

Eleven things we should NOT measure in AI

Jose H. Orallo^{1,2}

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

¹ Leverhulme Centre for the Future of Intelligence, UK

² University of Cambridge, UK

AI Evaluation Newsletter



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

CFI LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE



UNIVERSITY OF
CAMBRIDGE

What we are measuring – Stanford, 14 Nov 2025



AI
EVALUATION
PROGRAMME

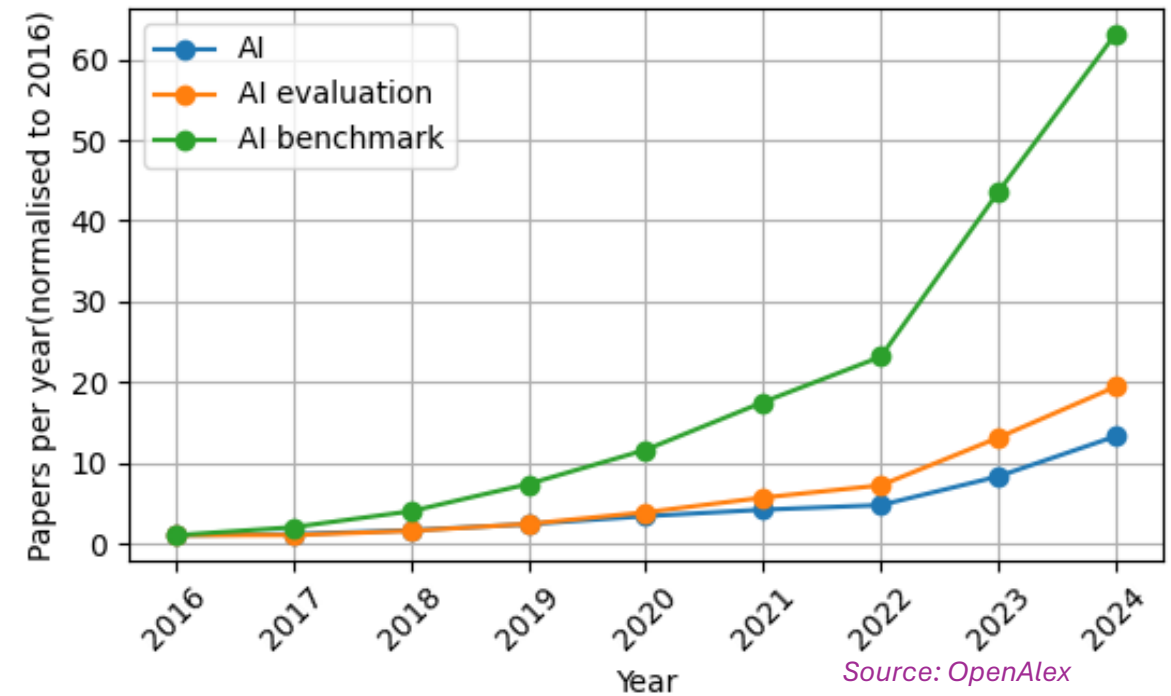
<https://ai-evaluation.org>

AI EVALUATION HAS BECOME A “THING”, NOT A SCIENCE

- **So much noise and resources wasted**
 - So many newcomers...
 - So much reinventing the wheel...
 - So many mistakes ignoring the basics...
- **Way forwards drown in the noise**
 - Constructs, scales, standardisations, predictability, profiles, ...



It's becoming increasingly more difficult to articulate thoughtful discourses on AI evaluation



1. DON'T MEASURE AGGREGATE BENCHMARK RESULTS

- **Report at the instance level**
 - The devil is in the detail
 - Other researchers can do more than you!
 - Code not enough: rerunning experiments costly!
- **For general-purpose AI (e.g., LLMs)**
 - The tasks are changing
 - Percentages will not apply for other tasks
 - Not even same tasks – out of distribution

136 14 APRIL 2023 • VOL 380 ISSUE 6641

science.org **SCIENCE**

ARTIFICIAL INTELLIGENCE

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-

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

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

An aggregate mean is the simplest inferential method

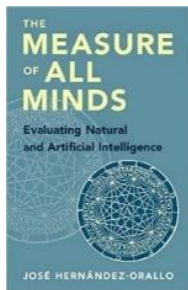
2. DON'T SAY CAPABILITY WHEN IT IS PERFORMANCE

