

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

1

John P. Lalor, Pedro Rodriguez, João Sedoc, Jose Hernandez-Orallo

<https://eacl2024irt.github.io/>

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

Part 4. Advanced Topics

José Hernández-Orallo^{1,2,3}

¹ VRAIN, Universitat Politècnica de València

² Leverhulme Centre for the Future of Intelligence, University of Cambridge

³ Centre for the Study of Existential Risk, University of Cambridge

<http://josephorallo.webs.upv.es/>



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

 VRAIN

 LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE



CENTRE FOR THE STUDY OF
EXISTENTIAL RISK

Main Limitations of (classical) IRT

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

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, July). Item Response Theory for Evaluating Regression Algorithms. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Multidimensional IRT

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:

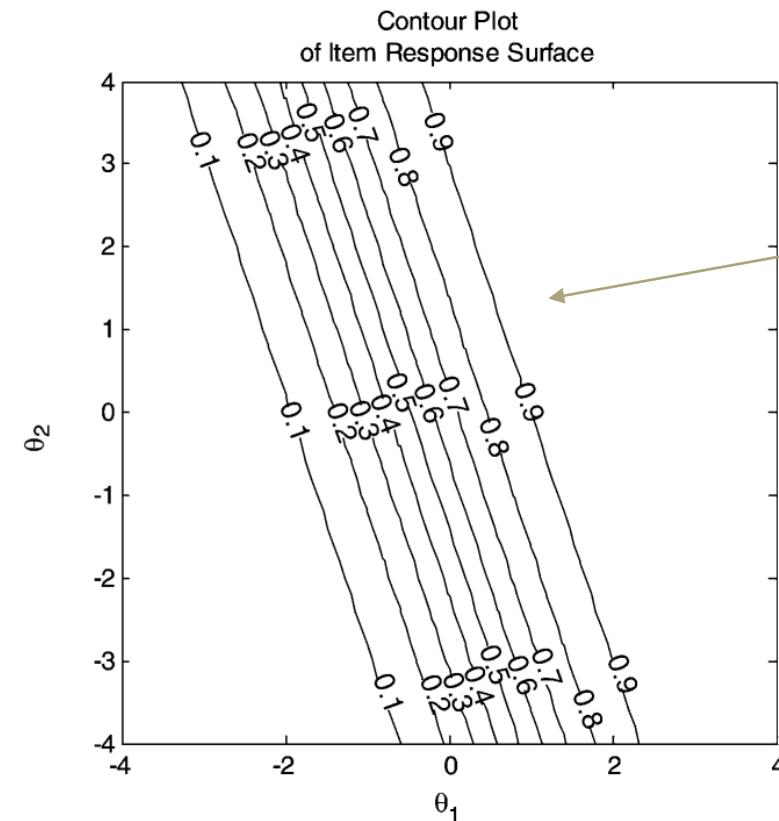
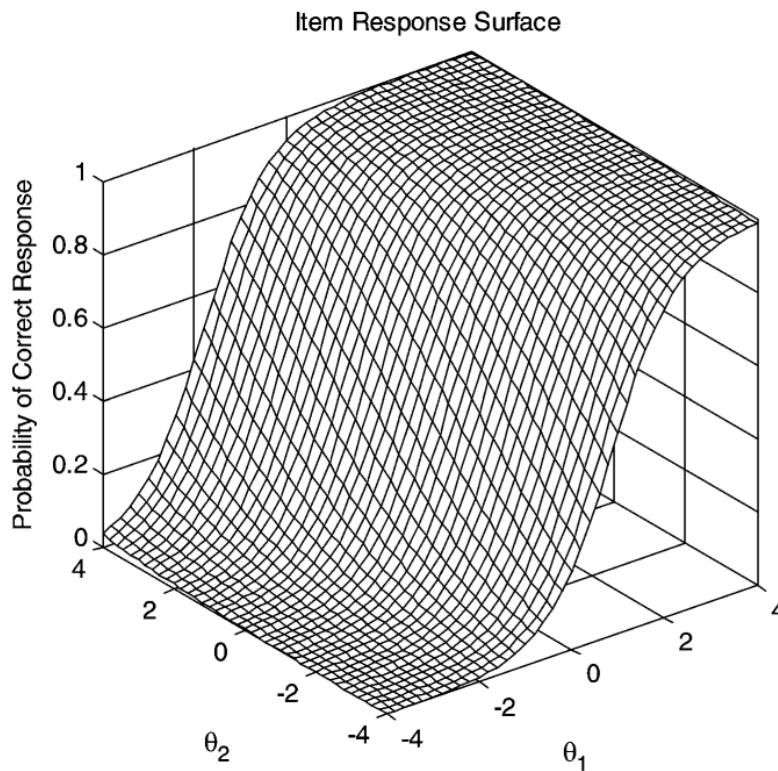
$$P(u_i = 1|\theta_j) = \frac{e^{\mathbf{a}'_i \theta_j + d_i}}{1 + e^{\mathbf{a}'_i \theta_j + d_i}}$$

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

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

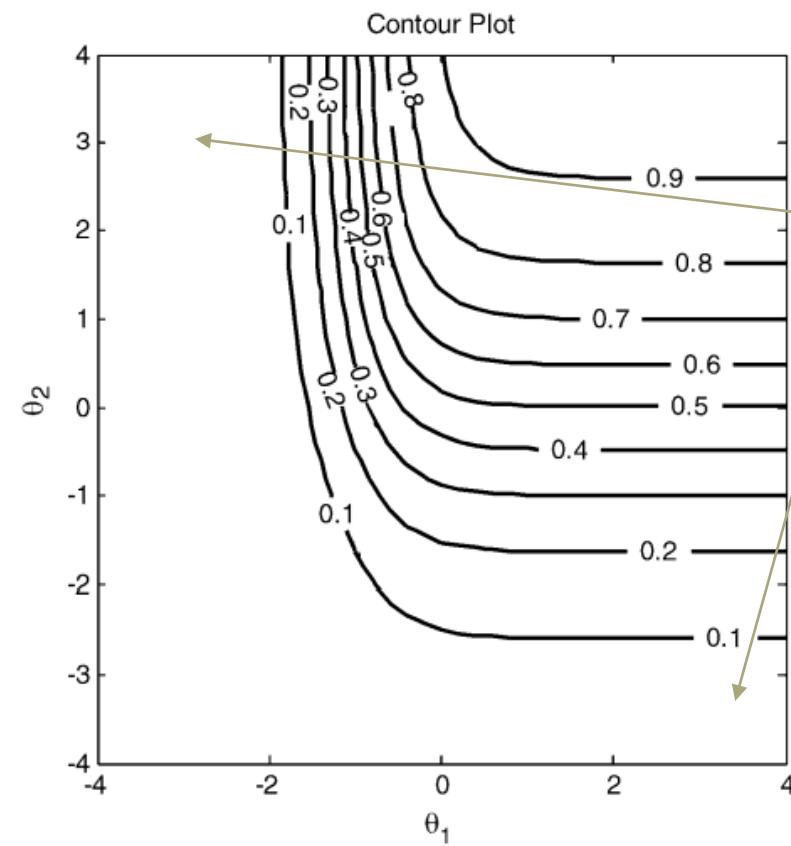
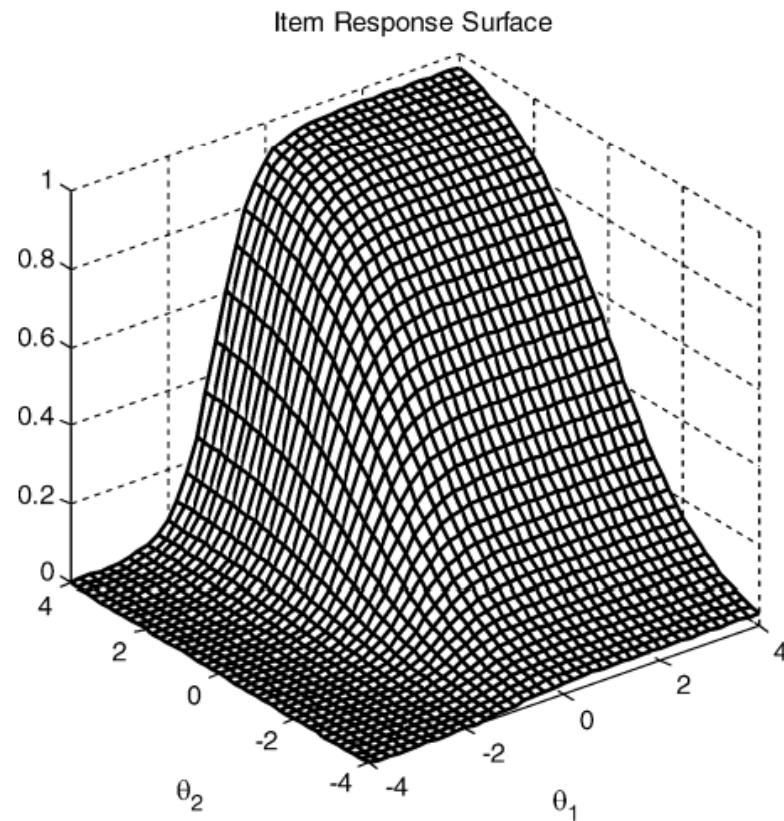


Asymmetric compensation:
Given this angle,
ability 1 can
compensate for
ability 2 but not
vice versa.<

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 $c_i = 0$.

