Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

Welcome

Welcome!

About Us

- John Lalor, University of Notre Dame
- Pedro Rodriguez, Meta AI - FAIR
- Joao Sedoc, New York University
- Jose Hernandez-Orallo, Universitat Politècnica de València and the Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK

Today's Schedule

- Evaluation in NLP
- Introduction to IRT
- Break (15 minutes)
- IRT in NLP
- Break (15 minutes)
- Advanced Topics and Opportunities for Future Work
- Conclusion

Next up

- Next section: Introduction to IRT

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

In this session

Motivation

Introducing IRT

IRT Models with Artificial Crowds

The py-irt Package

Motivation

Differences between Examples

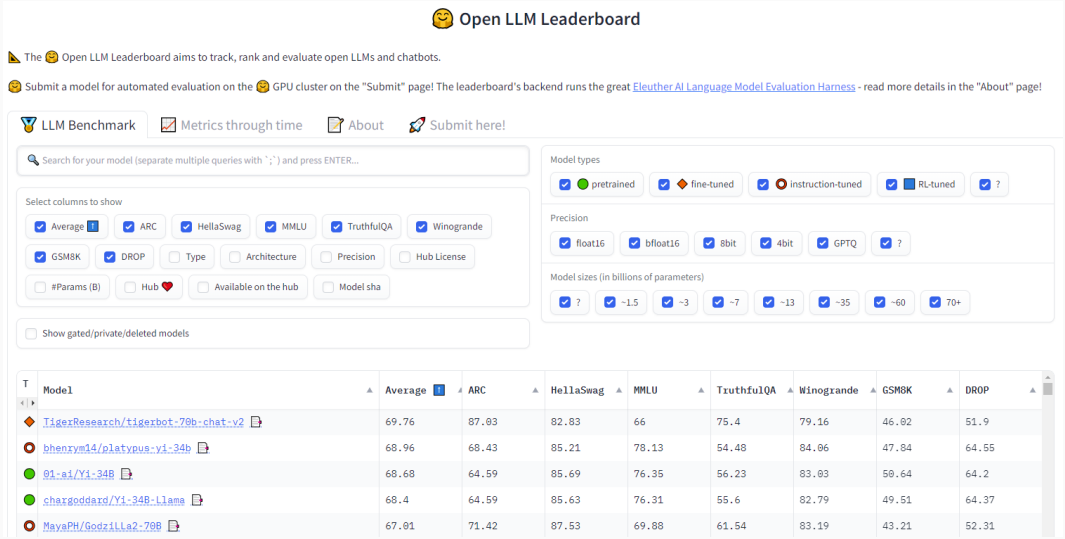
Natural language inference (NLI)

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	*easy*
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	*hard*
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	*easy*

Sentiment analysis (SA)

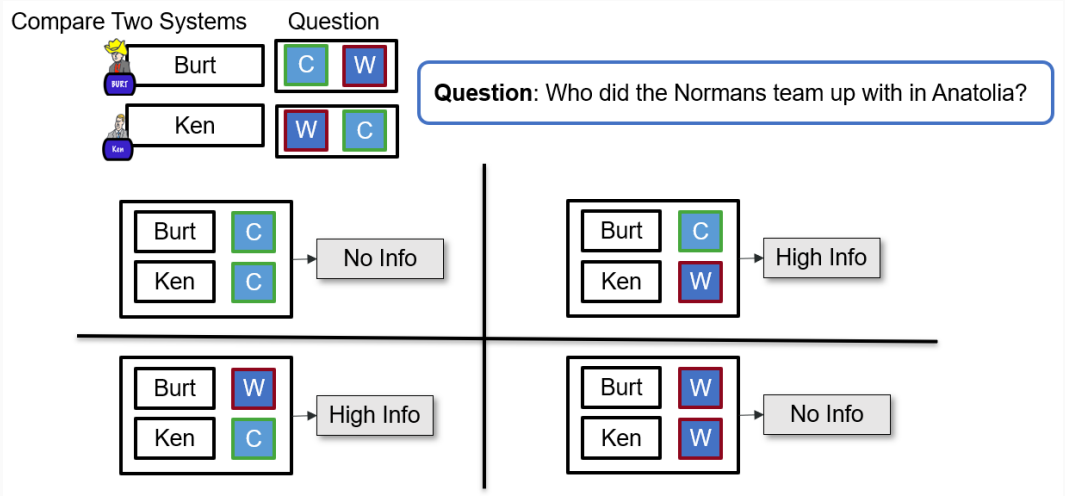
Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	*easy*
Still, it gets the job done - a sleepy afternoon rental.	Negative	*hard*
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	*easy*
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	*hard*