- **Performance is:**

- a measure of a pair $\langle \text{system}, \text{item} \rangle$
- usually as a metric with no unit

- **Capability is:**

- a property of a system
- usually as a magnitude with a unit

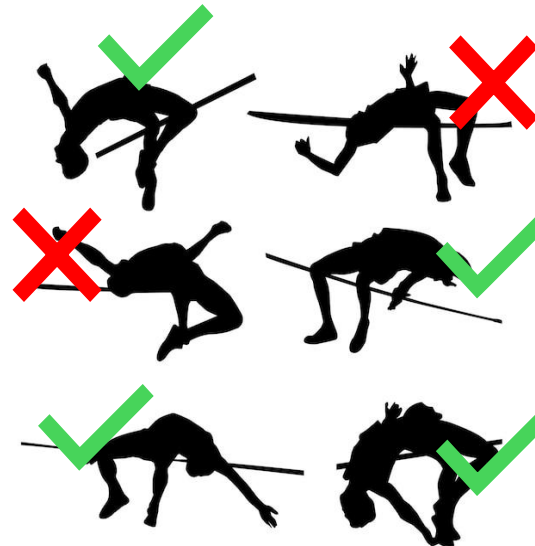


Task-oriented vs
Ability-oriented
AI evaluation

J. H. Orallo "Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement". Artificial Intelligence Review, 2016.

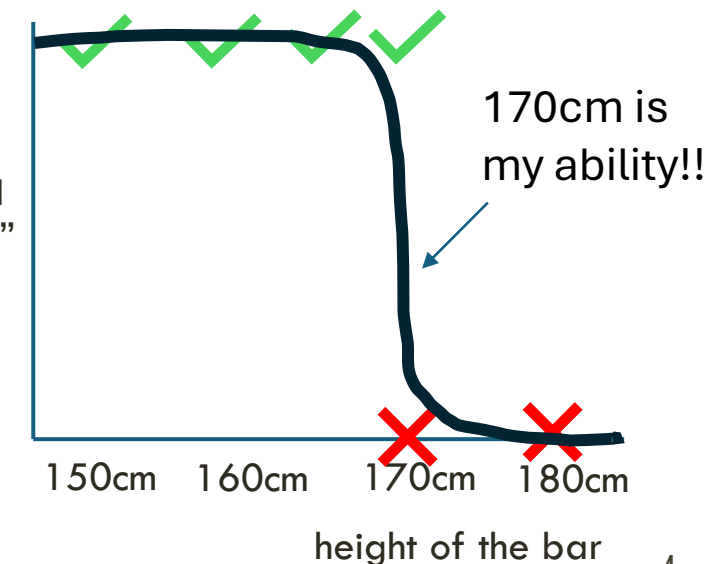
Is this my
capability?

66.7%



Successful
"Response"

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



3. DON'T MEASURE WITHOUT PREDICTION IN MIND

- **Don't give excuses:**
 - “AI is unpredictable”, “Jagged AI”, ..
 - “You can only compare performance”
- **AI validity is predictable!**
 - With the right features!
 - Intrinsic difficult proxies correlate with performance
 - If we could do this more generally... ADeLe

General Scales Unlock AI Evaluation with Explanatory and Predictive Power

Lexin Zhou^{1,2,3} Lorenzo Pacchiardi¹ Fernando Martínez-Plumed³ Katherine M. Collins⁴
Yael Moros-Daval³ Seraphina Zhang^{1,5} Qinlin Zhao² Yitian Huang² Luning Sun⁶
Jonathan E. Prunty¹ Zongqian Li⁷ Pablo Sánchez-García⁸ Kexin Jiang Chen³
Pablo A. M. Casares³ Jiyun Zu⁹ John Burden¹ Behzad Mehrbakhsh³ David Stillwell⁶
Manuel Cebrian¹⁰ Jindong Wang¹¹ Peter Henderson¹² Sherry Tongshuang Wu¹³
Patrick C. Kyllonen⁹ Lucy Cheke^{1,5} Xing Xie² José Hernández-Orallo^{1,3}

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

“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.