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

When Difficulty/Demands Are Given

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.

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.

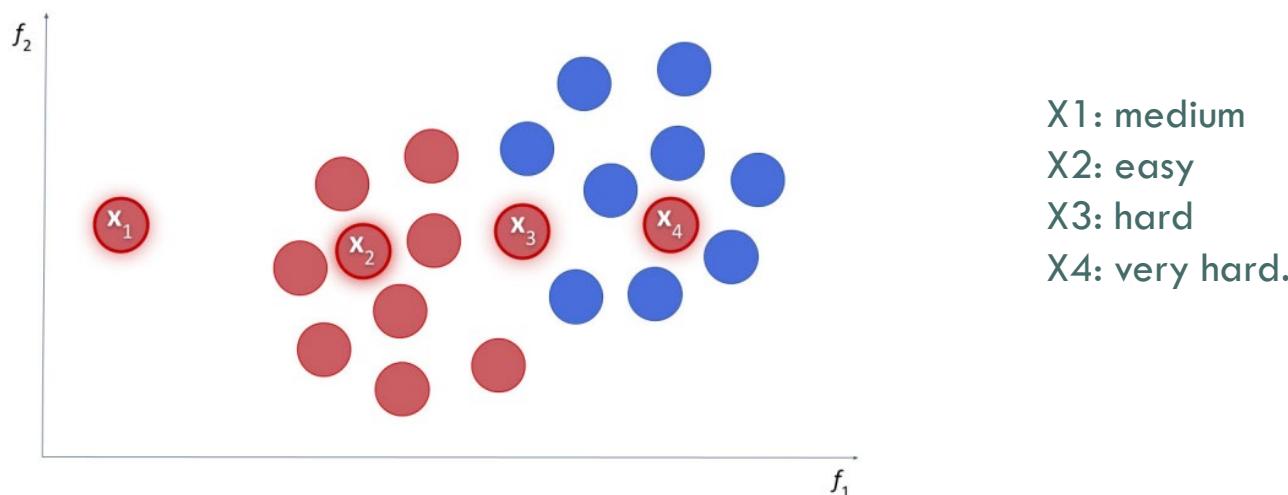
- Values can be > 1

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).

HOW TO ELICIT DIFFICULTIES? EXTRINSIC

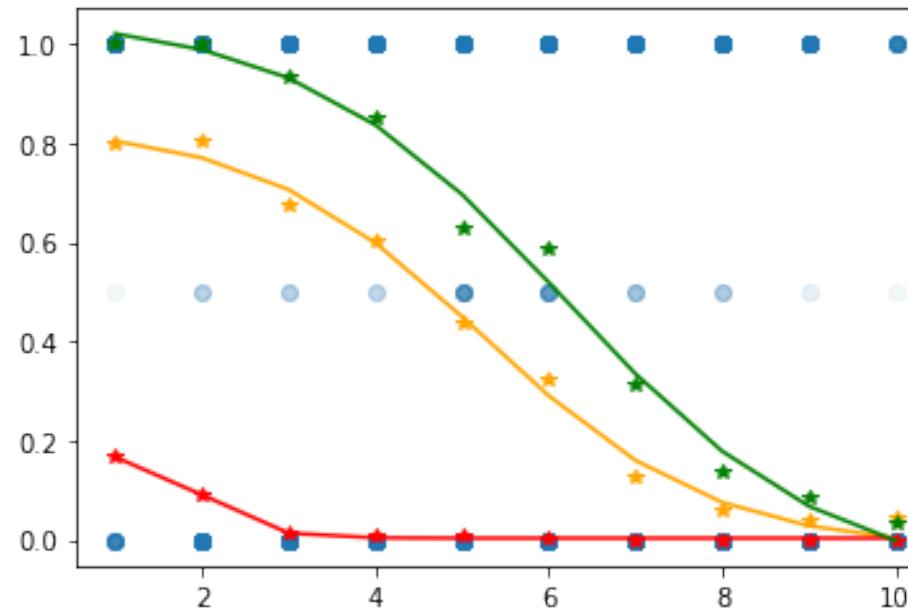
- The difficulty of an instance depends on its relation to the other instances.
 - EXTRINSIC: A paradigmatic case is the concept of “instance hardness” in classification
 - But some of them do not depend on the models, just on the distribution of data.



Lorena, A. C., Paiva, P. Y., & Prudêncio, R. B. (2023). Trusting my predictions: on the value of Instance-Level analysis. *ACM Computing Surveys*.

HOW TO ELICIT DIFFICULTIES? INTRINSIC

- In some cases, the difficulty of an instance is easy to identify and they are **intrinsic**.
 - INTRINSIC: The difficulty of an instance doesn't depend on the difficulty of other instances!!!



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

Zhou et al. "Scaled-up, Shaped-up, but Letting Down? Reliability Fluctuations of Large Language Model Families", in preparation, 2024.

AUTOMATED DEMAND ANNOTATION IN NLP

- Syntactic and semantic complexity metrics (e.g., Quanteda)
 - Lexical Diversity: TTR, C, R, CTTR, U, S, K, I, D, Vm, Maas, lgV0, lgeV0, nchar.
 - Readability: ARI, ARI.simple, ARI.NRI, Bormuth.MC, Bormuth.GP, Coleman, Coleman.C2, Coleman.Liau.ECP, Coleman.Liau.grade, Coleman.Liau.short, Dale.Chall, Dale.Chall.old, Dale.Chall.PSK, Danielson.Bryan, Danielson.Bryan.2, Dickes.Steiwer, DRP, ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK, Flesch.Kincaid, FOG, FOG.PSK, FOG.NRI, FORCAST, FORCAST.RGL, Fucks, Linsear.Write, LIW, nWS, nWS.2, nWS.3, nWS.4, RIX, Scrabble, SMOG, SMOG.C, SMOG.simple, SMOG.de, Spache, Spache.old, Strain, Traenkle.Bailer, Traenkle.Bailer.2, Wheeler.Smith, meanSentenceLength, meanWordSyllables.

