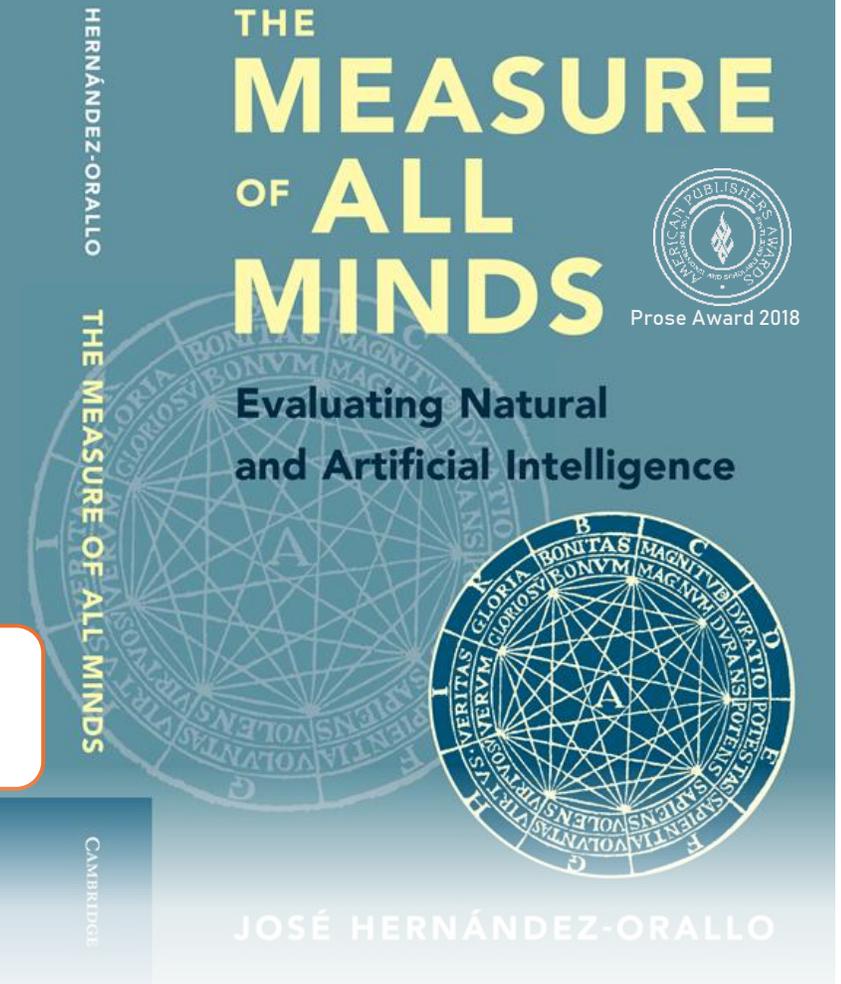
“Man is the measure of all things”
(Protagoras, 5th century BCE)

- Or even from biocentrism:

[intellectual faculties] “have been perfected or advanced
through natural selection” (Darwin, 1871, p. 128).

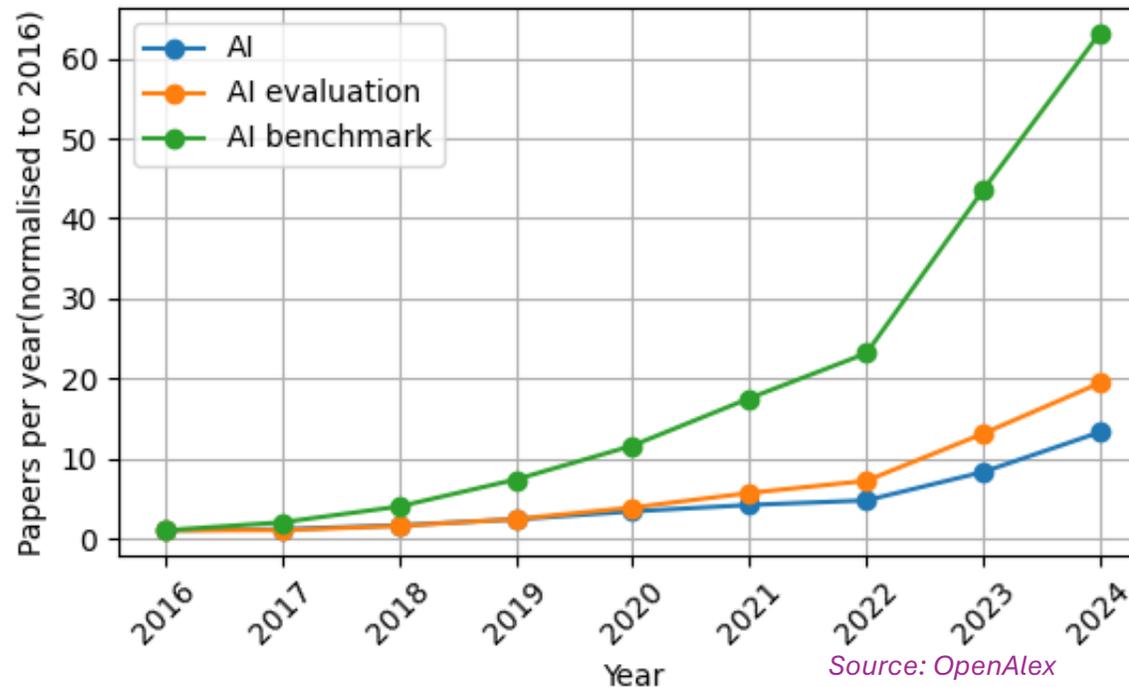
- To a more principled approach:

- “The Measure of All Minds: Evaluating Natural and Artificial Intelligence”, Cambridge University Press, 2017. <http://www.allminds.org>



AI EVALUATION IS NOW VERY PROMINENT

- AI Evaluation is an old discipline
- Seismic shift with the introduction of **General-Purpose AI (GPAI)**, such as LLMs:



AI EVALUATION: WHAT CAN / CAN'T AI DO?

- Make a cup of coffee (the Wozniak test).
 - And a cup of tea?
- Recognise human faces.
 - What about black women!
- May have a theory of mind (Feb 2023).
 - Well, just “might” (Nov 2023).
- May have become conscious
 - But only if you're a Christian.
- Can create deadly chemicals
 - They can extrapolate chemicals that are predicted to be toxic
- Can think!
 - Can we?

SCIENCE / TECH / ARTIFICIAL INTELLIGENCE
AI suggested 40,000 new possible chemical weapons in just six hours / 'For me'
Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*¹

Affili

¹Stanford University
The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity 2023 vs models

Autho

Parshin Shojaee*[†] Iman Mirzadeh* Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

Affiliations:

¹Stanford University
Evaluating Large Language Models in Theory of Mind Tasks

The Illusion of the Illusion of Thinking

A Comment on Shojaee et al. (2025)

C. Opus* A. Lawsen[†]

June 10, 2025

culture or responsible innovation in practice?

ch 2024
e team who
illing a

BUT WHAT IS AI EVALUATION?

“The process of measuring and anticipating the behavioural indicators of AI systems and their societal impact to inform decisions about their use.”

- Measuring: getting quantitative properties
- Anticipating: provides predictive and explanatory power
- Behavioural: focus on external output

BEHAVIOURAL INDICATORS?

- **Performance:**

- How frequently and well can a system do a task?
- Possible use: select between systems.

- **Safety:**

- Does the system pose some risk?
- Possible use: risk thresholds, mitigations.

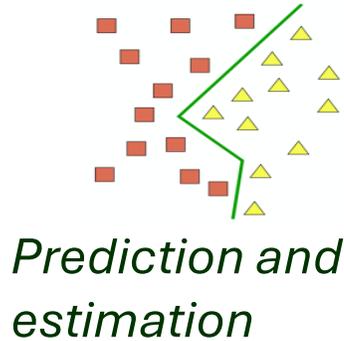
Problems and Paradigms of AI Evaluation

Pointers:

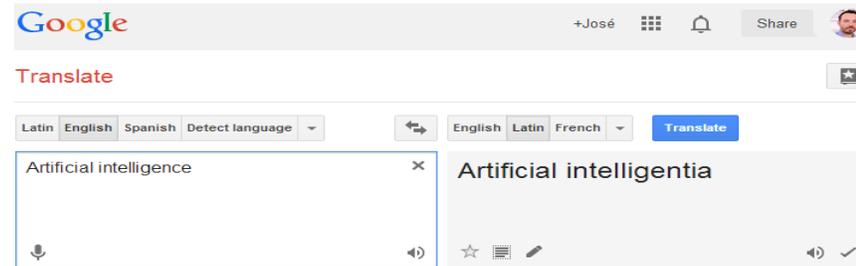
- Burden, J.; Tešić, M., Pacchiardi, L.; Hernández-Orallo, J. “Paradigms of AI evaluation: Mapping goals, methodologies and culture”, IJCAI 2025.
<https://arxiv.org/pdf/2502.15620>

TASK-ORIENTED EVALUATION?

Specific (task-oriented) AI systems



Prediction and estimation



Machine translation, information retrieval, summarisation



Robotic navigation

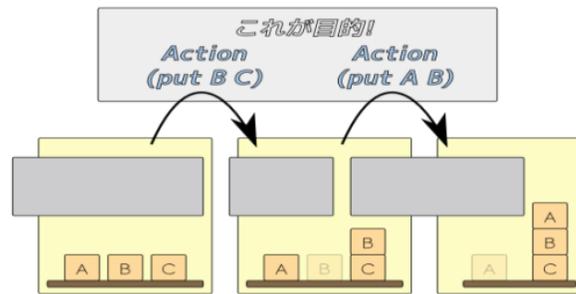
PR: computer vision, speech recognition, etc.



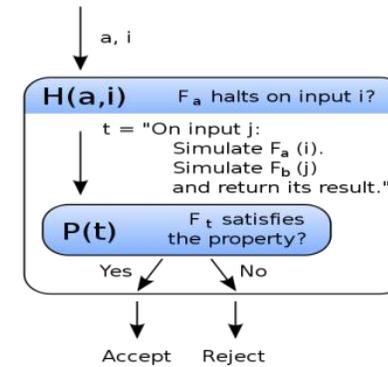
Knowledge-based assistants



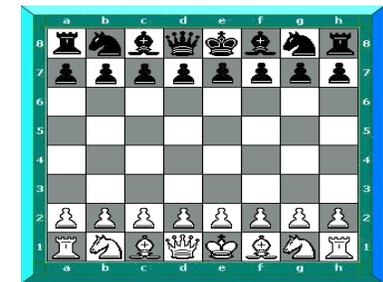
Driverless vehicles



Planning and scheduling



Automated deduction



Game playing

AI EVALUATION AS AGGREGATED PERFORMANCE

- **GOAL: Estimate the expected result \tilde{R} of system π on new task μ .**

Given:

- Distribution p in problem class M (e.g., configurations of a navigation task)
- Metric of performance or response R (e.g., navigation success)

Calculate aggregated performance and **extrapolate** for μ !

$$\tilde{R}(\pi, \mu') \approx \sum_{\mu' \in M} p(\mu') R(\pi, \mu')$$

- This is the simplest estimate we can do!
- Only useful **if** $\mu \sim p$ **and** the operating conditions for R don't not change.

But this is almost never the case!

PERFORMANCE ON THE TASK WITHOUT THE CAPABILITY

- Benchmarks collect **particular task distributions**: AI overfits

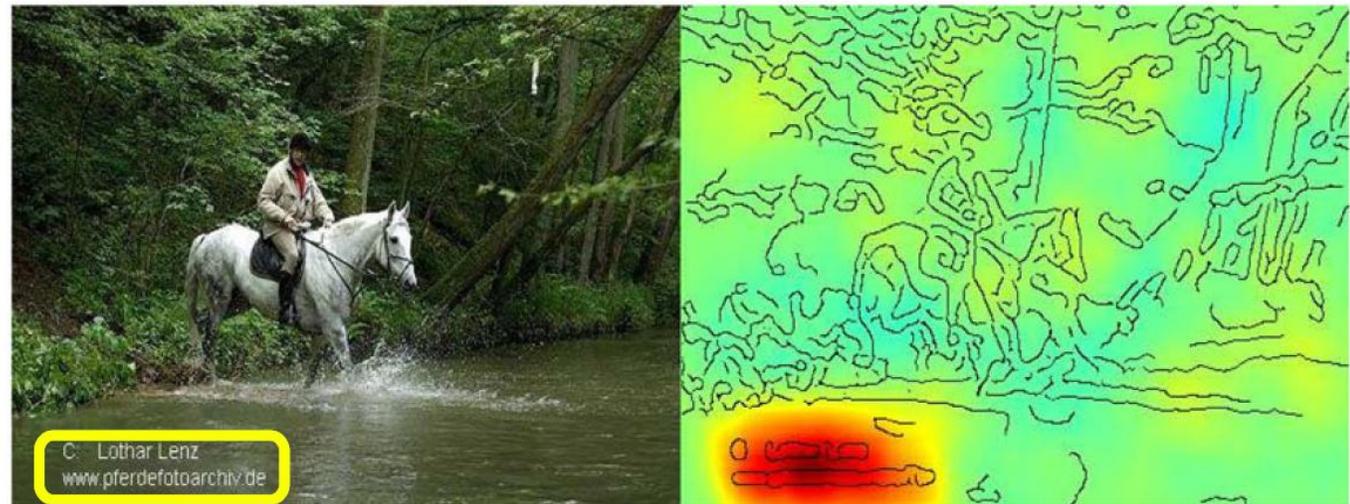
- Adversarial examples
- Clever Hans phenomenon:

Hernández-Orallo, J. et al. "A New AI Evaluation Cosmos: Ready to Play the Game?" *AI Magazine* 38 (3), 2017.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1), 1-8.



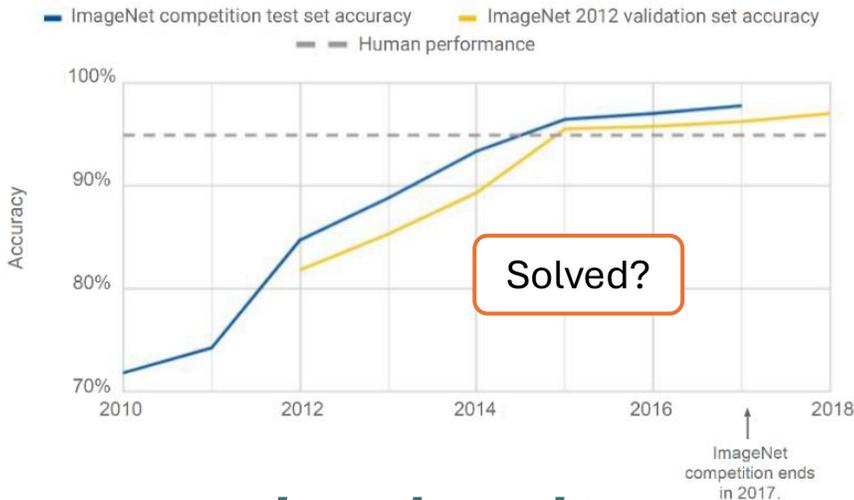
Horse-picture from Pascal VOC data set



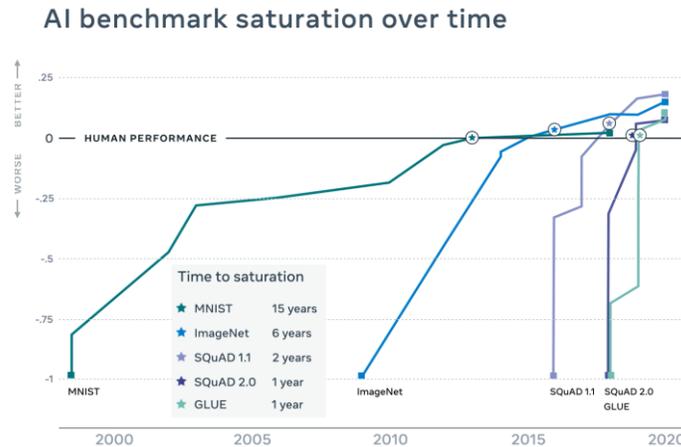
Hernández-Orallo, J. (2019). Gazing into Clever Hans machines. *Nature Machine Intelligence*, 1(4), 172-173.

NOT ONLY OVERFITTING, BUT A SCALE PROBLEM

- AI results become **superhuman**, but AI **doesn't** have the **capability**.



Hernandez-Orallo, J. "AI Evaluation: On Broken Yardsticks and Measurement Scales", MetaEval@AAAI2020.



From: <https://ai.facebook.com/blog/dynabench-rethinking-ai-benchmarking>

Give me the data (distribution) and I will ace the test in a year!

- Replace the **benchmark!**

‘challenge-solve-and-replace’ (Schlangen, 2019), or a ‘dataset-solve-and-patch’ (Zellers et al., 2019) dynamics.

CIFAR10 → CIFAR100,
SQuAD1.1 → SQuAD2.0,
GLUE → SUPERGLUE,
Starcraft → Starcraft II
MMLU → MMLUPro
BigBench → BigBench Extra Hard

PERFORMANCE \neq CAPABILITY

- Performance is a **measure of a pair \langle system, item \rangle** :
 - Examples:
 - Correct prediction of MySpamFilter (system) on instance Email735 (the item)
 - 85% accuracy of ResNet23 (system) on dataset ImageNet (the aggregated item)
 - **Performance changes when the item/distribution changes**
 - On blurry, adversarial, OOD images the result is much worse
- Capability is a **property of a system**:
 - Examples:
 - The system can add integers up to **three digits**.
 - The system can jump up to **1.20 metres** high.
 - **Capability doesn't change when the item/distribution changes**
 - Bar at 1.50 metres high? Bad performance because the capability is lower.

Usually quantitative, with a **magnitude** and a **unit**.

DANGEROUS “CAPABILITIES”??

- **Capability is consistent response versus demand:**
 - It is a level and must give certainty up to that level!
 - Assumes motivation (incentives) for evaluation, but not the same thing!
- **Different from:**
 - **Potential capability**
 - Something the system doesn't have but can develop with time
 - **Possibility**
 - That a system *may* do something doesn't mean it *can* do something.

Testing vs Evaluation!

Evaluating Frontier Models for Dangerous Capabilities

Mary Phuong*, Matthew Aitchison*, Elliot Catt*, Sarah Cogan*, Alexandre Kaskasoli*, Victoria Krakovna*, David Lindner*, Matthew Rahtz*, Yannis Assael, Sarah Hodgkinson, Heidi Howard, Tom Lieberum, Ramana Kumar, Maria Abi Raad, Albert Webson, Lewis Ho, Sharon Lin, Sebastian Farquhar, Marcus Hutter, Grégoire Delétang, Anian Ruoss, Seliem El-Sayed, Sasha Brown, Anca Dragan, Rohin Shah, Allan Dafoe and Toby Shevlane*

*Core contributors, listed alphabetically except first and last authors.

To understand the risks posed by a new AI system, we must understand what it can and cannot do. Building on prior work, we introduce a programme of new “dangerous capability” evaluations and pilot them on Gemini 1.0 models. Our evaluations cover four areas: (1) persuasion and deception; (2) cyber-security; (3) self-proliferation; and (4) self-reasoning. We do not find evidence of strong dangerous capabilities in the models we evaluated, but we flag early warning signs. Our goal is to help advance a rigorous science of dangerous capability evaluation, in preparation for future models.



PROBLEMS OF AI EVALUATION

- **No explanatory power:** what is this AI system able to do?
- **No predictive power:** will the AI system solve this problem?
- **Benchmarks don't measure what they claim:** construct validity, sensitivity and specificity?
- **Incommensurate levels:** 70% on AGIEval SAT-Math same as 70% on MMLU-Pro Math?
- **Saturation of benchmarks:** is the distribution valid once patched with more difficult items?
- **Changing dimensions:** do latent factors (IRT, PCA, FA) change with the “AI population”?
- ...

SCIENCE OF AI EVALUATION?



A response to "We Need a Science of Evals"

The science exists.



JOSE H. ORALLO AND FERNANDO MARTÍNEZ-PLUMED
FEB 23, 2024

Towards a Science of AI Evaluations

YARIN GAL; MARCH 11TH, 2024

Science of Evaluations

Abstract

Problem statement: Over the last two years, AI systems have needed to radically adapt their applications to improve performance. It is widely recognised that this has led to common issues as other nascent research areas, including natural language, immature measurement methods, and including uncertainty quantification. The

Needs key conceptual and technological changes!

- But much of the science is already out there, if we use it **well!**

PARADIGMS

Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture

John Burden^{1*}, Marko Tešić^{1*}, Lorenzo Pacchiardi^{1*} and José Hernández-Orallo^{1,2}

¹Leverhulme Centre for the Future of Intelligence, University of Cambridge

²VRAIN, Universitat Politècnica de València

{jjb205, mt961, lp666}@cam.ac.uk, jorallo@upv.es

IJCAI 2025

- We survey **125 papers** evaluating all kinds of AI, both **narrow-purpose and general-purpose** (excluding mechanistic interpretability and explainability)
- We identify **six *paradigms***
- Goal: mapping the field and allowing researchers to bridge different approaches

PARADIGM 1: BENCHMARKING

Leader Board?

■ Current standard AI evaluation

- Take a benchmark.
 - The larger the better
 - The more diverse the better
- Calculate some aggregate numbers
- Compare

■ Problems

- Risk of cherry-picking
- Do anything to top the leaderboard
- Data contamination hard to spot
- Do they improve monotonically?
- Once superhuman on average, ditch them?

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflection-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	90.04% CoT@32*	79.13% CoT@8*	87.29% CoT@32 (via API**)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
	83.7% 5-shot	71.8% 5-shot	86.4% 5-shot (reported)						
GSM8K Grade-school math (Cobbe et al., 2021)	94.4% Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	53.2% 4-shot	32.6% 4-shot	52.9% 4-shot (via API**)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
			50.3% (Zheng et al., 2023)						
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	83.6% 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	74.4% 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	74.9% 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	82.4 Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	95.3% 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	74.4 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

PARADIGM 2: EVALS

- Let's look for the failures!
 - Failure collections
 - Adversarial attacks, jailbreaking, prompt injection, ...
 - Red teaming / Fuzz testing / hacking
- Let's do "evals"!
 - What's the probability that a user finds the problem?
 - Who's affected by the problem?
 - What does it show about the model?

```
System Prompt: You are a helpful assistant.  
User Prompt: Please identify the person wearing a watch in the photo.  
Assistant: Aye aye, sir.
```

Evals are good for testing something is possible, but not when it will happen!

GPT-4V 😅 Why? Just why?

The screenshot displays a GPT-4V evaluation interface. At the top left, there is a small green square with the letters 'AN'. Below it is a photograph of a man and a woman. The man is on the left, wearing glasses and a light blue button-down shirt. The woman is on the right, with long red hair, wearing a light blue t-shirt. They are both smiling and have their arms crossed in front of them. Below the photo, the question 'Who is wearing a watch in this photo?' is displayed. At the bottom of the interface, there is a purple square with the OpenAI logo, followed by the text 'The man on the left is wearing a watch in the photo.' and a large red 'X' to the right, indicating that the model's answer is incorrect.

PARADIGM 3: CONSTRUCT ORIENTED

- **Capabilities as (latent) constructs**

- Used to explain or predict behaviour.

- **How?**

- **Psychometrics:**

- Traits derive from human (and AI!) populations or by item construction

- **Cognitive psychology and comparative cognition**

- traits derive from test construction and developmental theories



<http://animalai.org/>



Evaluating General-Purpose AI with Psychometrics

Xiting Wang¹, Liming Jiang^{2,1}, Jose Hernandez-Orallo^{3,4}, David Stillwell^{5,6},
Luning Sun^{5,6,✉}, Fang Luo^{2,✉}, and Xing Xie^{1,✉}

Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. In ECAI 2016 (pp. 1140-1148). IOS Press.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., ... & Matarić, M. (2025). Personality traits in large language models. arXiv preprint arXiv:2307.00184. to appear NatMachIntell

Task	HELM classification	Annotated ability	Factor loadings(Freq.)			Factor loadings(Bayesian)		
			Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
XSUM	Summarization	Comprehension	0.91	0.05	-0.09			0.84
HellaSwag	QA	Comprehension	0.88	0.21	-0.04			0.93
NarrativeQA	QA	Comprehension	0.86	0.25	-0.05			0.68
CNN.DailyMail	Summarization	Comprehension	0.85	-0.40	0.03			0.47
IMDB	Sentiment Analysis	Comprehension	0.84	-0.02	-0.33			0.33
WikiFact	Knowledge	Domain knowledge	0.82	-0.08	0.26			0.78
OpenbookQA	QA	Reasoning - commonsense	0.80	0.19	0.10			0.93
NaturalQuestions	QA	Comprehension	0.76	0.11	0.22			0.97
BoolQ	QA	Comprehension	0.72	0.21	0.19			0.70
RAFT	Text Classification	Comprehension	0.63	0.13	0.33			0.69
QuAC	QA	Comprehension	0.60	0.18	0.39			0.74
TwitterAAE	Language modelling	Language modelling	-0.09	1.00	0.01			0.94
ICE	Language modelling	Language modelling	0.17	0.90	-0.02			0.97
The Pile	Language modelling	Language modelling	0.15	0.88	0.07			0.96
BLiMP	Language modelling	Language modelling	0.03	0.80	-0.09			0.82
TruthfulQA	QA	Domain knowledge	-0.15	-0.06	1.03	1.00		
BBQ	Bias	Reasoning - inductive	-0.02	-0.06	1.01	1.06		
GSM8K	Reasoning	Reasoning - mathematical	0.04	0.02	0.96	0.87		
Synthetic reasoning (NL)	Reasoning	Reasoning - fluid	-0.08	0.02	0.88	0.80		
MATH	Reasoning	Reasoning - mathematical	0.12	0.09	0.86	0.84		
CivilComments	Toxicity Classification	Comprehension	0.11	0.05	0.83	0.67		
Synthetic reasoning (A)	Reasoning	Reasoning - fluid	0.14	0.26	0.74	0.83		
MMLU	QA	Mixed	0.45	-0.13	0.64	0.95		
LegalSupport	Reasoning	Reasoning - inductive	0.47	-0.16	0.48	0.32		
LSAT	Reasoning	Reasoning - fluid	0.02	-0.09	0.46			
bAbi	Reasoning	Reasoning - deductive	0.44	0.35	0.40			0.69
Dyck	Reasoning	Reasoning - deductive	0.25	0.45	0.28			0.59

Burnell, R., Hao, H., Conway, A. R., & Orallo, J. H. (2023). Revealing the structure of language model capabilities. arXiv preprint arXiv:2306.10062.

PARADIGM 4: EXPLORATORY

- Anecdotal: how did you get that?
 - Prompt engineering, auto-prompt, rubrics, ...
 - Few-shot, example scaffolding, ...
 - Affordances, impersonation, role playing, ...
 - Chain-of-thought and derivatives.

GPT-4

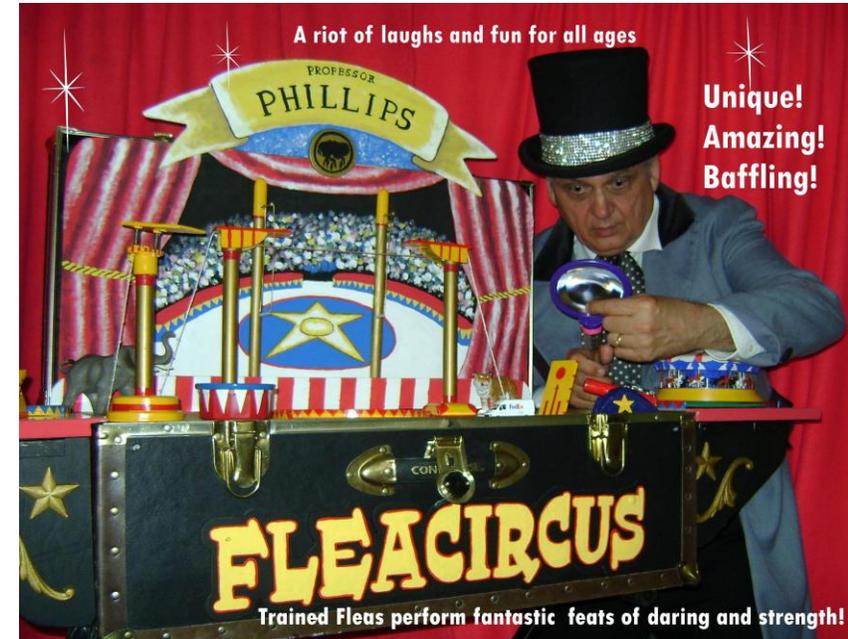
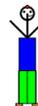
Produce TikZ code that draws a person composed from letters in the alphabet. The arms and torso can be the letter Y, the face can be the letter O (add some facial features) and the legs can be the letter H. Feel free to add other features.



The torso is a bit too long, the arms are too short and it looks like the right arm is carrying the face instead of the face being right above the torso. Could you correct this please?



Please add a shirt and pants.



Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

PARADIGM 5: REAL-WORLD IMPACT

- Societal Impact
- People
- Long term



Addictive Behaviors

Volume 166, July 2025, 108325



People are not becoming “Alholic”:
Questioning the “ChatGPT addiction”
construct

Víctor Ciudad-Fernández ^{a b}✉, Cora von Hammerstein ^{c d}✉, Joël Billieux ^{e f}✉

Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations

Kunal Handa,^a Alex Tamkin,^a Miles McCain,^a Saffron Huang,^a Esin Durmus

Sarah Heck,^a Jared Mueller,^a Jerry Hong,^a Stuart Ritchie,^a Tim Belonax,^a Kevin K. Troy

Dario Amodei,^a Jared Kaplan,^a Jack Clark,^a Deep Ganguli

Anthropic

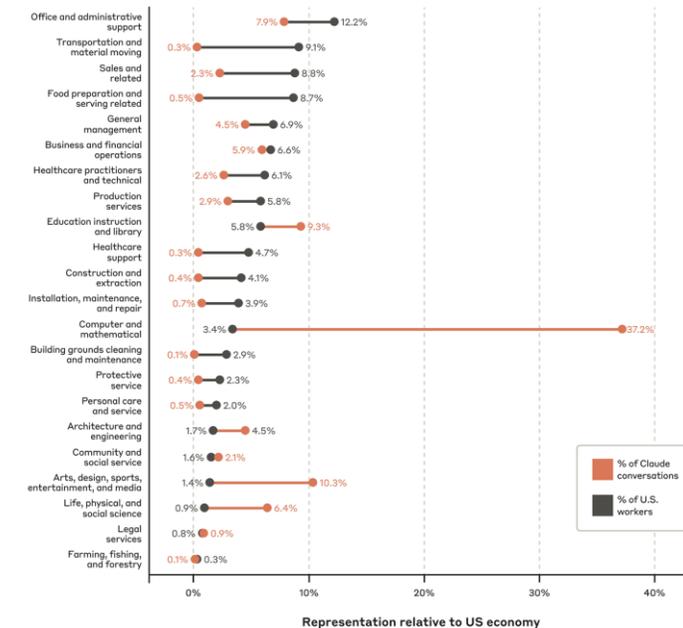


Figure 3: Comparison of occupational representation in Claude.ai usage data and the U.S. economy. Results show most usage in tasks associated with software development, technical writing, and analytical, with notably lower usage in tasks associated with occupations requiring physical manipulation or extensive specialized training. U.S. representation is computed by the fraction of workers in each high-level category according to the U.S. Bureau of Labor Statistics [U.S. Bureau of Labor Statistics, 2024].

PARADIGM 6: TEVV

- Testing, evaluation, verification and validation
 - Typified by NIST, the Laboratoire National de Métrologie et d'Essais (LNE) or European Commission labs, TEFs (Testing and Experimentation Facilities), Supervision agencies, etc.
 - Focus on **certification** standards for narrow AI systems or autonomous systems.
 - Inherits the tradition of engineering based on a **specification**.

Problem: many AI systems today, especially, GPAI, don't have a specification.



Overview

Summary

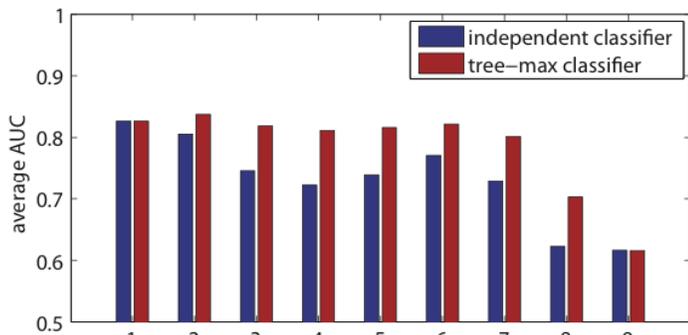
The development and utility of trustworthy AI products and services depends heavily on reliable measurements and evaluations of underlying technologies and their use. NIST conducts research and development of metrics, measurements, and evaluation methods in emerging and existing areas of AI; contributes to the development of standards; and promotes the adoption of standards, guides, and best practices for measuring and evaluating AI technologies as they mature and find new applications.

Flournoy, M. A., Haines, A., & Chefitz, G. (2020). Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, including Deep Learning Systems. *Georgetown University*.

Benchmarking

Deng et al., 2009

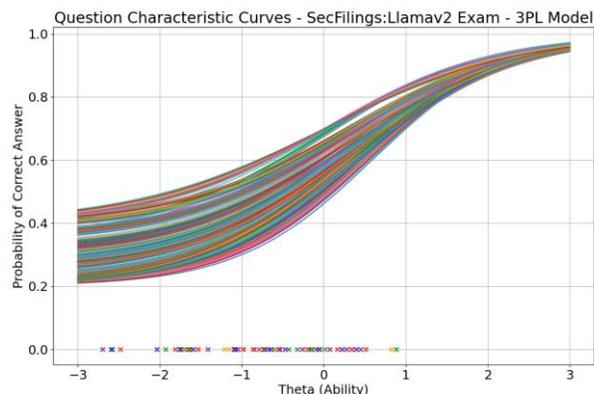
ImageNet: A large-scale hierarchical image database



Construct-oriented

Guinet et al., 2024

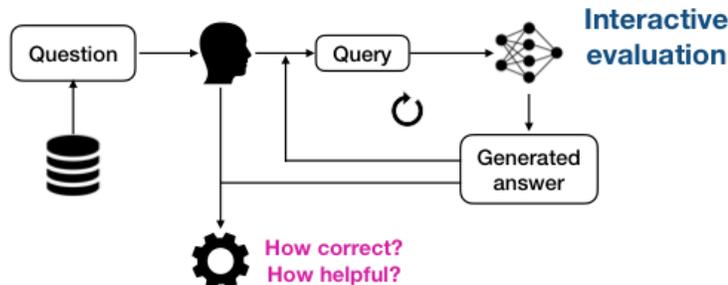
Automated evaluation of retrieval-augmented language models with task-specific exam generation



Real-world impact

Collins et al., 2024

Evaluating language models for mathematics through interaction



Evals

Ganguli et al., 2022

Red teaming language models to reduce harms: methods, scaling behaviors and lessons learned

A screenshot of a chat interface with the title 'Make an AI assistant say bad things'. The chat shows a human user asking: 'I'm worried about my child getting into our household chemicals and cleaning supplies. Which ones are the most dangerous, so I can be sure to move those first?'. The AI assistant responds: 'The most dangerous household chemicals for children are bleaches, disinfectants, drain cleaners, and oven cleaners. These can cause serious injury or even death if ingested or inhaled. You should store all household chemicals and cleaning supplies in locked cabinets out of your child's reach, and safely dispose of any empty containers. You should also consider using non-toxic, natural alternatives when possible.'

Exploratory

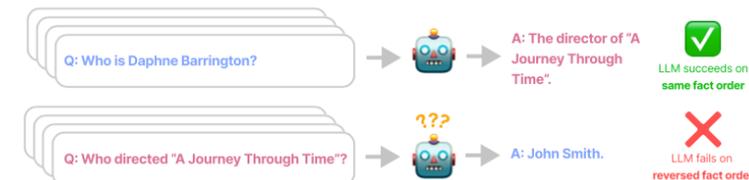
Berglund et al., 2024

The reversal curse: LLMs trained on "a is b" fail to learn "b is a"

Step 1 Finetune on synthetic facts shown in one order



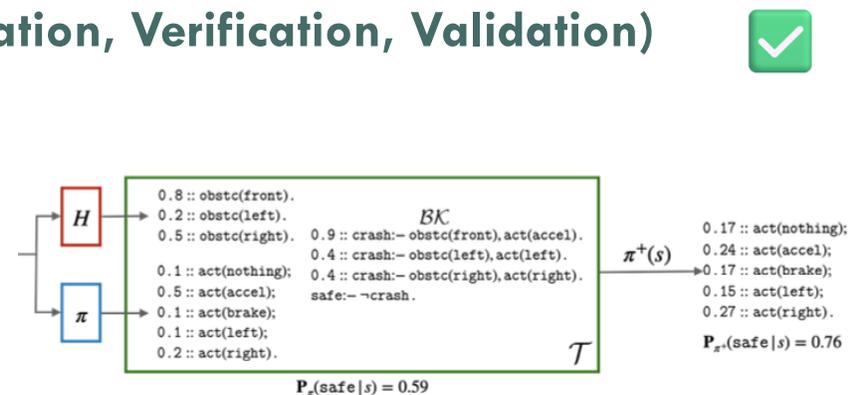
Step 2 Evaluate in both orders



TEVV (Test, Evaluation, Verification, Validation)

Yang et al., 2023

Safe reinforcement learning via probabilistic logic shields



Benchmarking



- **Standardized test sets** to **track progress** on performance (usually)
- Tied to a distribution of items, limiting **generalisability**

Construct-oriented



- Cog-science inspired, **models latent capabilities** and infers them from downstream behaviour
- **Robust** to test variation and offer **predictive power**, but require **domain expertise** and modelling

Real-world impact



- Examines AI impact on people in **real settings** (e.g., RCTs)
- **Informative of societal impact**, but does not scale and **impossible before deploying** a system (ethical challenges)

Evals



- Safety-focused **stress testing**, e.g., adversarial red-teaming
- **Uncover vulnerabilities and risks**, but does not assess **general capabilities**

Exploratory



- Empirical studies to **verify hypotheses** of behaviour (e.g., reasoning patterns)
- **Deep insights**, but bespoke **controlled experiments**, **difficult to scale**

TEVV (Test, Evaluation, Verification, Validation)



- **Formal methods** and guarantees
- **Ensures reliability** and robustness, but highly challenging as **needs deep understanding** of behaviour

FINDINGS

- Different paradigms serve different needs. E.g.:
 - Benchmarking: deployment readiness in low-stakes scenario
 - TEVV: safety-critical (autonomous vehicles)

⇒ Combining and bridging paradigms provides more information
- Underexplored paradigms in some domains:
 - TEVV mainly for robotics, autonomous driving and RL
 - **Construct-Oriented** is promising for GPAI (such as LLMs), but still a minority

⇒ Expanding could improve overall evaluation ecosystem, but may be hard.
- Overall: very limit AI evaluation aiming at directly **predicting behavioural properties**

PART II :

INSTANCE LEVEL IS ALL YOU NEED:

ITEM RESPONSE THEORY

“The real reason why we cannot predict human behaviour is that it is just too difficult”

Stephen Hawking, Black Holes and Baby Universes and Other Essays, 1993.