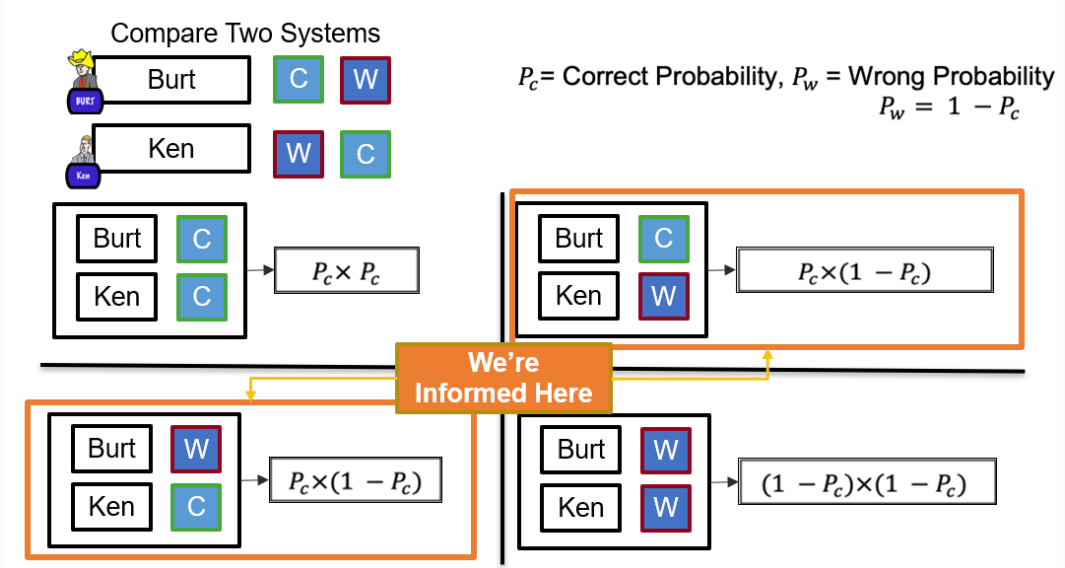
Leaderboards



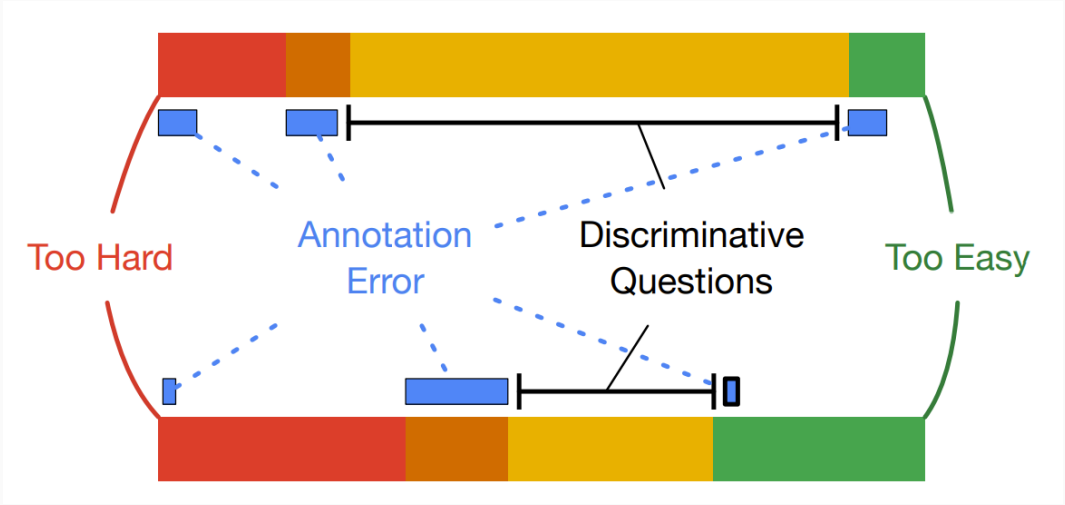
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Differences in Questions




Compare Two Systems
Burt
C
W
Ken
W
C
P_c= Correct Probability, P_w = Wrong Probability
$P_w = 1 - P_c$
Burt
C
Ken
C
$P_c \times P_c$
Burt
C
Ken
W
$P_c \times (1 - P_c)$
We're Informed Here
Burt
W
Ken
C
$P_c \times (1 - P_c)$
Burt
W
Ken
W
$(1 - P_c) \times (1 - P_c)$

Differences in Questions



Boyd-Graber J, Börschinger B. What Question Answering can Learn from Trivia Nerds. ACL 2020.