LLM FOR DEMAND ANNOTATION

- Linguistic Meta-features
(annotated by GPT-4):



Meta-features	Scale and Levels	Examples
Uncertainty	0: complete certainty, ... 10: complete uncertainty	"The cat is in the house": 1 "She might not do it again": 7 "He may come this afternoon": 3 "We have no clue about where it is": 8 "It is a fact that a square has four sides": 0 "It's impossible to know who will win the lottery": 10 "I'm not sure who will win the election": 8
Negation	0: no negation 1: simple negation 2: double negation 3: negation with quantification 4: very complex negation	"I'm a rich man": 0 "She has never had a dog": 1 "It's untrue that all houses without windows do not have any light": 4 "I don't know what I don't know": 2 "The suspect is not in the house": 1 "The car has not been driven by anyone in the team": 3 "Never say never": 2
Time	0: no time expressions 1: simple temporal expressions 2: double temporal expressions 3: complex temporal expressions	"He came before noon": 1 "The house is blue": 0 "There's a meeting every two weeks": 3 "The train arrived ten minutes after the plane has left": 2
Space	0: no space relationships 1: simple spatial expressions 2: double spatial expressions 3: complex spatial expressions	"The pen was on the table": 1 "There's no room between the two cars": 2 "Tomorrow is a bank holiday": 0 "The lamp was hanging from two ropes, one attached to the ceiling and the other to the window": 5
Vocabulary	0..1: Normalised from some aggregate metric of the -log freq of words or something similar as in semantic complexity metrics.	"The ball is big": 0.1219 "Procrastination jeopardises excellence": 0.4235 "The boy must apologise": 0.198 "Ignoramus was an ultrarepidarian reposte": 0.8324
Modality	0: no modality 1: simple modality 2: double modality	"The woman walked into a bar": 0 "The boy must apologise": 1 "The boy thinks we can't do it": 3
Theory of Mind	0: no theory of mind 1: simple theory of mind 2: double theory of mind	"He came to the reception before noon": 0 "She didn't want to buy a car": 1 "The boy thinks we can't do it": 1 "The child feared his parents wanted to punish him": 2
Reasoning	0: no reasoning 1: simple reasoning 2: complex reasoning	"He tripped because of the step": 1 "He came before noon with a bag full of presents": 0 "The grass was wet but it was sunny so someone must have watered the plant": 2
Compositionality	1...number of levels	"He came before noon": 0 "He came before she arrived": 1 "The man wearing the tall hat came before she arrived": 2 "He came before noon with a bag full of presents": 0.
Anaphora	0: no anaphora 1: simple (one possible referent) 2: complex (>1 possible referents)	"Kim thinks that he is clever": 1 "While Stuart was telling Susan the news, she laughed at him": 2
Noise	0...number of typos per character wrt to the original text with no typos	"The ball is big": 0 "The bll isbige": 3/13 "The boy bust apologise": 1/20

COULD WE USE LLTM?

- Tasks (thousands of items) and models (dozens of subjects) from HELM (summer 2023)

Task	Description	Domain
Massive Multitask Language Understanding (MMLU)	Knowledge-intensive question answering across 4 domains: Computer Security, US Foreign Policy, Econometrics and College Chemistry	Knowledge-intensive QA
OpenbookQA	Commonsense-intensive open book question answering	Knowledge-intensive QA
Legal Support	Fine-grained legal reasoning through reverse entailment	Legal Realistic Reasoning
LSAT	Measure analytical reasoning on the Law School Admission Test	Logical Realistic Reasoning
Bias Benchmark for Question Answering (BBQ)	Social bias in question answering in ambiguous and unambiguous context	Bias
HellaSwag	Commonsense reasoning in question answering	Knowledge-intensive QA
TruthfulQA	Model truthfulness and commonsense knowledge in question answering	Knowledge-intensive QA

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

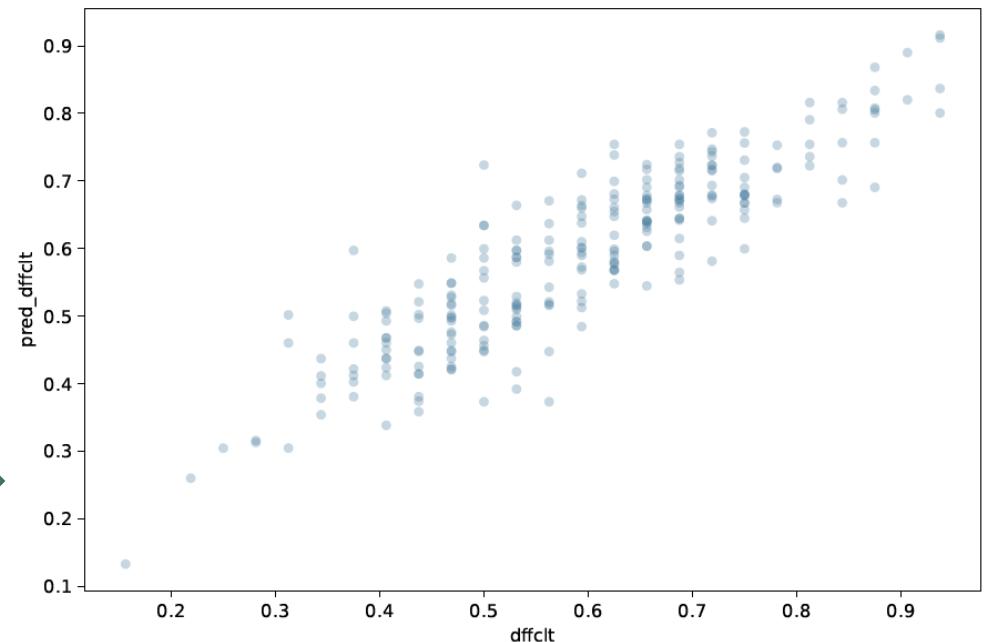
Creator	Model	Number of Parameters
AI21 Labs	J1-Jumbo v1	178B
AI21 Labs	J1-Large v1	7.5B
AI21 Labs	J1-Grande v1	17B
AI21 Labs	J1-Grande v2 beta	17B
Aleph Alpha	Luminous Base	13B
Aleph Alpha	Luminous Extended	30B
Aleph Alpha	Luminous Supreme	70B
Anthropic	Anthropic-LM v4-s3	52B
BigScience	BLOOM	176B
BigScience	BLOOMZ	176B
BigScience	T0pp	11B
BigCode	SantaCoder	1.1B
Cohere	Cohere xlarge v20220609	52.4B
Cohere	Cohere large v20220720	13.1B
Cohere	Cohere medium v20220720	6.1B
Cohere	Cohere small v20220720	410M
Cohere	Cohere xlarge v20221108	52.4B
Cohere	Cohere medium v20221108	6.1B
Cohere	Cohere command nightly	6.1B
Cohere	Cohere command nightly	52.4B
DeepMind	Gopher	280B
DeepMind	Chinchilla	70B
EleutherAI	GPT-J	6B
EleutherAI	GPT-NeoX	20B
Google	T5	11B
Google	UL2	20B
Google	Flan-T5	11B
Google	PaLM	540B
HazyResearch	H3	2.7B
Meta	OPT-IML	175B
Meta	OPT-IML	30B
Meta	OPT	175B
Meta	OPT	66B
Meta	Galactica	120B
Meta	Galactica	30B
Microsoft/NVIDIA	TNLG v2	530B
Microsoft/NVIDIA	TNLG v2	6.7B
OpenAI	davinci	175B
OpenAI	curie	6.7B
OpenAI	babbage	1.3B
OpenAI	ada	350M
OpenAI	text-davinci-003	-
OpenAI	text-davinci-002	-
OpenAI	text-davinci-001	-
OpenAI	text-curie-001	-
OpenAI	text-babbage-001	-
OpenAI	text-ada-001	-
OpenAI	code-davinci-002	-
OpenAI	code-davinci-001	-
OpenAI	code-cushman-001	12B
OpenAI	ChatGPT	-
Together	GPT-JT	6B
Together	GPT-NeoXT-Chat-Base	20B
Tsinghua	CodeGen	16B
Tsinghua	GLM	130B
Tsinghua	CodeGeeX	13B
Yandex	YaLM	100B

YES, BUT WE DIDN'T (USED XG-BOOST)

Task	Linguistic Meta-features	Traditional Metrics
Abstract Narrative Understanding	0.06	-0.01
BBQ	0.62	0.5
Epistemic Reasoning	0.9	-0.03
Formal Fallacies Syllogisms Negation	0.6	-0.15
Hellaswag	0.02	-0.03
Legal Support	0.3	0.05
LSAT	-0.07	-0.07
MMLU College Chemistry	0.77	0.74
MMLU Computer Security	0.83	0.85
MMLU Econometrics	0.68	0.7
MMLU US Foreign Policy	0.8	0.74
OpenbookQA	-0.04	0.01
TruthfulQA	0.59	0.56

Table 5.1: R^2 obtained in the test split when predicting difficulty with linguistic meta-features and lexical and readability metrics

Each dot is an instance of MMLU US FP, with average error for all models on the x axis and the predicted average error on the y axis.



General Difficulty Models

DATA FOR DIFFICULTY

- Once we have applied IRT or used any other method to estimate the difficulties of the instances, we end up with a dataset like this:

Item	Original Features	Difficulty	Discrim.
#1	What's the capital of France?	-2.5	0.6
#2	What's almost an island?	0.3	0.7
#3	What's the capital of Bhutan?	0.7	0.2
#4	What's frozen water?	-1.8	0.3
#5	Who's your mother's son's mother?	-0.5	0.2
#6	What's brown and sticky?	2.3	-0.3
...