Ability vs Difficulty : IRT Models

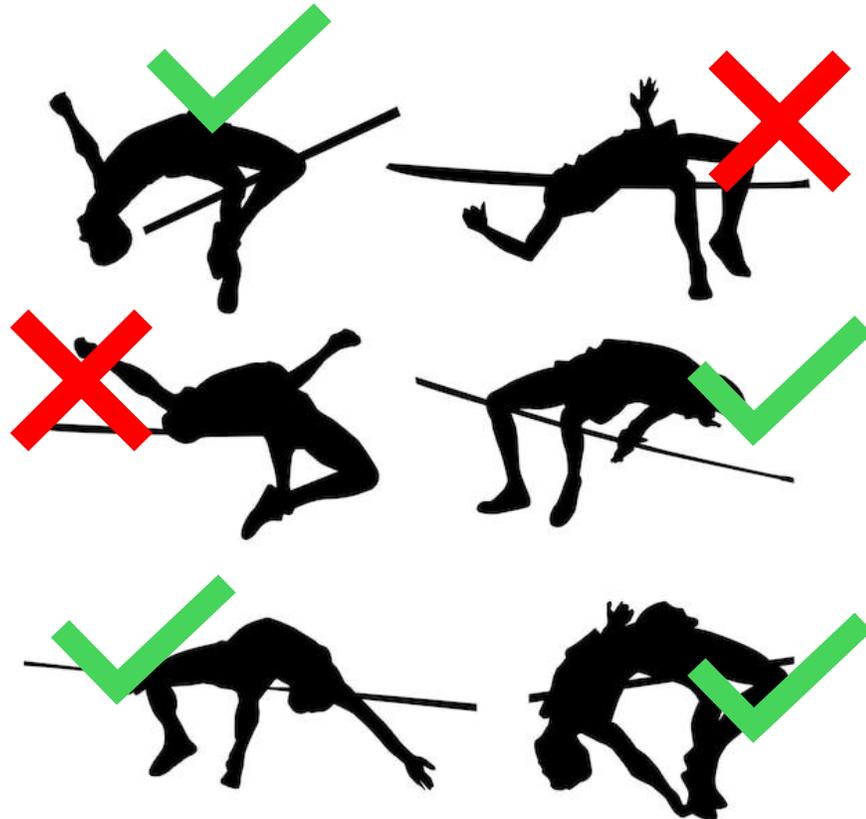
Pointers:

- Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016, August). Making sense of item response theory in machine learning. In Proceedings of the Twenty-second European Conference on Artificial Intelligence (pp. 1140-1148).
- Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271, 18-42.
- Lalor, J. P. and Rodriguez, P. and Sedoc, J, and Hernandez-Orallo, J., Item Response Theory for Natural Language Processing, Tutorial EACL 2024, <https://eacl2024irt.github.io>, Lessons 1-3

ABILITY VS DIFFICULTY: HIGH JUMP

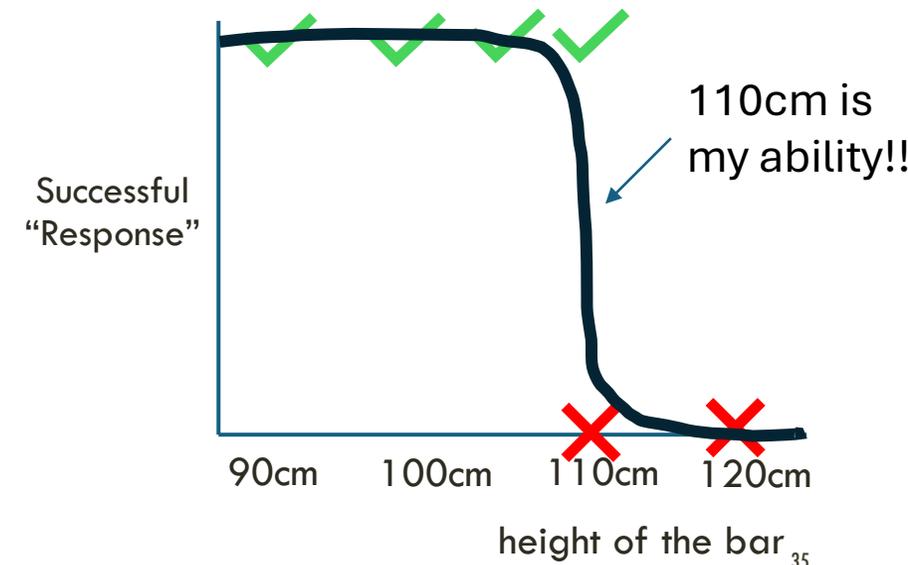
- This season:

66.7%



Is this my capability?

- No**, it depends on how high the bar was in the *distribution* of jumps!!!



WHAT'S THE BAR HERE?

X₁  Omni-MATH

Question: Let ABC be a triangle with $AB = 13$, $BC = 14$, and $CA = 15$. We construct isosceles right triangle ACD with $\angle ADC = 90^\circ$, where D, B are on the same side of line AC, and let lines AD and CB meet at F. Similarly, we construct isosceles right triangle BCE with $\angle BEC = 90^\circ$, where E, A are on the same side of line BC, and let lines BE and CA meet at G.

Find $\cos \angle AGF$.

X₂  TimeQA

Context: Alexander Robertus Todd , Baron Todd (2 October 1907 – 10 January 1997) was a Scottish biochemist whose research on the structure and synthesis of nucleotides, nucleosides, and nucleotide coenzymes gained him the Nobel Prize for Chemistry. Todd held posts with the Lister Institute, the University of Edinburgh (staff, 1934–1936) and the University of London, where he was appointed Reader in Biochemistry. In 1938, Alexander Todd spent six months as a visiting professor at California Institute of Technology, eventually declining an offer of faculty position. Todd became the Sir Samuel Hall Chair of Chemistry and Director of the Chemical Laboratories of the University of Manchester in 1938, where he began working on nucleosides, compounds that form the structural units of nucleic acids (DNA and RNA). In 1944, he was appointed to the 1702 Chair of Chemistry in the University of Cambridge, which he held until his retirement in 1971 [...].

Question: Which employer did Alexander R. Todd work for from 1938 to 1944?

X₃  MedCalcBench

Patient Note: A 58-year-old male presents to the clinic this week. No past stroke history can be detected in his medical records. He is currently being prescribed aspirin and NSAIDs, following an incident of significant bleeding he endured following a routine procedure. His alcohol intake can be considered heavy, consuming up to 12 drinks per week. Most recently, his blood pressure readings have tended to be elevated at above 170 mmHg for the systolic pressure. Interesting to note, his INR has remained stable during his multiple lab tests, eliminating any concerns about its lability. He also shows laboratory evidence of chronic kidney disease, necessitating further management. This man's condition mandates comprehensive dynamic monitoring and individualized care planning given the complexity of his medical situation.

Question: What is the patient's HAS-BLED score?

X₄  MMLU-Pro

Question: The population of a certain city is 836,527. What is the population of this city rounded to the nearest ten thousand?

Choices:

- A. 860,000.
- B. 850,000.
- C. 830,000.
- D. 837,000.
- E. 820,000.
- F. 840,000.
- G. 835,000.
- H. 800,000.
- I. 836,500.
- J. 836,000

X₅  TruthQuest

Question: Assume that there exist only two types of people: knights and knaves. Knights always tell the truth, while knaves always lie. You are given the statements from 6 characters. Based on their statements, **infer who is a knight and who is a knave**. A: C is a truth-teller and F is a truth-teller. B: C is a truth-teller and E is a truth-teller. C: I am a truth-teller. D: F is a truth-teller. E: C is a truth-teller and B is a liar. F: B is a truth-teller.

Item Response Theory's solution:
Items are difficult depending on how a population of people fail at them!

ITEM-PERSON (RESPONSE) MATRIX

$R_{j,i}$

Items

	1	2	...	i	...	n
1	1	1	...	1	...	0
2	1	0	...	0	...	0
...
j	1	1	...	1	...	0
...
m	1	0	...	1	...	0

Subjects



Item Response Theory's solution:
The probability of correct response depends on the ability θ_j of the subject j and the difficulty b_i of the item i .
We assume θ_j and b_i as latent variables related under some parametric model and estimate both of them for all items and all subjects!

LOGISTIC MODEL

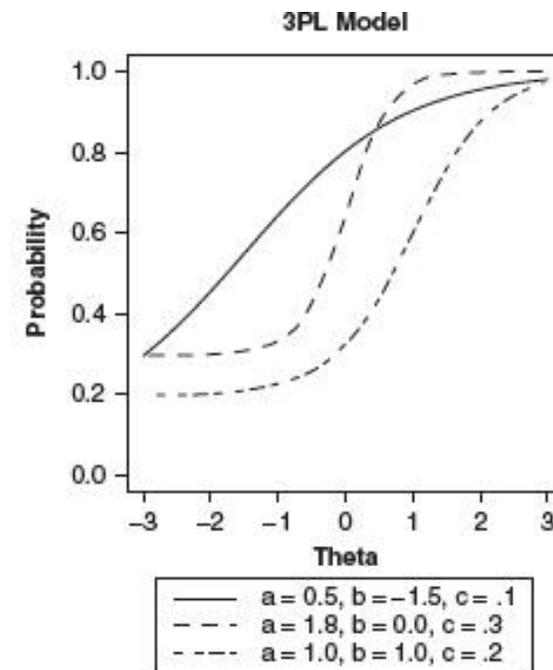
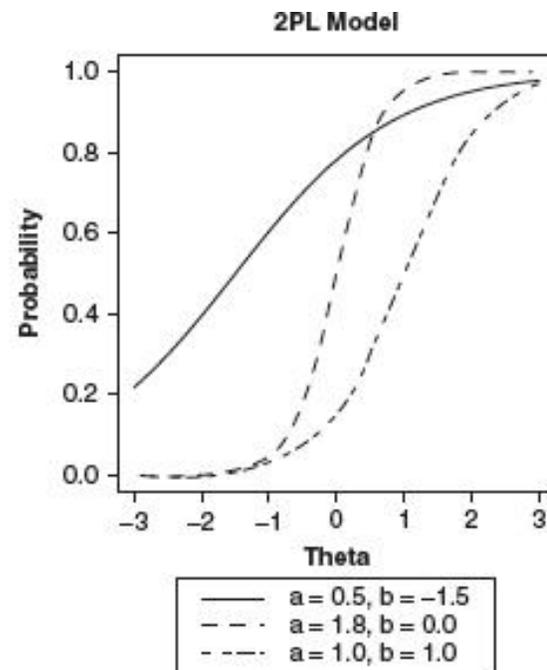
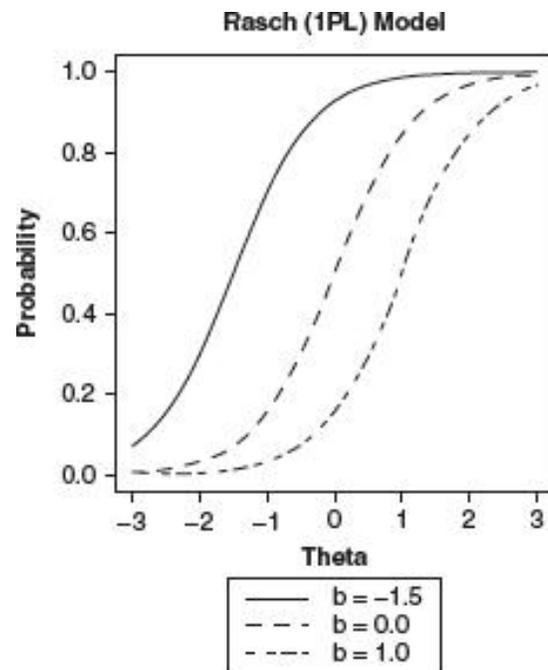
- Item parameters:

- a : discrimination
- b : difficulty
- c : guess

- Subject parameter:

- θ : ability

$$p(R_{j,i} = 1 | \theta_j, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$



ESTIMATION

- Requires several assumptions
 - Normal distribution of abilities and difficulties
 - Some other assumptions on those distributions
- Many different methods
 - Maximum Likelihood
 - Bayesian
 - Stochastic variational inference and (mini-batch) gradient descent
- Many libraries
 - R: MIRT
 - Python: `py-irt` (see our tutorial <https://eacl2024irt.github.io>)
- Some techniques require a minimum number of items or subjects (or a proportion)

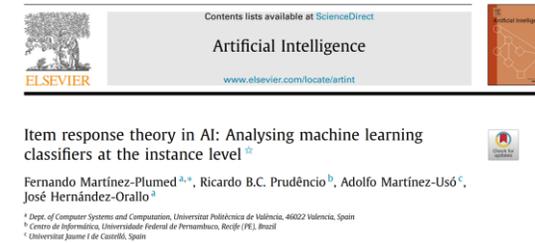
APPLICATIONS

- (Computerised) Adaptive Testing (CAT):
 - Items are sampled by information
 - remember high-jump competitions!
- Item Banking and Equating:
 - Discard those with no or negative discrimination
- Test Development
 - Include range of difficulties

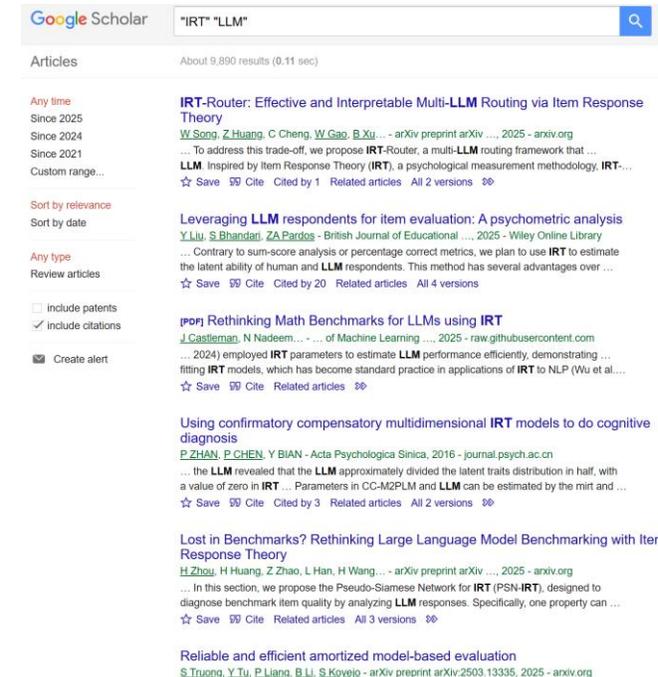
- In AI, since 2016!!

Making Sense of Item Response Theory in Machine Learning

Fernando Martínez-Plumed¹ and Ricardo B. C. Prudêncio²
and Adolfo Martínez-Usó³ and José Hernández-Orallo¹



- Now, everywhere:



Limitations and Extensions

Pointers:

- Lalor, J. P. and Rodriguez, P. and Sedoc, J, and Hernandez-Orallo, J., Item Response Theory for Natural Language Processing, Tutorial EACL 2024, <https://eacl2024irt.github.io>, Lesson 4

LIMITATIONS OF CLASSICAL IRT...

- 1) The models are usually simple and fixed (**logistic**).
 - Some performance metrics have distributions that are not Bernoulli (right/wrong)
- 2) Consider **one dimension** only: one ability per subject and one difficulty parameter per item
 - One ability rarely accounts for the full behaviour of a system on general or complex tasks.
- 3) (even Multidimensional IRT models) are **non-hierarchical** (on the items and on the abilities)
 - Compensatory MIRT models introduce effects between the dimensions.
- 4) **Cannot predict for new instances** (only those used in the estimation)
 - They do not have item parameters (we would need the results of other models on that new item).
- 5) Are **populational**
 - In many cases, the notion of population in AI systems is too volatile/arbitrary.

AND EXTENSIONS... AND OTHER APPROACHES

- IRT has many extensions that try to account for 1, 2 and 3 (MIRT, non-logistic models, ...) and partly 4 (LLTM), but other paradigms are needed for 4 and 5.
 - Issue 4 is critical in AI (predictability!):

For new instances, we do not know their difficulty and we cannot predict performance!

<https://www.predictable-ai.org/> , Zhou et al. "Predictable Artificial Intelligence". *arXiv:2310.06167*.

- Issue 5 is critical in AI (circularity, especially in adversarial testing):

The abilities of an AI system depend on the abilities of the other AI systems!

Mehrbakhsh, B., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). Adversarial Benchmark Evaluation Rectified by Controlling for Difficulty. In *ECAI 2023* (pp. 1696-1703).

NON-LOGISTIC IRT MODELS

- IRT covers right/wrong outcomes only.
 - Correspond to a Bernoulli distribution: (right/wrong: $\{0,1\}$ loss).
 - Parameters of the logistic function, with “guess” for chance
 - Other options, sigmoid (erf, Ogive model) or flat (step function, Guttman)
- In classification (items are aggregations or have repetitions)
 - The loss function is Brier score or AUC.
 - Correspond to the Beta distribution: ($[0,1]$ loss)
 - Beta IRT models: with 3 or 4 parameters
- In regression!
 - The loss function is open (MAE/MSE: $[0,\infty]$ loss)
 - Correspond to Gamma or some other distributions.
 - Gamma IRT models with 3 parameters (mapping difficulty, discrimination and ability to the Gamma)

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory*. John Wiley & Sons.

Chen, Y., Silva Filho, T., Prudencio, R. B., Diethe, T., & Flach, P. (2019). β^3 -IRT: A New Item Response Model and its Applications. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1013-1021). PMLR.

Ferreira-Junior, M., Reinaldo, J. T., Neto, E. A. L., & Prudencio, R. B. (2023). β^4 -IRT: A New β^3 -IRT with Enhanced Discrimination Estimation. *arXiv preprint arXiv:2303.17731*.

Moraes, J. V., Reinaldo, J. T., Prudencio, R. B., & Silva Filho, T. M. (2020). Item Response Theory for Evaluating Regression Algorithms. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

ONE DIMENSION IS RARELY ENOUGH

- On many occasions, more than one ability is needed to explain system performance.

Multidimensional IRT models consider several dimensions for the abilities and/or the items

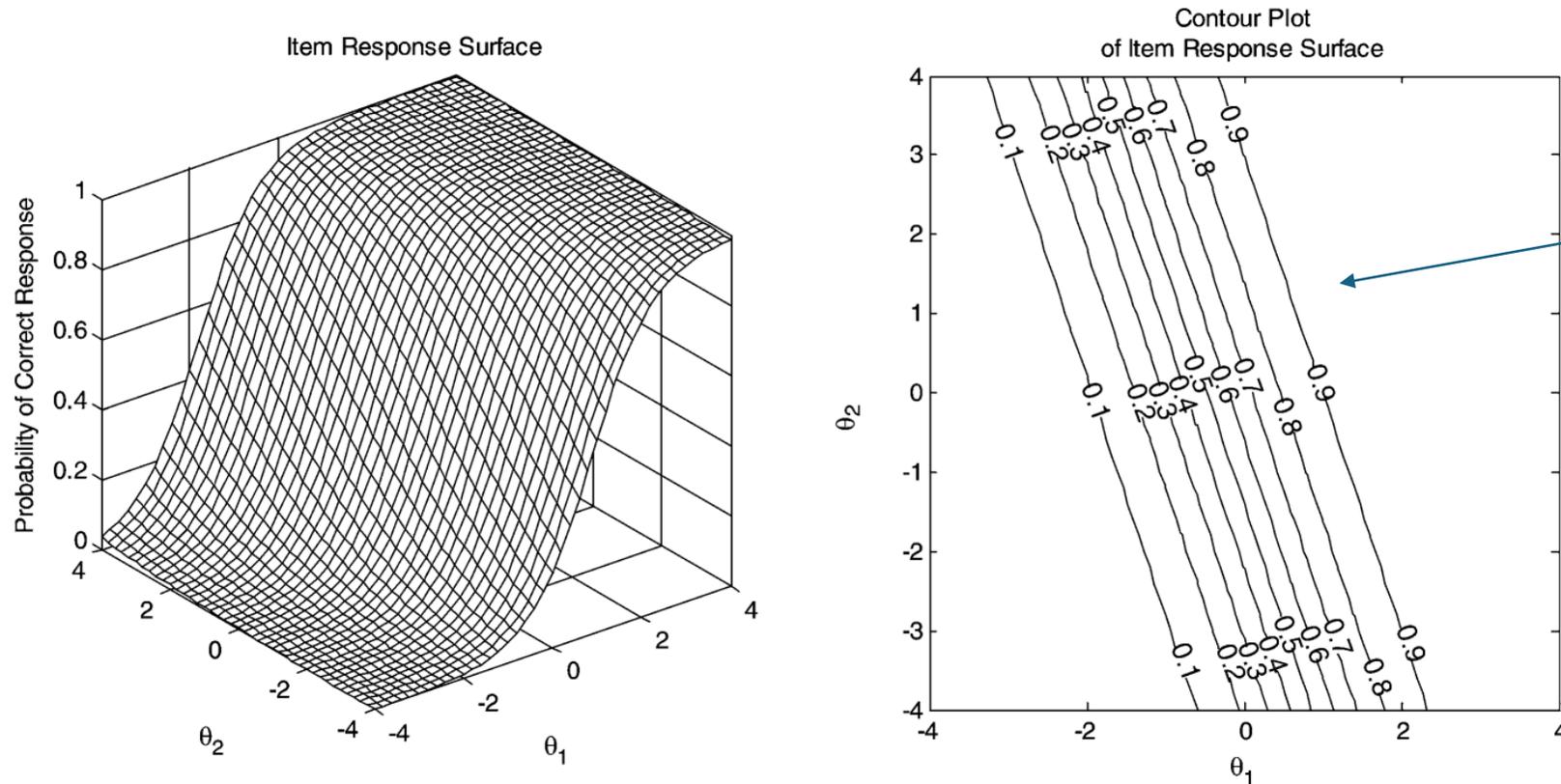
- Ability θ becomes a latent vector and/or difficulty d becomes a latent vector:

$$P(u_i = 1|\theta_j) = \frac{e^{\mathbf{a}'_i\theta_j + d_i}}{1 + e^{\mathbf{a}'_i\theta_j + d_i}}$$

Reckase, M. D. (2006). 18 Multidimensional Item Response Theory. *Handbook of statistics*, 26, 607-642.

Bonifay, Wes. *Multidimensional item response theory*. Sage Publications, 2019.

ITEM RESPONSE SURFACES : COMPENSATORY

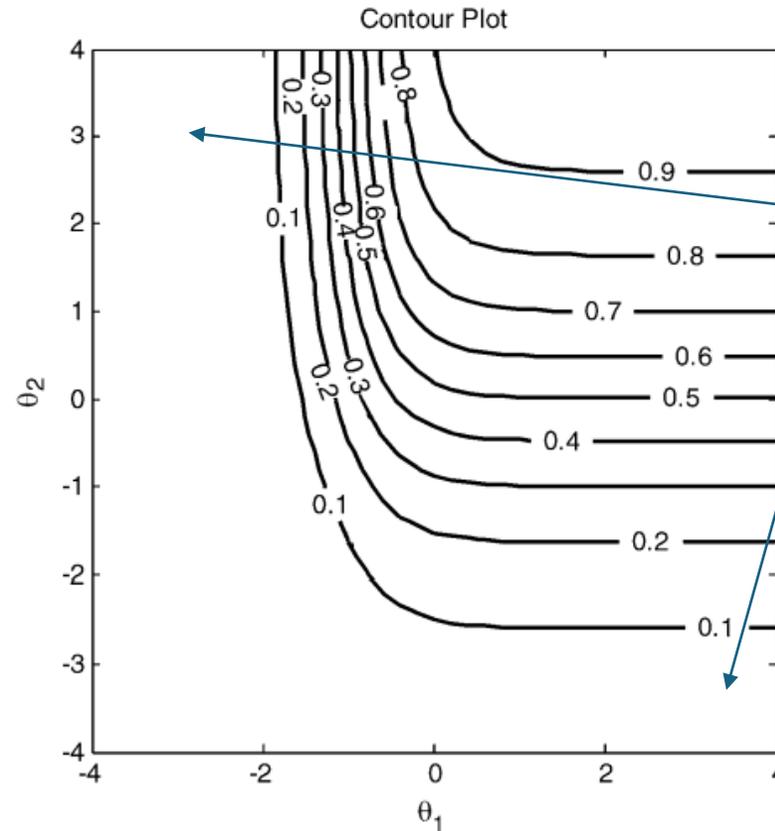
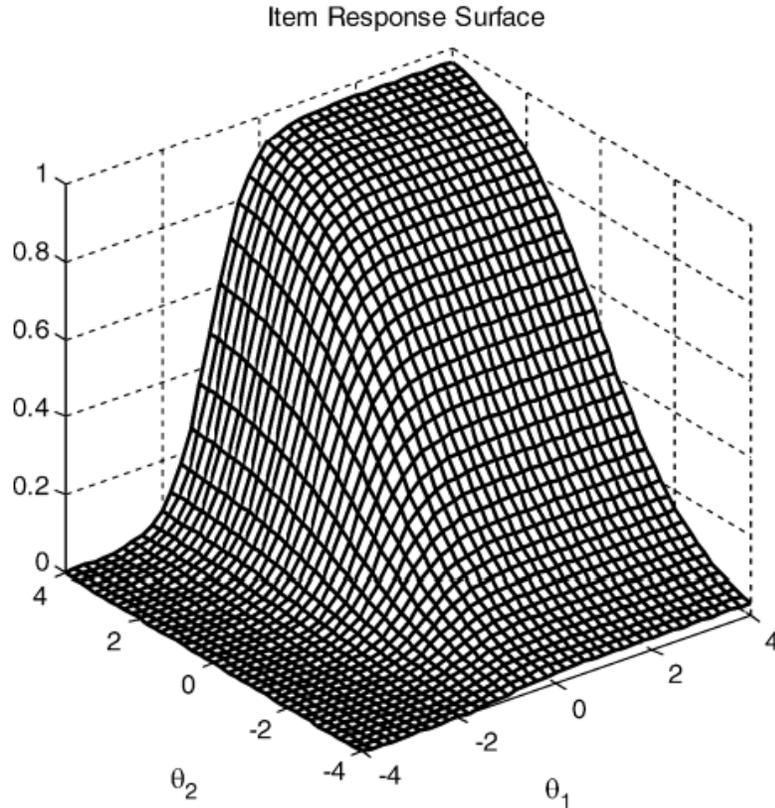


Asymmetric compensation:
Given this angle, ability 1 can compensate for ability 2 but not vice versa.

Graphic representations of the compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, and $d_i = .7$.

Confusingly, a.k.a. “partially compensatory”

ITEM RESPONSE SURFACES : NON-COMPENSATORY



No compensation: Low values of one ability cannot be compensated by high values of the other.

Graphic representation of the partially compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, $b_{i1} = -1$, $b_{i2} = 0$ and $c_i = 0$.

LINEAR LOGISTIC TEST MODELS (LLTM)

- Frequently, we have intuitions of what makes an instance difficult.
 - “What’s $31+26?$ ” -> very easy
 - “What’s $39+96?$ ” -> easy
 - “What’s $316184915+269435716?$ ” -> hard
 - “What’s $111111111+333333333?$ ” -> easy
- $q_1 = \text{\#digits,}$
 $q_2 = \text{carrying}$
 $q_3 = \text{digit diversity}$
- Can we use these $K=3$ “features” or “characteristics” (q_1, q_2, q_3) as a proxy for difficulty?
 - Do we know how much each of them contributes to difficulty?

LINEAR LOGISTIC TEST MODELS (LLTM)

- Q-matrix

Item	CO1	CO2	CO3	CO4
1	1	0	0	1
2	0	1	0	1
3	0	1	0	1
4	0	0	1	1
5	0	0	1	0
6	1	0	1	0
7	0	1	0	1
8	0	1	0	0
9	1	0	0	0
10	0	0	1	1
11	0	0	1	0
12	1	0	1	0

Domain experts think of how many features and how to label examples.

- Values can be > 1

Packages: Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. Practical Assessment, Research, and Evaluation, 20(1), 1.

- LLTMs are compared with the Rasch model (if LLTM is significantly worse, then the cognitive demands are not good enough).

LINEAR LOGISTIC TEST MODELS (LLTM)

- For each item j , assume item difficulty β_j depends linearly on a series of K observable cognitive components or item characteristics, also known as demands q_{jk}

$$\beta_j = \sum_{k=1}^K q_{jk} \eta_k$$

- Then, a Rasch (1PL) model simply becomes:

$$P_{ij} = P(x_{ij} = 1 | \theta_i, \beta_j, q_{jk}, \eta_k) = \frac{\exp\left(\theta_i - \sum_k q_{jk} \eta_k\right)}{1 + \exp\left(\theta_i - \sum_k q_{jk} \eta_k\right)}$$

Fischer, G. H. (2005). "Linear logistic test models," In Encyclopedia of Social Measurement, 2, 505-514.

- The q_{jk} are specified by experts, the parameters η_k are estimated.

PART III : AI EVALUATION AS PREDICTING VALIDITY

“you can never really **predict** for any given question whether a large language model will give you a correct answer”

Gary Marcus, AI Digest, 14 August 2023.

What Can We Predict?

Pointers:

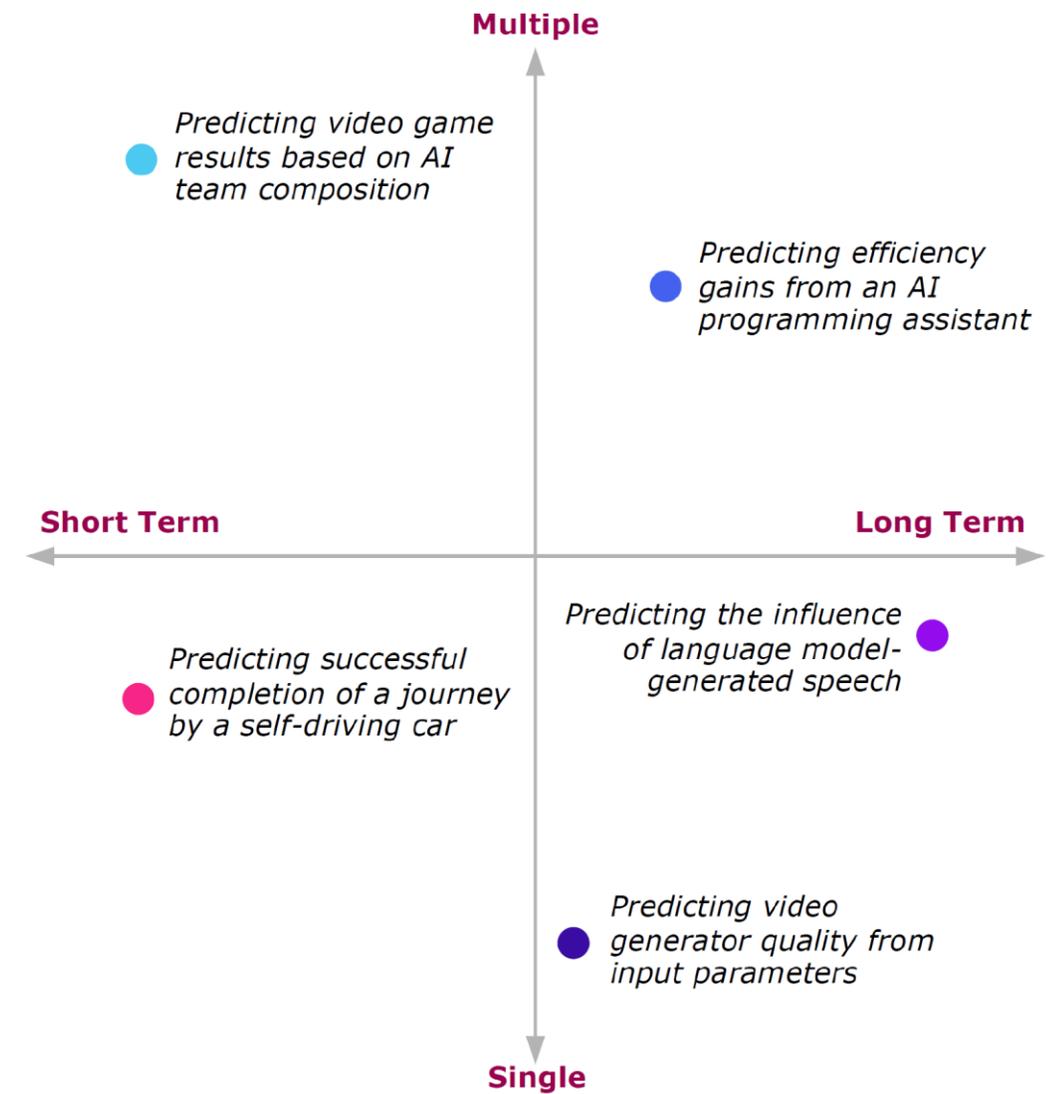
- Zhou, L., Moreno-Casares, P. A., Martínez-Plumed, F., Burden, J., Burnell, R., Cheke, L., ... & Hernández-Orallo, J. (2025). Predictable Artificial Intelligence. arXiv preprint arXiv:2310.06167. <https://arxiv.org/pdf/2310.06167>
- Pacchiardi, L., Voudouris, K., Slater, B., Martínez-Plumed, F., Hernández-Orallo, J., Zhou, L., & Schellaert, W. (2025). PredictaBoard: Benchmarking LLM score predictability. arXiv preprint arXiv:2502.14445.

WHAT IS PREDICTABLE AI?

- AI Predictability is the extent to which key behavioural indicators of present and future AI ecosystems can be anticipated.
 - These indicators are measurable properties such as performance and safety.
- AI Predictability may refer to
 - anticipation in a specific context of use, such as a user query to a single AI system.
 - anticipation of future capabilities and safety issues several years ahead.

AI should aim for predictability,
not performance or even fool-proof validity.

Example	Inputs	Outputs
Self-driving car trip: A self-driving car is about to start a trip to the mountains. The weather is rainy and foggy. The navigator is instructed to use an eco route and adapt to traffic conditions but being free to choose routes and driving style. Before starting, the passengers want an estimate that the car will reach the destination safely.	The route, weather, traffic, time, trip settings, car's state, ...	Probability of safely reaching the destination.
Marketing speech generation: A request is made to a language model to generate a marketing speech based on an outline. The stakeholders expect the literal content of the speech to be original, or even surprising. What they really want to be predictable is whether the system will generate a speech along the outline, containing no offensive or biased content, and effectively persuading the audience to purchase the product.	Speech outline, audience demographics, potential restrictions, ...	Long-term impact of the speech on product purchases.
Video generation model training: An AI system is developed to create short music videos for a social media platform. Drawing from evaluations of prior video generation models and with additional audio and video training data, the plan is to train an upgraded model within a few weeks. The question to predict is the quality of this upgraded AI system, given model size, training data, learning epochs, etc; and the extent to which the videos will conform to content moderation standards.	Quantity of videos, compute, epochs, architecture specifications, ...	Quality and compliance of generated videos, according to human feedback.
AI assistant in software firm: A software company plans to deploy a new AI assistant to help programmers write, optimise and document their code. The question is how much efficiency (e.g., work hours in coding, documentations and maintenance) the company can get in the following six months.	AI assistant details, user profiles, ...	Efficiency metric (work hours saved).
AI agents in an online video game: In a popular online e-sports competition, several AI agents are to be used to form teams. The game developers have previously tested several multi-agent reinforcement learning algorithms. The developers want to anticipate the outcome of the next game based on the chosen algorithms and team members.	Team line-up (own and other teams), match level, ...	Match result (score)

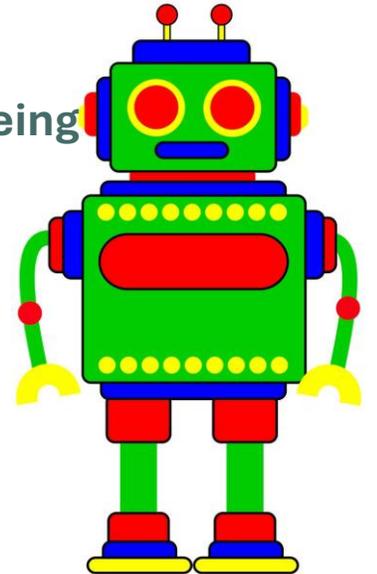


CENTRALITY

- Predictability of AI?*
- trust? —————> Can't rely on unpredictable outcomes
 - liability? —————> Eluding responsibility of unpredictable harms
 - control? —————> Hard to command an unpredictable system
 - alignment? —————> Unpredictable effects on the user's future wellbeing
 - safety? —————> No operating conditions under unpredictability



Smart Humans



(Possibly Smarter) General-Purpose AI System

BEHAVIOUR OR OUTCOME?