Introducing IRT

Psychometrics

Psychometrics: study of quantitative measurement practices

- Building instruments for measurement
- Development of theoretical approaches to measurement

Item Response Theory (IRT): measure latent traits of test-takers and test questions (“items’ ’)

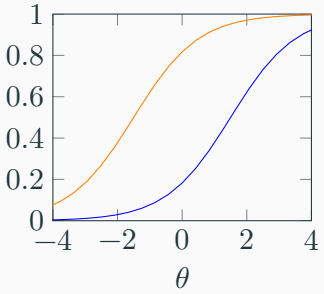
IRT: 1 Parameter Logistic Model (1PL)

Also known as *Rasch model*

$$p(y_{ij} = 1 | b_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$$

θ_j: latent ability

b_i: difficulty



Parameter Estimation

$$p(y_{ij} = 1|b_i, \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$
$$p(y_{ij} = 0|b_i, \theta_j) = 1 - p(y_{ij} = 1|b_i, \theta_j)$$

$$L = \prod_{j=1}^{J} \prod_{i=1}^{I} p(Y_{ij} = y_{ij}|b_i, \theta_j)$$
$$q(\Theta, B) = \prod_j \pi_j^{\theta}(\theta_j) \prod_i \pi_i^{b}(b_i)$$

Natesan, Prathiba, et al. "Bayesian prior choice in IRT estimation using MCMC and variational Bayes.'' Frontiers in psychology 2016

Let’s look at the code

Intro to IRT notebook 1 – 2_IntroToIrt.ipynb

Evaluating DNN Performance with IRT

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	-2.74
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction	0.08

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	-2.46
Still, it gets the job done - a sleepy afternoon rental.	Negative	1.78
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	-2.27
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	2.05

IRT for NLP: Human Annotations

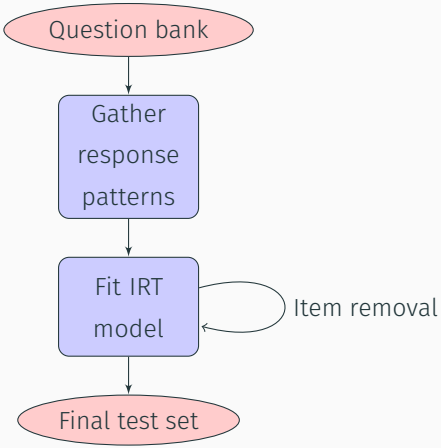
Item Set	Ability Score	Percentile	Test Acc.
“Easier”			
Entailment	-0.133	44.83%	96.5%
Contradiction	1.539	93.82%	87.9%
Neutral	0.423	66.28%	88%
“Harder”			
Contradiction	1.777	96.25%	78.9%
Neutral	0.441	67%	83%

Human Bottleneck

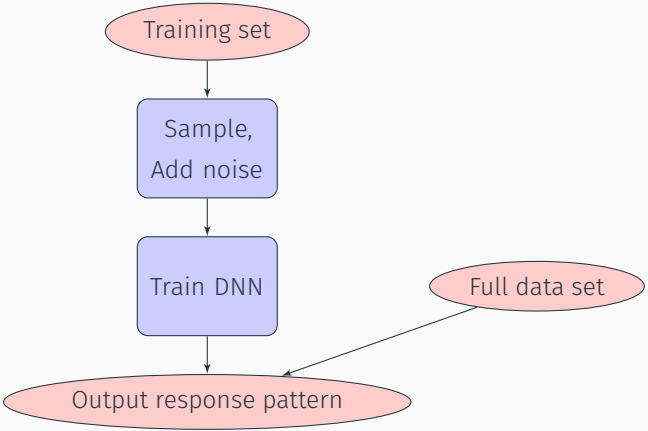
- Gathering human response patterns is expensive
- Can we use ensembles of models to gather response patterns?
- Even if we can, should we?