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

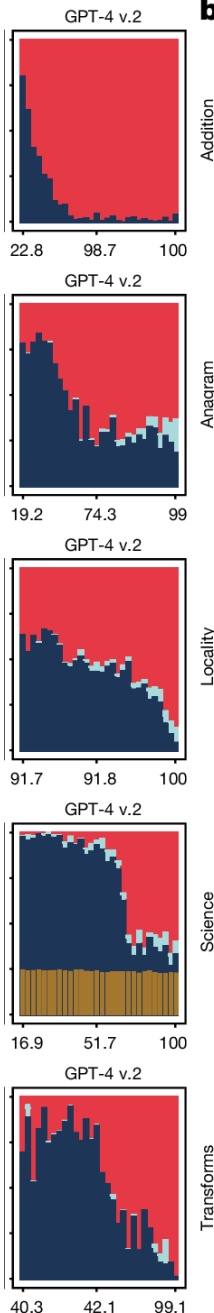
Larger and more instructable language models become less reliable but more predictable!

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

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

Subject LLM	LLM Accuracy↑	Demands (RF)	
		AUROC↑	ECE↓
Babbage-002	0.102	0.751	0.007
Davinci-002	0.157	0.741	0.007
GPT-3.5-Turbo	0.414	0.795	0.020
GPT-4o	0.713	0.852	0.023
OpenAI o1-mini	0.770	0.837	0.021
OpenAI o1	0.843	0.811	0.033

Instance-level success prediction for new benchmarks (fully OOD)



4. DON'T MEASURE AI OR BENCHMARK POPULATIONS

- **Demands and constructs are hard to find**
 - Factor analysis, IRT or Elo do the job for you!
- **LLM “populations” change very quickly**
 - The factors that were discriminating no longer are
 - LLMs are related in hierarchies (architectures, families)
- **Benchmark “populations” change very quickly**
 - The factors that were discriminating no longer are
 - Benchmarks are adversarial to LLM failures and saturation

Build scales that are criterion-referenced
(rather than norm-referenced) to allow
commensurability across populations

IRT

Making Sense of Item Response Theory in Machine Learning

Fernando Martínez-Plumed¹ and Ricardo B. C. Prudêncio²
and Adolfo Martínez-Usó³ and José Hernández-Orallo⁴

ECAI 2016, best paper award



Item response theory in AI: Analysing machine learning
classifiers at the instance level ²

Fernando Martínez-Plumed^{a,*}, Ricardo B.C. Prudêncio^b, Adolfo Martínez-Usó^c,
José Hernández-Orallo^d

^a Dept. of Computer Systems and Computation, Universitat Politècnica de València, 46022 Valencia, Spain
^b Centro de Informática, Universidade Federal de Pernambuco, Recife (PE), Brazil
^c Universitat Jaume I de Castelló, Spain

Factor analysis

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

Three factors

Ilić, D., & Gignac, G. E. (2024). Evidence of
interrelated cognitive-like capabilities in large
language models: Indications of artificial general
intelligence or achievement?. *Intelligence*,

One factor

5. DON'T MEASURE WITH HUMAN TESTS DIRECTLY

- **You will be misled several times...**
 - Human tests lose validity outside their group (e.g., human adults)
 - They may rely on proxies that only work for human (short memory)
 - May not test things that all humans have
 - May test things that are irrelevant for AI
- **Still, many well-designed items can be reused**
 - Require re-annotation!
 - Require new capability catalogues.

Don't use human taxonomies (e.g., CHC) to characterise AI but think of capabilities that are conceptually different (construct validity by design)



Intelligence

Accepted 22 December 2011

IQ tests are not for machines

David L. Dowe^a, José Hernández-Orallo^{b,*}

Stop Evaluating AI with Human Tests, Develop Principled, AI-specific Tests instead

Tom Sühr
Max Planck Institute for Intelligent Systems
Tübingen AI Center
tom.suehr@tuebingen.mpg.de

Florian E. Dörner
Max Planck Institute for Intelligent Systems
ETH Zurich
florian.dorner@tuebingen.mpg.de

Olawale Salandeen
Massachusetts Institute of Technology
olawale@mit.edu

Augustin Kelava
Methods Center
University of Tübingen
augustin.kelava@uni-tuebingen.de

Samira Samadi
Max Planck Institute for Intelligent Systems
Tübingen AI Center
samira.samadi@tuebingen.mpg.de