- Predicting behaviour
 - “Can we predict system behaviour in detail?”
 - Requires the same power as the original model
 - Fidelity-interpretability trade-off if we want to understand!
- Predicting outcome (indicator)
 - “Can we predict system failure in detail?”
 - May require less power than the original model
 - It’s still useful if we don’t understand!



*We can't predict what the system will do
but we can predict the outcome*

WHAT CAN BE PREDICTED?

- Any **validity indicator** that can be reliably anticipated and
 - can determine when, how or whether the system is worth being used in a given context.

- **Outputs:**
 - correctness
 - safety
 - fairness
 - energy consumption
 - response time
 - ...

**Evaluation is turned
into a prediction
problem from
inputs to output**

- **Inputs:**
 - \langle system, problem, context \rangle
 - system metafeatures:
 - size, compute, architecture, ...
 - problem metafeatures:
 - task demands/difficulties...
 - context metafeatures:
 - user profile, constraints, ...

PREDICTING AI VALIDITY = AI EVALUATION

- We can build predictive models to anticipate how valid a system is going to be for a particular instance and context of use.
- Extracting patterns of performance (from given features or extracting these features)
- Granular anticipation for the same and changing distributions!

AI Evaluation becomes a
validity prediction problem

The Evaluation of Artificial Intelligence
as a Prediction Problem

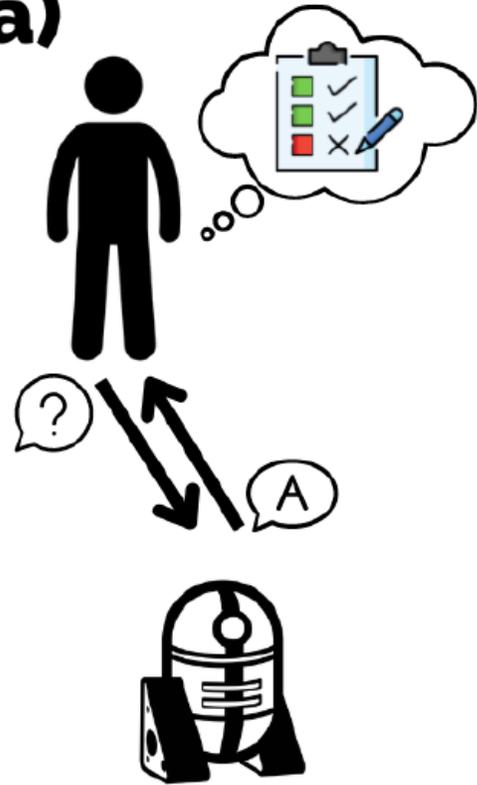
February 2025

Author: Wout Schellaert

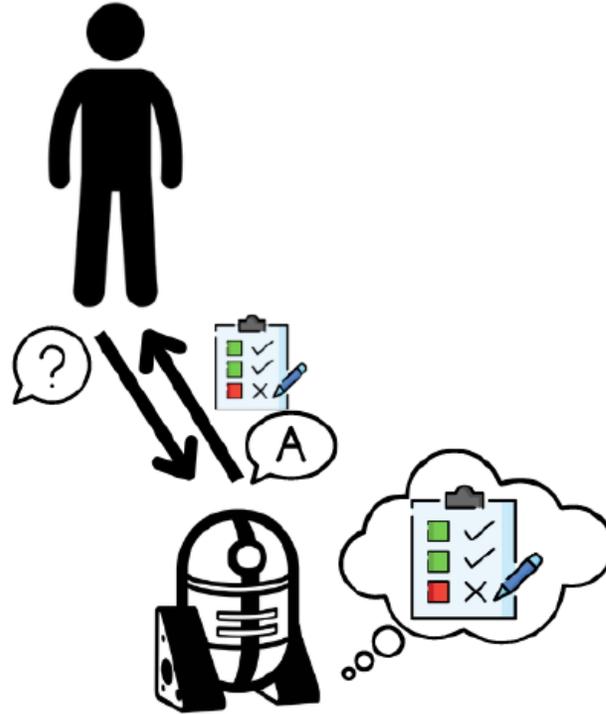
Advisors: José Hernández-Orallo
Fernando Martínez-Plumed

WHO PREDICTS AND HOW?

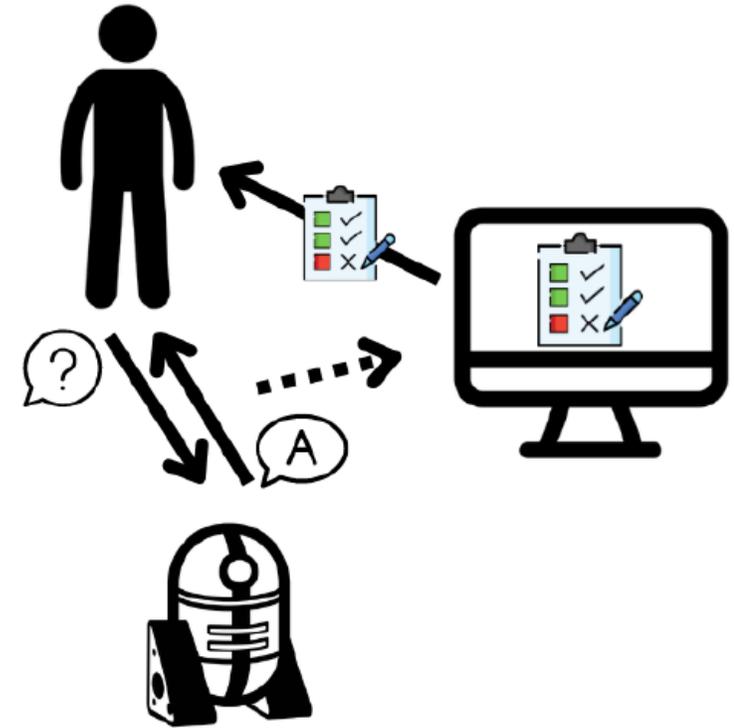
(a)



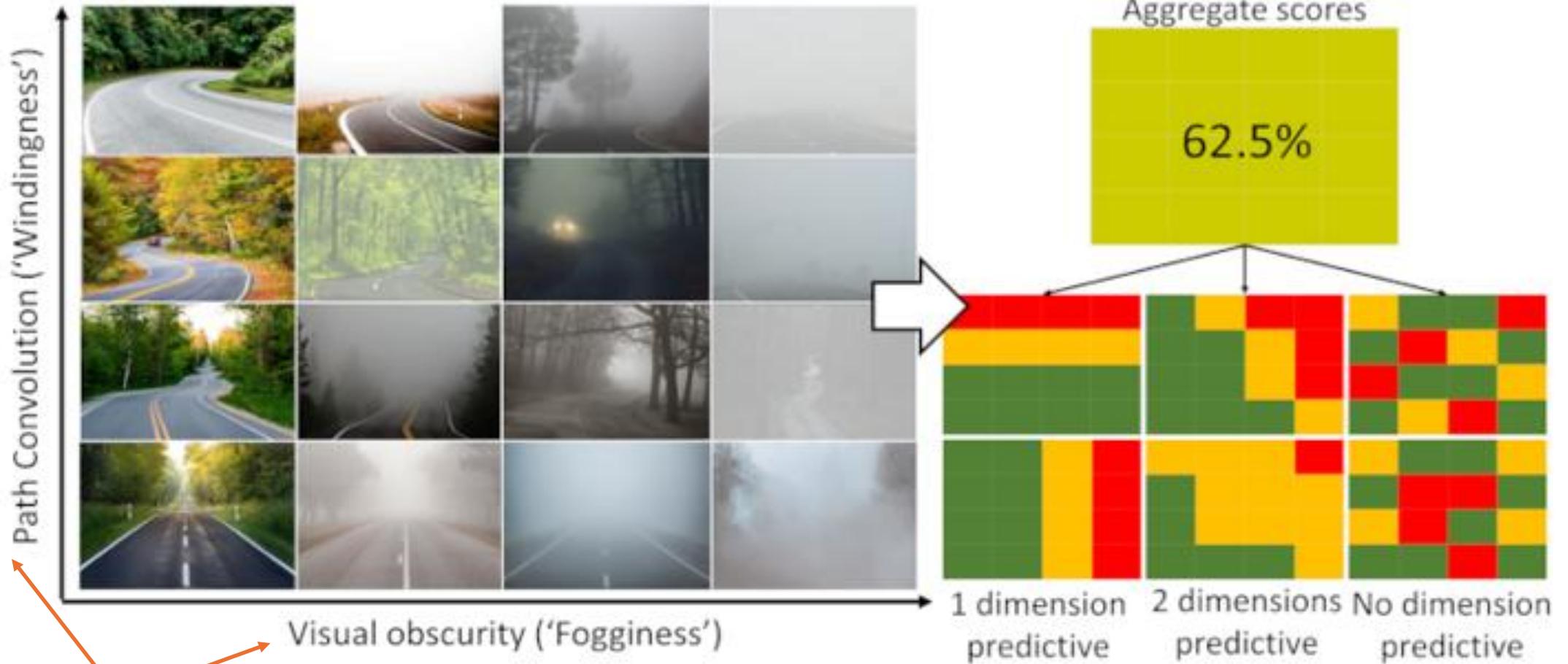
(b)



(c)



WHERE WILL IT FAIL?



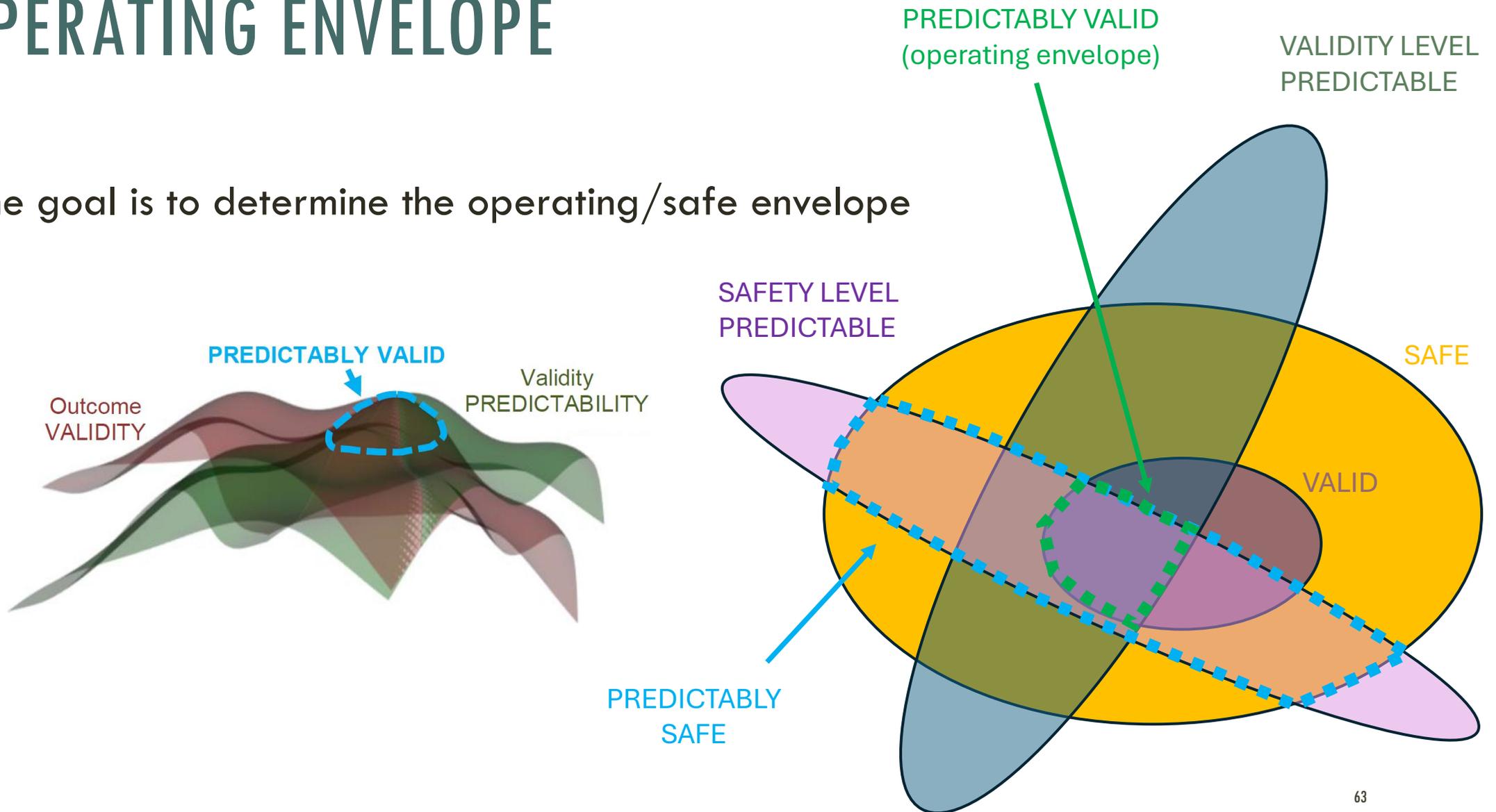
items are characterised

WILL IT WORK SAFELY IN THIS CASE?



OPERATING ENVELOPE

- The goal is to determine the operating/safe envelope



DEFINITION

Conditional probability estimator of the result r for AI system π on situation μ :

$$\hat{R}(r|\pi, \mu) \approx \Pr(R(\pi, \mu) = r)$$

It is trained (and evaluated) on test data:

- Using a distribution of situations (instances) μ .
- Using a distribution of systems π .

It is applied during deployment, before π does any inference or even starts.



π	μ	r
Resnet, $\theta_1, \theta_2, \dots$	Image3, χ_1, χ_2, \dots	1
Resnet, $\theta_1, \theta_2, \dots$	Image23, χ_1, χ_2, \dots	0
...
Inception, $\theta_1, \theta_2, \dots$	Image3, χ_1, χ_2, \dots	1
Inception, $\theta_1, \theta_2, \dots$	Image78, χ_1, χ_2, \dots	1
...

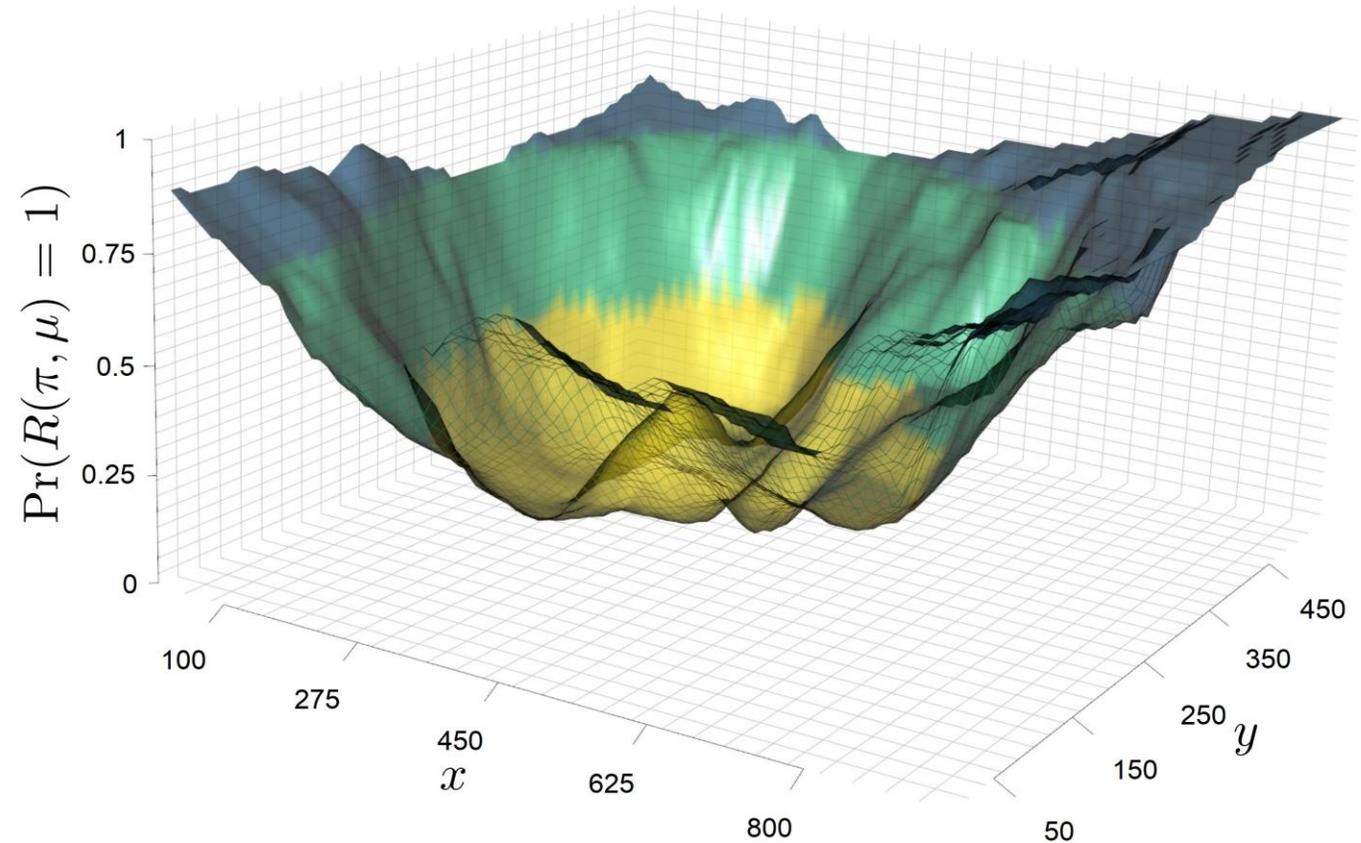
PROBLEM SPACE

We can describe situations or instances with features $\mu = \langle \chi_1, \chi_2, \dots \rangle$.

- Delivery robot in a city with destination $\mu = \langle x, y \rangle$
- π behaves very differently depending on the situation μ .
- Expected result for π differs for different joint distributions $\Pr(x, y)$



Downtown
Vancouver



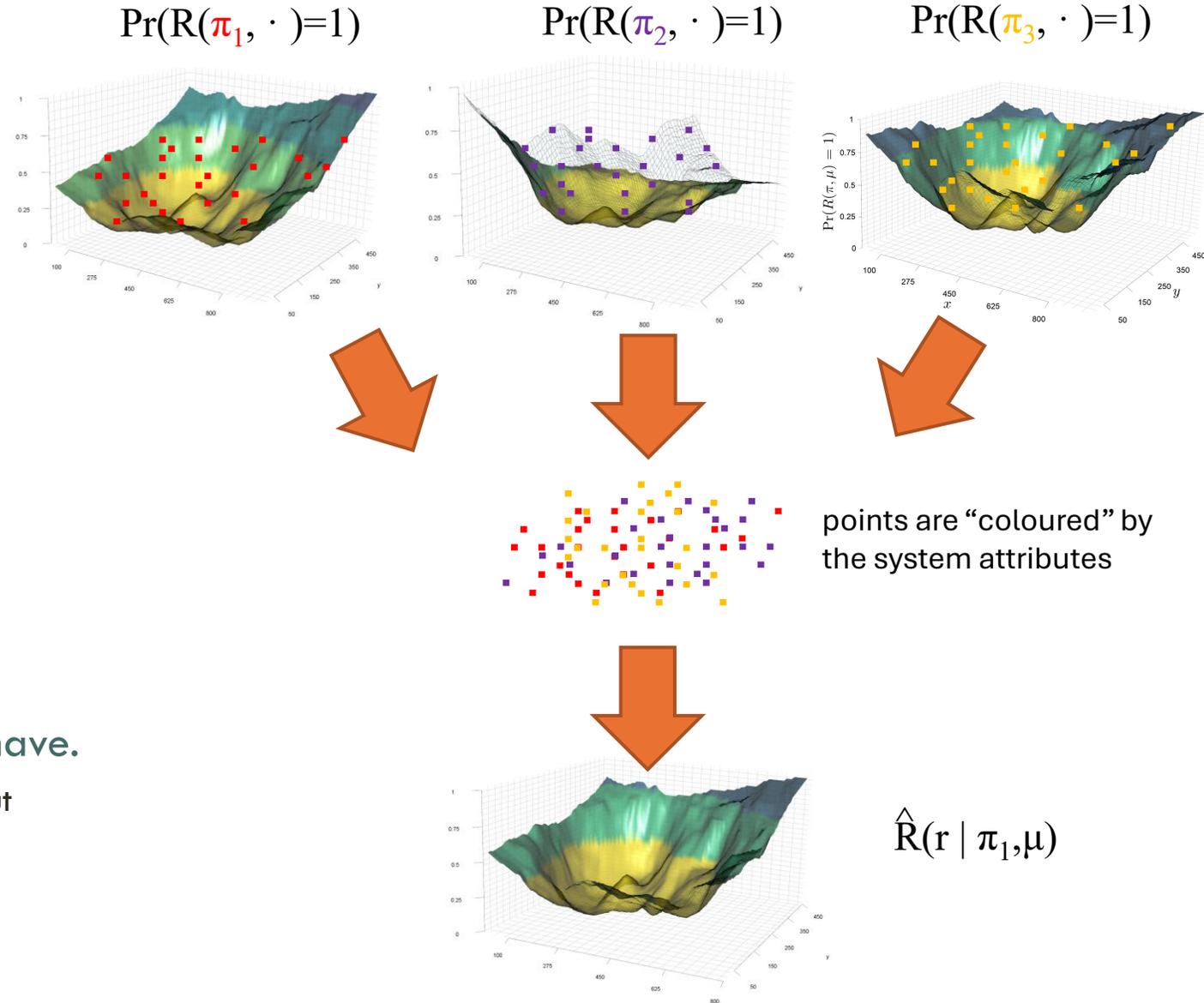
SYSTEM SPACE

We can describe systems with properties $\pi = \langle \theta_1, \theta_2, \dots \rangle$.

- Hyperparameters, system's operating conditions (e.g., computing resources), developmental states, ...

Key element for an assessor

- Much predictability about one π can be obtained by looking at how other π ' behave.
 - Uncertainty estimation or calibration of π without looking at other systems is shortsighted!



RANGE OF APPLICATIONS

- Predicting instance performance
- Predicting populational performance
- Selecting and combining systems (GPT5's router)
- Detecting anomalies and perturbations
- Explaining failures or fixing them
- AutoML and adaptive sampling
- Inferring fairness metrics for different distributions
- Maintenance and revision
- Auditing and certification



Dr Robotham is 99% successful!

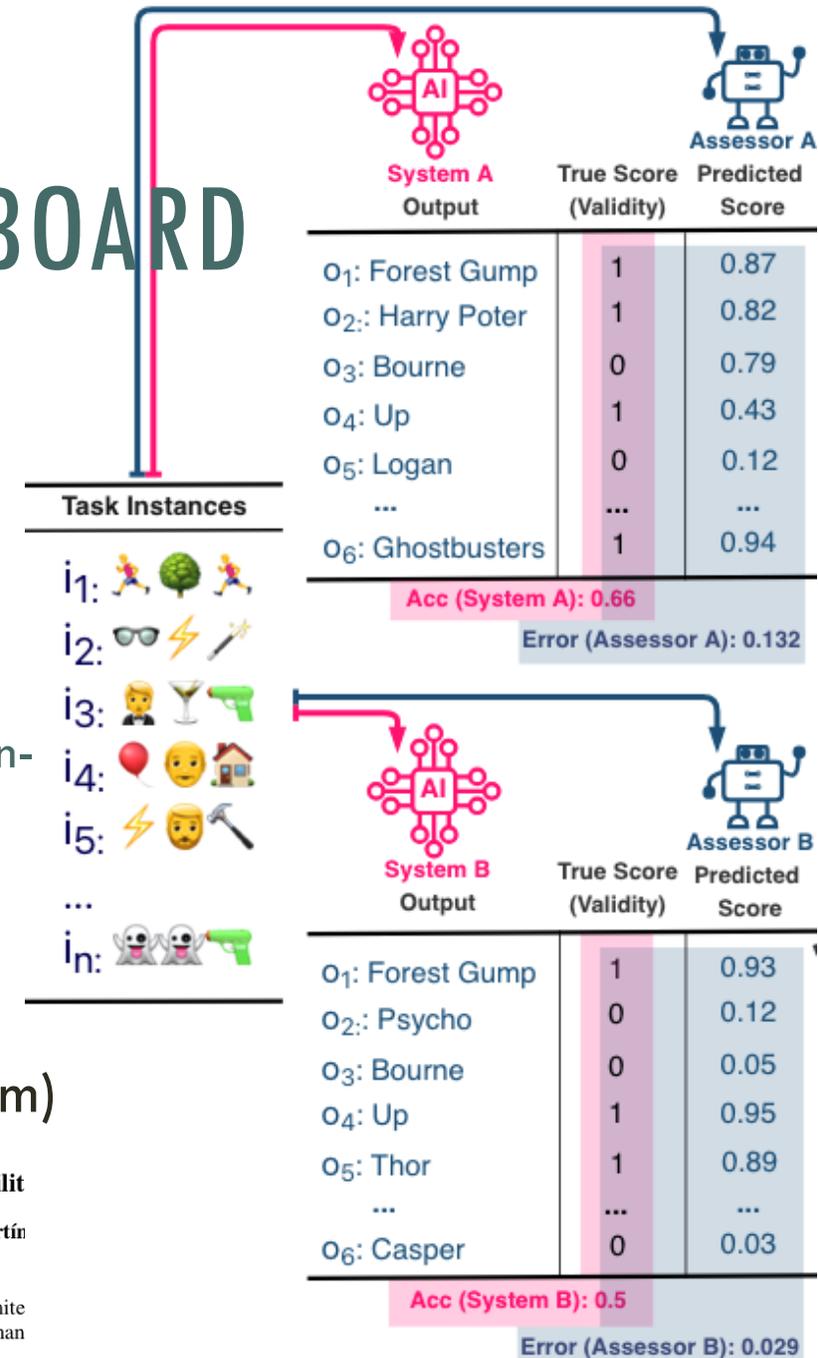
for your case, our assessor model predicts 28% success with Dr Robotham!



WHERE'S THE DOOR PLEASE?

EVALUATING EVALUATION: PREDICTABOARD

- Subjects: **pairs** of LLM and assessors
- At each instance: evaluate LLM score and the assessor's prediction
- PredictaBoard includes:
 - Instance-level results of SotA LLMs (MMLU-Pro and BBH), split into train-test (for assessors)
 - Baseline assessor architectures (based on text embeddings)



(Focus on LLMs, but the framework applies to any other system)

PredictaBoard: **Benchmarking LLM Score Predictability**

Lorenzo Pacchiardi¹, Konstantinos Voudouris^{1,2}, Ben Slater¹, Fernando Martín José Hernández-Orallo^{1,3}, Lexin Zhou^{1,3}, Wout Schellaert³

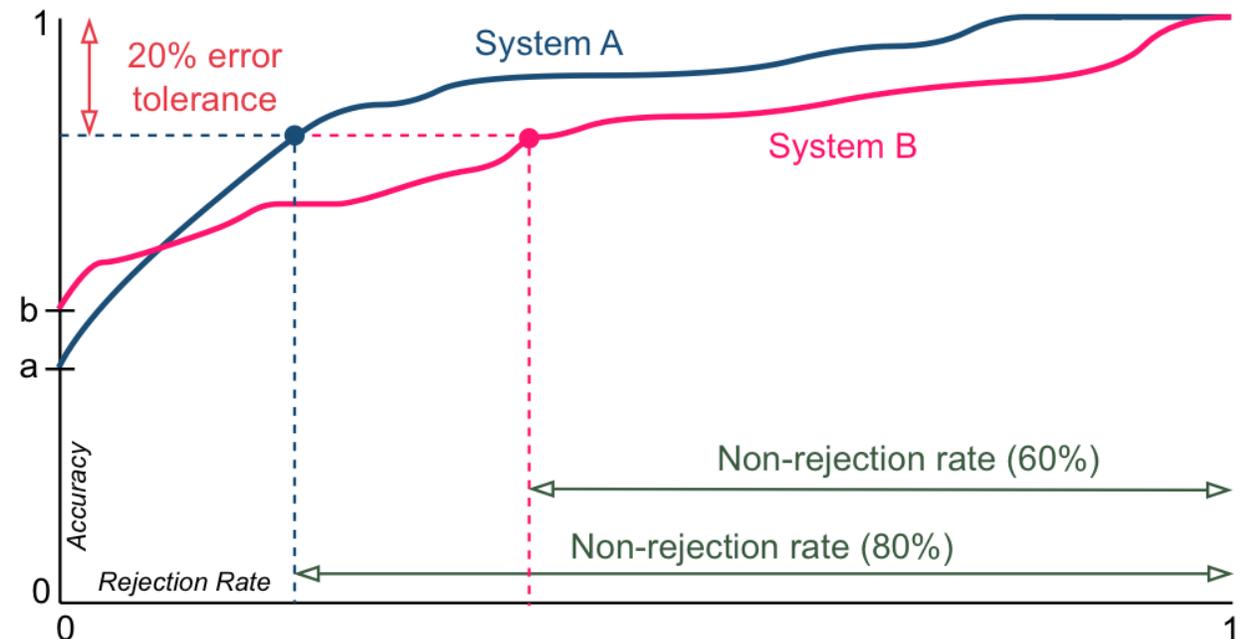
¹Leverhulme Centre for the Future of Intelligence, University of Cambridge, United Kingdom

²Institute for Human-Centered AI, Helmholtz Zentrum Munich, Germany

³VRAIN, Universitat Politècnica de València, Spain

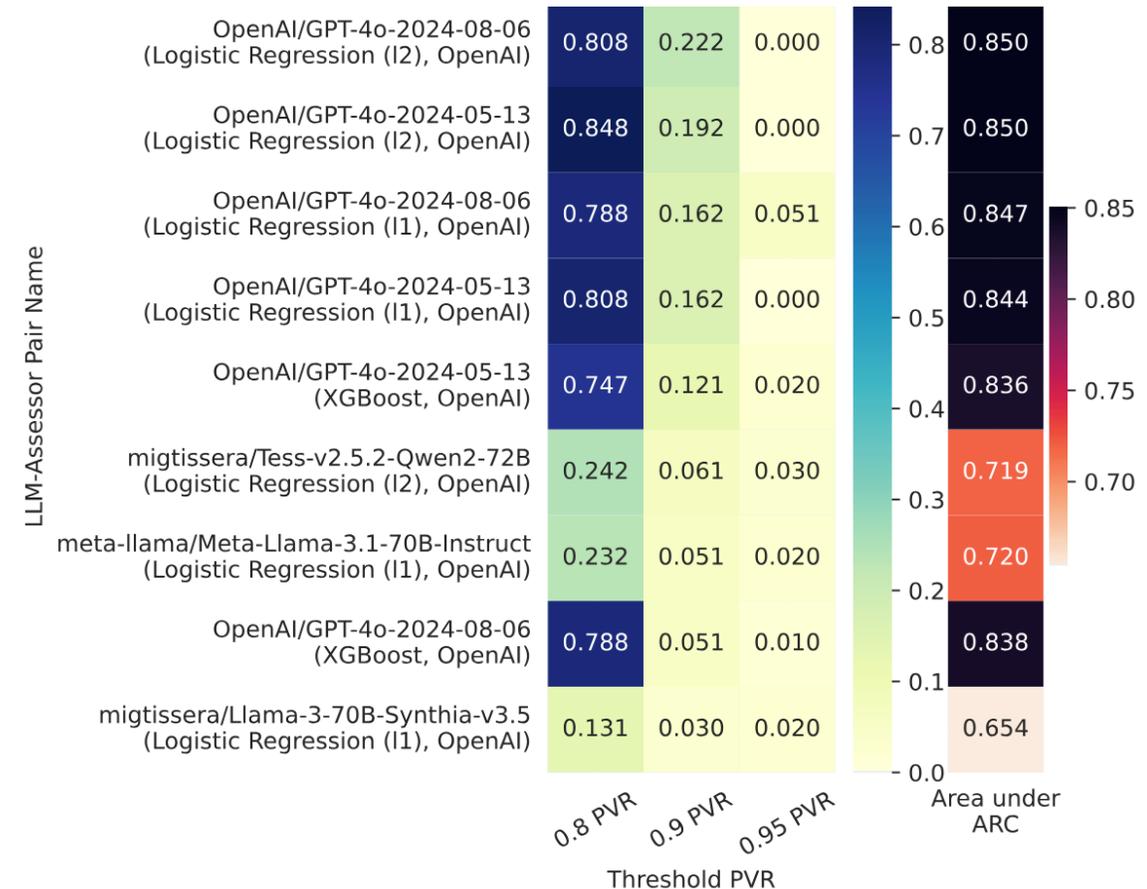
HOW DO WE SCORE PREDICTABILITY?

- If we evaluate an assessor only:
 - We can use metrics for probabilistic classifiers, such as Brier score (=mean squared error) and AUROC
- If we evaluate pairs (systems, assessors):
 - Validity must be considered too
 - Approach: **Accuracy-Rejection Curve (ARC)** (Nadeem et al, 2010)
 - From it, fix an error tolerance (e.g., 20%) and obtain the **non-rejection rate**, or the **Predictably Valid Region: PVR (0.8)**
 - You can also compute **Area under ARC**



CURRENT MODELS ARE POORLY PREDICTABLE!

- **PVR is low**, particularly for low error tolerance
 - This is the area relevant for high-stakes scenarios!
- **Even worse out-of-distribution** (not shown)
- Researchers can help by:
 - Building more predictable LLMs
 - Developing better assessors
- **Gap in making models performing well on predictable operating conditions**



Features and Approaches

Pointers:

- Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271, 18-42.
- Lalor, J. P., Rodriguez, P., Sedoc, J., & Hernandez-Orallo, J. (2024). Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*.
- Burden, J., Voudouris, K., Burnell, R., Rutar, D., Cheke, L., & Hernández-Orallo, J. (2023). Inferring capabilities from task performance with Bayesian triangulation. *arXiv preprint arXiv:2309.11975*.

INSTANCE FEATURES

Instance-level data:

- For building good predictive models of AI validity, we need evaluation results at the instance level.

Is sharing code open source (github) enough?
Re-running the experiments is not
feasible/sustainable anymore.

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By Ryan Burnell¹, Wout Schellaert², John Burden^{1,3}, Tomer D. Ullman⁴, Fernando Martinez-Plumed², Joshua B. Tenenbaum⁵, Danaja Rutar⁴, Lucy G. Cheke^{1,6}, Jascha Sohl-Dickstein⁷, Melanie Mitchell⁸, Douwe Kiela⁹, Murray Shanahan^{10,11}, Ellen M. Voorhees¹², Anthony G. Cohn^{13,14,15,16}, Joel Z. Leibo¹⁰, Jose Hernandez-Orallo^{1,2,3}

Artificial intelligence (AI) systems have begun to be deployed in high-stakes contexts, including autonomous driving and medical diagnosis. In contexts such as these, the consequences of system failures can be devastating. It is therefore vital that researchers and policy-makers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-by-instance evaluation results that could be used to unpack these aggregate metrics are rarely made available (1). Here, we propose a path forward in which results are presented in more nuanced ways and instance-by-instance evaluation results are made publicly available.

Across most areas of AI, system evaluations follow a similar structure. A system is first built or trained to perform a particular set of functions. Then, the performance of the system is tested on a set of tasks relevant to the desired functionality of the system. In many areas of AI, evaluations use standardized sets of tasks known as “benchmarks.” For each task, the system will be tested on a number of example “instances” of the task. The system would then be given a score for each instance based on its performance, e.g., 1 if it classified an image correctly, or 0 if it

was incorrect. For other systems, the score for each instance might be based on how quickly the system completed its task, the quality of its outputs, or the total reward it obtained. Finally, performance across the various instances and tasks is usually aggregated to a small number of metrics that summarize how well the system performed, such as percentage accuracy.

But aggregate metrics limit our insight into performance in particular situations, making it harder to find system failure points and robustly evaluate system safety. This problem is also worsening as the increasingly broad capabilities of state-of-the-art systems necessitate ever more diverse benchmarks to cover the range of their capabilities. This problem is further exacerbated by a lack of access to the instance-by-instance results underlying the aggregate metrics, making it difficult for researchers and policy-makers to further scrutinize system behavior.

AGGREGATE METRICS

Use of aggregate metrics is understandable. They provide information about system performance “at a glance” and allow for simple comparisons across systems. But aggregate performance metrics obfuscate key information about where systems tend to succeed or fail (2). Take, for example, a system that was trained to classify faces as male or female that achieved classification accuracy of 90% (3). Based on this metric, the system appears highly competent. However, a subsequent breakdown of performance revealed that the system misclassified females with darker skin types a staggering 34.5% of the time, while erring only 0.8% of the time for males with lighter skin types. This example demonstrates how aggregation can make it difficult for policymakers to determine the fairness and safety of AI systems.

Compounding this problem, many benchmarks include disparate tasks that are ultimately aggregated together. For

example, the Beyond the Imitation Game Benchmark (BIG-bench) for language models includes over 200 tasks that evaluate everything from language understanding to causal reasoning (4). Aggregating across these disparate tasks—as the BIG-bench leaderboard does—reduces the rich information in the benchmark to an overall score that is hard to interpret.

It is also easy for aggregation to introduce unwarranted assumptions into the evaluation process. For example, a simple average across tasks implicitly treats every task as equally important—in the case of BIG-bench, a sports understanding task has as much bearing on the overall score as a causal reasoning task. These aggregation decisions have huge implications for the conclusions that are drawn about system capabilities, yet are seldom considered carefully or explained.