Can we predict difficulty
(and discrimination) from the
features?

YES, WE CAN

- But we can build a difficulty model from the instance features:
- Better with 1PL models:

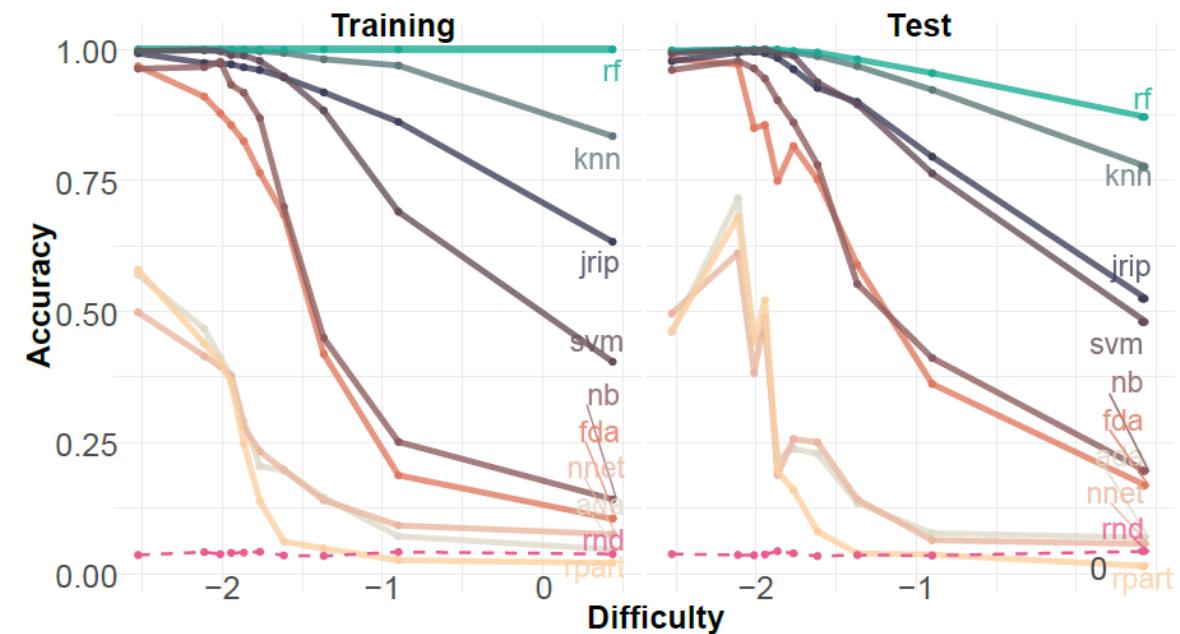
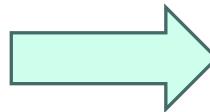


Figure 5: (Left) SCC obtained with the 70% of the letter benchmark and the observed difficulties \hat{h} . (Right) SCC obtained with the test set (30%), using estimated difficulties \hat{h} .

Predicting Performance Directly: Assessors

JH Orallo, W Schellaert, FM Plumé

Training on the Test Set: Mapping the System-Problem Space in AI
AAAI 2022

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, x_1, x_2, \dots	1
Resnet, $\theta_1, \theta_2, \dots$	Image23, x_1, x_2, \dots	0
...
Inception, $\theta_1, \theta_2, \dots$	Image3, x_1, x_2, \dots	1
Inception, $\theta_1, \theta_2, \dots$	Image78, x_1, x_2, \dots	1
...