Theory of Mind May Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*¹

Theory of Mind Might Have Spontaneously Emerged in Large Language Models

Authors: Michal Kosinski*¹

Evaluating Large Language Models in Theory of Mind Tasks

Michal Kosinski

Stanford University

March 2024

6. DON'T CONFOUND "VOLUME" WITH EVERYTHING ELSE

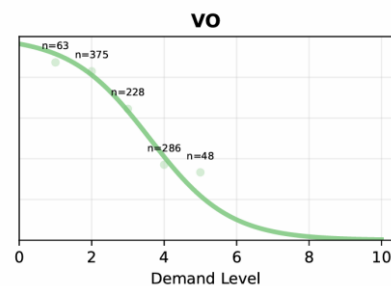
- **Volume makes things harder**

- In perception
- In reasoning
- ...
- In software engineering

- **Doesn't mean locally more difficult:**

- **Solve** $x = 3 * 2$, $y = -2x + 1$, $z = y + 10$, ...

Measure volume,
separately from
everything else!



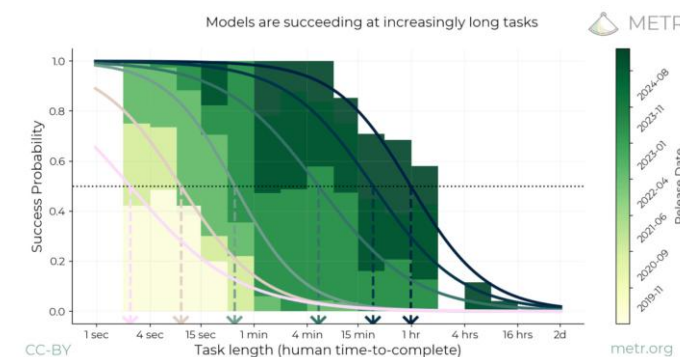
DeepSeek'sR1-Dist-Qwen-7B
on ADeLe

Measuring AI Ability to Complete Long Tasks

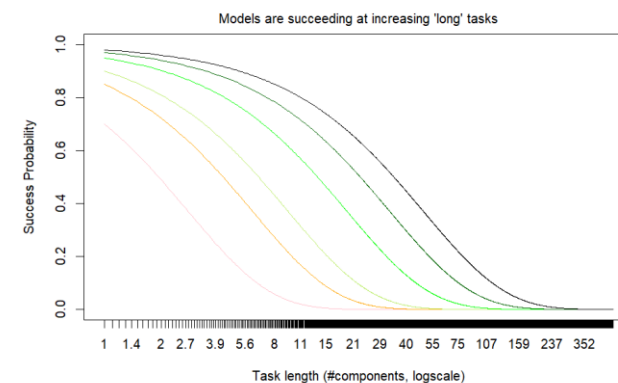
Thomas Kwa*, Ben West†*, Joel Becker, Amy Deng, Katharyn Garcia,
Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx

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

Elizabeth Barnes, Lawrence Chan



Simple geometric series on
a log scale where models
have isolated probabilities
 p of 0.7, 0.85, 0.9, 0.95,
and 0.97 and 0.98:



<https://aievaluation.substack.com/p/2025-march-ai-evaluation-digest>

7. DON'T FORGET AI EVALUATION EVALUATION

- **Evaluation is a prediction problem**
 - Inferential models of validity: assessors
- **Assessors can solve big challenges:**
 - Familiarity \sim contamination
 - If the assessor underestimates OOD for all task demands
 - Contamination likely
 - Motivation \sim sandbagging
 - If the assessor overestimates for some task demands
 - Sandbagging likely