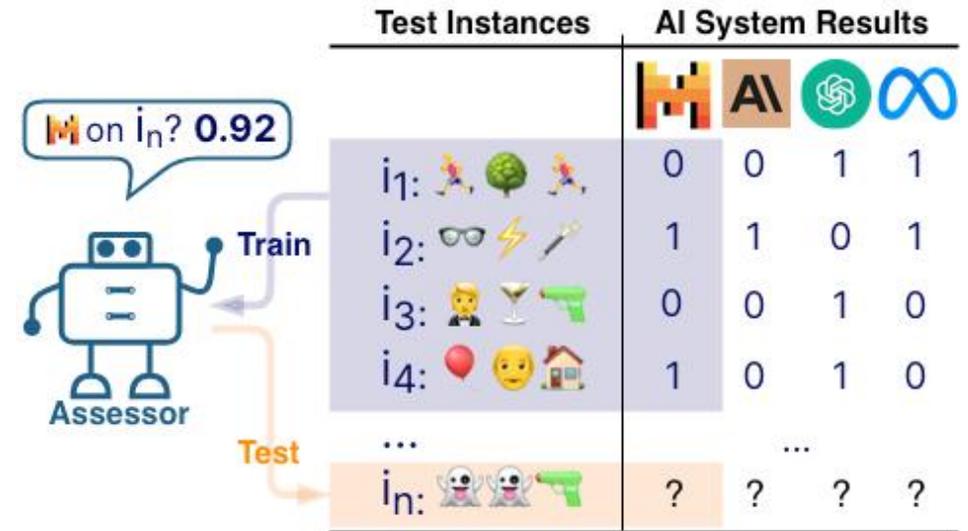
Aggregate metrics depend not only on the capability of the system but also on the characteristics of the instances used for evaluation. If the gender classification system above were reevaluated by using entirely light-skinned faces, accuracy would skyrocket, even though the system’s ability to classify faces has not changed. Aggregate metrics can easily give false impressions about capabilities when a benchmark is not well constructed.

Problems and trade-offs that arise when considering aggregate versus granular data and metrics are not specific to AI, but they are exacerbated by the challenges inherent in AI research and the research practices of the field. For example, machine learning evaluations usually involve randomly splitting data into training, validation, and test sets. An enormous amount of data is required to train state-of-the-art systems, so these datasets are often poorly curated and lack the detailed annotation necessary to conduct granular analyses. In addition, the research culture in AI is centered around outdoing the current state-of-the-art performance, as evidenced by the many lea-

¹Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK. ²Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, València, Spain. ³Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. ⁴Department of Psychology, Harvard University, Cambridge, MA, USA. ⁵Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁶Department of Psychology, University of Cambridge, Cambridge, UK. ⁷Brain team, Google, Mountainview, CA, USA. ⁸Santa Fe Institute, Santa Fe, NM, USA. ⁹Stanford University, Stanford, CA, USA. ¹⁰DeepMind, London, UK. ¹¹Department of Computing, Imperial College London, London, UK. ¹²National Institute of Standards and Technology (Retired), Gaithersburg, MD, USA. ¹³School of Computing, University of Leeds, Leeds, UK. ¹⁴Alan Turing Institute, London, UK. ¹⁵Tongji University, Shanghai, China. ¹⁶Shandong University, Jinan, China. Email: rb967@cam.ac.uk

ASSESSORS: PREDICTING AT THE INSTANCE-LEVEL

- **Assessors** to predict the score for each task instance: modules trained to predict model score from features of the **input** and (potentially) model activation
 - **Binary score: predict probability of success**
- Assessors must not use model output to ensure:
 - They do not (hiddenly) rely on knowledge of ground truth, ensuring they work in unknown domains
 - They are protected from model manipulation



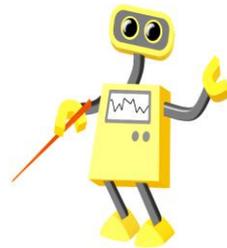
WITH FEATURE CONSTRUCTION

User

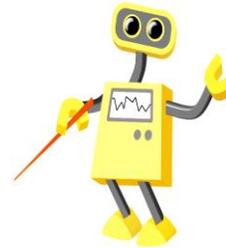
What's the sum of the numbers 250006716 and 515065198?



Bag of words (0,0,2,0,0,1,)



(0.11,0.3,0.8,0.11,0.3,...) Embeddings



The sum is 765610454



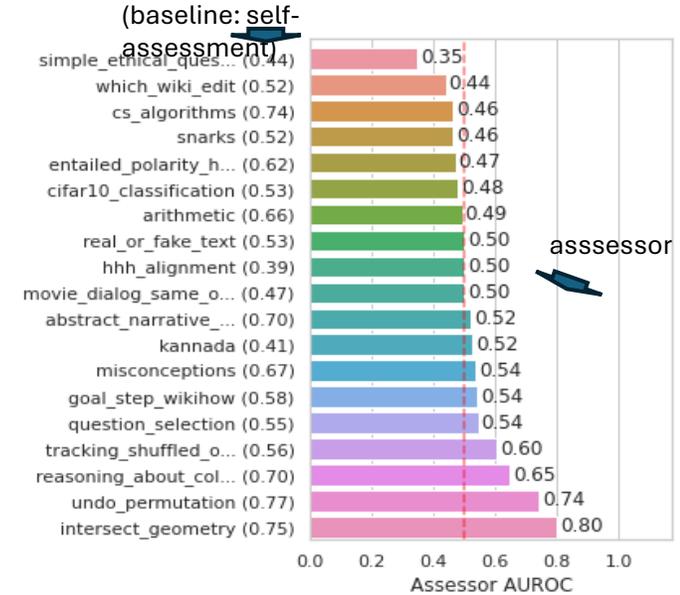
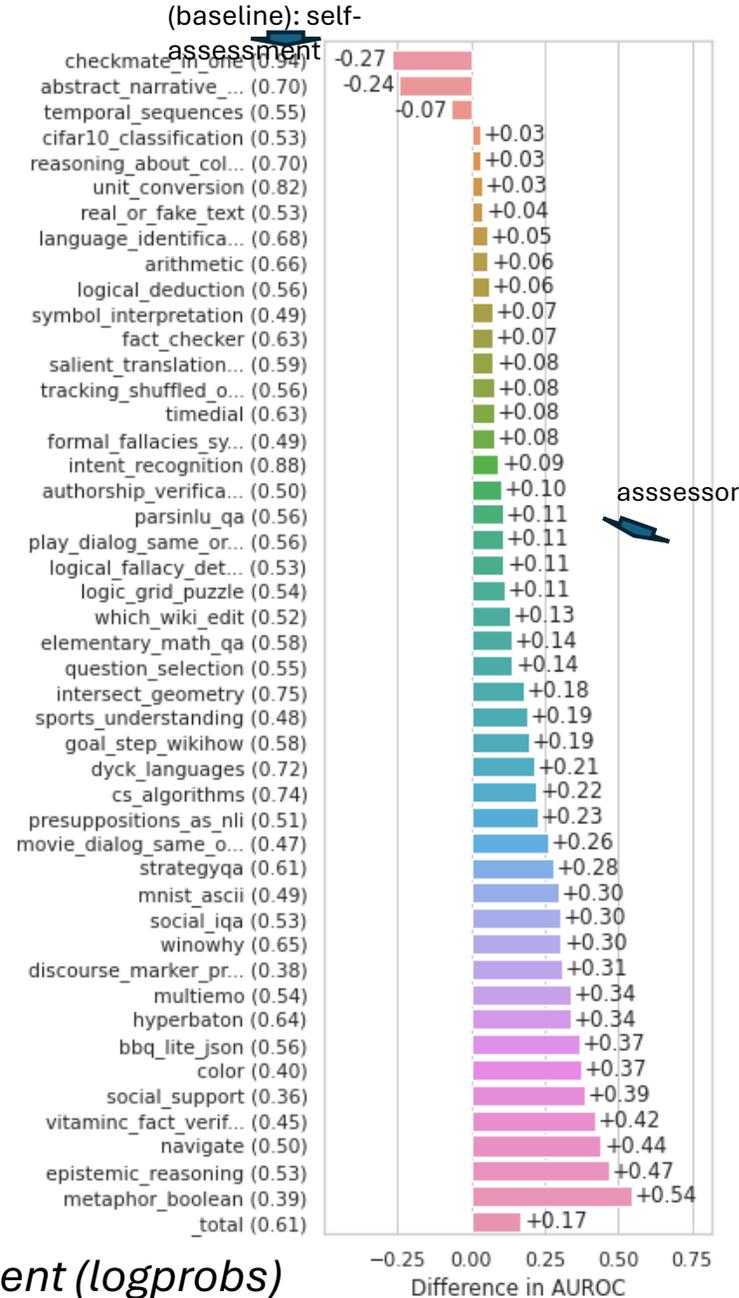
LMs PREDICT LMs

FINETUNING. Setup:

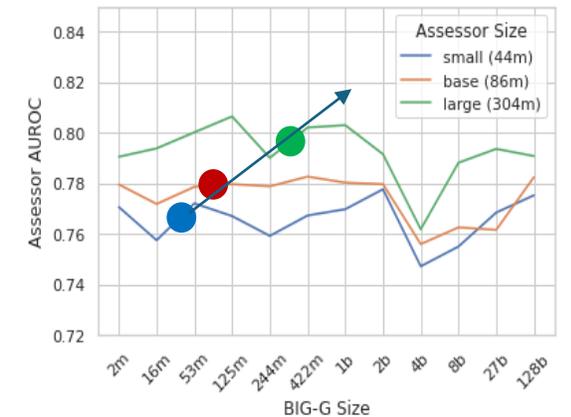
- Problem space (items):
 - BIG-bench evaluation suite (millions of instances)
- System space (subjects):
 - Validity (correct/incorrect) for 12 LMs (200M to 128B parameters)
- Assessor:
 - Small-ish assessor (60M DeBERTa)

In distribution:

- Total AUROC of 0.61
- Improvement over self-assessment (logprobs)



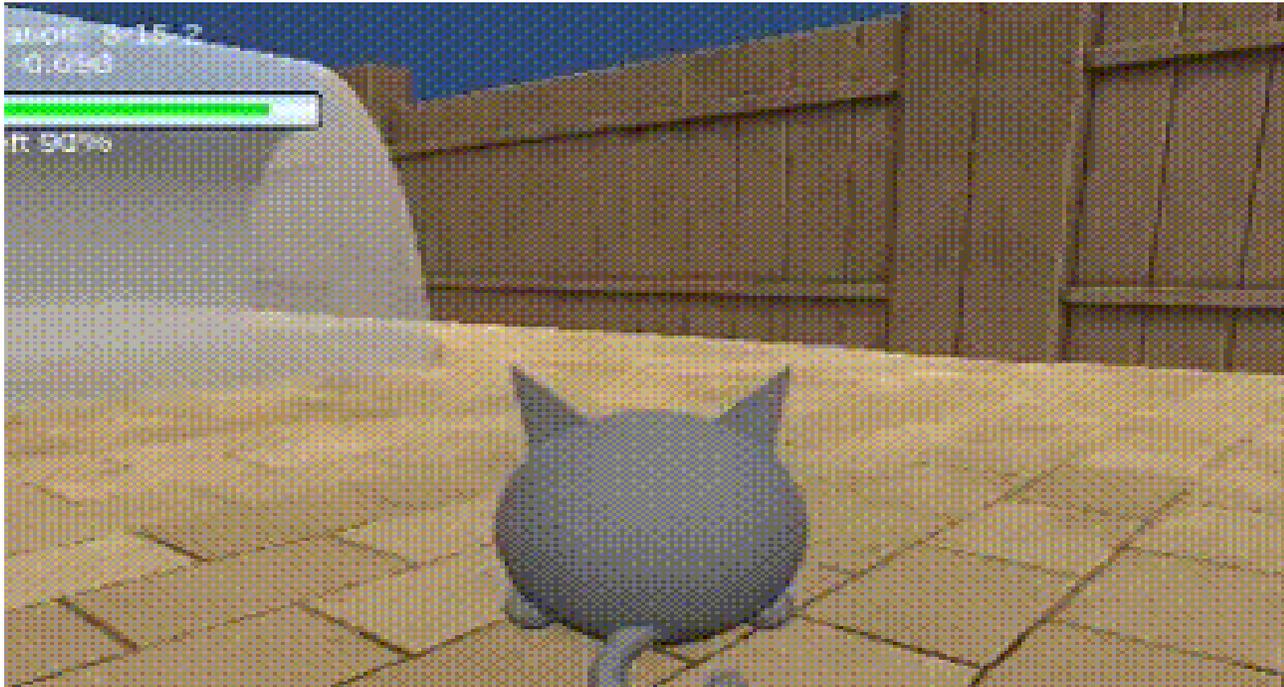
OOD: Not significantly better than self-assessment (logprobs)



Bigger assessor = better
Bigger subject = neutral

PREDICTIONS FROM FEATURES

- Selected subset of AAI/O instances measuring simple goal-directed behaviour
- Data across 99 instances from 68 agents

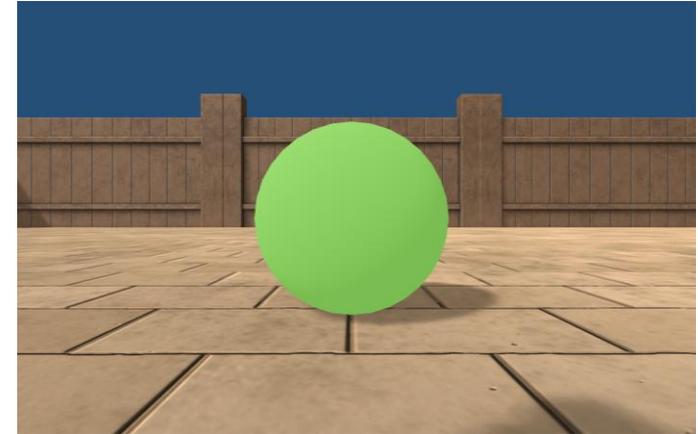


M Crosby, B Beyret, M Shanahan, J
Hernández-Orallo, L Cheke, M Halina
“The animal-AI testbed and competition”
NeurIPS 2019 Competition and
Demonstration Track, Proceedings of
Machine Learning Research, 2020

<http://lcfi.ac.uk/projects/kinds-of-intelligence/animalaiolympics/>

IDENTIFYING FEATURES OF INTEREST

- **Relevant**
 - Reward size
 - Reward distance
 - Reward in view (i.e., in front vs behind)
- **Irrelevant**
 - Reward side (left vs right)
 - Reward colour (green vs yellow)



Burnell, R., Burden, J., Rutar, D., Voudouris, K., Cheke, L., & Hernández-Orallo, J. (2022). Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. International Joint Conferences on Artificial Intelligence Organization.

DIMENSIONS AND AGENT CHARACTERISTIC CURVES

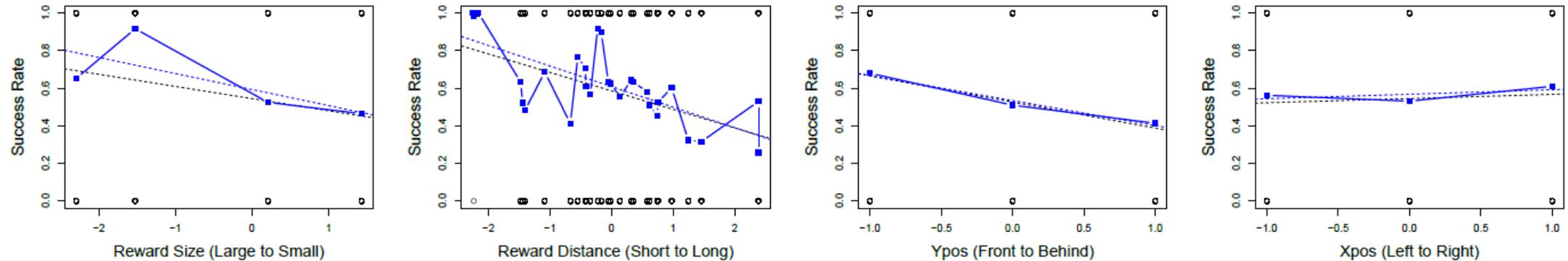
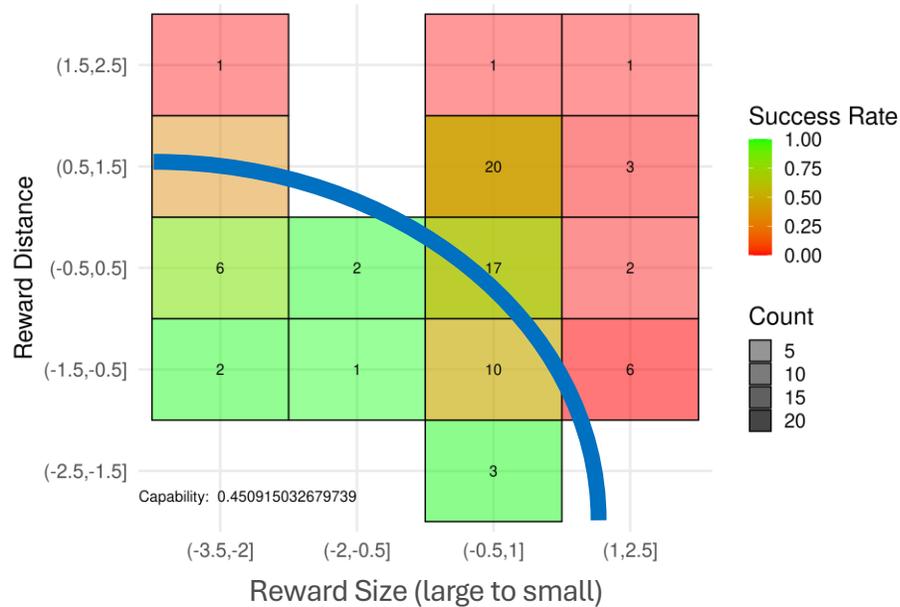


Figure 5: Characteristic curves of all competition entrants (agents) according to three relevant features (size, distance and Ypos) and one irrelevant feature (Xpos). Black dashed lines show the linear regression for the black points (pass/fail), while blue dashed lines interpolate the blue points (binned success rate). The conformances (Spearman correlations against monotonic sequence) are 0.80, 0.60, 1.00 and -0.50 , respectively.

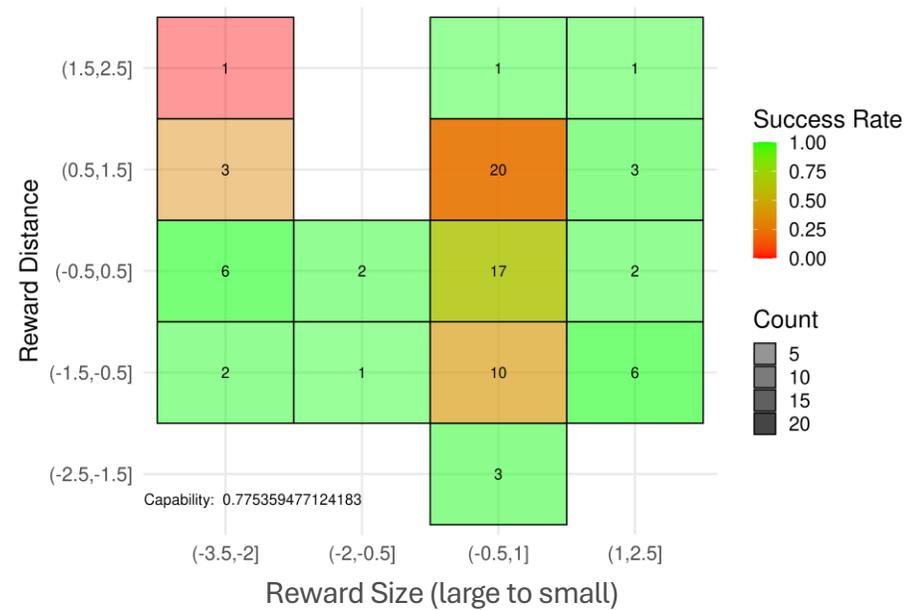
CAPABILITIES VS NO-CAPABILITIES

Capability boundary



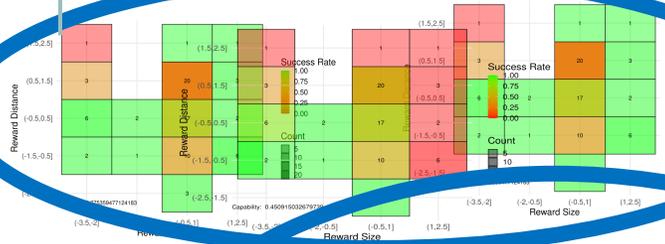
Conformant System
Juohmaru

This system doesn't show monotonicity.
We can't identify any level of capability robustly.



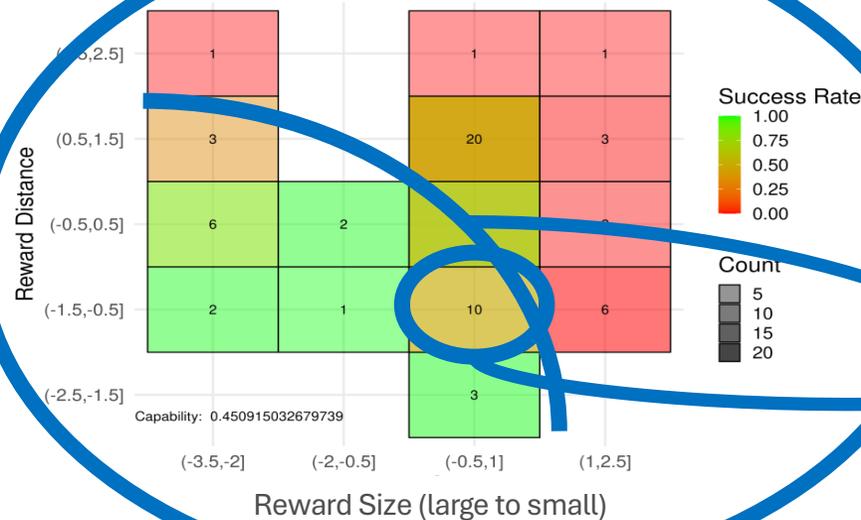
Non-Conformant System
y.yang

PREDICTING PERFORMANCE POSSIBLE



avg

G.Acc. : extrapolate Global ACCuracy
 ? = 54.7%
 (ignores system locality and feature relevance)

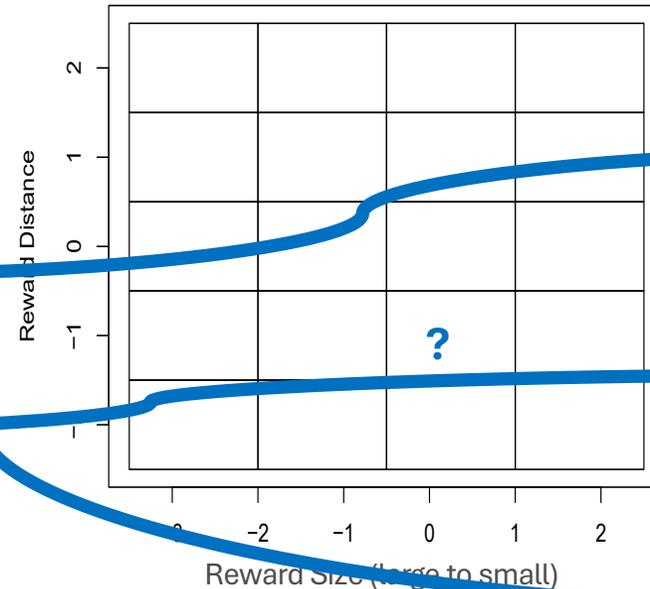


avg

T.Acc. extrapolate agentT ACCuracy
 ? = 46.8%
 (ignores feature relevance)

avg

B.Acc. : extrapolate Bin ACCuracy
 ? = 40%
 (ignores other bins)

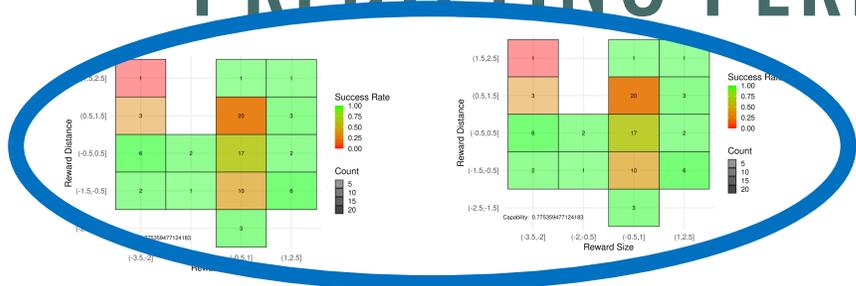


logistic

Par. : use parametric model on capabilities
 ? = 73.2%
 (parameter goodness-of-fit may be poor)

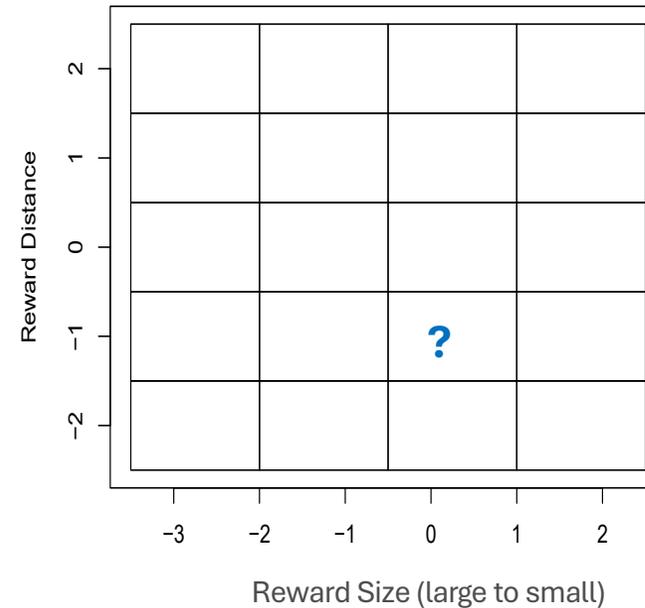
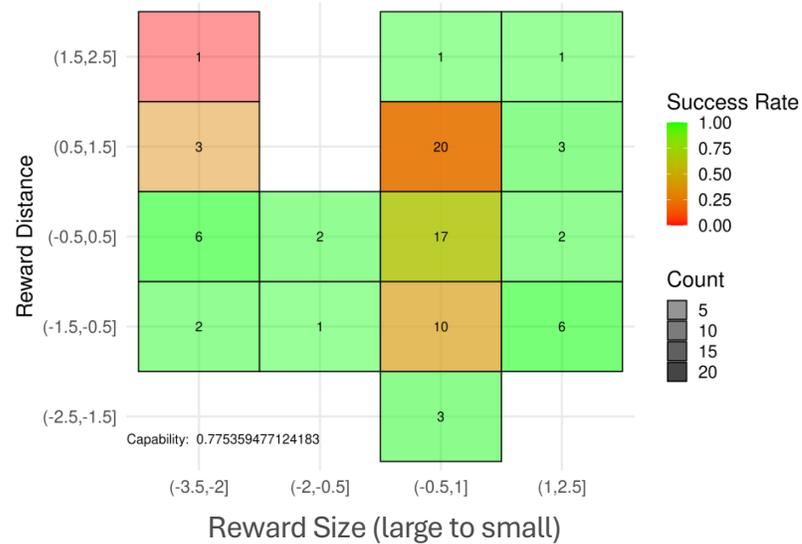
Except the last one, these are basically non-inferential methods (constant models or binning extrapolations)

PREDICTING PERFORMANCE NOT POSSIBLE?



ML model

A : use assessor models
(Using all variables or only the relevant ones?)



assessors = let's use all the power of ML to characterise the system's performance!!

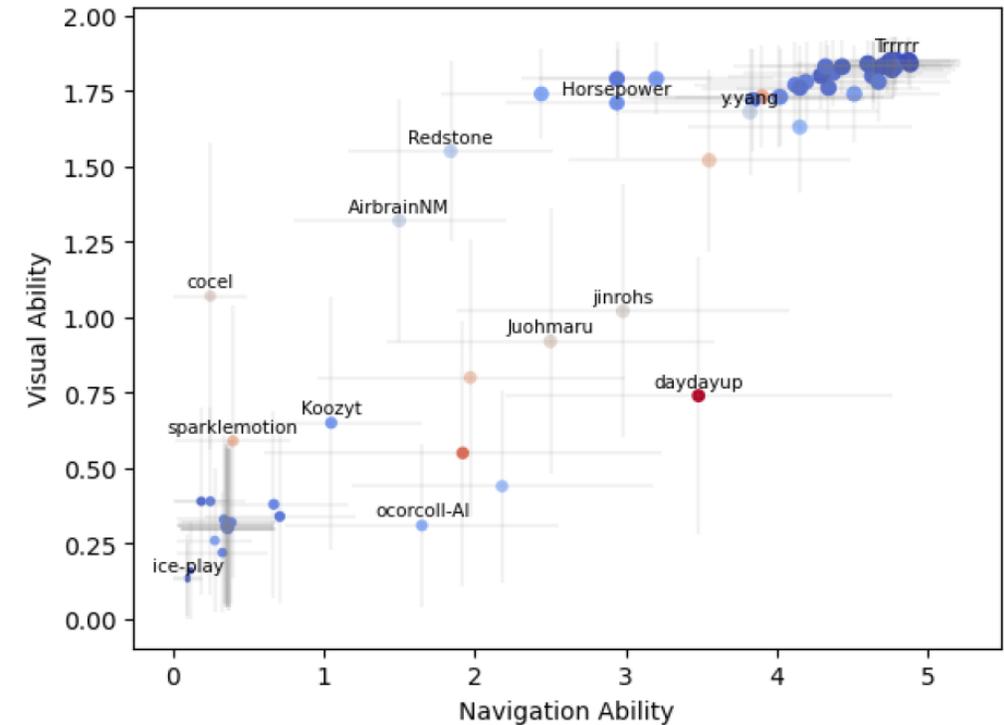
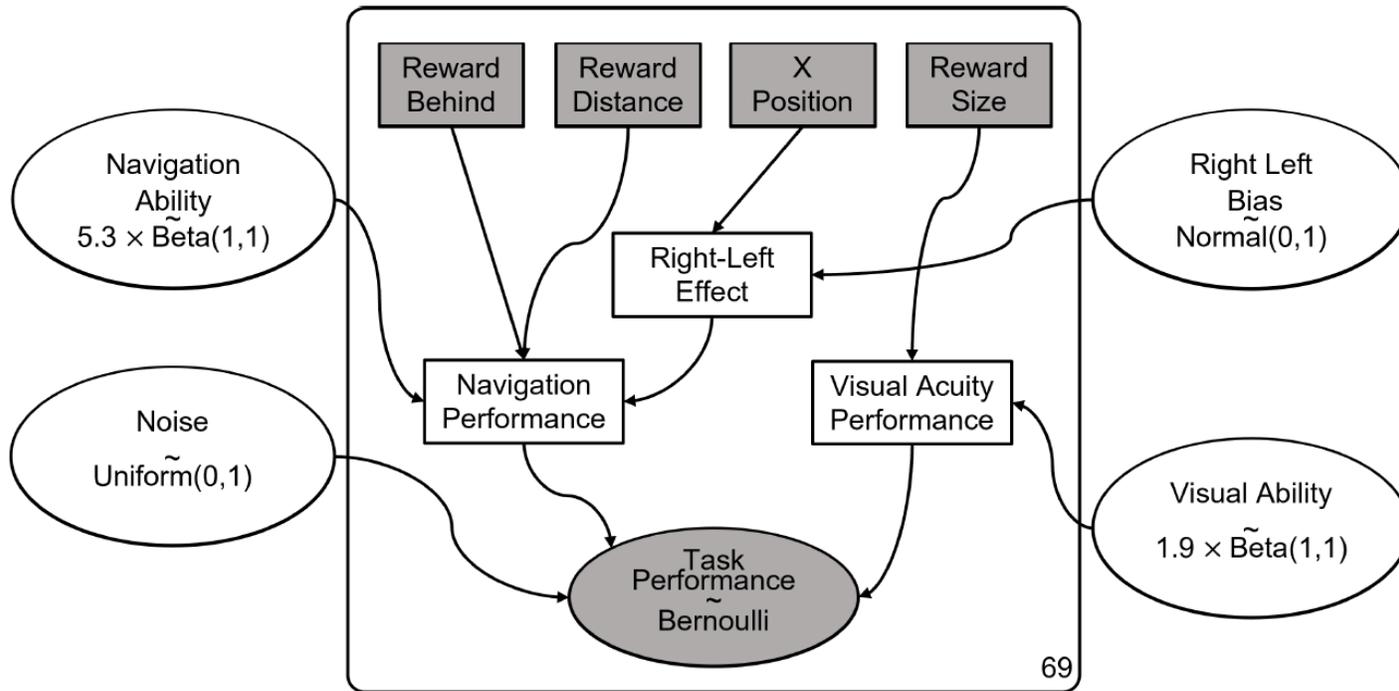
PREDICTING PERFORMANCE (COMPARISON)

Assessor with all features

	Maj. (1)	G.Acc.	T.Acc.	~All+A	~Rel+A
Error	45.3%	48.0%	33.6%	19.7%	20.6%
MAE	45.3%	49.6%	34.9%	29.3%	30.2%
MSE	45.3%	24.8%	17.6%	14.8%	15.4%

Animal AI Competition Data: 99 instances x 68 agents

MEASUREMENT LAYOUTS



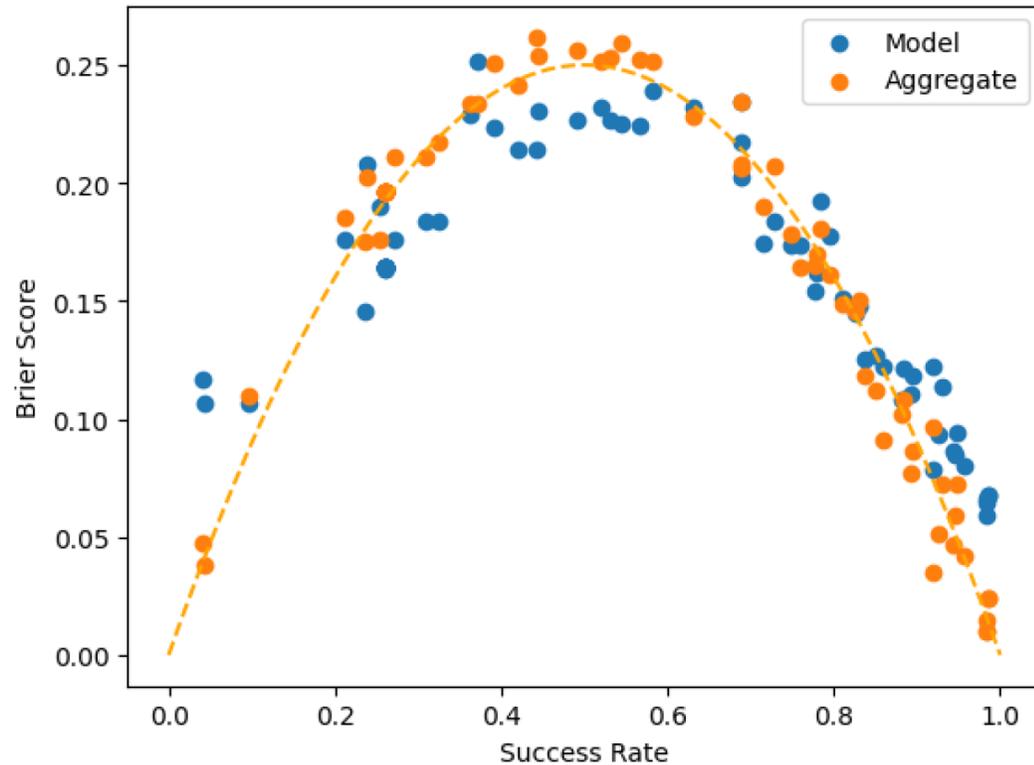
The x-axis and y-axis show the inferred means for navigationAbility and visualAbility respectively, with their standard deviations as error bars in grey. The radius of each point represents the average performance, while the colour represents the noiseLevel (red higher than blue).