IRT Models with Artificial Crowds

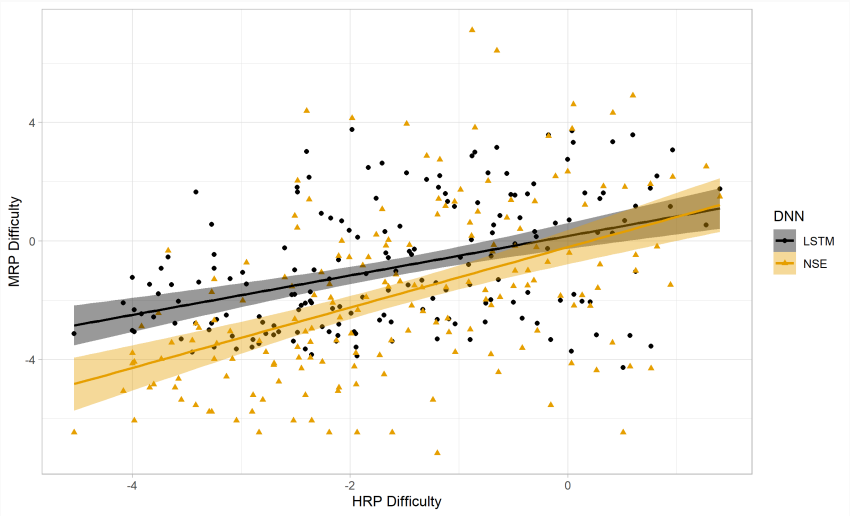
Building IRT Test Sets



Artificial Crowd Construction

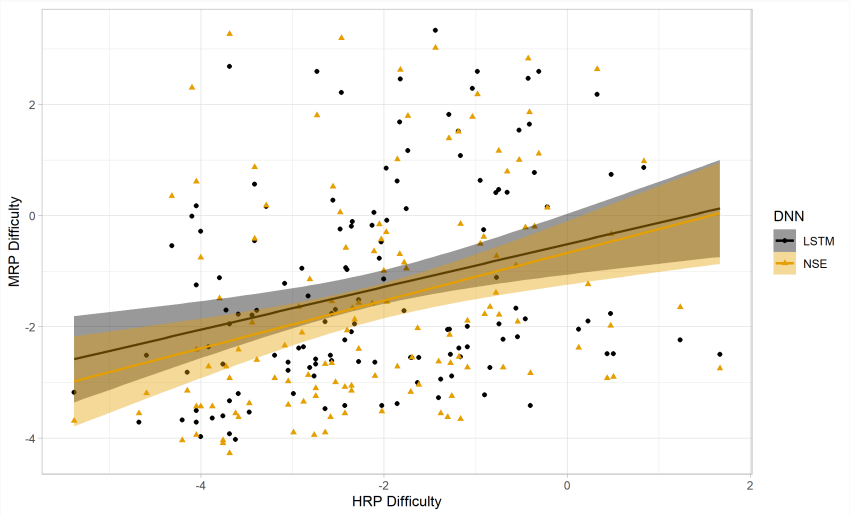


Human-Machine Correlation



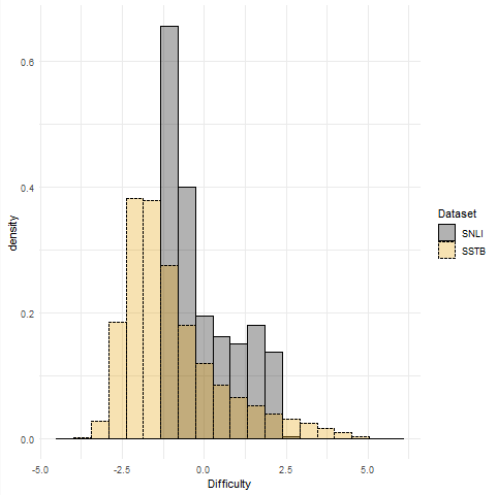
- Spearman ρ (NLI): 0.409 (LSTM) and 0.496 (NSE).