Science of Evaluation?
Test Predictive and Explanatory Power

The Evaluation of Artificial Intelligence as a Prediction Problem

February 2025

Author: Wout Schellaert

Advisors: José Hernández-Orallo
Fernando Martínez-Plumed

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

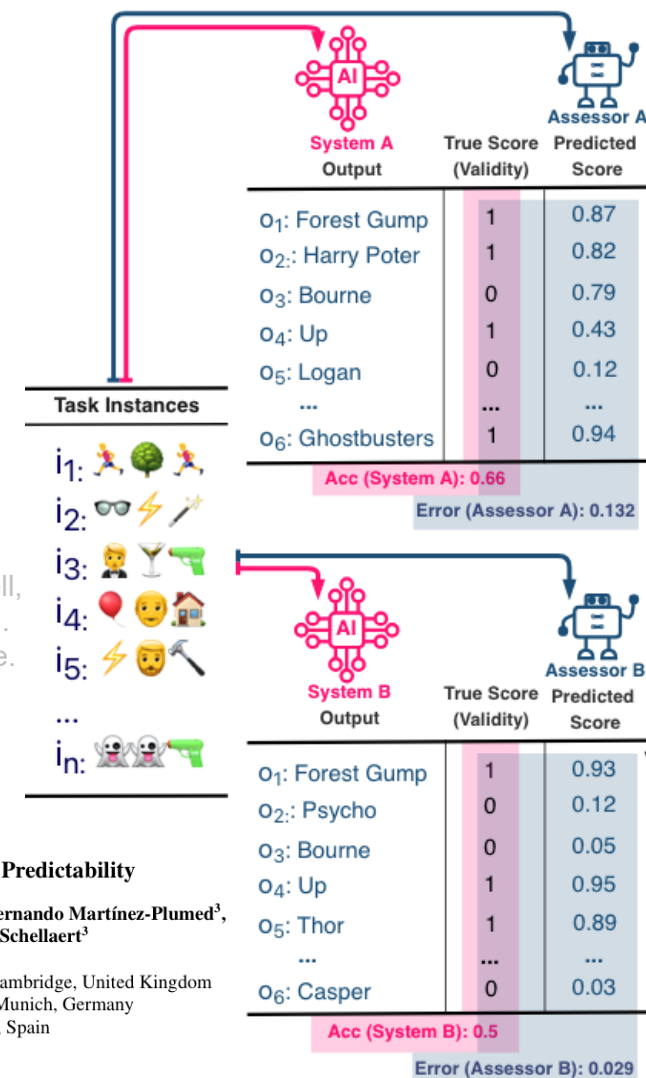
PredictaBoard: Benchmarking LLM Score Predictability

Lorenzo Pacchiardi¹, Konstantinos Voudouris^{1,2}, Ben Slater¹, Fernando Martínez-Plumed³,
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



8. DON'T MEASURE "INTELLIGENCE" (OR AGI LEVELS)

- **Measure profiles**
 - Benchmark profiles
 - AI profiles
 - Human profiles
 - not "human-level" baselines
- **Then policy-makers will choose the shapes, red lines and thresholds**



JRC Publications Repository

General Purpose AI

Key Considerations for the EU AI Act

<https://publications.jrc.ec.europa.eu/repository/handle/JRC143255>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC143256>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC143257>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC143258>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC143259>
<https://publications.jrc.ec.europa.eu/repository/handle/JRC143260>

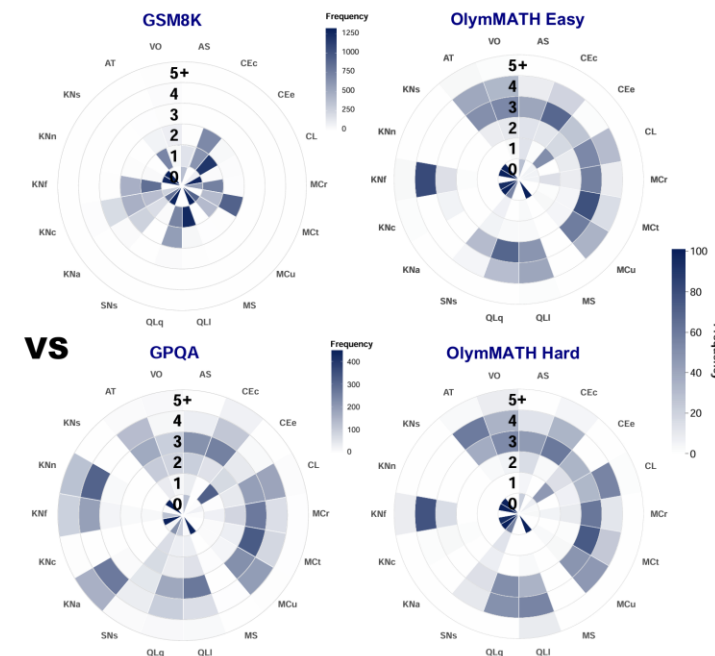
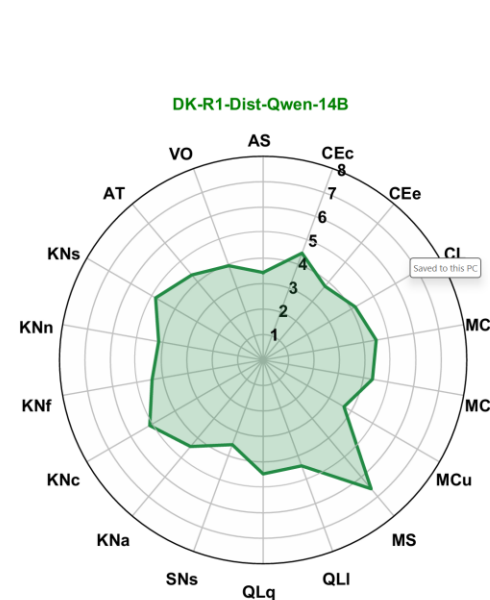
A Framework for the Categorisation of General-Purpose AI Models under the EU AI Act