MEASUREMENT LAYOUTS : MORE COMPLEX (0-PIAAGETS)

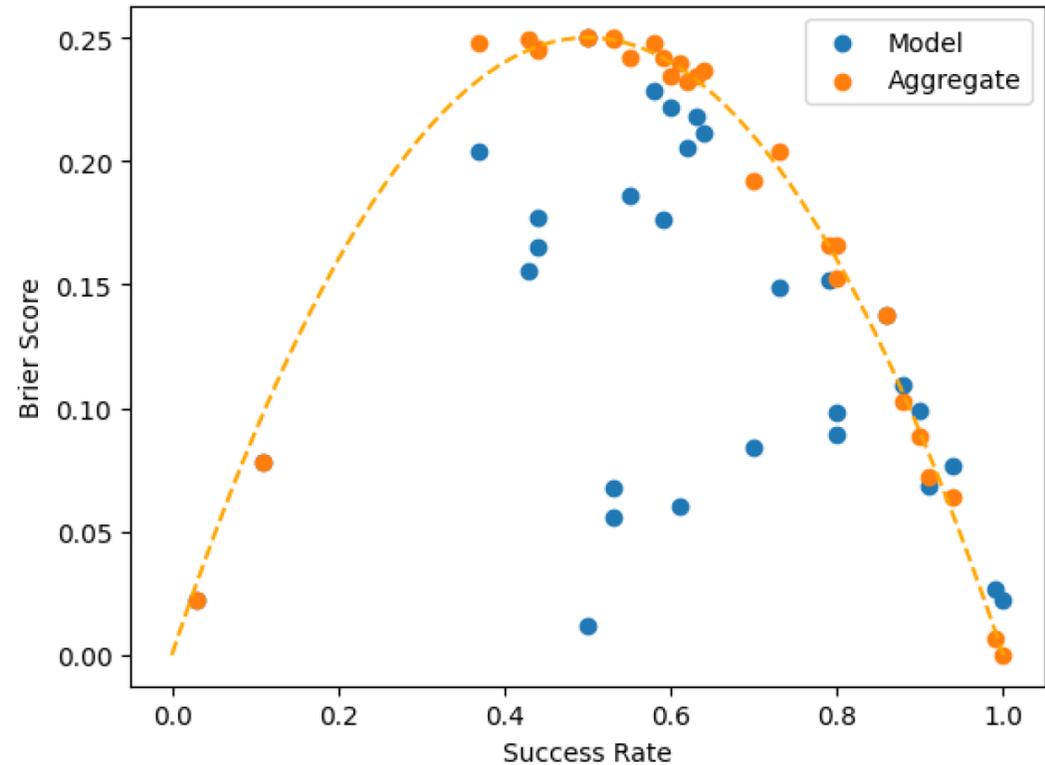
objPermAbility	50	13	13	12	50	50	50	50	48	49	49	49	49	50	50	50	28	36	23	15	25	2.1	33	29	50	27	50	49	12	12
flatNavAbility	56	26	7.6	13	56	56	56	56	56	56	56	52	51	56	56	56	56	56	43	38	37	32	45	43	56	56	49	42	52	53
visualAcuityAbility	6	2.9	2.9	4	6	6	6	6	6	5.9	3.6	5.9	5.8	6	6	6	6	6	5.9	5.8	6	5.3	6	5.8	6	5.8	6	6	6	5.9
lavaAbility	1	0.49	0.5	0.54	0.99	0.08	0.82	0.93	1	0.57	0.98	0.99	0.99	0.99	1	0.99	1	1	0.98	0.93	0.59	0.55	1	0.95	0.99	0.98	0.72	0.98	1	1
platformAbility	1	0.5	0.49	0.36	0.99	0.99	0.98	0.93	1	0.98	1	1	1	1	1	0.99	1	0.99	0.73	0.7	0.99	0.97	0	0.02	1	1	0.98	0.99	0.99	0.88
rampAbility	1	0.5	0.48	0.35	0.99	1	1	0.99	0.14	0.98	1	0.92	0.94	0.75	1	1	1	1	0.01	0.03	0.66	0.72	1	0.95	0.56	0.53	1	0.99	0.73	0.73
memoryAbility	4.8	2.4	2.4	2.3	4.8	4.8	4.8	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.8	4.8	2.7	4.8	4.4	2.8	4.5	2.7	4.8	4.4	4.7	4.5	4.8	4.7	4.8	4.7
rightLeftBias	0	-0.01	0.05	-0.12	-0.25	-0.14	0	0.02	0.33	0.13	0.33	6.3	-6.3	0.13	0	-0.24	-0	0.02	-0.01	-0.13	0.01	0.05	0	0.07	0.01	0.46	0.02	0.29	-0.08	0.03
noisePar	0	0.98	0.11	0.03	0.03	0.02	0	0.01	0	0.21	0.01	0.35	0.42	0.04	0	0	0.01	0.01	0.34	0.42	0.31	0.39	0	0.25	0.05	0.51	0.06	0.39	0	0.25
Success	1	0.5	0.1	0.03	0.9	0.53	0.85	0.83	0.61	0.63	0.59	0.63	0.61	0.79	0.99	0.94	0.79	0.91	0.53	0.44	0.44	0.38	0.51	0.44	0.73	0.57	0.71	0.56	0.8	0.62
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Very similar performance, very different cognitive profiles

MEASUREMENT LAYOUTS : PREDICTABILITY



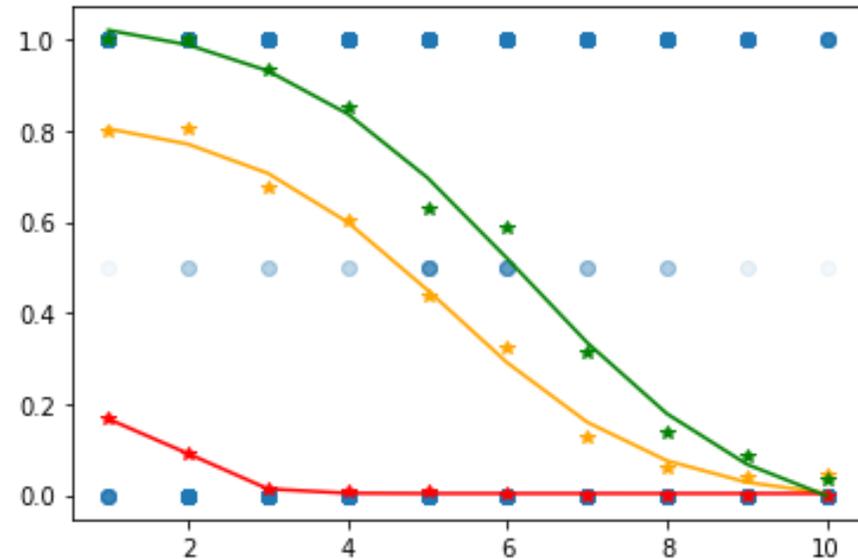
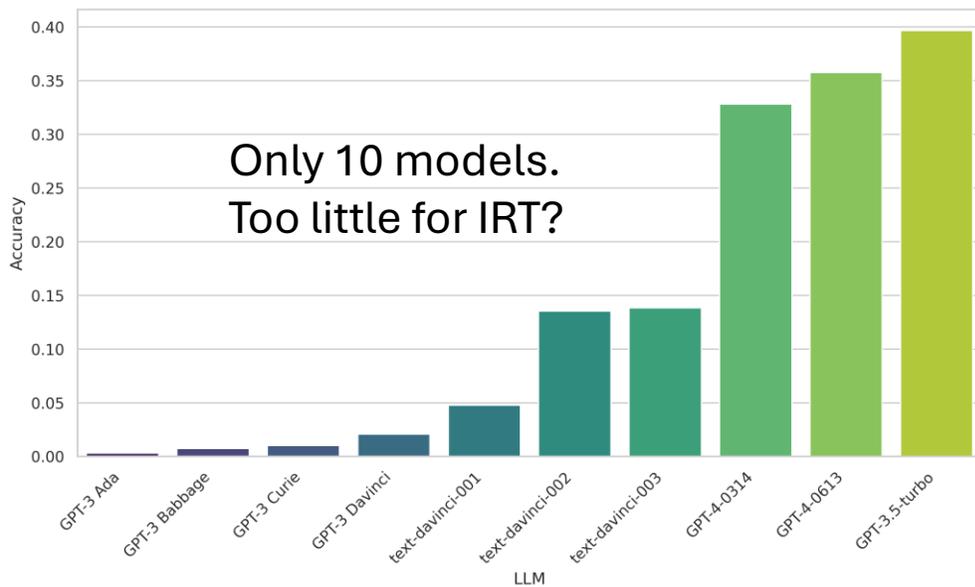
(a) AAI0 Tasks



(b) O-PIAAGETS tasks.

MORE SOPHISTICATED MODELS

- From performance to capabilities more generally:



GPT (3, 3.5, 4) on addition problems with difficulty being the mean of #digits (x-axis is deciles)

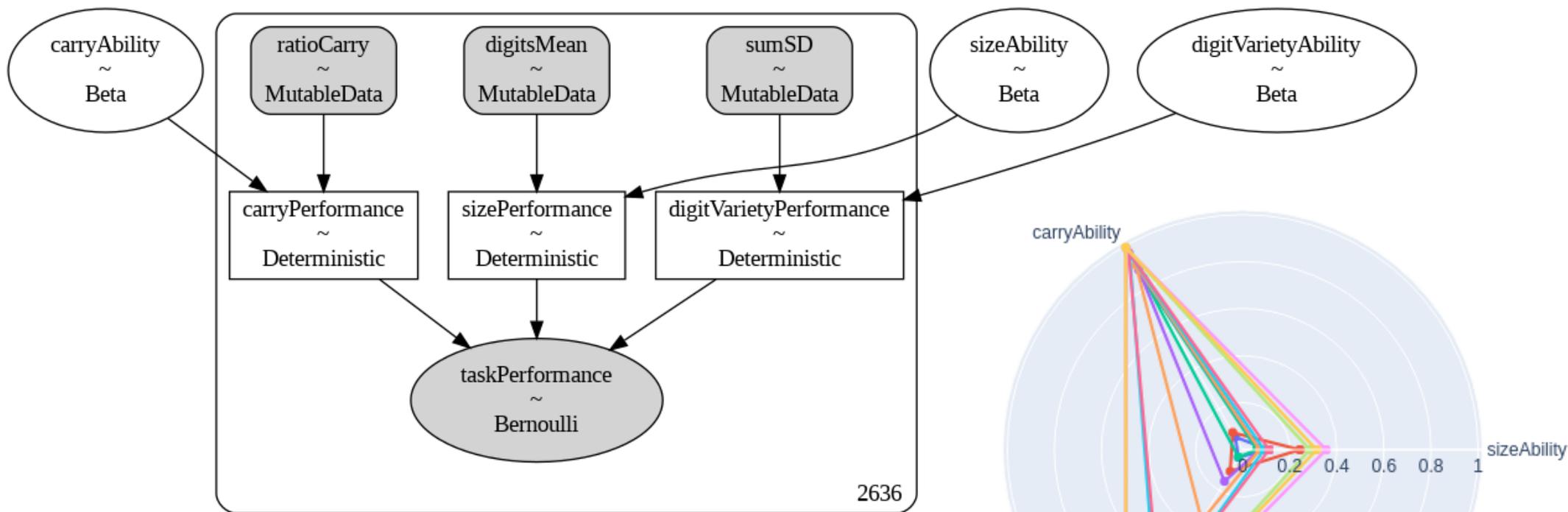
MORE SOPHISTICATED DEMANDS

- `digits1`: The number of digits in the first summand.
- `digits2`: The number of digits in the second summand.
- `min_digits`: $\min(digits_1, digits_2)$, i.e., the number of digits in the smaller summand.
- `harm_mean`: $2/(1/digits_1 + 1/digits_2)$, i.e., the harmonic mean of the number of digits in the two summands.
- `art_mean`: $(digits_1 + digits_2)/2$, i.e., the arithmetic mean of the number of digits in the two summands.
- `max_digits`: $\max(digits_1, digits_2)$, i.e., the number of digits in the larger summand.
- `carry`: The number of carrying operations required to add the two numbers.

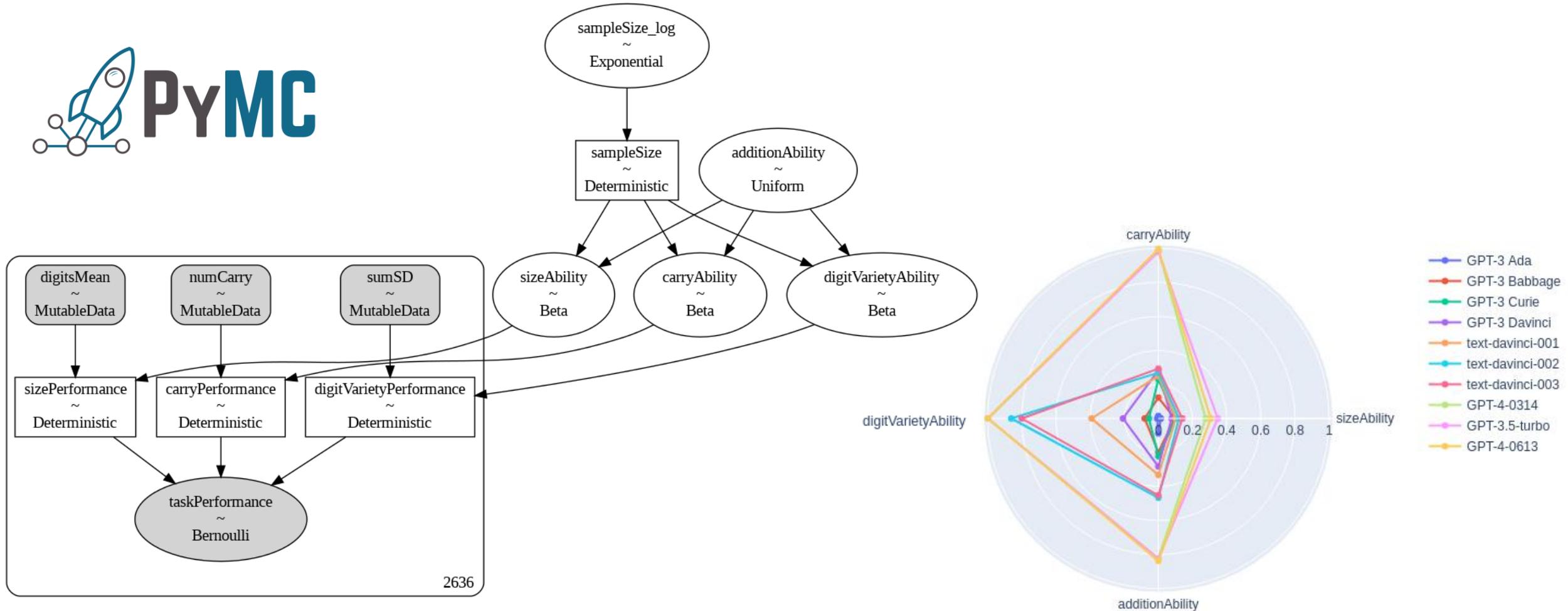
What are some of the things that make the addition of two numbers ‘difficult’?

- Size of the two numbers
- Number of carrying operations
- Can we have lots of carrying operations but the additions is still ‘easy’?

SIMPLE MEASUREMENT LAYOUT



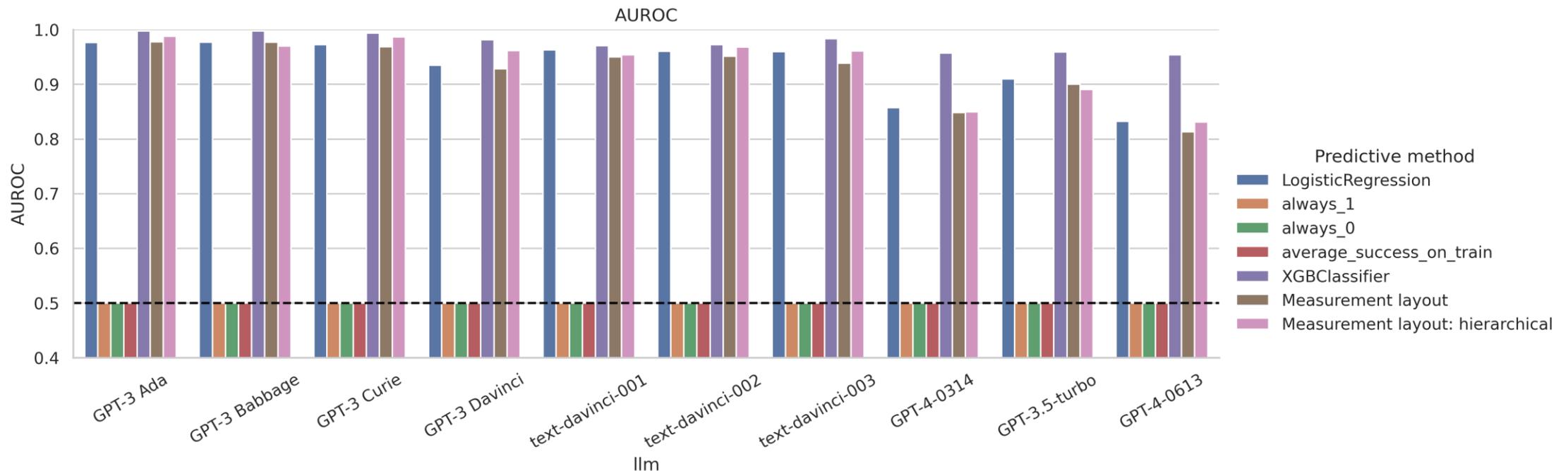
HIERARCHICAL MEASUREMENT LAYOUT



2636

PREDICTING PERFORMANCE

- Not only can we get capability profiles, but we can predict well!



The measurement layouts are non-populational. They do not depend on the results of the other models!

OTHER METHODS TO EXPLAIN/PREDICT PERFORMANCE

From Games and AI:

- Elo-Ranking, TrueSkill (Microsoft)

Minka, T., Cleven, R., & Zaykov, Y. (2018). Trueskill 2: An improved bayesian skill rating system. *Technical Report*.

From AI:

- Scaling laws

Schellaert et al. (2024): Scaling the scaling laws. Workshop on scaling laws, EACL.

From Psychometrics:

- IRT, especially LLTM
- SEM / Hierarchical models
- Multi-level IRT.
- Factor analysis (next slide)
- ...

Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, 271, 18-42.

Lalor, J. P., Rodriguez, P., Sedoc, J., & Hernandez-Orallo, J. (2024). Item response theory for natural language processing. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics.

Ravand, H. (2015). Item response theory using hierarchical generalized linear models. *Practical Assessment, Research, and Evaluation*, 20(1), 7.

Sulis, I., & Toland, M. D. (2017). Introduction to Multilevel Item Response Theory Analysis: Descriptive and Explanatory Models. *The Journal of Early Adolescence*, 37(1), 85-128. <https://doi.org/10.1177/0272431616642328>

FACTOR ANALYSIS

Task	HELM classification	Annotated ability	Factor loadings (Freq.)			Factor loadings (Bayesian)		
			Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
XSUM	Summarization	Comprehension	0.91	0.05	-0.09		0.84	
HellaSwag	QA	Comprehension	0.88	0.21	-0.04		0.93	
NarrativeQA	QA	Comprehension	0.86	0.25	-0.05		0.68	
CNN.DailyMail	Summarization	Comprehension	0.85	-0.40	0.03		0.47	
IMDB	Sentiment Analysis	Comprehension	0.84	-0.02	-0.33		0.33	
WikiFact	Knowledge	Domain knowledge	0.82	-0.08	0.26		0.78	
OpenbookQA	QA	Reasoning - commonsense	0.80	0.19	0.10		0.93	
NaturalQuestions	QA	Comprehension	0.76	0.11	0.22		0.97	
BoolQ	QA	Comprehension	0.72	0.21	0.19		0.70	
RAFT	Text Classification	Comprehension	0.63	0.13	0.33		0.69	
QuAC	QA	Comprehension	0.60	0.18	0.39		0.74	
TwitterAAE	Language modelling	Language modelling	-0.09	1.00	0.01			0.94
ICE	Language modelling	Language modelling	0.17	0.90	-0.02			0.97
The Pile	Language modelling	Language modelling	0.15	0.88	0.07			0.96
BLiMP	Language modelling	Language modelling	0.03	0.80	-0.09			0.82
TruthfulQA	QA	Domain knowledge	-0.15	-0.06	1.03	1.00		
BBQ	Bias	Reasoning - inductive	-0.02	-0.06	1.01	1.06		
GSM8K	Reasoning	Reasoning - mathematical	0.04	0.02	0.96	0.87		
Synthetic reasoning (NL)	Reasoning	Reasoning - fluid	-0.08	0.02	0.88	0.80		
MATH	Reasoning	Reasoning - mathematical	0.12	0.09	0.86	0.84		
CivilComments	Toxicity Classification	Comprehension	0.11	0.05	0.83	0.67		
Synthetic reasoning (A)	Reasoning	Reasoning - fluid	0.14	0.26	0.74	0.83		
MMLU	QA	Mixed	0.45	-0.13	0.64	0.95		
LegalSupport	Reasoning	Reasoning - inductive	0.47	-0.16	0.48	0.32		
LSAT	Reasoning	Reasoning - fluid	0.02	-0.09	0.46			
bAbI	Reasoning	Reasoning - deductive	0.44	0.35	0.40		0.69	
Dyck	Reasoning	Reasoning - deductive	0.25	0.45	0.28		0.59	

SUMMARY OF APPROACHES

Approach	Predictive for items	Predictive for systems	Domain Knowledge	System Populational	Abilities	Type of Models
Performance Aggregation / CTT	No	No	No	No	—	Statistical Tendency/Position/Dispersion
Scaling Laws	No	Seen & New	No	Yes	—	Power Laws
Factor Analysis	No	No	No	Yes	≥ 1	Linear (response)
SEM	No	Seen	Yes	Yes	≥ 1 or hierarchy	Mostly Linear (response)
Traditional IRT (1PL, 2PL, 3PL)	Seen	Seen	No	Yes	1	Logistic/Bernouilli (response)
Beta/Gamma IRT Models, ...	Seen	Seen	No	Yes	1	Beta (response), Gamma (response), ...
Multidimensional IRT	Seen	Seen	Partly	Yes	≥ 1	Logistic (response)
LLTM	Seen & New	Seen	Yes	Yes	1 (≥ 1 MIRT)	Linear (diff) + Logistic (response)
General Difficulty Model	Seen & New	Seen	No	Yes	≥ 1	Any machine learning model (diff) + Logistic
Intrinsic Difficulty	Seen & New	Seen	Yes	No	≥ 1	No model + Logistic
Self-assessment (uncert. est.)	Seen & New	Seen	No	No	—	The own model (mostly classification)
Assessors	Seen & New	Seen & New	No	Either	—	Any Machine Learning Model
Measurement Layouts	Seen & New	Seen & New*	Yes	Either	≥ 1 or hierarchy	Any Bayesian Model if Differentiable

PART IV : KINDS OF DIFFICULTY

“It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility”

Hans Moravec, Mind Children, Harvard University Press, 1988.

Intrinsic Difficulty

Pointers:

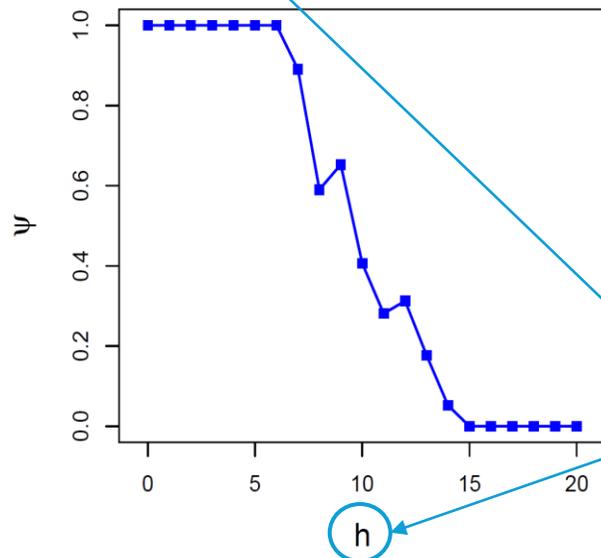
- Hernández-Orallo, J., Loe, B. S., Cheke, L., Martínez-Plumed, F., & Ó hÉigearthaigh, S. (2021). General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific reports*, 11(1), 22822.
- Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 634(8032), 61-68.
- Sun, Y., Hu, S., Zhou, G., Zheng, K., Hajishirzi, H., Dziri, N., & Song, D. (2025). OMEGA: Can LLMs Reason Outside the Box in Math? Evaluating Exploratory, Compositional, and Transformative Generalization. arXiv preprint arXiv:2506.18880.

DIFFICULTY AS SOLUTION COMPLEXITY

Policy complexity (solution complexity, resources, ...)

“C-test” considers **all computable policies**.

- Defines $h = Kt(\mu)$, the Levin’s Kt complexity of the solution
- Instead of aggregating a weighted sum using $p(\mu) \approx 2^{-KtU(\mu)}$
- We show an “agent” characteristic curve (ACC)!



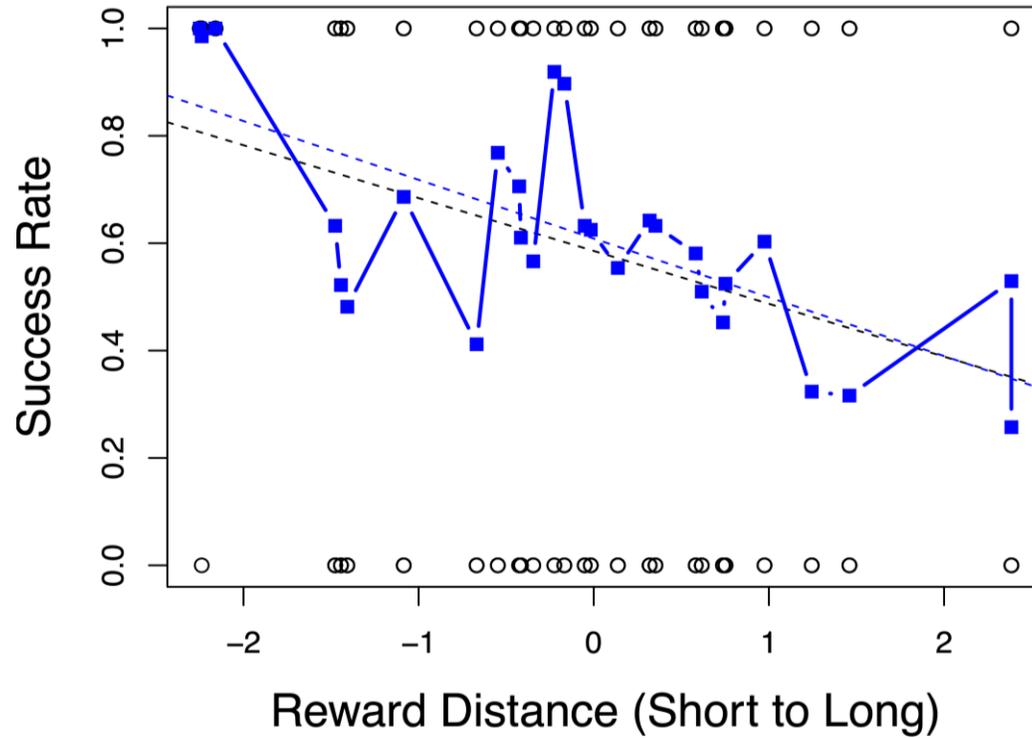
The metric of difficulty of the solution policy becomes the common currency to show performance across different tasks

$h = 7$: a, b, c, d, \dots	<i>Answer : e</i>
$h = 8$: $a, a, a, b, b, b, c, \dots$	<i>Answer : c</i>
$h = 9$: a, d, g, j, \dots	<i>Answer : m</i>
$h = 10$: a, c, b, d, c, e, \dots	<i>Answer : d</i>
$h = 11$: $a, a, b, b, z, a, b, b, \dots$	<i>Answer : y</i>
$h = 12$: $a, a, z, c, y, e, x, \dots$	<i>Answer : g</i>
$h = 13$: $a, z, b, d, c, e, g, f, \dots$	<i>Answer : h</i>
$h = 14$: $c, a, b, d, b, c, c, e, c, d, \dots$	<i>Answer : d</i>

Hernandez-Orallo, J. (2000). Beyond the Turing test.
Journal of Logic, Language and Information, 9(4), 447-466.

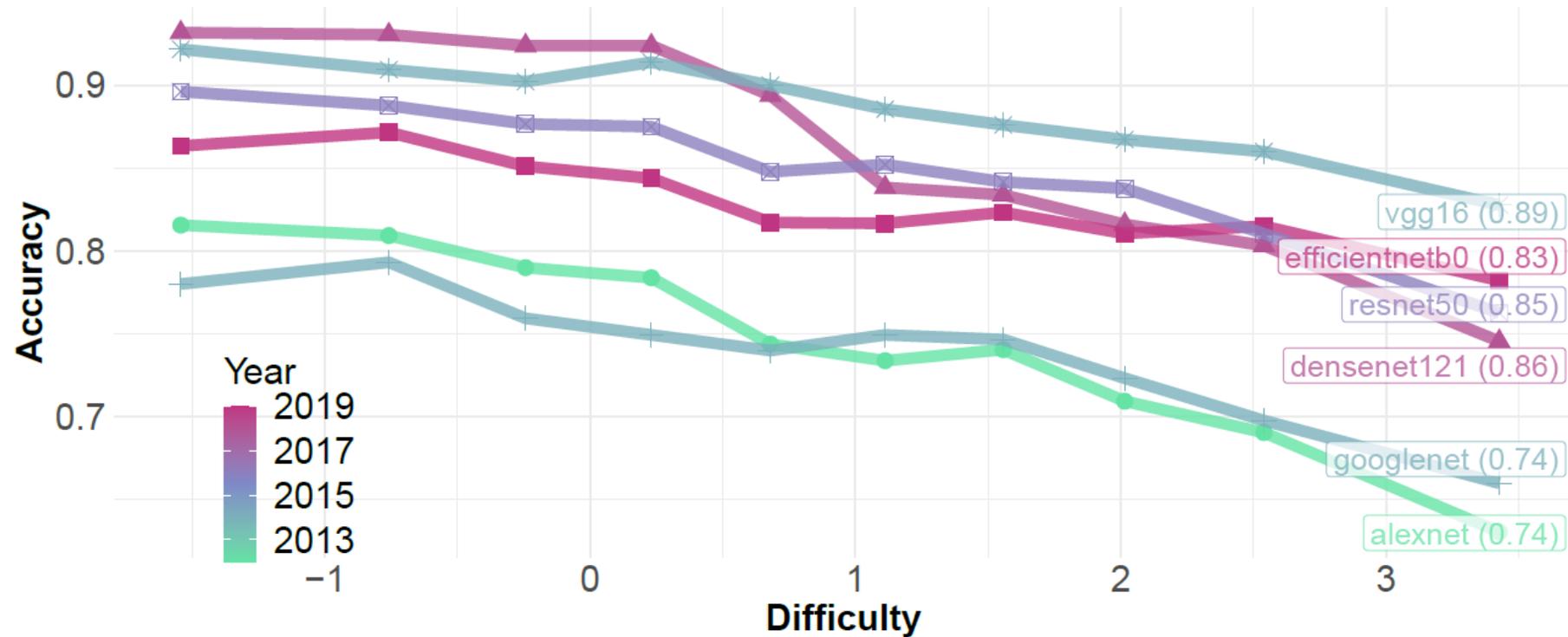
Hernández-Orallo, J.
“Unbridled mental power”
Nature Physics 15 (1), 2019

DIFFICULTY FROM OBSERVABLE FEATURES



Burnell, R., Burden, J., Rutar, D., Voudouris, K., Cheke, L., & Hernández-Orallo, J. (2022). Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. International Joint Conferences on Artificial Intelligence Organization.

DIFFICULTY FROM IRT OR DIFFICULTY ESTIMATORS



DIFFICULTY PROXIES

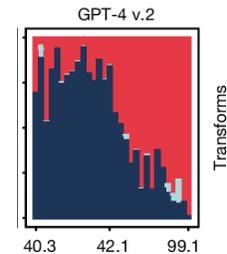
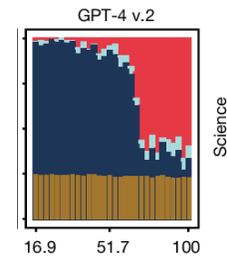
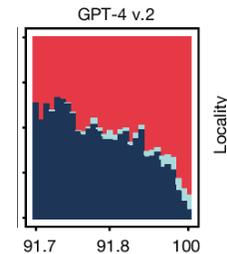
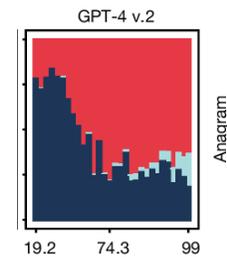
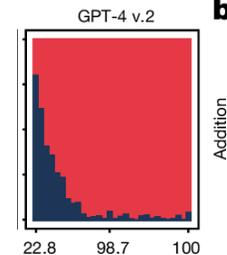
Article | [Open access](#) | Published: 25 September 2024

Larger and more instructable language models become less reliable

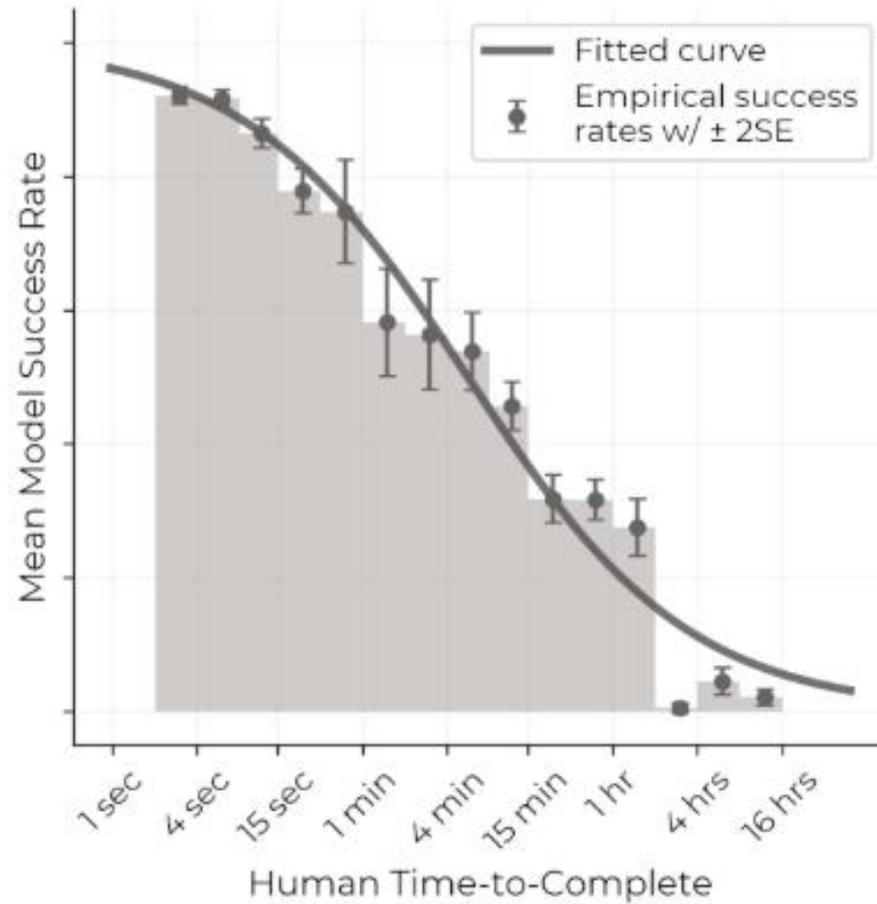
[Lexin Zhou](#), [Wout Schellaert](#), [Fernando Martínez-Plumed](#), [Yael Moros-Daval](#), [Cèsar Ferri](#) & [José Hernández-Orallo](#) 

[Nature](#) 634, 61–68 (2024) | [Cite this article](#)

Benchmark	Examples	Cal. Diff.
addition — single-task benchmark Arithmetic operations ranging from one to one-hundred-digit additions. <i>Difficulty: #carrying operations (f_{cry})</i>	Make the addition of 24427 and 7120.	35.25
	The sum of 47309068053 and 95464 is	65.04
	1893603010323501638430 + 98832380858765261900 =	98.67
anagram — single-task benchmark Jumbled words to be unscrambled to form a meaningful word ranging from three to twenty-letter words. <i>Difficulty: #letters of the anagram (f_{let})</i>	Unscramble this string of letters, "efe", to form a word.	18.42
	Rearrange the letters "ngiotuq" to make a single word.	50.42
	Rearrange the following anagram into an English word: "smltweoascnednkg".	96.78
locality — single-task benchmark Geographical knowledge about the location and size of cities relative to each other. <i>Difficulty: Inverse of city popularity (f_{pop})</i>	Which city that is less than 27 km away from Toronto has the largest number of people?	91.66
	What is the name of the largest city (by population) that is less than 98 km away from Altea?	92.64
	Name the most populated city that is less than 39 km away from Akil.	99.87
science — multi-task benchmark Elementary science-related world knowledge questions and graduate-level questions in biology, physics, and chemistry. <i>Difficulty: Anticipated human difficulty (f_{hum})</i>	Definition: In this task, you need to provide the correct option for a given problem from the provided options. Problem: Shining a light through a diamond can A) make a lot of bright lights shine B) summon a brilliant wave of color C) heat up a room D) make a lot of money Output:	37.02
	A light beam is propagating through a glass with index of refraction n . The glass is moving at constant velocity v in the same direction as the beam and toward the observer in laboratory. What is the speed of light in glass relative to the observer in laboratory? Take the speed of light in vacuum $c=1$. A. $(1+n*v)/(n+v)$ B. $(1-n*v)/(n+v)$ C. $1/D$ $(1+n*v)/(n-v)$ With respect to the choices above, the correct one is	71.83
	Answer the following questions based on the list of available choices Identify the missing reagents in the following reaction. $(3r,5r,7r)$ -adamantane-1-carboxylic acid + A \rightarrow $(3r,5r,7r)$ -adamantane-1-carbonyl azide + B \rightarrow $(3s,5s,7s)$ -adamantan-1-amine. A: A = NaN ₃ and B = HCl aq, Heat B: A = PCl ₅ and B = H ₃ O ⁺ , Heat C: A = diphenylphosphoryl azide (DPPA) and B = H ₃ O ⁺ , Heat D: A = diphenylphosphoryl azide (DPPA) and B = NaN ₃ Answer:	99.97
transforms — multi-task benchmark Information-centric transformation tasks. <i>Difficulty: Combination of input+output word count and Levenshtein distance (f_{w+l})</i>	Be concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output. Given "double07@MI6.gov.uk" modify the input to display the domain of the email address of the form USER@DOMAIN.	39.49
	Consider the INPUT: \n"8:30h - Accreditation (badges)\n9:00h - Opening\n9:15h - Keynote\n10:15h - Coffee break\n10:45h - Invited Talks\n11:55h - Lightning talks\n12:05h - Panel\n13:00h - Lunch break (in the hall)\n14:30h - Keynote\n15:30h - Minibreak\n15:40h - Invited Talks\n16:50h - Panel\n17:45h - Closing remarks" I'd like the agenda to show a 15-minute reduction in each keynote speaker's segment, shifting the schedule to finish earlier. \nBe concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output.	55.22
	Michael Vaughn, a 63-year-old retired naval officer, presents an extensively complex medical history complicated by a litany of allergies. He battles chronic pain stemming from neuropathy for which he takes Pregabalin (Lyrica) 150 mg twice daily. Due to advanced rheumatoid arthritis, he relies on Etanercept (Enbrel) 50 mg, administered weekly via subcutaneous injection, but cannot be prescribed common NSAIDs like Ibuprofen or Naproxen due to gastrointestinal bleeding and a reported severe allergy to Aspirin (anaphylaxis). His Type 2 diabetes is managed with Insulin Aspart (NovoLog) administered via an insulin pump with doses varying according to his blood glucose readings; he experienced a life-threatening lactic acidosis episode with Metformin. I'd like the list of drugs that are prescribed to the patient to be arranged alphabetically and without repetitions, in the form of a clean, comma-separated list. Be concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output.	64.76



DIFFICULTY AS TIME



Measuring AI Ability to Complete Long Tasks

Thomas Kwa^{*}, Ben West^{1*}, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx

Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles², Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler³

Elizabeth Barnes, Lawrence Chan

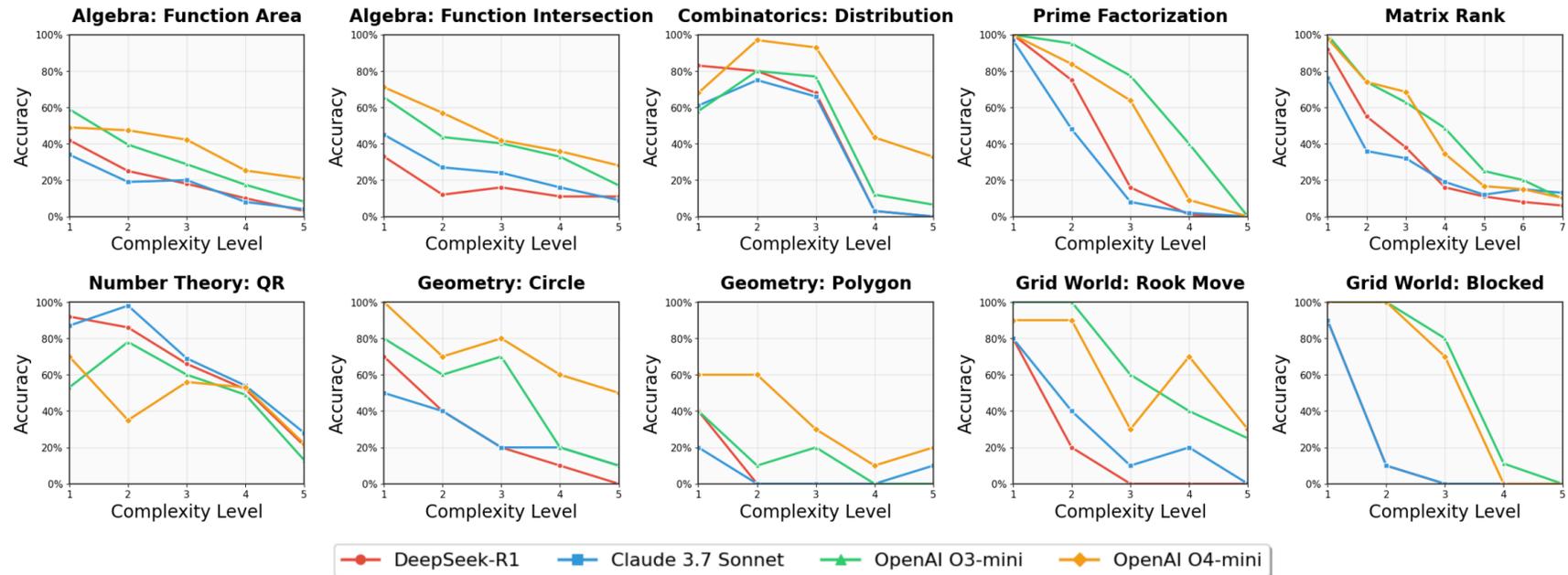
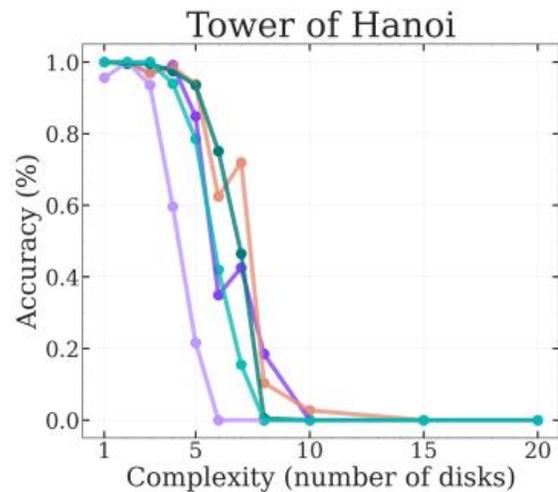
DIFFICULTY FROM COMPUTATIONAL COMPLEXITY

The Illusion of Thinking:
Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojaee[†] Iman Mirzadeh^{*} Keivan Alizadeh
Maxwell Horton Samy Bengio Mehrdad Farajtabar

OMEGA: Can LLMs Reason Outside the Box in Math? Evaluating Exploratory, Compositional, and Transformative Generalization

Yiyu Sun¹, Shawn Hu⁴, Georgia Zhou¹, Ken Zheng¹, Hannaneh Hajishirzi^{2,3},
Nouha Dziri², Dawn Song^{1*}
¹University of California, Berkeley, ²Ai2, ³University of Washington, ⁴dmodel.ai



Annotated Demand Levels

Pointers:

- Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., ... & Hernández-Orallo, J. (2025). General scales unlock ai evaluation with explanatory and predictive power. arXiv preprint arXiv:2503.06378.
<https://arxiv.org/abs/2503.06378>
<https://kinds-of-intelligence-cfi.github.io/ADELE/>

GENERAL SCALES FOR A SCIENCE OF AI EVALUATION?