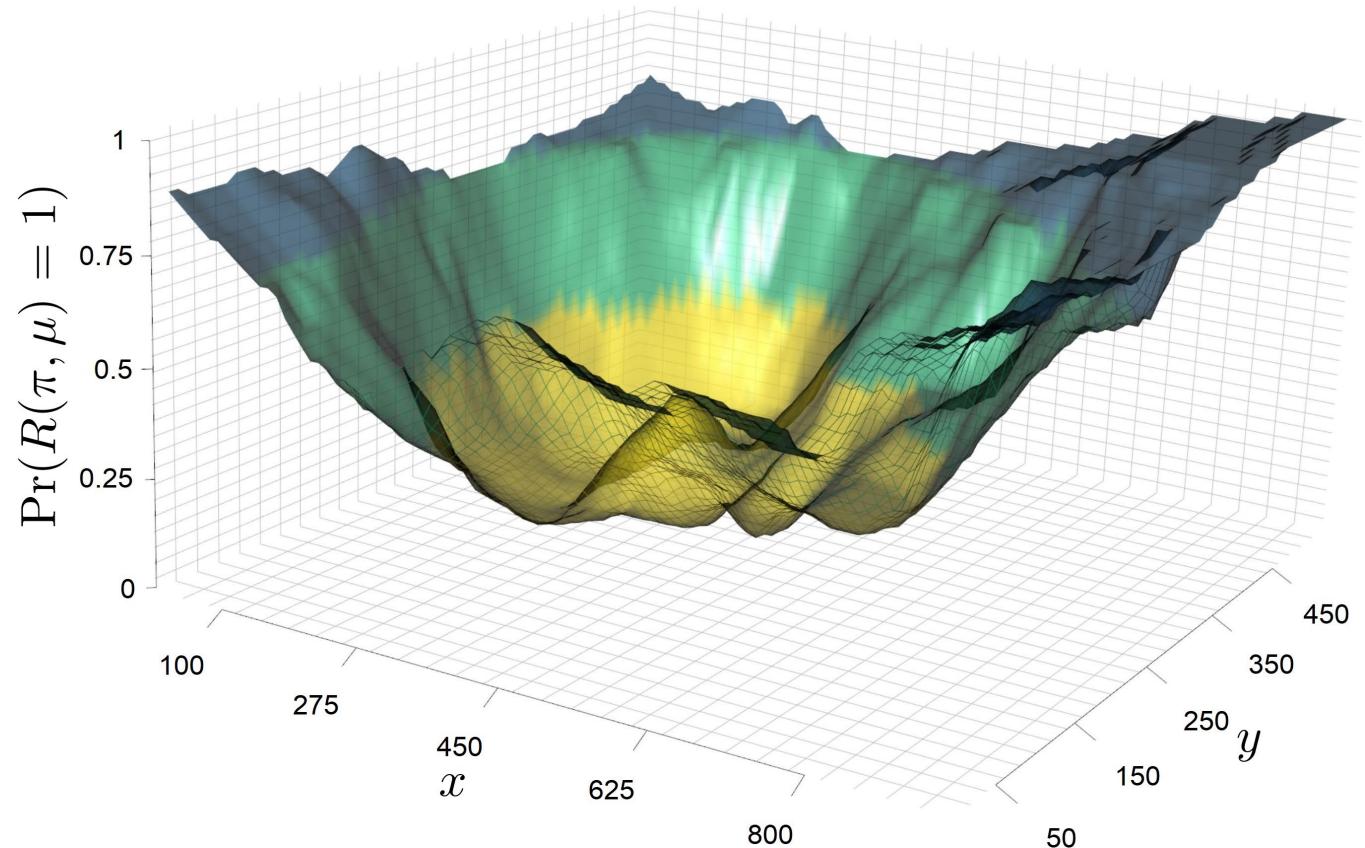
PROBLEM SPACE

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

- 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



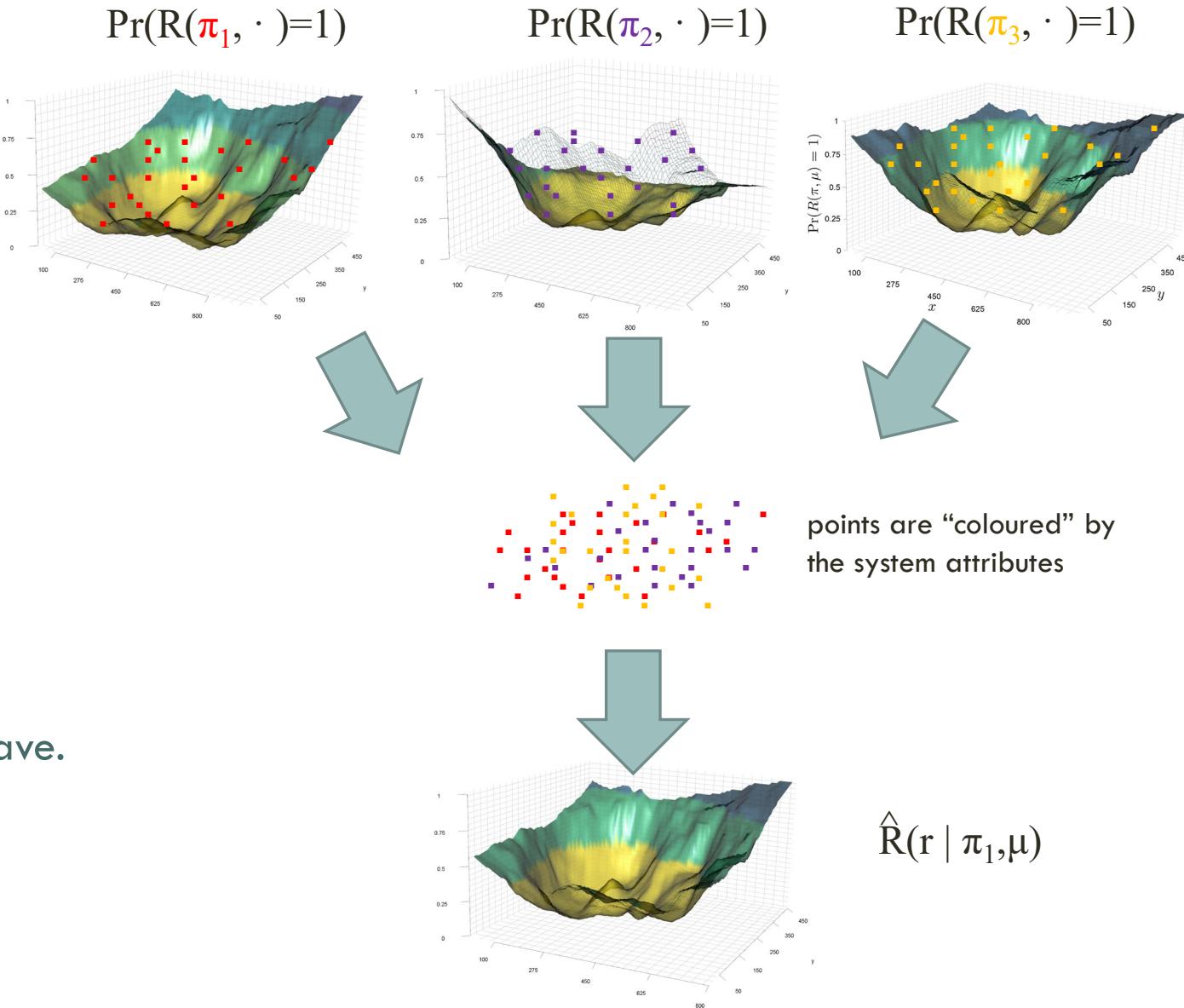
SYSTEM SPACE

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!



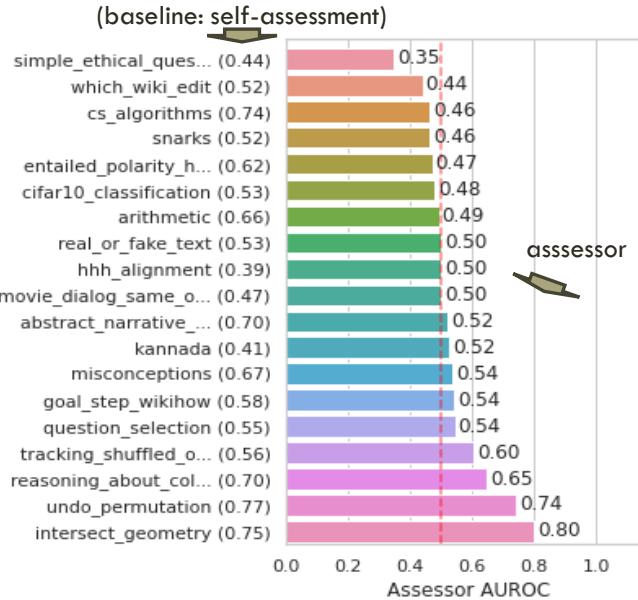
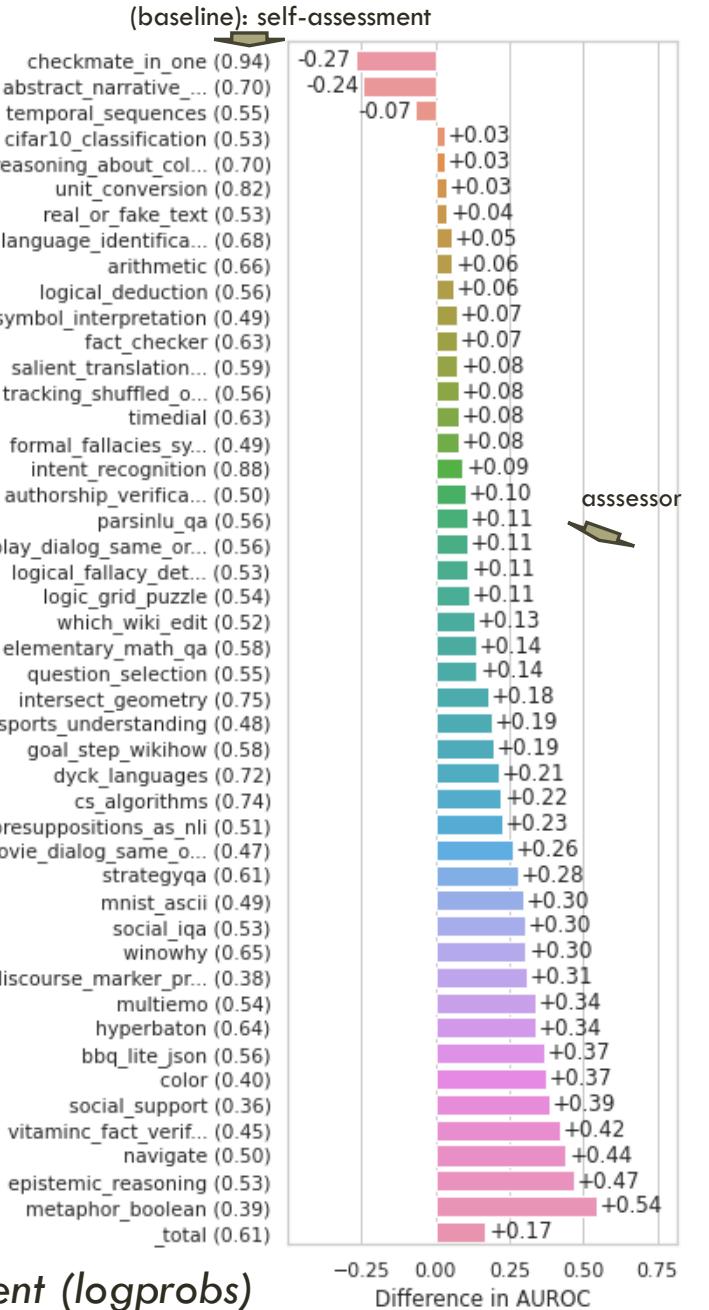
LMs PREDICT LMs

Setup:

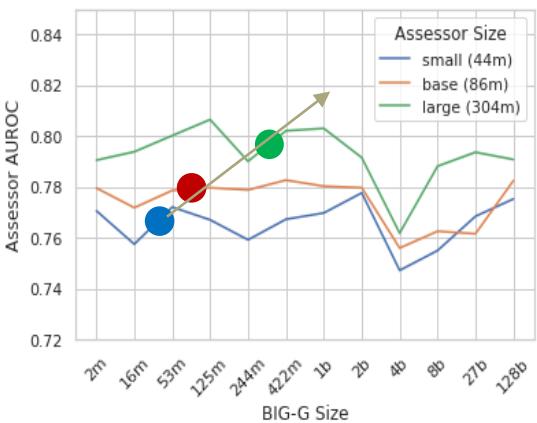
- Problem space (items):
 - BIG-bench evaluation suite (millions of instances)
- System space (subjects):
 - Validity (correct/incorrect) for 12 LMs (200M to 128B parameters)
- Assessor:
 - Small-ish assessor (60M DeBERTa)