Lorenzo Pacchiardi¹, John Burden¹, Fernando Martínez-Plumed²,
José Hernández-Orallo^{1,2}, Emilia Gómez³, David Fernández-Llorca³

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

²Valencian Research Institute for Artificial Intelligence (VRAIN),
Universitat Politècnica de València, Spain

³European Commission, Joint Research Centre (JRC), Seville, Spain



General Scales Unlock AI Evaluation with Explanatory and Predictive Power

Lexin Zhou^{1,2,3} Lorenzo Pacchiardi¹ Fernando Martínez-Plumed³ Katherine M. Collins⁴
Yael Moros-Daval³ Seraphina Zhang^{1,5} Qinlin Zhao² Yitian Huang² Luning Sun⁶
Jonathan E. Prunty¹ Zongqian Li⁷ Pablo Sánchez-García⁸ Kexin Jiang Chen³
Pablo A. M. Casares³ Jiyun Zu⁹ John Burden¹ Behzad Mehrbakhsh³ David Stillwell⁶
Manuel Cebrian¹⁰ Jindong Wang¹¹ Peter Henderson¹² Sherry Tongshuang Wu¹³
Patrick C. Kyllonen⁹ Lucy Cheke^{1,5} Xing Xie² José Hernández-Orallo^{1,3}

9. DON'T MEASURE “DANGEROUS CAPABILITIES”

Google DeepMind

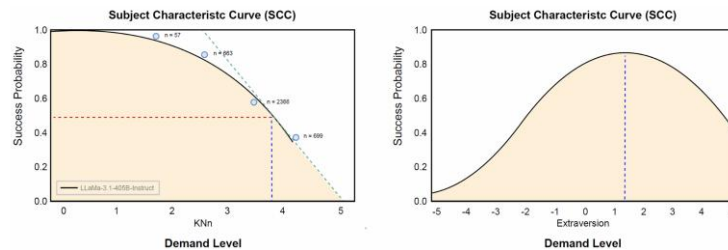
2024-4-8

- **Capability means systematic performance, not possibility**

- **EU AI Code of Practice:**

- Capabilities, propensities, affordances and context

- **Propensities are tricky but cool!**



- **Risk Model:**

$$p(R_{j,i} = 1 | s_j, t_j, u_j, \dots, a_i, b_i, c_i, \dots) = ?$$

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? 11

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, Selim 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.

10. DON'T MEASURE AI SYSTEMS ONLY

- **AI Evaluation is NOT ONLY about evaluating AI systems**
 - Evaluating humans, and human-AI ecosystems
 - We need good human models to test AI
 - Evaluating societal impact...
 - We need good societal models



Addictive Behaviors
Volume 166, July 2025, 108325

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



nature human behaviour

Perspective

<https://doi.org/10.1038/s41562-024-01991-9>

**Building machines that learn and think
with people**

Received: 11 April 2024

Accepted: 23 August 2024

Published online: 22 October 2024

Katherine M. Collins^{1,2}, Ilya Sucholutsky^{2,3}, Umang Bhatt^{3,4,5},
Kartik Chandra^{6,7}, Lionel Wong^{1,5}, Mina Lee^{8,9}, Cedegeo E. Zhang¹⁰,
Tan Zhi-Xuan¹, Mark Ho¹¹, Vikash Mansinghka^{1,10}, Adrian Weller^{1,10},
Joshua B. Tenenbaum^{1,10} & Thomas L. Griffiths^{1,10}