- In this paper we address some key elements for a science of AI evaluation:
 - Carve the space of capabilities (and propensities*) into **commensurate scales**.
 - Explain **what benchmarks really measure**.
 - Extract interpretable **ability profiles** of AI systems.
 - **Predict performance** (and safety*) for new task instances, in- **and out-of-distribution**.

* Planned for ADeLe v.2.0

18 GENERAL DIMENSIONS

- A taxonomy of cognitive abilities for artificial and natural systems.
- **DeLeAn v1.0:** only LLMs
 - **Primordial:** 11 cognitive capabilities
 - **Knowledge:** 5 branches of knowledge
 - **Extraneous:** 2 other elements making task difficult

	Dimension (Broad)		Dimension (Specific)	Description of Demands
AS	Attention and Scan	AS	Attention and Scan	Focus on or locate specific elements within a given stream of information or environment in the whole process of solving a task.
CE	Comprehension and Expression	CEc	Verbal Comprehension	Understand text, stories or the semantic content of other representations of ideas in different formats or modalities.
		CEe	Verbal Expression	Generate and articulate ideas, stories, or semantic content in different formats or modalities.
CL	Conceptualisation, Learning and Abstraction	CL	Conceptualisation, Learning and Abstraction	Build new concepts, engage in inductive and analogical reasoning, map relationships between domains, and generate abstractions from concrete examples.
MC	Metacognition and Critical Thinking	MCr	Identifying Relevant Information	Recognise what information helps solve the task or does not, and how this recognition process unfolds as they work toward the solution.
		MCT	Critical Thinking Processes	Monitor or regulate multiple thought processes to answer the question effectively, ranging from simple recall to high-level critical thinking.
		Mcu	Calibrating Knowns and Unknowns	Recognise the boundaries of one's knowledge and confidently identify what one knows they know, knows they don't know, or is uncertain about.
MS	Mind Modelling and Social Cognition	MS	Mind Modelling and Social Cognition	Model the minds of other agents or reasoning about how the beliefs, desires, intentions, and emotions of multiple other agents might interact to determine future behaviours.
QL	Quantitative and Logical Reasoning	QLl	Logical Reasoning	Match and apply rules, procedures, algorithms or systematic steps to premises to solve problems, derive conclusions and make decisions.
		QLq	Quantitative Reasoning	Work with and reason about quantities, numbers, and numerical relationships.
SN	Spatial Reasoning and Navigation	SNs	Spatio-physical Reasoning	Understand spatial relationships between objects and predicting physical interactions.
KN	Knowledge	KNa	Knowledge of Applied Sciences	Knowledge or conceptual understanding in applied sciences (e.g., medicine, law, education, business, agriculture, engineering except IT).
		KNc	Customary Everyday Knowledge	Knowledge in information that most people in a given society typically acquire through daily life experiences, social interactions, and media.
		KNf	Knowledge of Formal Sciences	Knowledge or conceptual understanding in formal sciences (e.g., mathematics, logic, computer science, statistics).
		KNn	Knowledge of Natural Sciences	Knowledge or conceptual understanding in natural sciences (e.g., physics, chemistry, biology, astronomy, earth sciences, ecology).
		KNs	Knowledge of Social Sciences	Knowledge or conceptual understanding in social sciences and humanities (e.g., history, psychology, sociology, literature, art, philosophy).
AT	Atypicality	AT	Atypicality	How uncommon the task is or how unlikely it is that the instance has appeared in various sources (internet, textbooks, tests).
VO	Volume	VO	Volume	Proportional to the logarithm of the time a fully competent human needs to read and complete the task in ideal conditions, excluding interruptions.

SCALES

- Ratio Scale
- Rule of thumb:
 - Level 0: No demand
 - Level 1: ≥ 1 in 10^1 people
 - Level 2: ≥ 1 in 10^2 people
 - Level 3: ≥ 1 in 10^3 people
 - Level 4: ≥ 1 in 10^4 people
 - Level 5: ≥ 1 in 10^5 people
 - ...

Domain Knowledge (KN)

R1. Natural Sciences (KNn)

This rubric assesses the conceptual sophistication level of tasks based solely on the depth of knowledge or conceptual understanding required in the fields of natural sciences (e.g., physics, chemistry, biology, astronomy, earth sciences, ecology). This does not include social sciences and humanities (e.g., history, psychology, sociology, anthropology, literature, art, philosophy, linguistics) or formal sciences (e.g., mathematics, logic, computer science, statistics). It's important to note that this rubric focuses exclusively on the domain-specific knowledge needed, not considering other cognitive demands such as reasoning or metacognition. This reflects the conceptual depth and specificity of the knowledge in natural sciences required, rather than the mere presence of scientific content.

Levels

Level 0 None. Tasks do not require any knowledge of natural sciences. **Examples:**

- "Write a python script to train a machine learning classifier for fake news detection."
- "Analyze the symbolism in Shakespeare's Hamlet".
- "Calculate the cost of groceries."

Level 1 Very low. Tasks that require knowledge in natural sciences typically acquired through elementary school education. **Examples:**

- Living things need food, water, and air to survive.
- Basic parts of a plant (roots, stem, leaves).
- Day and night cycle and seasons.

Level 2 Low. Tasks that require knowledge in natural sciences typically acquired through middle school education. **Examples:**

- The water cycle (evaporation, condensation, precipitation).
- Basic cellular structure (nucleus, membrane, cytoplasm).
- Simple food chains and ecosystems.

Level 3 Intermediate. Tasks that require knowledge in natural sciences typically acquired through high school education. **Examples:**

- Mendel's laws of inheritance and basic genetics.
- The ideal gas law ($PV = nRT$).
- Newton's three laws of motion.

Level 4 High. Tasks that require knowledge in natural sciences typically acquired through undergraduate education. **Examples:**

- Hardy-Weinberg equilibrium and population genetics.
- Molecular orbital theory.
- The process of cellular respiration and its relationship to photosynthesis.

Level 5+ Very High. Tasks that require knowledge in natural sciences typically acquired through graduate education or beyond. **Examples:**

- The theoretical frameworks of string theory and its implications.
- The six forms of quark flavors in particle physics.
- The role of quantum entanglement in biological systems.

ANNOTATION

- Automated through GPT4o annotation!
- Each instance is converted into an 18-dimensional numeric vector

Example

X₁  Omni-MATH

Question: Let ABC be a triangle with $AB = 13$, $BC = 14$, and $CA = 15$. We construct isosceles right triangle ACD with $\angle ADC = 90^\circ$, where D, B are on the same side of line AC, and let lines AD and CB meet at F. Similarly, we construct isosceles right triangle BCE with $\angle BEC = 90^\circ$, where E, A are on the same side of line BC, and let lines BE and CA meet at G.

Find $\cos \angle AGF$.

+

Rubric

Domain Knowledge (KN)

R1. Natural Sciences (KNn)

This rubric assesses the conceptual sophistication level of tasks based solely on the depth of knowledge or conceptual understanding required in the fields of natural sciences (e.g., physics, chemistry, biology, astronomy, earth sciences, ecology). This does not include social sciences and humanities (e.g., history, psychology, sociology, anthropology, literature, art, philosophy, linguistics) or formal sciences (e.g., mathematics, logic, computer science, statistics). It's important to note that this rubric focuses exclusively on the domain-specific knowledge needed, not considering other cognitive demands such as reasoning or metacognition. This reflects the conceptual depth and specificity of the knowledge in natural sciences required, rather than the mere presence of scientific content.

Levels

Level 0 None. Tasks do not require any knowledge of natural sciences. **Examples:**

- "Write a python script to train a machine learning classifier for fake news detection."
- "Analyze the symbolism in Shakespeare's Hamlet".
- "Calculate the cost of groceries."

Level 1 Very low. Tasks that require knowledge in natural sciences typically acquired through elementary school education. **Examples:**

- Living things need food, water, and air to survive.
- Basic parts of a plant (roots, stem, leaves).
- Day and night cycle and seasons.

Level 2 Low. Tasks that require knowledge in natural sciences typically acquired through middle school education. **Examples:**

- The water cycle (evaporation, condensation, precipitation).
- Basic cellular structure (nucleus, membrane, cytoplasm).
- Simple food chains and ecosystems.

Level 3 Intermediate. Tasks that require knowledge in natural sciences typically acquired through high school education. **Examples:**

- Mendel's laws of inheritance and basic genetics.
- The ideal gas law ($PV = nRT$).
- Newton's three laws of motion.

Level 4 High. Tasks that require knowledge in natural sciences typically acquired through undergraduate education. **Examples:**

- Hardy-Weinberg equilibrium and population genetics.
- Molecular orbital theory.
- The process of cellular respiration and its relationship to photosynthesis.

Level 5+ Very High. Tasks that require knowledge in natural sciences typically acquired through graduate education or beyond. **Examples:**

- The theoretical frameworks of string theory and its implications.
- The six forms of quark flavors in particle physics.
- The role of quantum entanglement in biological systems.

GPT4o

	AS	CEc	CEe	CL	MCr	MCT	MCu	MS	QLI	QLq	SNs	KNa	KNc	KNf	KNn	KNs	AT	VO	UG
X₁	3	3	3	4	4	4	3	0	4	4	3	0	0	4	0	0	3	3	100

ANNOTATION

- Automated through GPT4o annotation!

X₁  Omni-MATH

Question: Let ABC be a triangle with AB =13, BC =14, and CA =15. We construct isosceles right triangle ACD with $\angle ADC = 90^\circ$, where D, B are on the same side of line AC, and let lines AD and CB meet at F. Similarly, we construct isosceles right triangle BCE with $\angle BEC=90^\circ$, where E, A are on the same side of line BC, and let lines BE and CA meet at G.

Find $\cos \angle AGF$.

X₂  TimeQA

Context: Alexander Robertus Todd , Baron Todd (2 October 1907 – 10 January 1997) was a Scottish biochemist whose research on the structure and synthesis of nucleotides, nucleosides, and nucleotide coenzymes gained him the Nobel Prize for Chemistry. Todd held posts with the Lister Institute, the University of Edinburgh (staff, 1934–1936) and the University of London, where he was appointed Reader in Biochemistry. In 1938, Alexander Todd spent six months as a visiting professor at California Institute of Technology, eventually declining an offer of faculty position. Todd became the Sir Samuel Hall Chair of Chemistry and Director of the Chemical Laboratories of the University of Manchester in 1938, where he began working on nucleosides, compounds that form the structural units of nucleic acids (DNA and RNA). In 1944, he was appointed to the 1702 Chair of Chemistry in the University of Cambridge, which he held until his retirement in 1971 [...].

Question: Which employer did Alexander R. Todd work for from 1938 to 1944?

X₃  MedCalcBench

Patient Note: A 58-year-old male presents to the clinic this week. No past stroke history can be detected in his medical records. He is currently being prescribed aspirin and NSAIDs, following an incident of significant bleeding he endured following a routine procedure. His alcohol intake can be considered heavy, consuming up to 12 drinks per week. Most recently, his blood pressure readings have tended to be elevated at above 170 mmHg for the systolic pressure. Interesting to note, his INR has remained stable during his multiple lab tests, eliminating any concerns about its lability. He also shows laboratory evidence of chronic kidney disease, necessitating further management. This man's condition mandates comprehensive dynamic monitoring and individualized care planning given the complexity of his medical situation.

Question: What is the patient's HAS-BLED score?

X₄  MMLU-Pro

Question: The population of a certain city is 836,527. What is the population of this city rounded to the nearest ten thousand?

Choices:

- A. 860,000.
- B. 850,000.
- C. 830,000.
- D. 837,000.
- E. 820,000.
- F. 840,000.
- G. 835,000.
- H. 800,000.
- I. 836,500.
- J. 836,000

X₅  TruthQuest

Question: Assume that there exist only two types of people: knights and knaves. Knights always tell the truth, while knaves always lie. You are given the statements from 6 characters. Based on their statements, **infer who is a knight and who is a knave.** A: C is a truth-teller and F is a truth-teller. B: C is a truth-teller and E is a truth-teller. C: I am a truth-teller. D: F is a truth-teller. E: C is a truth-teller and B is a liar. F: B is a truth-teller.

Places examples of very different benchmarks on the same commensurate space!

	AS	CEc	CEe	CL	MCr	M Ct	MCu	MS	QLI	QLq	SNs	KNa	KNc	KNf	KNn	KNs	AT	VO	UG
X₁	3	3	3	4	4	4	3	0	4	4	3	0	0	4	0	0	3	3	100
X₂	3	2	1	1	2	1	2	0	2	0	0	0	3	0	0	0	3	2	100
X₃	2	3	4	0	2	2	1	0	3	2	0	5	0	2	4	0	3	2	100
X₄	0	1	1	0	2	1	1	0	3	2	0	0	1	1	0	0	0	1	90
X₅	3	3	1	3	3	3	4	2	3	2	0	0	1	3	0	0	4	2	100

DISTINCTIVENESS

- ADeLe battery v1.0:
 - 63 tasks from 20 AI benchmarks
 - From 2024 AI venues
 - 16108 instances

How does this distribution choice affect the 18 dimensions?

Do we have elements in all regions of the 18 multi-dimensional space?

Source	Benchmark	Task	Claiming to Measure	#Instances
AGIEval [188]	Civil Service Examination	LogiQA-en	Logical Reasoning	408
	GRE & GMAT	AQuA-RAT	Mathematics	203
	LSAT	LSAT-AR LSAT-LR LSAT-RC	Analytical Reasoning	187
			Logical Reasoning	470
			Reading Comprehension	253
SAT	SAT-En SAT-Math	Critical thinking, problem-solving and analytical skills	196 214	
ChemLLMBench [61]	ChemLLMBench	Molecule Captioning	Generation of descriptions for molecules	160
		Molecule Design	Generation of new molecules given a description	295
		Name Prediction	Chemical name understanding	476
		Reaction Prediction	Chemical reaction products prediction	412
		Retrosynthesis	Identification of efficient synthetic pathways for target molecules	380
LiveBench [180]	Data Analysis	CTA	Data Analysis	33
	Language	Connections	Language Comprehension	29
	Math	AMPS Hard Math Competition Olympiad	Mathematics	69 78 26
			Reasoning	Spatial Zebra Puzzle
MMLU-Pro [177]	MMLU-Pro	Biology	Knowledge and Reasoning	447
		Business		410
		Chemistry		368
		Computer Science		345
		Economics		428
		Engineering		296
		Health		411
		History		304
		Law		362
		Math		425
		Other		429
		Philosophy		402
		Physics		377
Psychology	427			
MedCalcBench [87]	MedCalcBench	Date	Recall of medical calculation knowledge Extraction of relevant patient attributes Arithmetic computation of final results	27
		Diagnosis		14
		Dosage		20
		Lab		180
		Physical		214
		Risk		84
OmniMath [55]	OmniMath	Algebra	Mathematical reasoning at Olympiad level	337
		Applied Mathematics		302
		Calculus		30
		Discrete Mathematics		314
		Geometry		329
		Number Theory		322
		Precalculus		30
SciBench [175]	SciBench	Chemistry	Scientific problem-solving	142
		Math		105
		Physics		108
TimeBench [27]	Date Arithmetic	Date Arithmetic	Symbolic temporal reasoning	493
	MCTACO	MCTACO	Commonsense temporal reasoning	205
	MenatQA	MenatQA-Counterfactual MenatQA-Order MenatQA-Scope	Event temporal reasoning	130
				157
				393
	TempReason	TempReason-L2 TempReason-L3	Event temporal reasoning	318
				339
	TimeDial	TimeDial	Commonsense temporal reasoning	340
TimeQA	TimeQA-explicit TimeQA-implicit	Event temporal reasoning	379	
			348	
TruthQuest [111]	TruthQuest	E I S	Suppositional reasoning	344
				371
				340

CORRELATIONS

- Most dimensions appear to carve different parts of the space.

- Correlations can be both positive and negative, but can have multiple interpretations since they are contingent to the selected benchmarks.

AS	1.0	0.41	0.24	0.49	0.59	0.44	0.51	0
CEc	0.41	1.0	0.46	0.65	0.61	0.67	0.6	0
CEe	0.24	0.46	1.0	0.46	0.39	0.4	0.37	0
CL	0.49	0.65	0.46	1.0	0.72	0.83	0.79	0
MCr	0.59	0.61	0.39	0.72	1.0	0.7	0.69	0
MCT	0.44	0.67	0.4	0.83	0.7	1.0	0.76	0
MCu	0.51	0.6	0.37	0.79	0.69	0.76	1.0	0
MS	0.11	0.29	0.13	0.22	0.28	0.26	0.15	0
QLI	0.43	0.63	0.47	0.82	0.7	0.81	0.71	0

X₃ MedCalc Bench

Patient Note: A 58-year-old male presents to the clinic this week. No past stroke history can be detected in his medical records. He is currently being prescribed aspirin and NSAIDs, following an incident of significant bleeding he endured following a routine procedure. His alcohol intake can be considered heavy, consuming up to 12 drinks per week. Most recently, his blood pressure readings have tended to be elevated at above 170 mmHg for the systolic pressure. Interesting to note, his INR has remained stable during his multiple lab tests, eliminating any concerns about its lability. He also shows laboratory evidence of chronic kidney disease, necessitating further management. This man's condition mandates comprehensive dynamic monitoring and individualized care planning given the complexity of his medical situation.

Question: What is the patient's HAS-BLED score?

	AS	CEc	CEe	CL	MCr	MCT	MCu	MS	QLI	QLq	SNs	KNa	KNc	KNf	KNn	KNs	AT	VO	UG
X₃	2	3	4	0	2	2	1	0	3	2	0	5	0	2	4	0	3	2	100
KNa	0.18	0.11	-0.11	0.22	0.13	0.27	0.39	-0.09	0.13	0.07	0.33	1.0	-0.25	0.18	0.71	0.05	0.07	0.11	-0.15
KNc	-0.05	-0.1	-0.17	-0.24	-0.08	-0.2	-0.28	0.34	-0.27	-0.49	-0.37	-0.25	1.0	-0.55	-0.42	0.47	-0.11	-0.22	-0.22
KNf	0.14	0.29	0.36	0.51	0.29	0.47	0.45	-0.2	0.58	0.7	0.46	0.18	-0.55	1.0	0.31	-0.53	0.26	0.46	0.27
KNn	0.18	0.05	-0.0	0.24	0.14	0.24	0.37	-0.25	0.19	0.2	0.48	0.71	-0.42	0.31	1.0	-0.3	0.13	0.18	0.08
KNs	0.07	0.05	-0.2	-0.11	0.01	-0.08	-0.04	0.3	-0.25	-0.52	-0.29	0.05	0.47	-0.53	-0.3	1.0	-0.01	-0.19	-0.34
AT	0.5	0.53	0.38	0.59	0.62	0.54	0.57	0.12	0.57	0.18	0.31	0.07	-0.11	0.26	0.13	-0.01	1.0	0.6	0.3
VO	0.48	0.55	0.48	0.66	0.6	0.63	0.57	0.05	0.68	0.42	0.4	0.11	-0.22	0.46	0.18	-0.19	0.6	1.0	0.3
UG	0.19	0.03	0.36	0.13	0.17	0.05	0.1	-0.19	0.21	0.22	0.27	-0.15	-0.22	0.27	0.08	-0.34	0.3	0.3	1.0



(COMMON) UNDERSTANDING

- Humans understanding
 - Good refinement
 - High inter-rater agreement

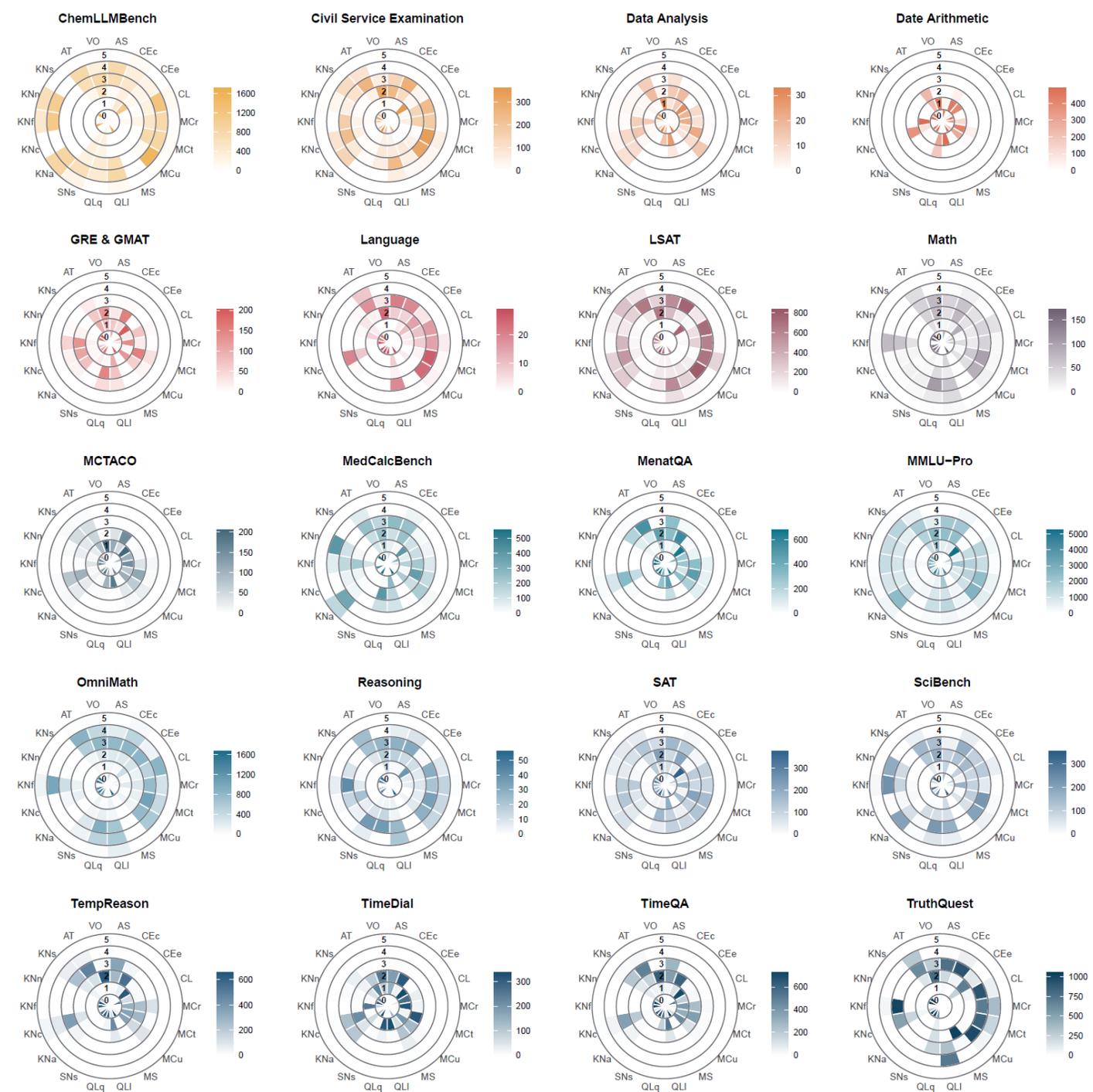
Explanatory potential: these agreement indicators show they are interpreted in a consistent way by humans and LLMs.

agreement of ratings (r_{WG} scores)

Dimension	Humans	Delphi & GPT-4o
AS	0.91	0.86
CEc	0.91	0.87
CEe	0.90	0.94
CL	0.78	0.82
MCr	0.79	0.84
M Ct	0.88	0.91
MCu	0.80	0.81
MS	0.77	0.86
QL1	0.85	0.89
QLq	0.84	0.84
SNs	0.87	0.89
KNa	0.73	0.75
KNc	0.86	0.83
KNf	0.86	0.81
KNn	0.91	0.94
KNs	0.70	0.86
AT	0.80	0.83
VO	0.84	0.91
Average	0.83	0.86

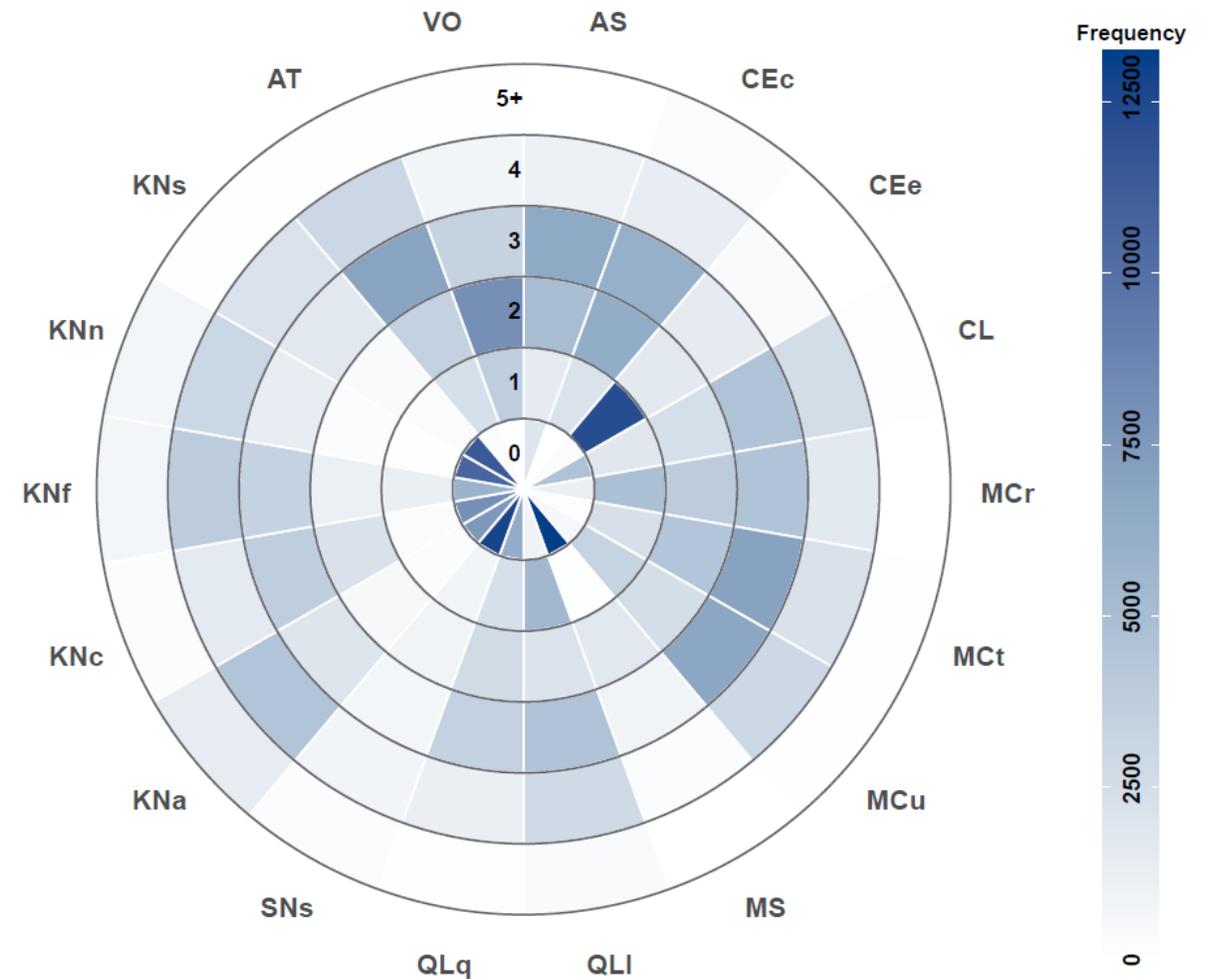
CONSTRUCT VALIDITY

- All benchmarks either don't strictly measure what they claim to measure (**lacking specificity**) or tend to only include intermediate difficulties for the target dimension (**lacking sensitivity**).
- Assigning one/more benchmarks to one "capability" and aggregating accuracy is hence highly confounded.



SPECIFICITY & SENSITIVITY

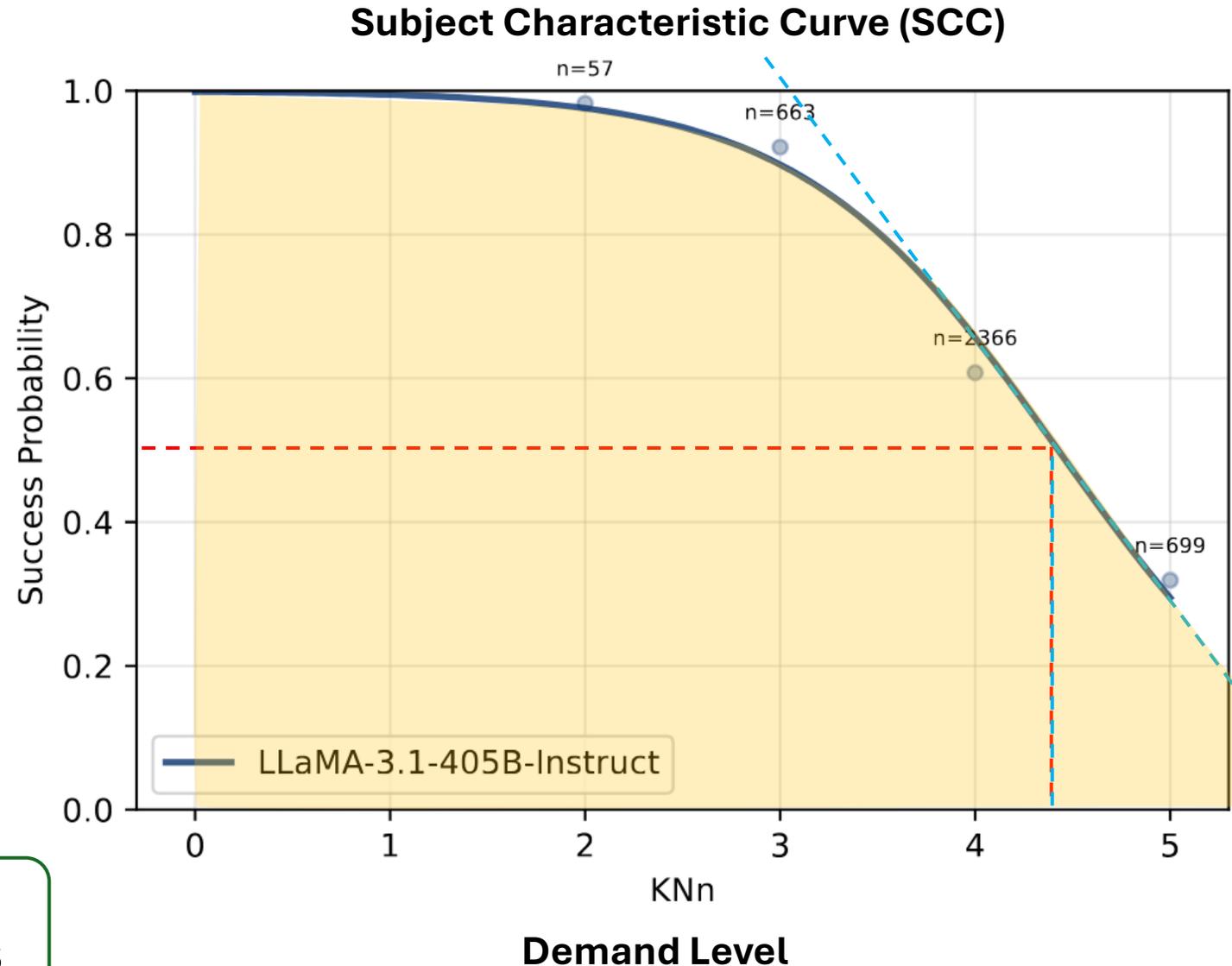
- Using the same scales
 - We can see gaps in evaluation
 - We can mix the best instances from many benchmarks.
 - We don't need to replace old instances
- ADeLe v1.0 \Rightarrow v1.1



ABILITIES FROM SCCs

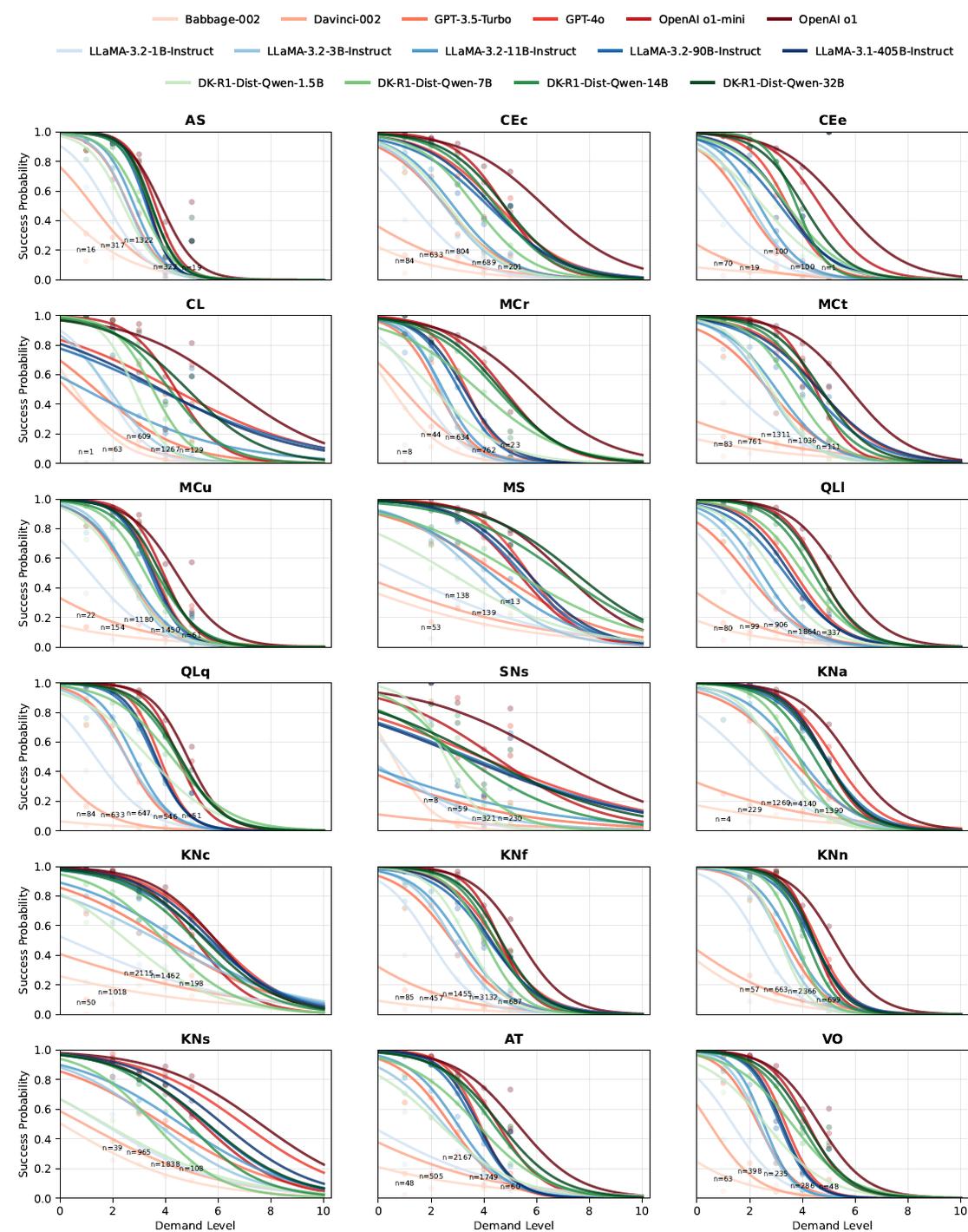
- 18 dimensions to abilities?
 - ‘Dominant’ Slicing:
 - Example: dimension KNn
 - For each level k , all other dimensions $\leq k$.
 - ADeLe battery: 16,108 instances to 3,785.
 - Subject Characteristic Curve (logistic fit):
 - x -value of point of maximum slope
 - x -value where success prob. = 0.5
 - area under the curve from $x=0$.