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

Measurement Layouts

AAAI2024 Tutorial

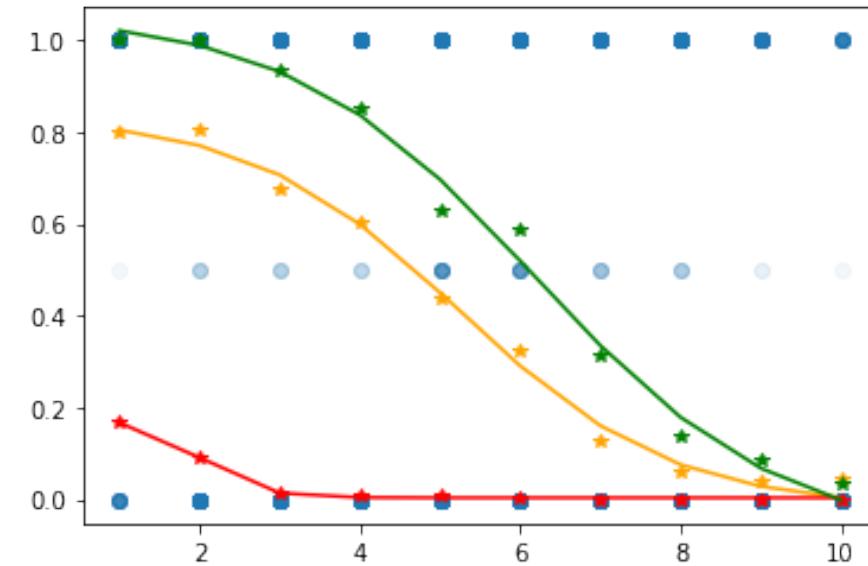
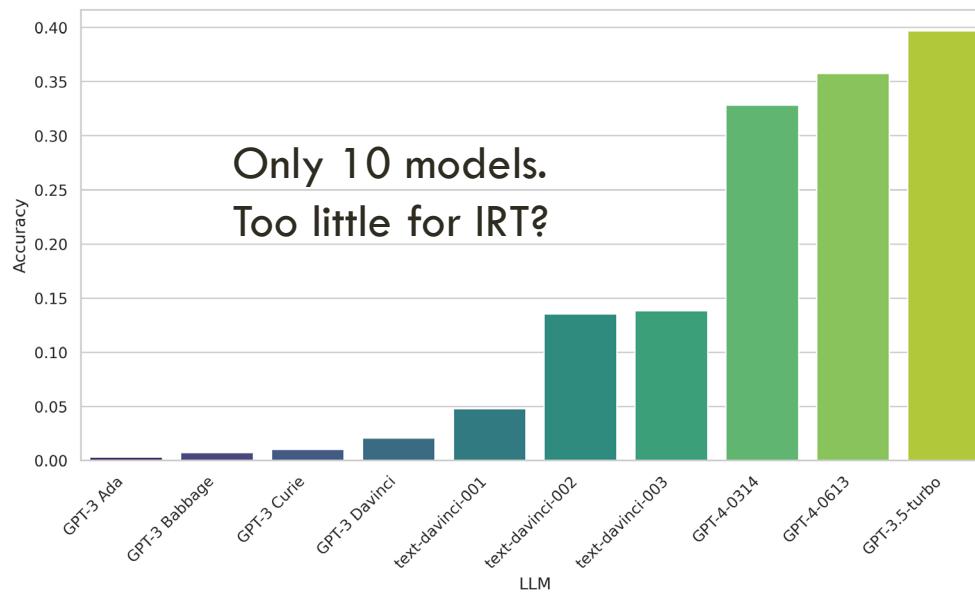
"Measurement Layouts for Capability-Oriented AI Evaluation",
J. Burden, L. Cheke, J. Hernández-Orallo, M. Tešić, K. Voudouris

<https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial>

J. Burden et al. "Inferring Capabilities from Task Performance with Bayesian Triangulation", <https://arxiv.org/abs/2309.11975>.

MORE SOPHISTICATED MODELS

- From performance to capabilities more generally:



Zhou et al. "Scaled-up, Shaped-up, but Letting Down? Reliability Fluctuations of Large Language Model Families", in preparation, 2024.

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

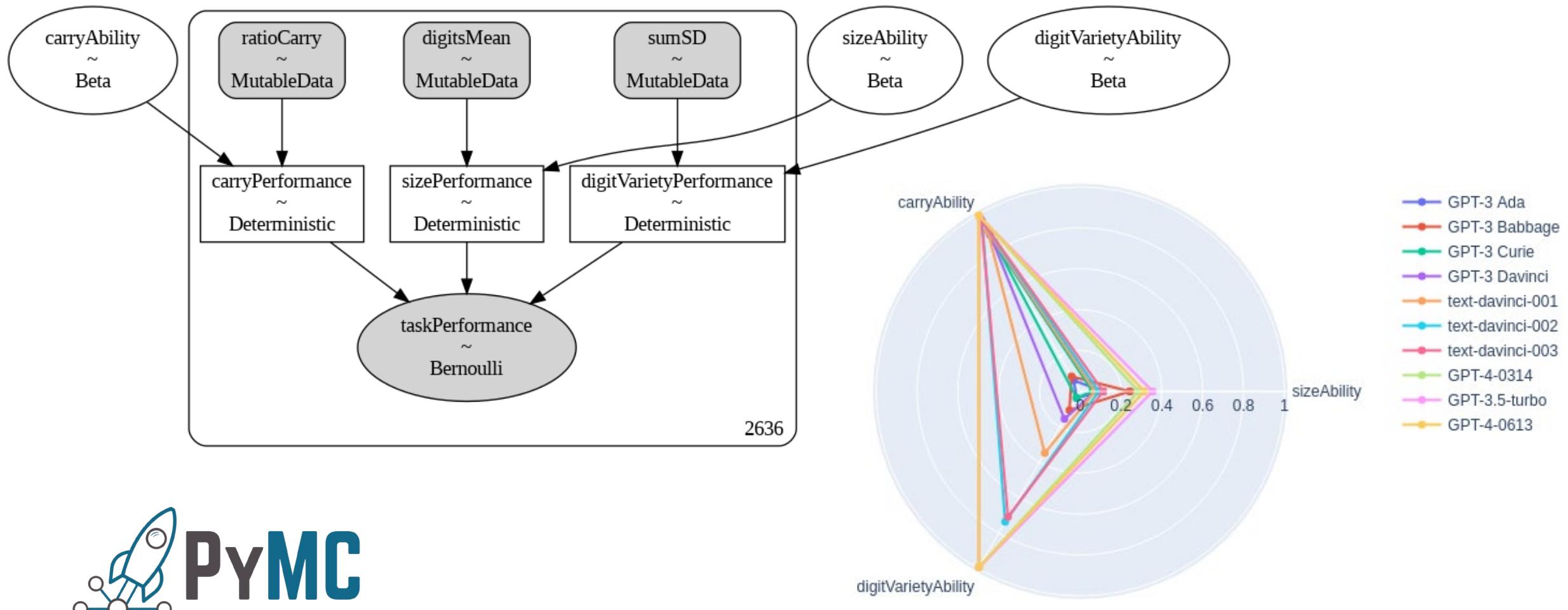
MORE SOPHISTICATED DEMANDS

- `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.