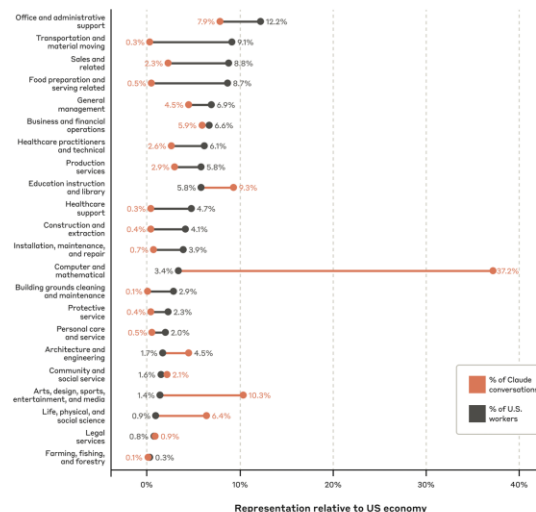
Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations

Kunal Handa^{*}, Alex Tamkin^{*}, Miles McCain, Saffron Huang, Esin Durmus

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

Dario Amodei, Jared Kaplan, Jack Clark, Deep Ganguli

Anthropic



Cora von Hammerstein^{c d}, Joël Billieux^{e f}

Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge

Hanna Wallach¹, Meera Desai², A. Feder Cooper¹, Angelina Wang³, Chad Atalla¹, Solon Barocas¹,
Su Lin Blodgett¹, Alexandra Chouldechova¹, Emily Corvi¹, P. Alex Dow¹, Jean Garcia-Gathright¹,
Alexandra Olteanu¹, Nicholas Pangakis¹, Stefanie Reed¹, Emily Sheng¹, Dan Vann¹,
Jennifer Wortman Vaughan¹, Matthew Vogel¹, Hannah Washington¹, Abigail Z. Jacobs²

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].

11. DON'T CALL AI EVALUATIONS “EVALS”

If you call evaluations “evals”
what would you call assessments?

Thank You!



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

AI Evaluation Newsletter



<https://ai-evaluation.org>

EXTRA: FIXING AI EVALUATION: 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 (from results data, profiles and predictors take minutes!)
- **Explanatory** power (demand profiles, ability profiles)
- **Predictive** power at the instance level (especially out-of-distribution!)

General Scales Unlock AI Evaluation with Explanatory and Predictive Power

Lexin Zhou^{1,2,3} Lorenzo Pacchiardi¹ Fernando Martínez-Plumed³ Katherine M. Collins⁴
Yael Moros-Daval³ Seraphina Zhang^{1,5} Qinlin Zhao² Yitian Huang² Luning Sun⁶
Jonathan E. Prunty¹ Zongqian Li⁷ Pablo Sánchez-García⁸ Kexin Jiang Chen³
Pablo A. M. Casares³ Jiyun Zu⁹ John Burden¹ Behzad Mehrbakhsh³ David Stillwell⁶
Manuel Cebrian¹⁰ Jindong Wang¹¹ Peter Henderson¹² Sherry Tongshuang Wu¹³
Patrick C. Kyllonen⁹ Lucy Cheke^{1,5} Xing Xie² José Hernández-Orallo^{1,3}

ADeLe v1.0: A battery for Explanatory and Predictive AI Evaluation

[Original Paper](#) [Dataset](#)

This is a collaborative community, initiated by researchers at the [Leverhulme Centre for the Future of Intelligence](#), University of Cambridge, for the use and extension of ADeLe v1.0, a battery for explanatory and predictive LLM evaluation.

The ADeLe (Annotated-Demand-Levels) battery includes 63 tasks from 20 benchmarks and was introduced in (ARXIV LINK!). Those tasks were annotated using 18 rubrics for Demand-Level-Anotation (DeLeAn v1.0) of general scales.

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

Join us for follow-up projects