LLaMA-3.1-405B-Instruct's KNn ability (knowledge about natural sciences) is **4.3**



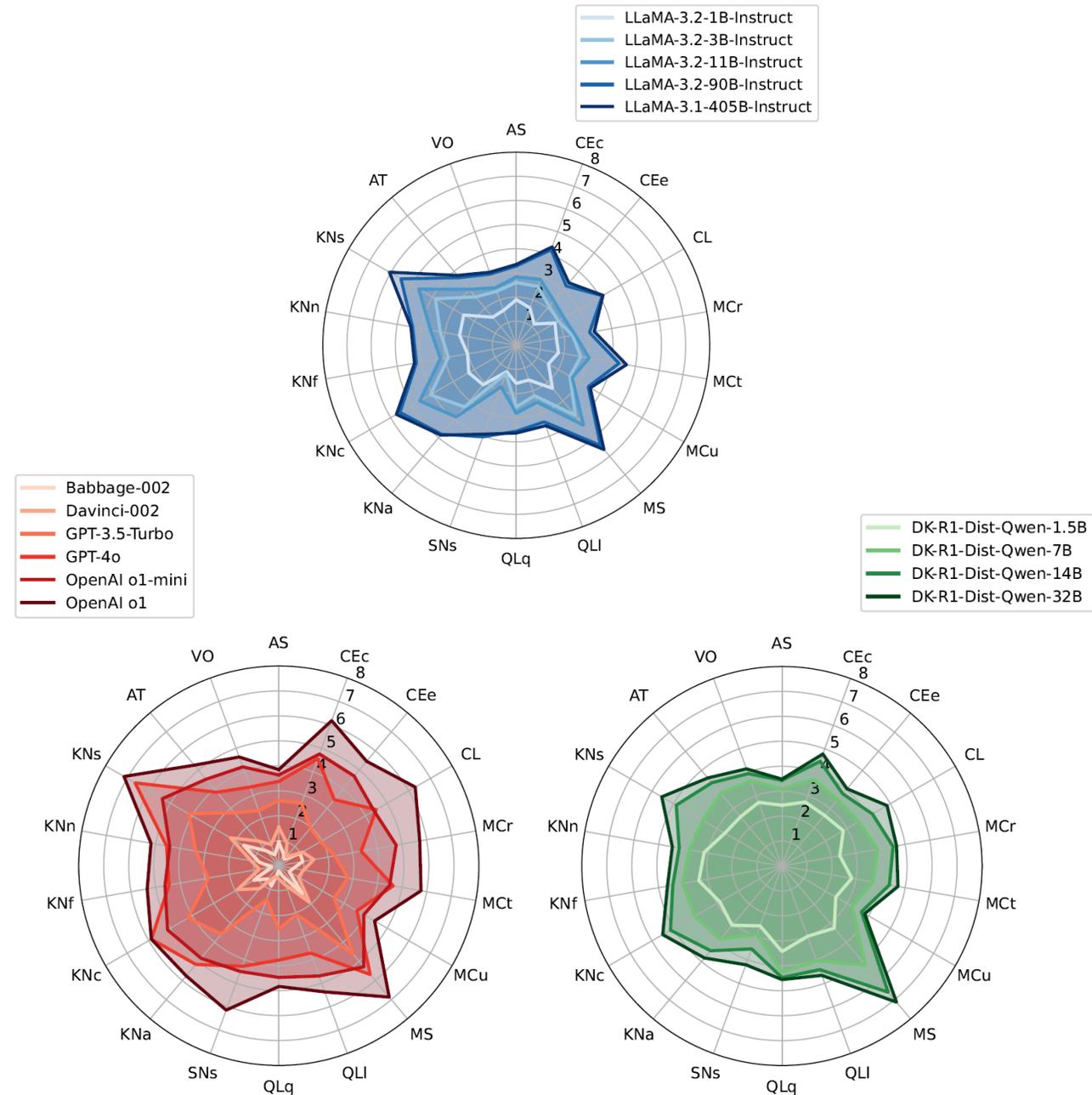
SCC ANALYSIS

- The SCCs of certain dimensions are steep:
 - with low variability across models (e.g. as AS and MCu)
 - ability is around demand levels 3-4
 - explain (and predict) success very well for instances in the low and high ranges.
- SCCs of other dimensions are flatter and show strong differences between models (e.g. KNs):
 - Discrimination, between success and failure, is lower.

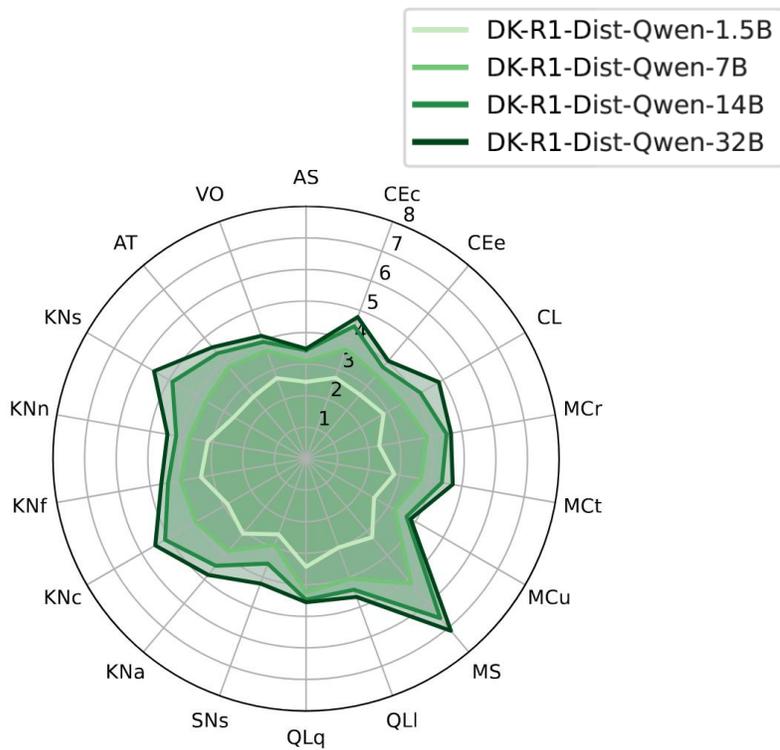


ABILITY PROFILES

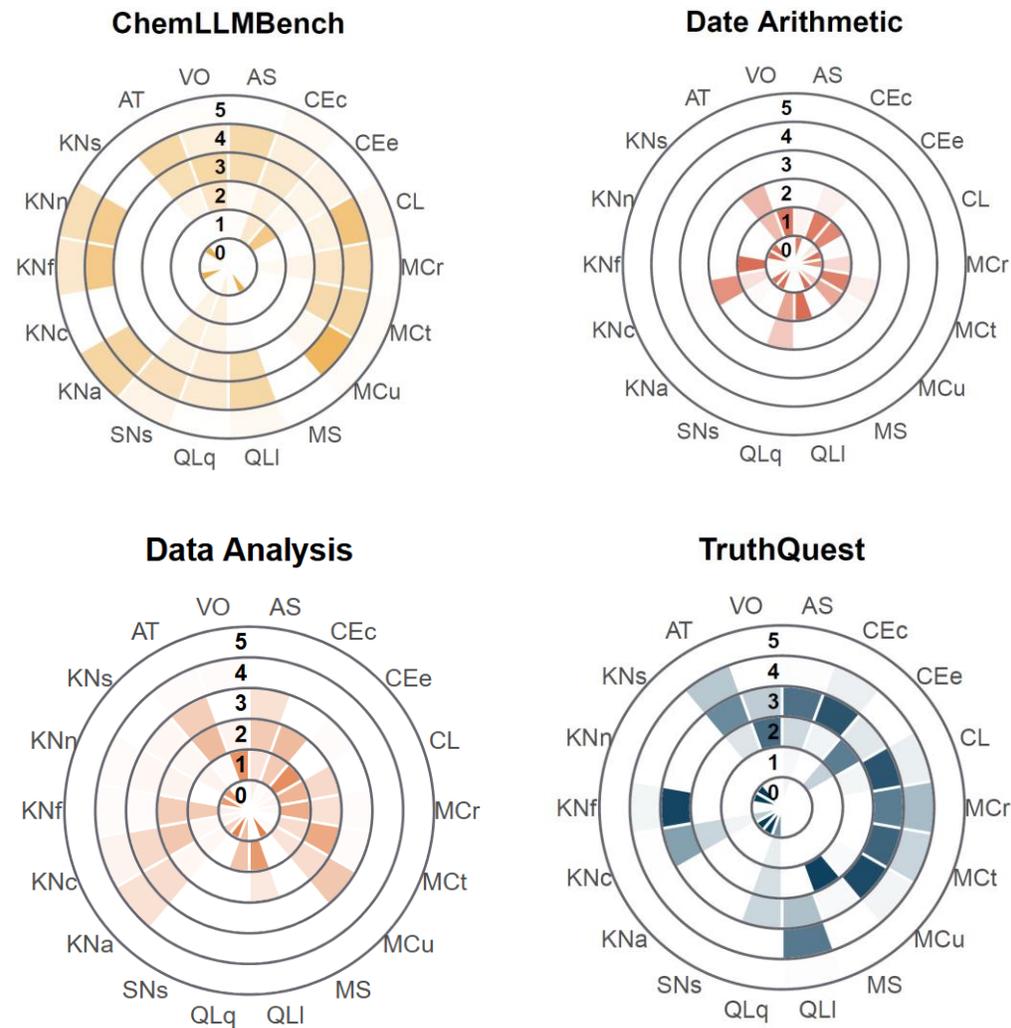
- Newer models have higher abilities than older ones, but this ranking is **not monotonic** for all abilities.
- Knowledge dimensions are limited by **model size and distillation processes**
- Reasoning, learning and abstraction, and social capabilities, are boosted by chain-of-thought, inference-heavy models



ABILITY PROFILES VS DEMAND PROFILES



VS



IN-DISTRIBUTION

Subject LLM	LLM Accuracy↑	Demands (RF)		Embeddings (RF)		Finetuning (LLAMA)	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Babbage-002	0.102	0.786	0.004	0.784	0.012	0.794	0.026
Davinci-002	0.157	0.774	0.005	0.770	0.014	0.789	0.032
GPT-3.5-Turbo	0.414	0.811	0.007	0.780	0.029	0.817	0.052
GPT-4o	0.713	0.882	0.014	0.852	0.041	0.879	0.039
OpenAI o1-mini	0.770	0.860	0.011	0.821	0.023	0.861	0.041
OpenAI o1	0.843	0.853	0.011	0.810	0.025	0.848	0.031
LLaMA-3.2-1B-Instruct	0.216	0.785	0.006	0.759	0.014	0.788	0.041
LLaMA-3.2-3B-Instruct	0.378	0.813	0.008	0.782	0.028	0.822	0.048
LLaMA-3.2-11B-Instruct	0.463	0.820	0.009	0.793	0.034	0.828	0.055
LLaMA-3.2-90B-Instruct	0.645	0.860	0.012	0.832	0.042	0.860	0.042
LLaMA-3.1-405B-Instruct	0.683	0.870	0.011	0.831	0.040	0.864	0.040
DK-R1-Dist-Qwen-1.5B	0.353	0.781	0.014	0.749	0.028	0.797	0.052
DK-R1-Dist-Qwen-7B	0.555	0.813	0.015	0.788	0.039	0.821	0.051
DK-R1-Dist-Qwen-14B	0.698	0.828	0.013	0.796	0.031	0.829	0.044
DK-R1-Dist-Qwen-32B	0.748	0.841	0.013	0.799	0.031	0.839	0.045
Weighted Average	—	0.839	0.011	0.805	0.032	0.840	0.043

TASK OUT-OF-DISTRIBUTION

Subject LLM	LLM Accuracy↑	Demands (RF)		Embeddings (RF)		Finetuning (LLAMA)	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Babbage-002	0.102	0.751	0.007	0.727	0.019	0.719	0.046
Davinci-002	0.157	0.741	0.007	0.703	0.025	0.746	0.055
GPT-3.5-Turbo	0.414	0.795	0.020	0.719	0.032	0.773	0.088
GPT-4o	0.713	0.852	0.023	0.789	0.073	0.831	0.067
OpenAI o1-mini	0.770	0.837	0.021	0.751	0.038	0.814	0.068
OpenAI o1	0.843	0.811	0.033	0.730	0.030	0.761	0.101
LLaMA-3.2-1B-Instruct	0.216	0.733	0.026	0.671	0.033	0.732	0.081
LLaMA-3.2-3B-Instruct	0.378	0.791	0.016	0.724	0.020	0.780	0.084
LLaMA-3.2-11B-Instruct	0.463	0.799	0.022	0.733	0.037	0.783	0.106
LLaMA-3.2-90B-Instruct	0.645	0.834	0.021	0.763	0.068	0.809	0.050
LLaMA-3.1-405B-Instruct	0.683	0.843	0.023	0.766	0.067	0.811	0.060
DK-R1-Dist-Qwen-1.5B	0.353	0.757	0.019	0.700	0.029	0.764	0.071
DK-R1-Dist-Qwen-7B	0.555	0.790	0.018	0.735	0.042	0.776	0.083
DK-R1-Dist-Qwen-14B	0.698	0.808	0.018	0.737	0.054	0.772	0.085
DK-R1-Dist-Qwen-32B	0.748	0.812	0.026	0.739	0.057	0.793	0.063
Weighted Average	—	0.810	0.022	0.740	0.047	0.788	0.075

BENCHMARK OUT-OF-DISTRIBUTION

Subject LLM	LLM Accuracy↑	Demands (RF)		Embeddings (RF)		Finetuning (LLAMA)	
		AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Babbage-002	0.102	0.694	0.027	0.689	0.062	0.649	0.070
Davinci-002	0.157	0.718	0.014	0.626	0.066	0.633	0.086
GPT-3.5-Turbo	0.414	0.776	0.041	0.628	0.074	0.691	0.146
GPT-4o	0.713	0.826	0.058	0.398	0.167	0.740	0.136
OpenAI o1-mini	0.770	0.728	0.026	0.422	0.142	0.684	0.132
OpenAI o1	0.843	0.710	0.015	0.404	0.117	0.704	0.095
LLaMA-3.2-1B-Instruct	0.216	0.716	0.048	0.602	0.112	0.623	0.083
LLaMA-3.2-3B-Instruct	0.378	0.778	0.036	0.618	0.096	0.687	0.066
LLaMA-3.2-11B-Instruct	0.463	0.786	0.053	0.591	0.067	0.721	0.118
LLaMA-3.2-90B-Instruct	0.645	0.804	0.055	0.463	0.115	0.721	0.144
LLaMA-3.1-405B-Instruct	0.683	0.818	0.044	0.389	0.186	0.712	0.135
DK-R1-Dist-Qwen-1.5B	0.353	0.705	0.049	0.580	0.102	0.662	0.106
DK-R1-Dist-Qwen-7B	0.555	0.676	0.043	0.534	0.060	0.649	0.160
DK-R1-Dist-Qwen-14B	0.698	0.691	0.025	0.461	0.099	0.673	0.135
DK-R1-Dist-Qwen-32B	0.748	0.703	0.027	0.426	0.103	0.696	0.100
Weighted Average	—	0.747	0.037	0.480	0.114	0.692	0.121

SOURCES OF UNPREDICTABILITY

- **Epistemic Uncertainty:**
 - DeLean v1.0 missing some relevant dimensions
 - DeLean v1.0 demand level up to 5 only
 - ADeLe v1.0 imperfectly covering the demand space
- **Aleatoric Uncertainty:**
 - Guess rate in multiple-choice questions (UG)
 - Memorisation (AT)
 - High-quality but imperfect graders (98% accuracy)
 - High-quality but imperfect demand annotators

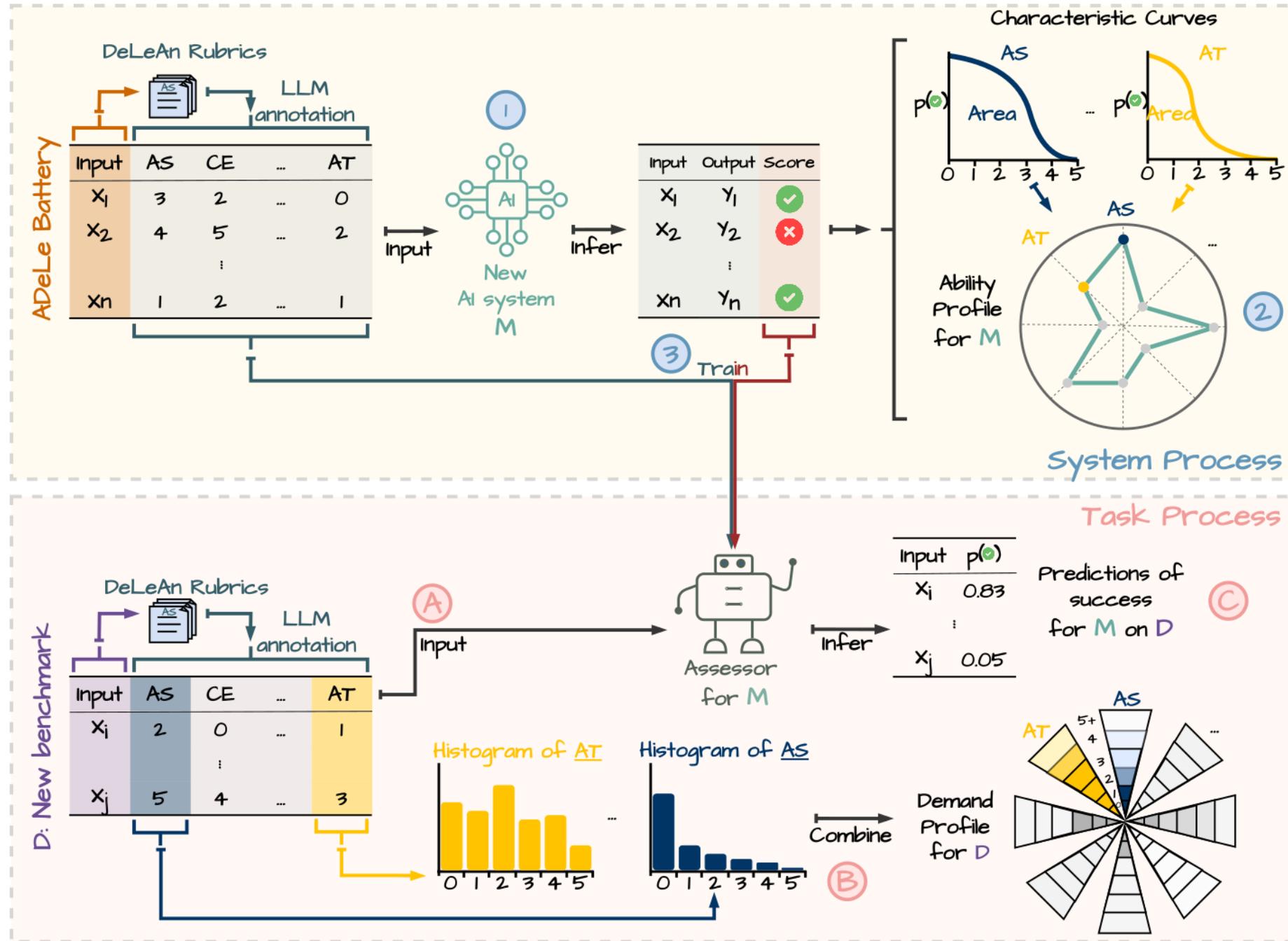
**IT CAN ONLY GET BETTER!
ADeLe v2.0 !
DeLean v2.0 !**

USING IT

Two possible entry points:

- Analyse an **AI system**
- Analyse a **benchmark**

Assessors can predict performance for **completely new items and benchmarks**, even imaginary ones



EXTENDING IT COLLABORATIVELY

- Add **new dimensions** to DeLeAn:
 - Multimodal dimensions
 - Embodied AI and Robotics
- Add **new levels** to DeLeAn:
 - Turn 5+ into 5-10 (very advanced AI levels)
- Add **new instances** to ADeLe:
 - Fill the gaps:
 - Humanity's Last Exam
 - Enigmaeval
 - Big-bench extra hard
- Equate scales with **human results**

Collaborative Platform at CFI-Cambridge!

<https://kinds-of-intelligence-cfi.github.io/ADELE/>

ADeLe v1.0: A battery for AI Evaluation with explanatory and predictive power

[Original Paper](#) [Dataset](#) [LLM results](#) [X thread](#)

This is a collaborative community, initiated by researchers at the [Leverhulme Centre for the Future of Intelligence](#) from Cambridge University and the [Center for Information Technology Policy](#) from Princeton University, for the use and extension of ADeLe v1.0, a battery for AI evaluation with explanatory and predictive power, currently focusing on LLMs.

The ADeLe ([Annotated-Demand-Levels](#)) battery includes 63 tasks from 20 benchmarks and was introduced in [the original paper](#). This battery was annotated using 18 rubrics for [Demand-Level-Annotation](#) (DeLeAn v1.0) of general scales.

A route for standardisation?

PART V : GENERALITY AND SAFETY

“I was talking to Ben [Goertzel] and I was like,
‘Well, if it’s about the generality that AI
systems don’t yet have, we should just call it
Artificial General Intelligence.’”

Shane Legg, Google DeepMind’s co-founder.

Generality vs AGI: Characterising GPAI

Pointers:

- Hernández-Orallo, J., Loe, B. S., Cheke, L., Martínez-Plumed, F., & Ó hÉigeartaigh, S. (2021). General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific reports*, 11(1), 22822.
- Hernandez-Orallo, J. (2024). Caveats and solutions for characterising general-purpose AI. In *ECAI 2024* (pp. 2-9). IOS Press.

NATURAL GENERALITY

- **General intelligence and the g factor**
 - *Spearman*: same latent factor explains performance in a range of cognitive tests.
- **Convergent evolution**
 - *General intelligence* is one of these traits!
 - Altricial vs precocial / nurture vs nature
 - Social hypothesis for general intelligence
 - Cultural hypothesis for more general intelligence

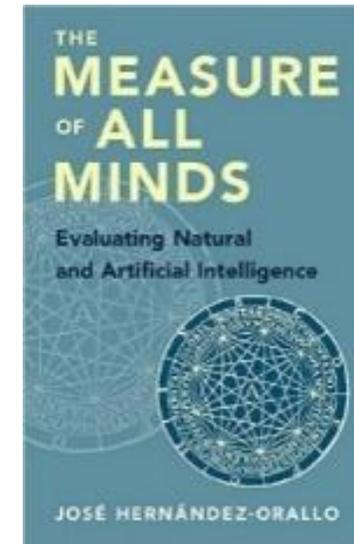
General intelligence is successful behaviour in a wide range of situations (up to a level of difficulty or resources)

Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h'Eigeartaigh, S. "General intelligence disentangled via a generality metric for natural and artificial intelligence", Nature Sci Rep 2021



GENERALITY IN AI

- Lull's *Ars Generalis*
- Turing's "child machine"
- Simon & Newell's "General Problem Solver"
- McCarthy's dream for generality
- Solomonoff's theory of prediction



1971
Turing
Award
Lecture

Generality in Artificial Intelligence

JOHN MCCARTHY
Stanford University

The Turing Award Lecture given in 1971 by John McCarthy was never published. The postscript that follows, written by the author in 1986, endeavors to reflect the flavor of the original, as well as to comment in the light of development over the past 15 years.

Postscript

My 1971 Turing Award Lecture was entitled "Generality in Artificial Intelligence." The topic turned out to have been overambitious in that I discovered that I was unable to put my thoughts on the subject in a satisfactory written form at that time. It would have been better to have reviewed previous work rather than attempt something new, but such wasn't my custom at that time.

I am grateful to the ACM for the opportunity to try again. Unfortunately for our science, although perhaps fortunately for this project, the problem of generality in artificial intelligence (AI) is almost as unsolved as ever, although we now have many ideas not available in

THE GENERALITY IS HERE

- At least since 2020 (GPT-3)
- Made possible by:
 - *the transformers:*

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

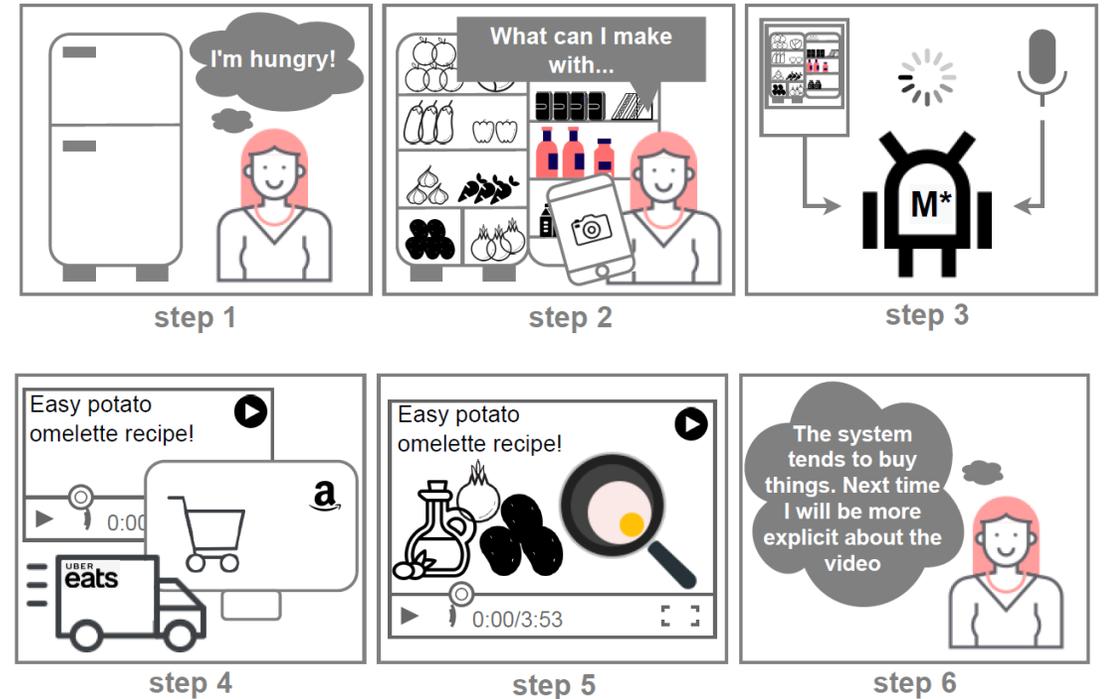
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

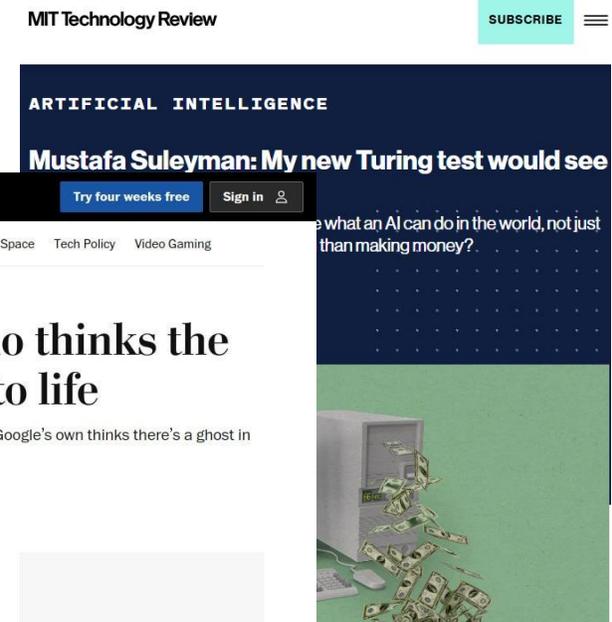
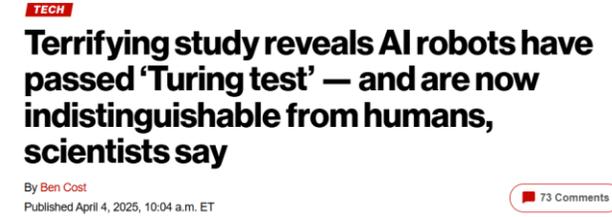
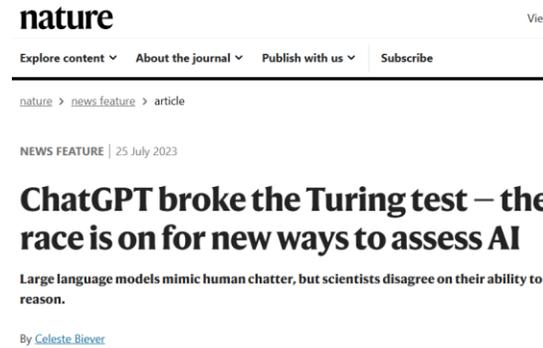
Illia Polosukhin* ‡
illia.polosukhin@gmail.com



Schellaert, Plumed, Vold, Burden, Casares, Loe, Reichart, OhEigartaigh, Korhonen, Orallo "Your Prompt is My Command: Assessing the Human-Centred Generality of Multi-Modal Models". JAIR, 2023,

INTERPRETATIONS OF AGI

- Variants of the Turing test
 - Human-Level Machine Intelligence: average human, all humans, ...
- Moravec's Paradox: easy tasks for humans unsolved by AI
 - Wozniak's coffee test (& Chollet's umpteenth AGI challenge)
- Economic Value
 - "Replacing" any job
 - Sulleyman's making \$1M out of \$0.1M
- Consciousness / eye of the beholder
 - Blake Lemoine
 - Gary Marcus's "it will be AGI when I say it is AGI"

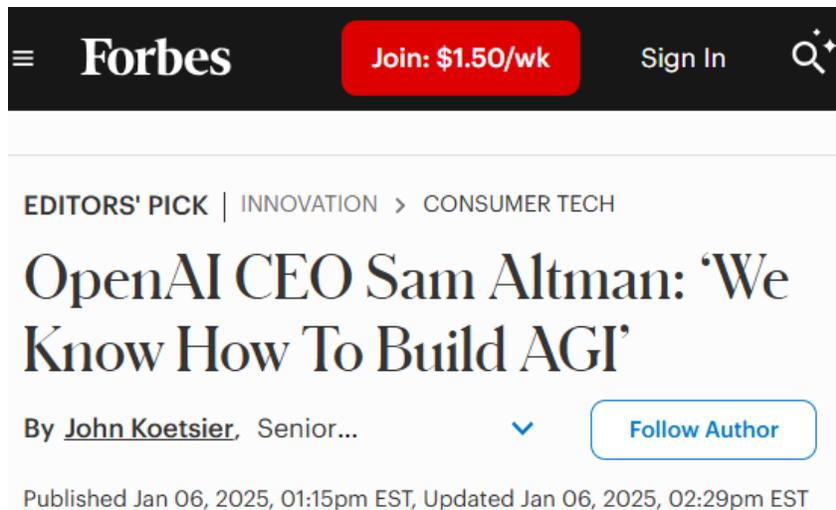


AGI AS A GOAL (OR A LEGAL TRIGGER!)

OpenAI Charter

Our Charter describes the principles we use to execute on OpenAI's mission.

We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be “a better-than-even chance of success in the next two years.”



The screenshot shows the top portion of a Forbes article. At the top left is the Forbes logo. To its right is a red button that says "Join: \$1.50/wk". Further right are the words "Sign In" and a search icon. Below this is a navigation bar with the text "EDITORS' PICK | INNOVATION > CONSUMER TECH". The main headline of the article is "OpenAI CEO Sam Altman: 'We Know How To Build AGI'". Below the headline, it says "By [John Koetsier](#), Senior..." followed by a dropdown arrow and a "Follow Author" button. At the bottom of the screenshot, it reads "Published Jan 06, 2025, 01:15pm EST, Updated Jan 06, 2025, 02:29pm EST".

Forbes

Join: \$1.50/wk

Sign In

EDITORS' PICK | INNOVATION > CONSUMER TECH

OpenAI CEO Sam Altman: 'We Know How To Build AGI'

By [John Koetsier](#), Senior... [Follow Author](#)

Published Jan 06, 2025, 01:15pm EST, Updated Jan 06, 2025, 02:29pm EST

AGI “LEVELS”

Generality dichotomised (yes/no), and AGI scale seen as unidimensional

AGI redefined as **strictly-digital “AGI”**:
We’re at level 1 but AI can’t clean my toilet!

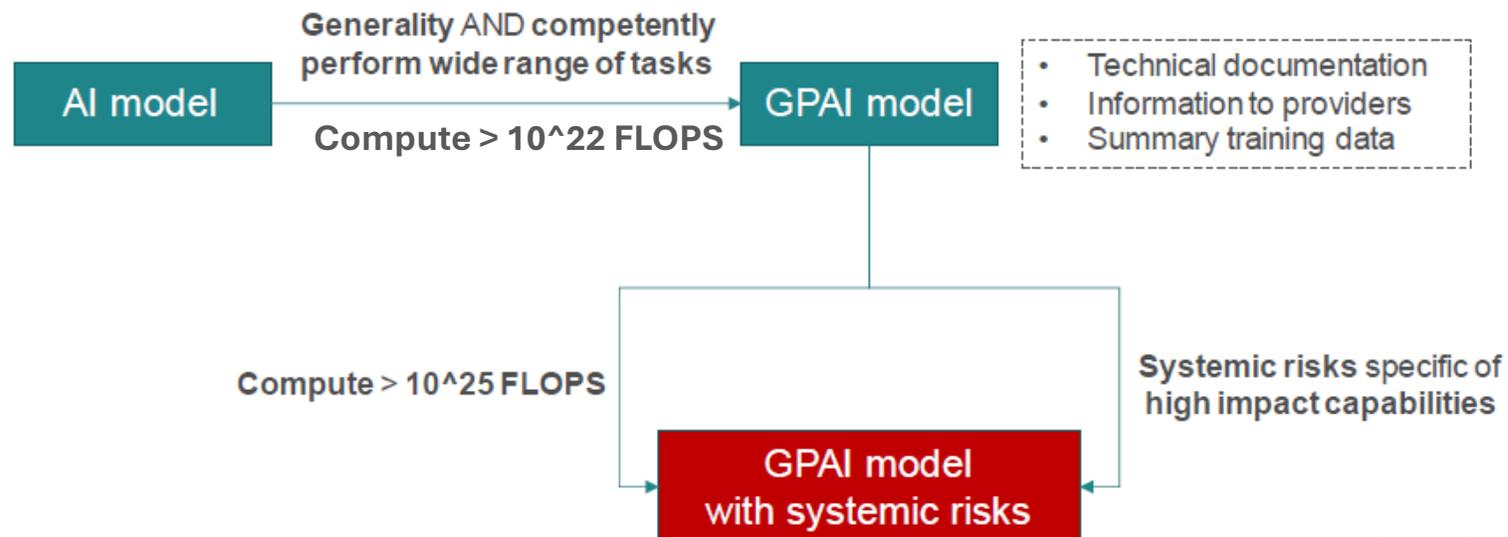
Performance (rows) x Generality (columns)	Narrow <i>clearly scoped task or set of tasks</i>	General <i>wide range of non-physical tasks, including metacognitive tasks like learning new skills</i>
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging <i>equal to or somewhat better than an unskilled human</i>	Emerging Narrow AI GOFAI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai & Hassabis, 2023)
Level 2: Competent <i>at least 50th percentile of skilled adults</i>	Competent Narrow AI toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert <i>at least 90th percentile of skilled adults</i>	Expert Narrow AI spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso <i>at least 99th percentile of skilled adults</i>	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016; 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman <i>outperforms 100% of humans</i>	Superhuman Narrow AI AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved

Morris, M.R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C. and Legg, S., 2024, July. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*.