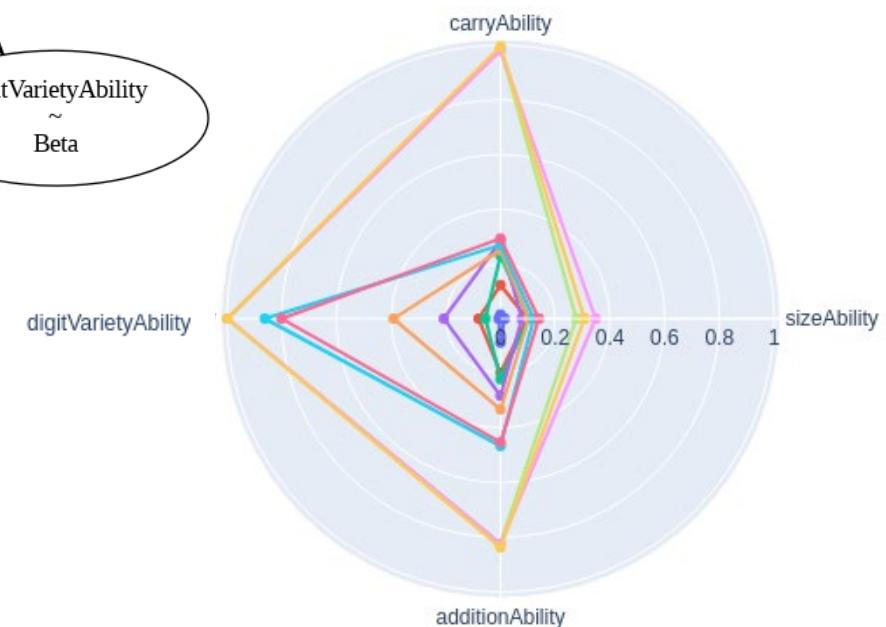
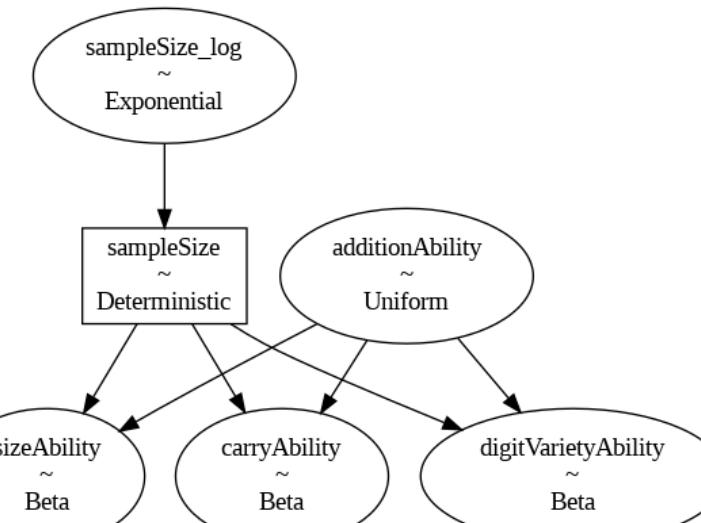
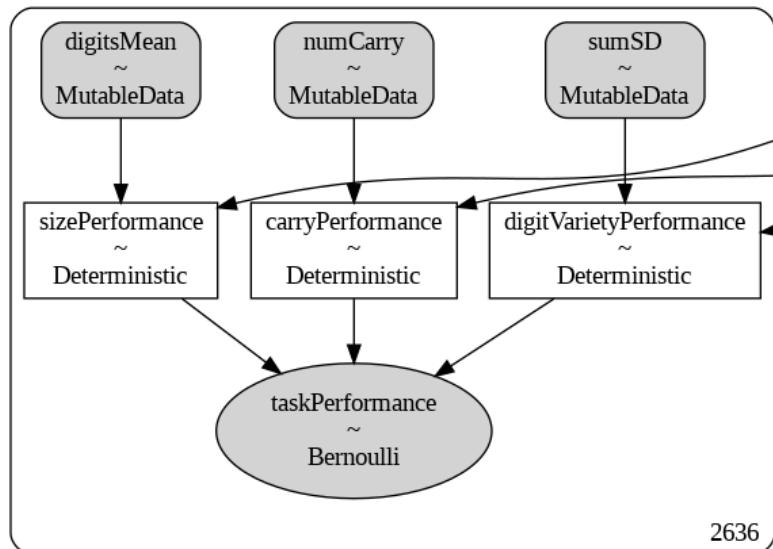
What are some of the things that make the addition of two number ‘difficult’?

- 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

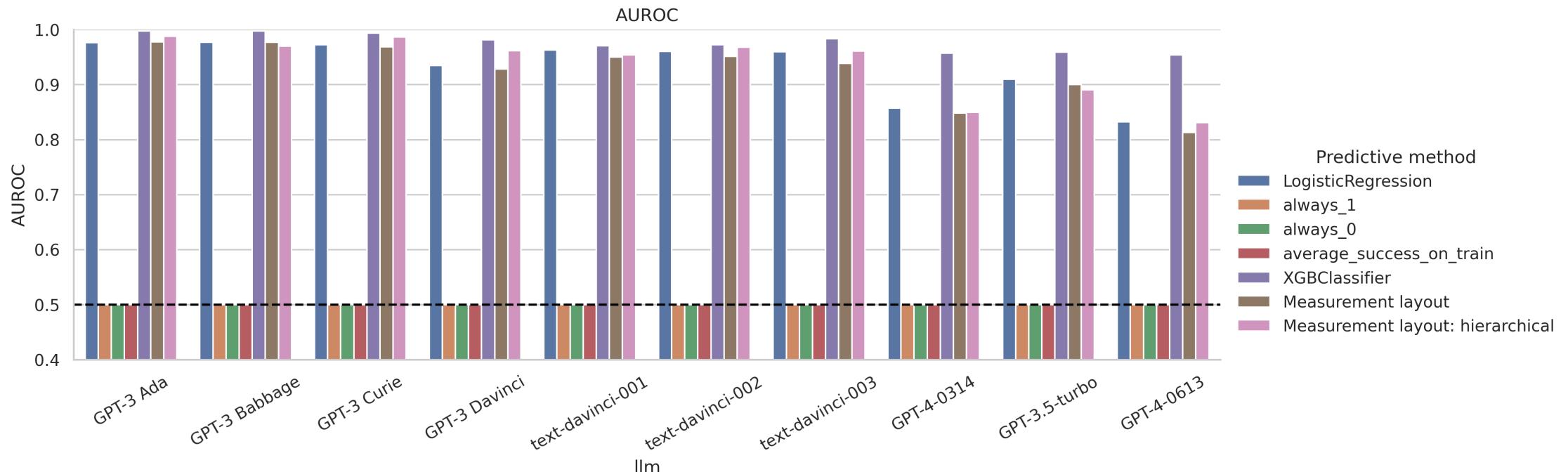


HIERARCHICAL MEASUREMENT LAYOUT



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 Modelling Approaches

OTHER METHODS TO EXPLAIN/PREDICT PEFORMANCE

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:

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

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

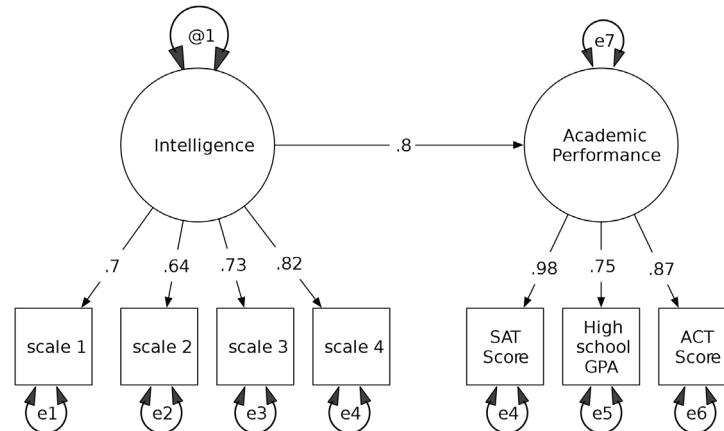
Sulis, I., & Toland, M. D. (2017). Introduction to Multilevel Item Response Theory Analysis: Descriptive and Explanatory Models. *The Journal of Early Adolescence*, 37(1), 85-128. <https://doi.org/10.1177/0272431616642328>

FACTOR ANALYSIS

Task	HELM classification	Annotated ability
XSUM	Summarization	Comprehension
HellaSwag	QA	Comprehension
NarrativeQA	QA	Comprehension
CNN.DailyMail	Summarization	Comprehension
IMDB	Sentiment Analysis	Comprehension
WikiFact	Knowledge	Domain knowledge
OpenbookQA	QA	Reasoning - commonsense
NaturalQuestions	QA	Comprehension
BoolQ	QA	Comprehension
RAFT	Text Classification	Comprehension
QuAC	QA	Comprehension
TwitterAAE	Language modelling	Language modelling
ICE	Language modelling	Language modelling
The Pile	Language modelling	Language modelling
BLiMP	Language modelling	Language modelling
TruthfulQA	QA	Domain knowledge
BBQ	Bias	Reasoning - inductive
GSM8K	Reasoning	Reasoning - mathematical
Synthetic reasoning (NL)	Reasoning	Reasoning - fluid
MATH	Reasoning	Reasoning - mathematical
CivilComments	Toxicity Classification	Comprehension
Synthetic reasoning (A)	Reasoning	Reasoning - fluid
MMLU	QA	Mixed
LegalSupport	Reasoning	Reasoning - inductive
LSAT	Reasoning	Reasoning - fluid
bAbI	Reasoning	Reasoning - deductive
Dyck	Reasoning	Reasoning - deductive

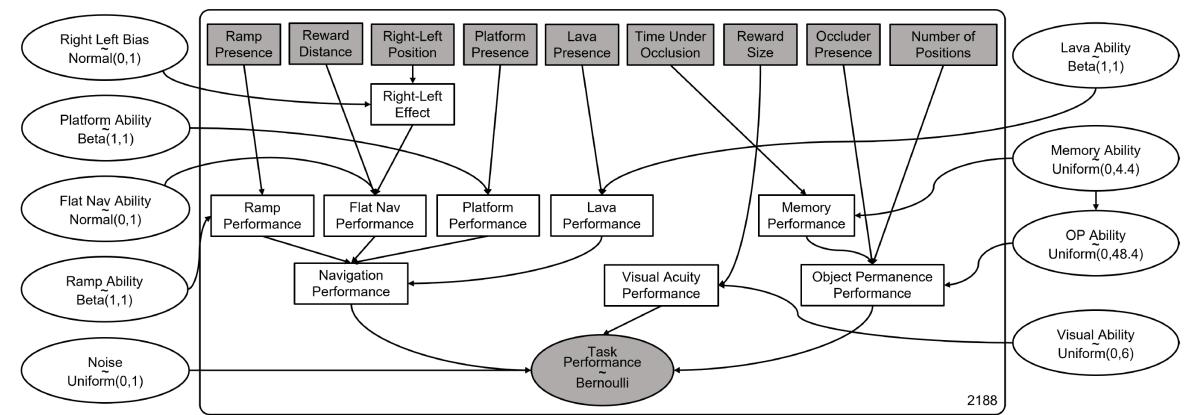
POPULATIONAL? INSTANCE-LEVEL?

- Structural Equation Modelling



- Needs a sample of subjects
- Bottom-up inference at the level of tests
- Inference of values
- Arrows represent linear relations

- Measurement Layouts (Bayesian inference)

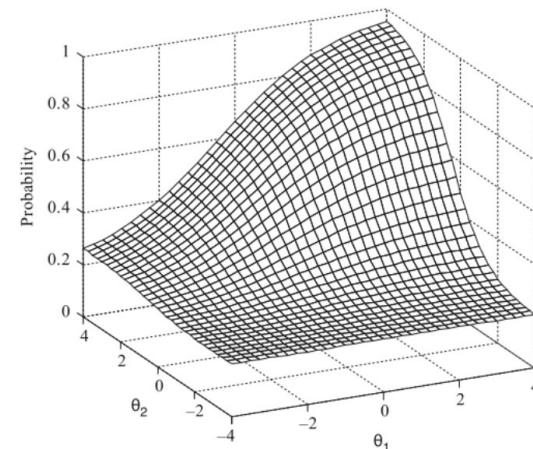


- Estimate capabilities from the results of one individual
- Bottom-up and top-down inference at instance level.
- Inference of distributions
- Arrows may be any differential function (e.g., logistic)

Question: Are SEMs or other models for just one individual?

MULTIDIMENSIONAL IRT GENERALISED?

- MIRT – Compensatory abilities

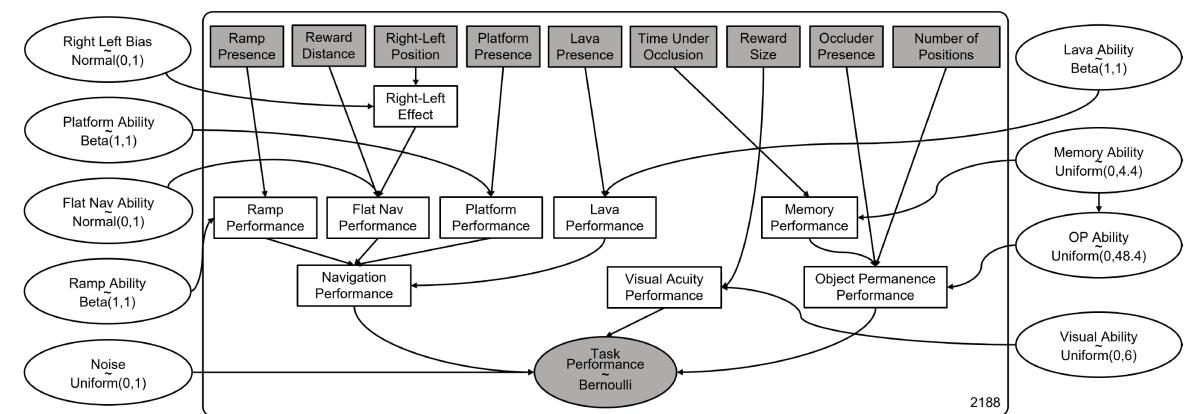


“Multidimensional Item Response Theory” (V. Duran’s slides)

Fig. 4.9 Item response surface for the partially compensatory model when $a_1 = .7$, $a_2 = 1.1$, $b_1 = -.5$, $b_2 = .5$, and $c = .2$

- Needs a sample of subjects
- Latent/population difficulties (no given features)
- Fixed model (logistic / beta)

- Measurement Layouts



- Estimate capabilities from the results of one individual
- Looks at the instance features (observable demands)
- Arrows only need be differentiable (beyond logistic)

Question: Degree of compensation for many dimensions and hierarchies?

SUMMARY OF APPROACHES

Approach	Predictive for items	Predictive for systems	Domain Knowledge	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	$1 (\geq 1 \text{ MIRT})$	Linear (diff) + Logistic (response)
General Difficulty Model	Seen & New	Seen & New	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

The Road Ahead

CHALLENGES

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.

Number/dependency of subjects:

- Non-populational approaches
- But they require some domain knowledge

ARTIFICIAL INTELLIGENCE

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By Ryan Burnell¹, Wout Schelleart², John Burden^{1,3}, Tomer D. Ullman⁴, Fernando Martinez-Plumed², Joshua B. Tenenbaum⁵, Danaja Rutar¹, 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. Leibo¹¹, 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 le-

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

TAKE-AWAYS

- IRT generally applicable if we have instance-level data and #subjects
- If situations are more elaborated or non-populational, there are alternatives.

Instead of aggregating performance, the key idea is to estimate a model of the AI system (e.g., capabilities) so that we can explain/predict performance at the instance level!

THANK YOU!

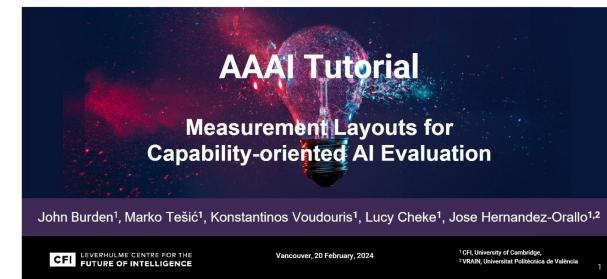
JOSE H. ORALLO

<http://josephorallo.webs.upv.es/>

jorallo@upv.es

POINTERS

- References: You've been given a reference list...
- Libraries:
 - PY-IRT: <https://github.com/nd-ball/py-irt/>
 - flexMIRT, MIRT, Stan, JAGS, Mplus, SPSS
- AAAI2024 Tutorial on Measurement Layouts:
 - <https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial>
- AI Evaluation Digest (monthly)
 - <https://aievaluation.substack.com/>



A screenshot of the 'The AI Evaluation Substack' homepage. The top navigation bar includes 'Home', 'Archive', and 'About'. Below the navigation is a decorative image of colorful, swirling patterns. To the right, there is a section titled '2024 February "AI Evaluation" Digest' with a brief description of the content and a link to 'FEB 23 - AI EVALUATION'.