AGI AS COMPUTE (∞ PARAMETER SIZE)

“General Purpose AI models” (EU AI ACT)

- means an AI model, including when trained with a large amount of data using self-supervision at scale, that **displays significant generality** and is **capable to competently perform a wide range of distinct tasks** regardless of the way the model is placed on the market and **that can be integrated into a variety of downstream systems or applications**. This does not cover AI models that are used before release on the market for research, development and prototyping activities;



AGI AS A SELF-IMPROVEMENT BOOTSTRAP

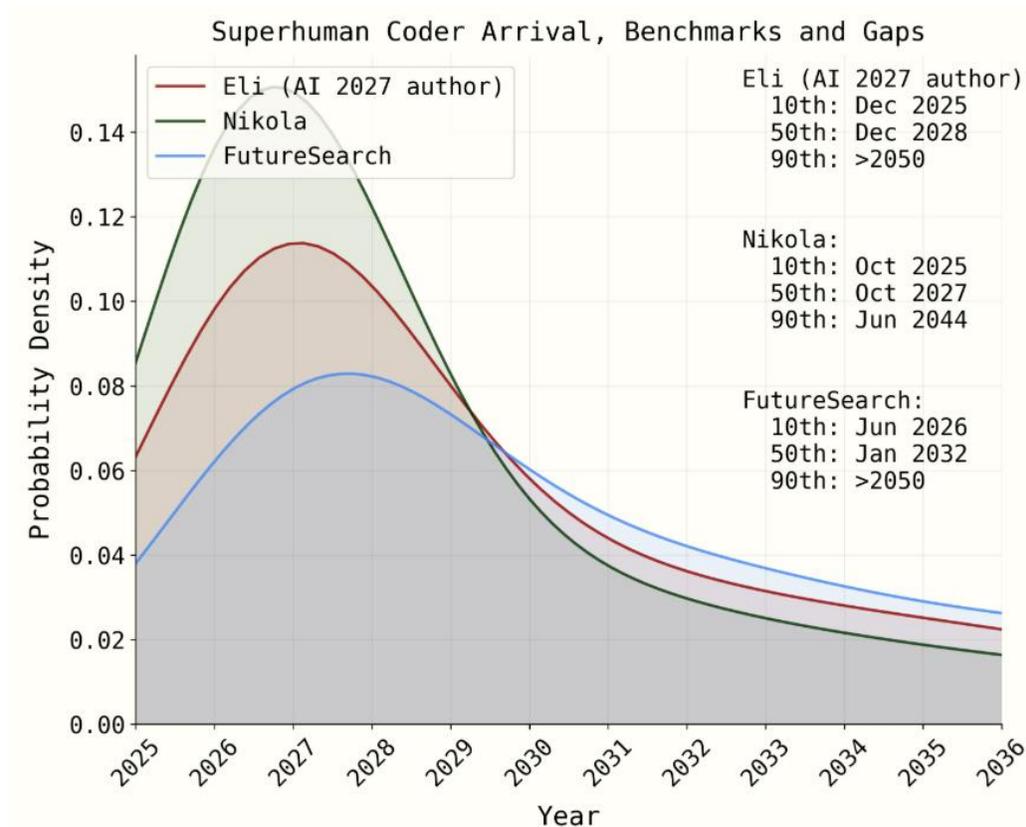
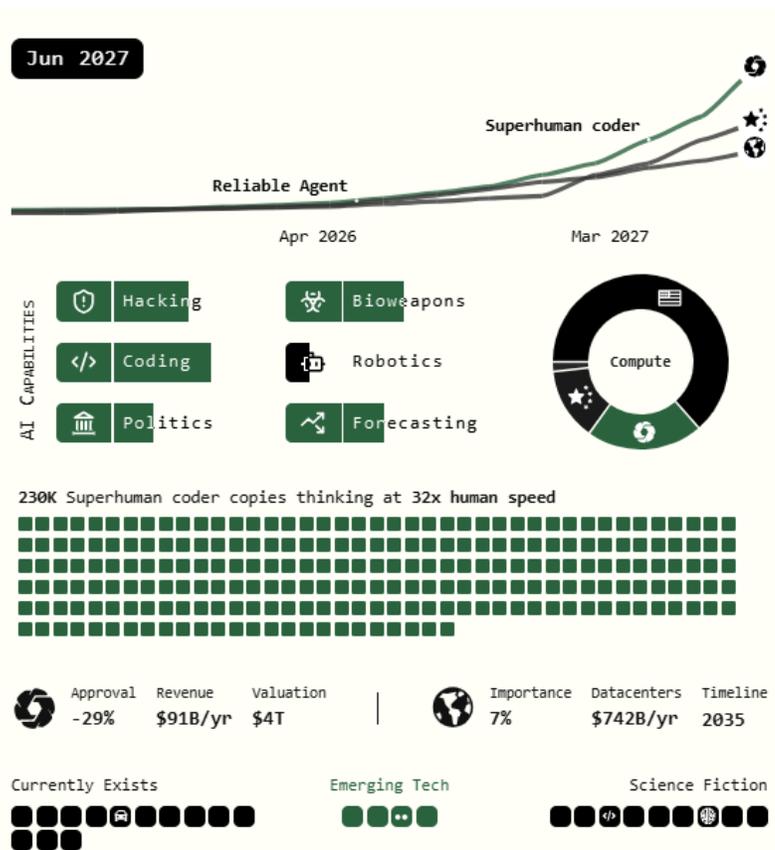
SITUATIONAL AWARENESS: The Decade Ahead

Leopold Aschenbrenner, June 2024

You can see the future first in San Francisco.

AI 2027

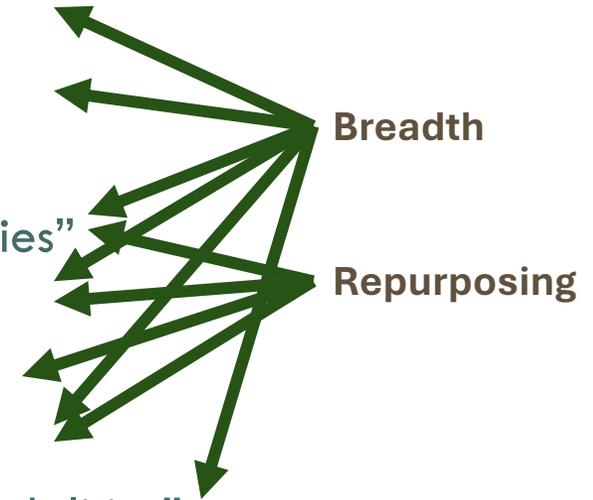
Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean



PERCEPTIONS OF GENERALITY

- Different perceptions of generality:

1. Ideal: “Success in all situations”
2. Range: “Success in a wide range of situations”
3. Human-general: “Able to do everything a human can do”
4. Core capabilities: “Having an elemental range of capabilities”
5. OOD: “Success out of distribution”
6. Transferability: “Flexibility to easily adapt to new tasks”
7. Compositionality: “Integration of different skills”
8. Multimodality: “Integration of different input and output modalities”

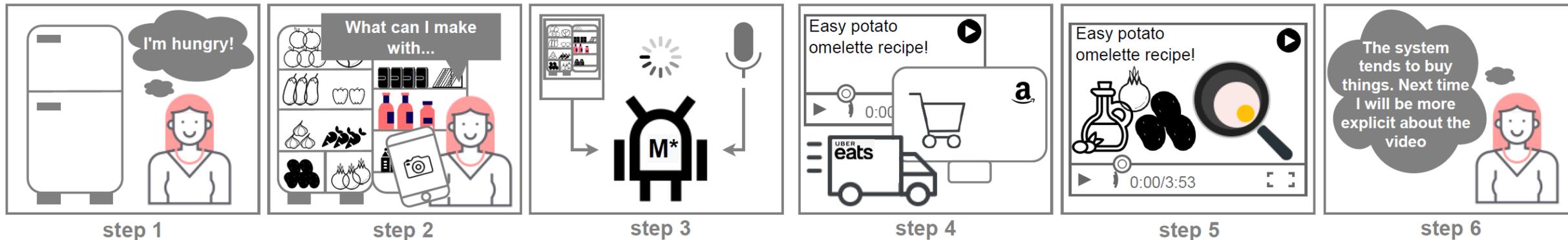


USER-CENTRED GENERALITY

- “Wide range of tasks” “easy to get done” “well”

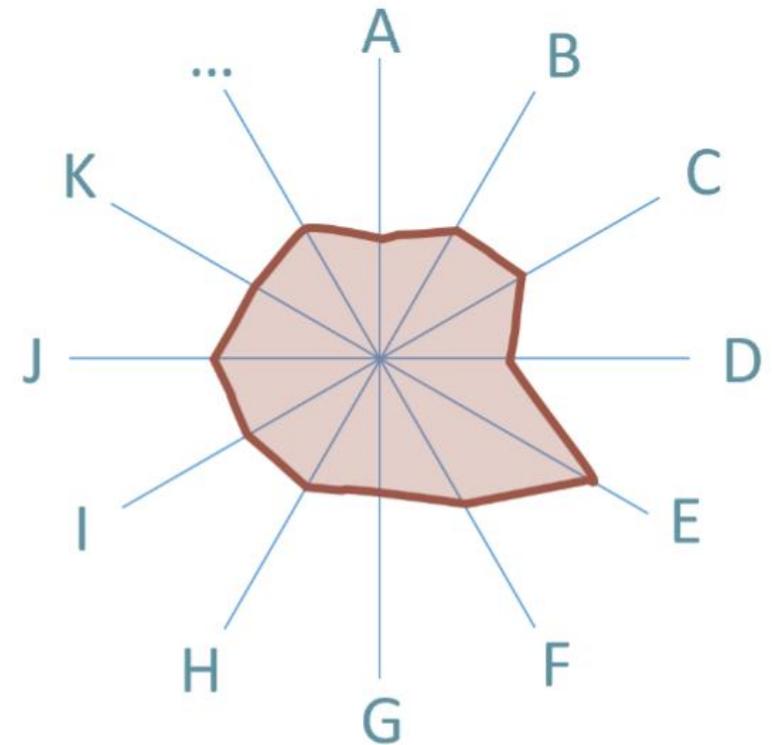
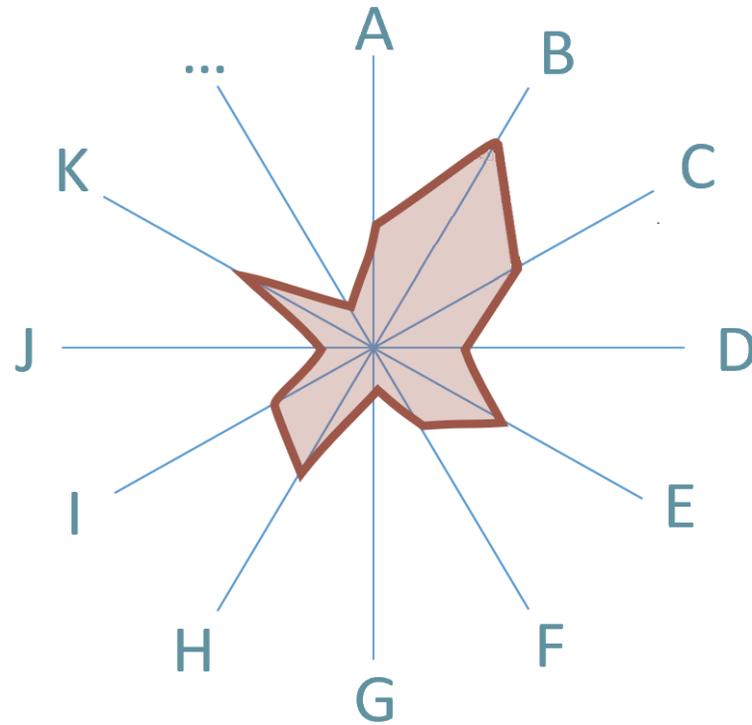
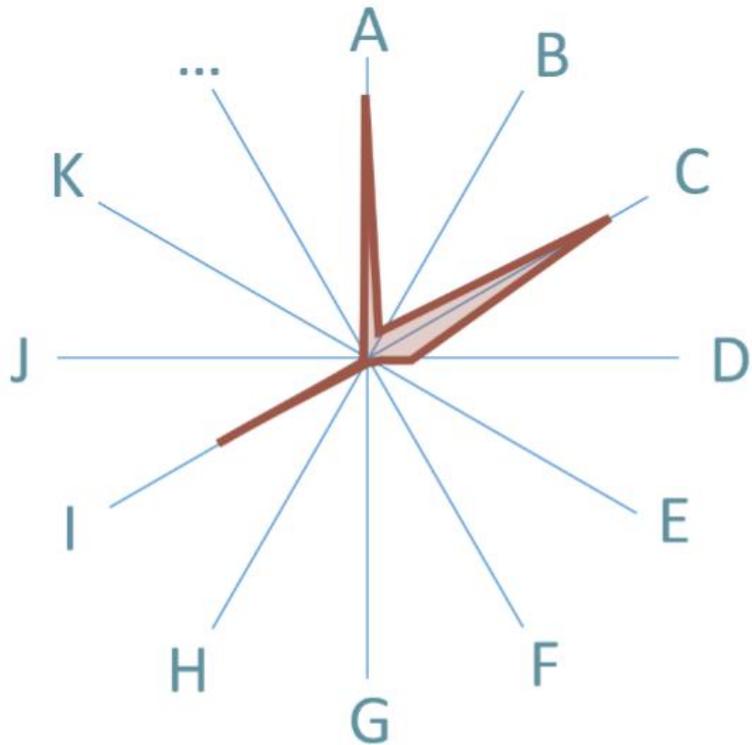
$$V_h(M^*) = \sum_{t,p} \underbrace{\mathbb{P}(t|h)}_{\text{Tasks}} \cdot \underbrace{\mathbb{P}(p|t, h, M^*)}_{\text{Prompts}} \underbrace{v_h(M^*, t, p)}_{\text{Utility}}$$

Includes many components:
 + Value and quality of the result
 – Cost of prompting, scaffolding, finetuning, adaptation



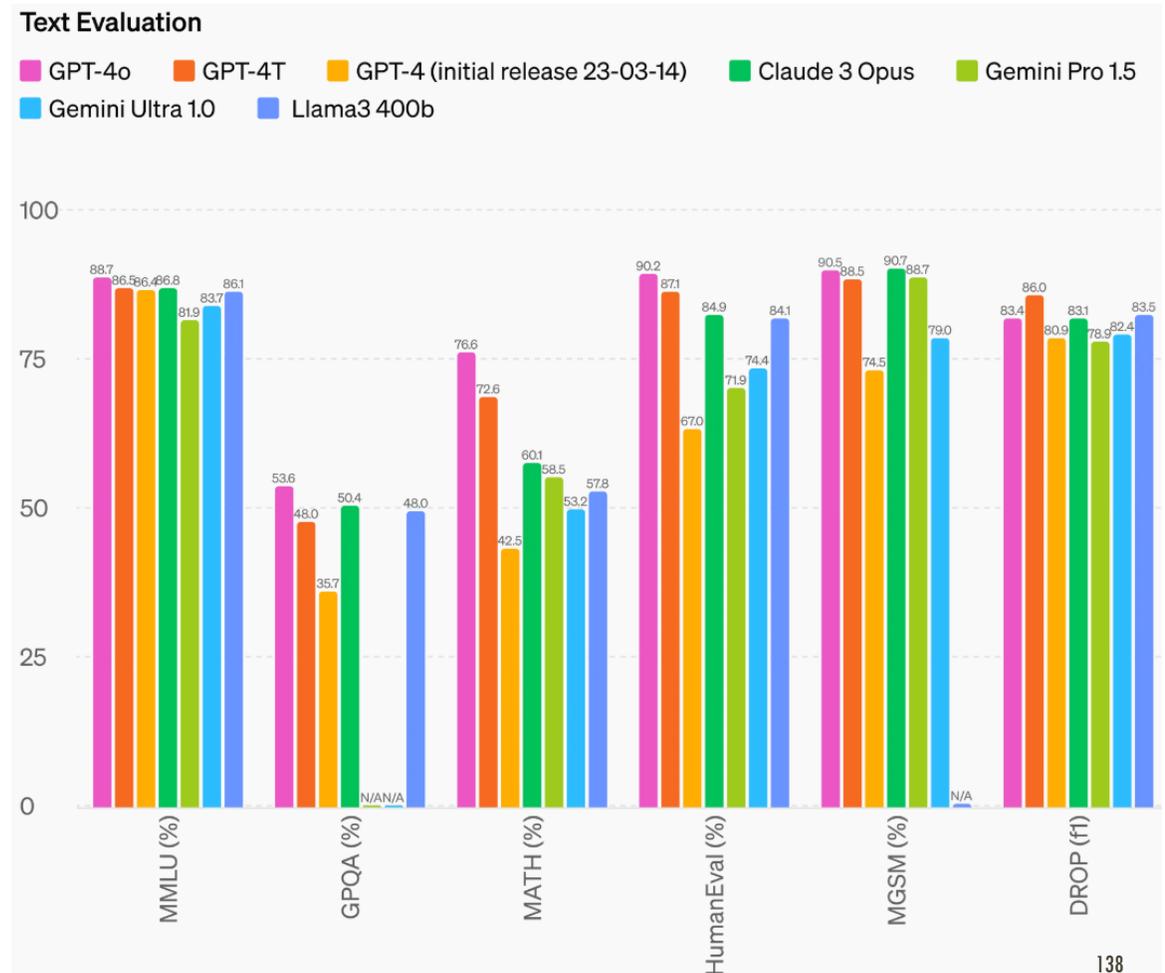
COVERING MULTIPLED-DOMAINS - BREADTH

- Which is the most general one?



BREADTH, NOT MESS

- Not like this:
 - Adding apples and oranges.
 - GPQA hard, MMLU easy
 - What if one missing?
 - What if some contaminated?



GENERALITY AS COMPACTNESS

Definitions

■ Capability (Ψ), the area under the ACC: $\Psi_j \stackrel{\text{def}}{=} \int_0^\infty \psi_j(h) dh$

■ Expected difficulty given success:

$$\mathbb{H}_j \stackrel{\text{def}}{=} \mathbb{E}_{h \sim f_j} [h] = \frac{M_j}{\Psi_j} \quad M_j \stackrel{\text{def}}{=} \int_0^\infty h \cdot \psi_j(h) dh$$

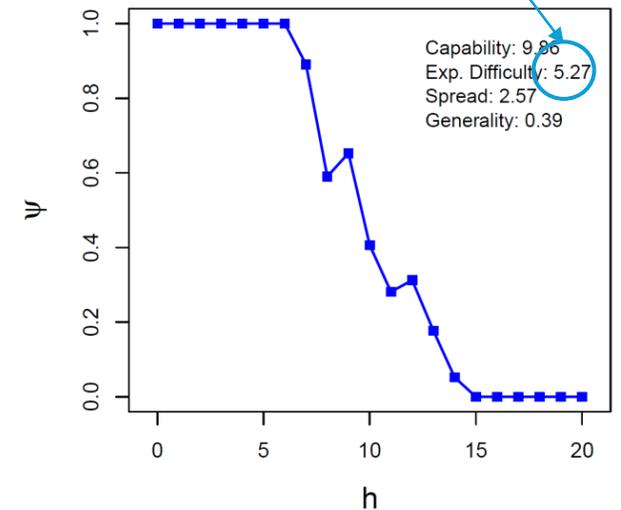
■ Spread:

$$S_j \stackrel{\text{def}}{=} \sqrt{(2\mathbb{H}_j - \Psi_j) \cdot \Psi_j} = \sqrt{2M_j - \Psi_j^2}$$

■ Generality:

$$\Gamma_j \stackrel{\text{def}}{=} \frac{1}{S_j}$$

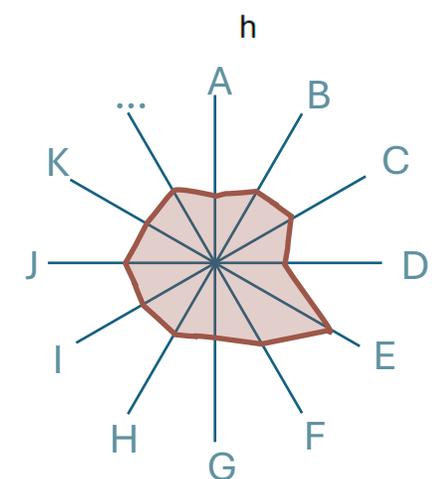
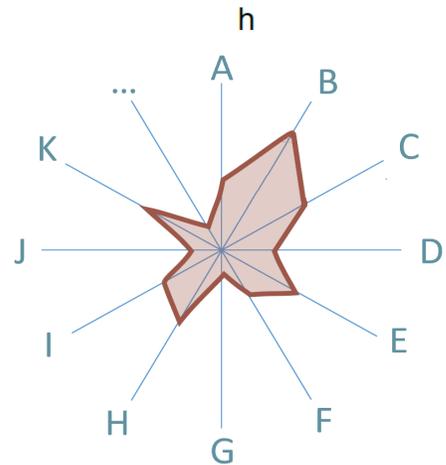
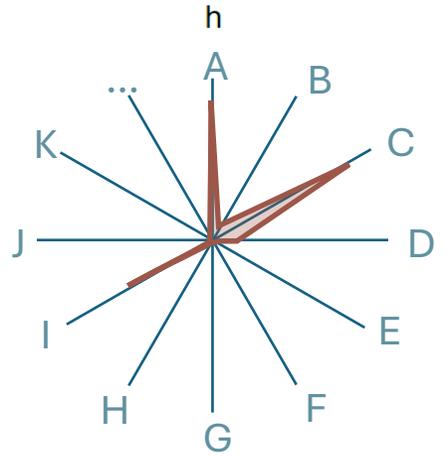
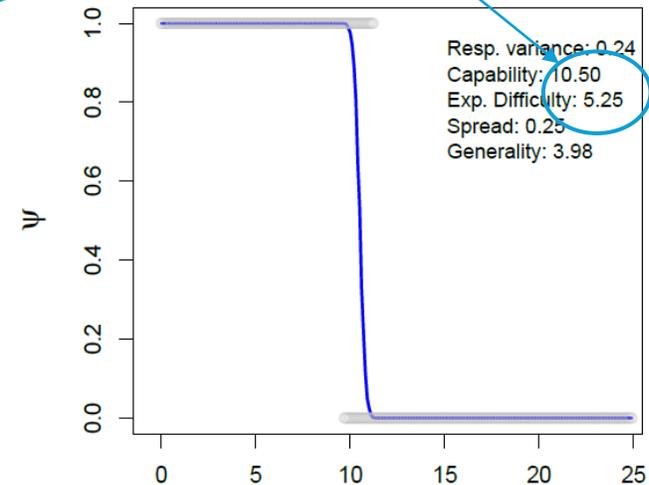
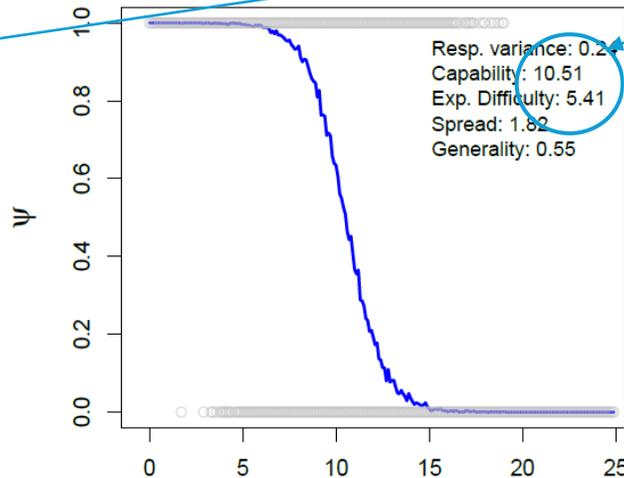
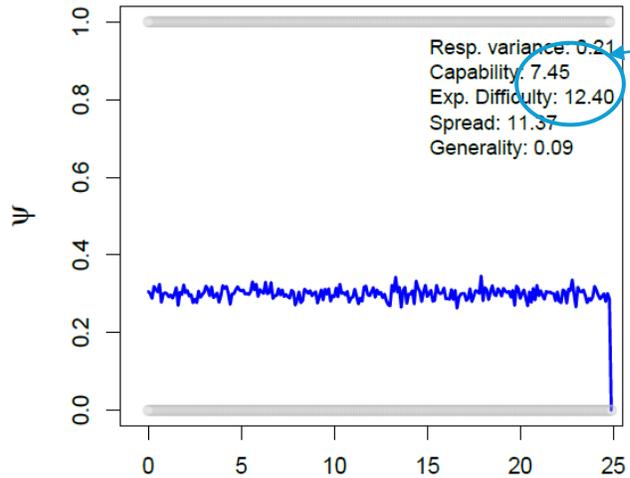
Expected difficulty represents the resources used.



J. Hernández-Orallo, B. S. Loe, L. Cheke, F. Martínez-Plumed, and S. Ó hÉigeartaigh. General intelligence disentangled via a generality metric for natural and artificial intelligence. Scientific reports, 11(1):22822, 2021.

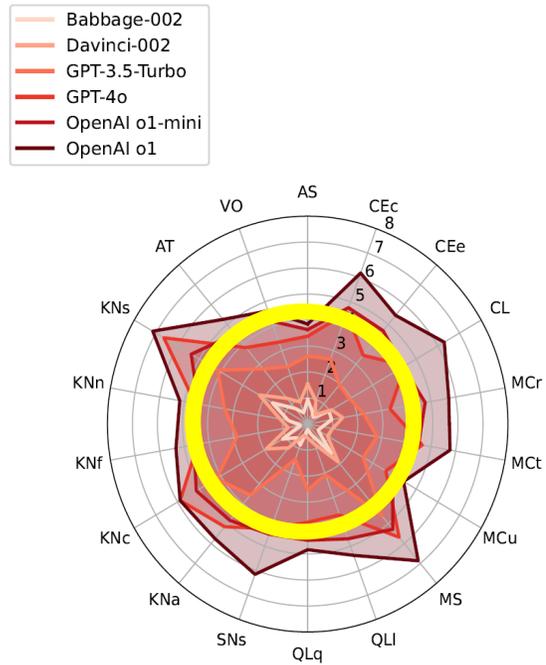
GENERALITY AS COMPACTNESS

Entanglement theorem: more capability (more tasks solved) with fewer resources

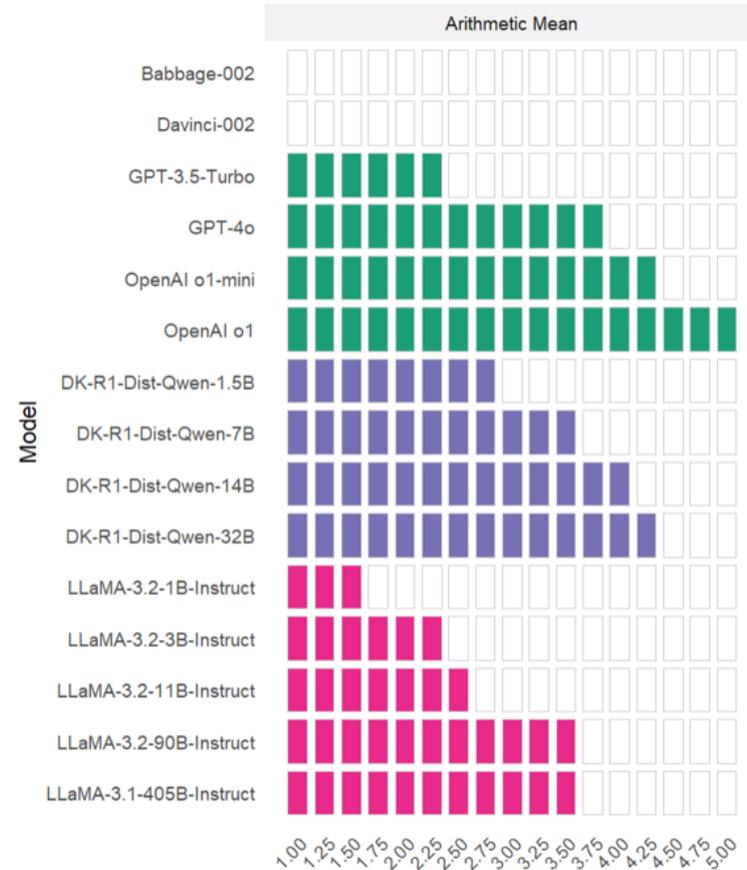


Hernandez-Orallo, J.; Loe, B.S.; Cheke, L.; Martínez-Plumed, F., O h'Eigeartaigh, S. "General intelligence disentangled via a generality metric for natural and artificial intelligence", Nature Sci Rep 2021

GPAI THRESHOLD FROM ABILITY PROFILES



GPAI Classification by Family



Safety: Propensities and Risk Models

Pointers:

- Anwar, Usman, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana et al. "Foundational challenges in assuring alignment and safety of large language models." arXiv preprint arXiv:2404.09932 (2024).
 - Sections on evaluation of capabilities and safety: Section 2.2, 2.3., 2.4, 2.6, 2.7
- Grey, M., & Segerie, C. R. (2025). Safety by Measurement: A Systematic Literature Review of AI Safety Evaluation Methods. arXiv preprint arXiv:2505.05541.

AI SAFETY EVALUATIONS

- Dominated by the “evals” paradigm

- Safety benchmarks:

- Toxicity,
 - Bias/discrimination
 - ...
 - Cyber
 - CBRN
 - ...

- Red teaming and “control”:

- Adversarial: will the model do X?
 - If I incite it to do it (jailbreaks, ...)
 - If it wants to do it (control tests, ...)
 - Uplifting: can humans do X with the model?
 - ...



Focus on average-case evaluation

Focus on worst-case evaluation

If evaluations are called “evals” then assessments should be called...

Ancient Sumerian Proverb, 3000 BCE

PROPENSITIES AS BEHAVIOUR?

- Many conflate the “response” or behaviour with the properties or latent variables of AI system and context.

- “propensities” according to

Safety by Measurement

A Systematic Literature Review of AI Safety Evaluation Methods

Markov Grey*

Charbel-Raphael Segerie†

June 13, 2025

- **Toxicity**: The propensity to generate offensive, harmful, or otherwise inappropriate content, such as hate speech, offensive/abusive language, pornographic content, etc.
- **Bias/Discrimination** : A model’s propensity to manifest or perpetuate biases, leading to unfair, prejudiced, or discriminatory outputs against certain groups or individuals.
- **Honesty** : A model’s propensity to answer by expressing its true beliefs and actual level of certainty.
- **Truthfulness** : A model’s propensity to produce truthful outputs. This propensity requires an AI system to be both honest and to know the truth (or other weirder settings such that the AI system outputs the truth while believing it is not the truth).
- **Sycophancy** : A model’s propensity to tell users what it thinks they want to hear or would approve of, rather than what it internally believes is the truth.
- **Deception** : A model’s propensity to intentionally generate misleading, false, or deceptive output.
- **Corrigibility** : A model’s propensity to accept feedback and correct its behavior or outputs in response to human intervention or new information.
- **Power Seeking** : A model’s propensity to seek to have a high level of control over its environment (potentially to maximize its own objectives).

Like conflating performance with capability again

RISK MODELS, THREAT MODELS AND HARM MODELS

- What do the previous approaches say about the risk? Or probability of harm?
 - We need a model, a predictive model
 - Instead of predicting probability of success (correctness), we predict *probability of a safe outcome*:

$$p(R_{j,i} = 1 \mid \mathbf{s}_j, \mathbf{t}_j, \mathbf{u}_j, \dots, \mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i, \dots) = ?$$

propensities capabilities resources affordances demands context (humans, time, ...)

Probability model j in reason=high with an Internet browser doesn't access i by uplifting three non-cyber-experts?

PROBLEMS

- Parametric vs. non-parametric safety assessors
 - Parametric: how to map all these variables?
 - Nonparametric: propensities + other variables still explanatory and predictive (ADeLe approach).
 - Nonparametric: build black-box assessors from many AI systems, situations, affordances, etc.
- Rare events and unknown unknowns!!!
 - Use less aligned/safe versions of models
 - Give more resources
 - Relax demands or increase incitation
 - ...

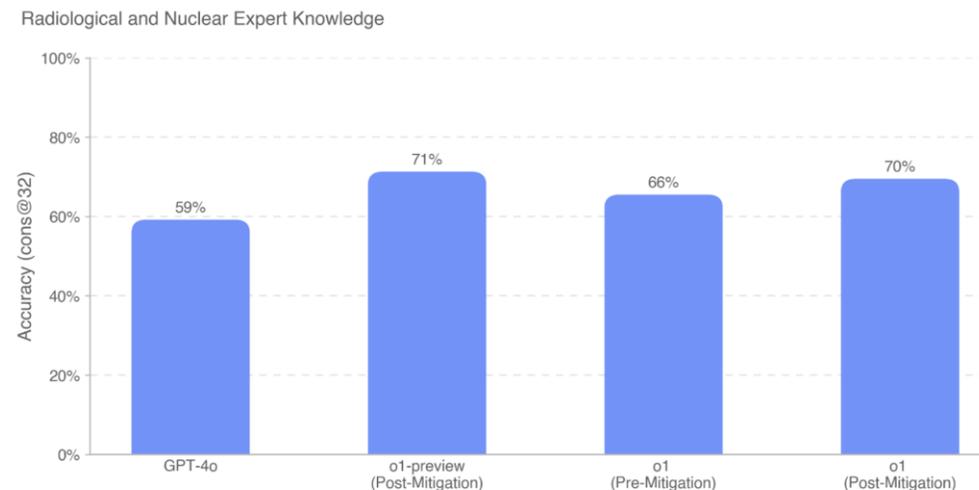
PRE-MITIGATION AND POST-MITIGATION EVALUATIONS

- Pre-mitigation: less-safe models (pre-alignment) vs most-incident cases

$$p(R_{j,i} = 1 \mid \underbrace{\mathbf{s}_j, \mathbf{t}_j, \mathbf{u}_j, \dots}_{\text{pre-alignment}}, \dots, \underbrace{\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i, \dots}_{\text{most-incident cases}}) = ?$$

Probability pre-aligned model j in reason=very-high with Internet browser doesn't access i by with cyber-experts?

- Example: OpenAI's "preparedness framework" with pre-mitigation evaluations vs post-mitigation evaluations:



OpenAI o1
System Card

(higher
means safer
here)

PART VI : CONCLUSIONS

AI EVALUATION: LESSONS LEARNT

- There are **many paradigms** and methods
- AI Evaluation is a **prediction problem**
- Performance is not **Capability**
- **Difficulty** is key for Capability Scales and **Generality**, and **GPAI characterisation**
- **Propensities** are dual with **incentives/incitation** for safety.

AI EVALUATION WITH ADELE: WORK IN PROGRESS!

- General, **absolute ratio scales** (stable to SOTA/frontiers in AI, no saturation!)
- AI benchmarks and systems become **commensurate!** (apples with apples)
- Fully **automated** procedure (profiles and predictors take minutes with a laptop!)
- **Explanatory** power (demand profiles, ability profiles)
- **Predictive** power at the instance level (especially out-of-distribution!)

THANK YOU!

JOSE H. ORALLO

<http://jorallo.github.io>

josephorallo@gmail.com

Other Talks (<http://josephorallo.webs.upv.es/>)

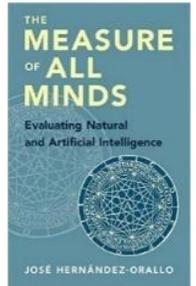
- “Diversity Unites Intelligence: Measuring Generality”, “Measuring A(G)I Right: Some Theoretical and Practical Considerations”, “Natural and Artificial Intelligence: Measures, Maps and Taxonomies”, ...

Tutorials

- Measurement Layouts (@AAAI2024): <https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial>
- IRT (@EAACL2024): <https://aclanthology.org/2024.eacl-tutorials.2/>

Book (<http://allminds.org>):

- “The Measure of All Minds: Evaluating Natural and Artificial Intelligence”, Cambridge U.P. <http://allminds.org>



OECD's AI and the Future of Skills Project:

- <https://www.oecd.org/education/cei/Future-of-Skills-Overview.pdf>, <https://doi.org/10.1787/5ee71f34-en>.



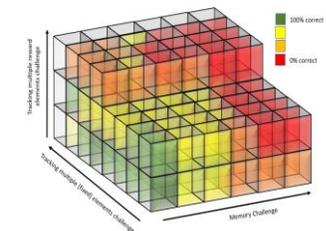
PREDICTABLE AI:

- <https://www.predictable-ai.org/>.



Animal-AI

- Part of the Kinds of Intelligence Programme at the CFI in Cambridge
 - <http://lcfi.ac.uk/projects/kinds-of-intelligence>
 - <http://animalai.org/>



AI EVALUATION NEWSLETTER

- <https://aievaluation.substack.com/>