Human-Machine Correlation

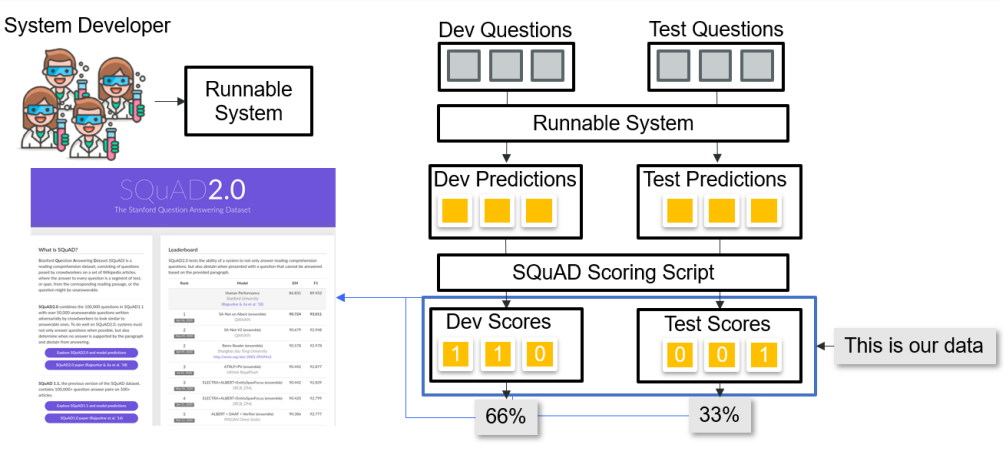


- Spearman ρ (SA): 0.332 (LSTM) and 0.392 (NSE).

Difficulty Distribution

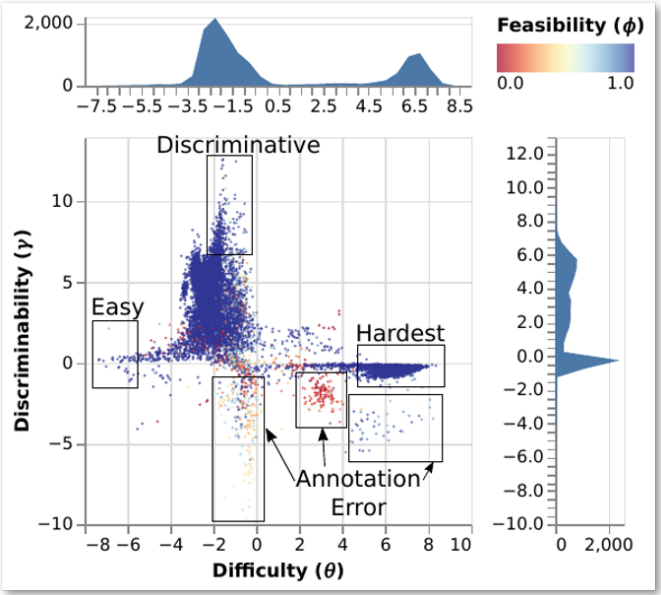


IRT for Leaderboards (SQuAD)

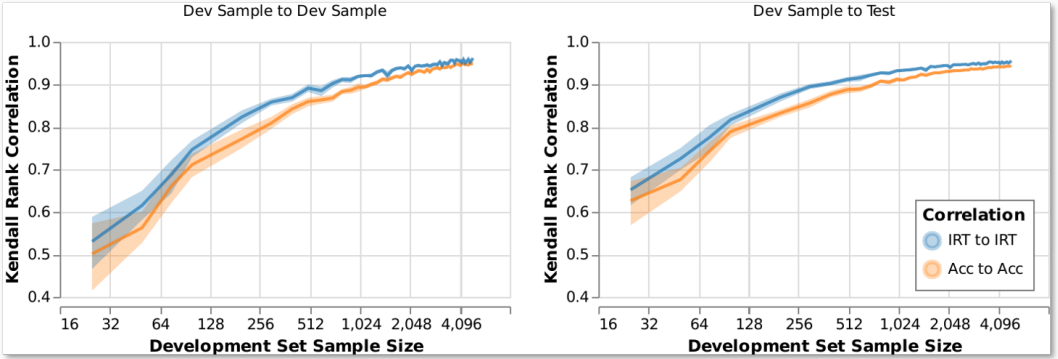


- 1.9 million subject-item pairs

IRT for SQuAD



Ranking Performance



The py-irt Package

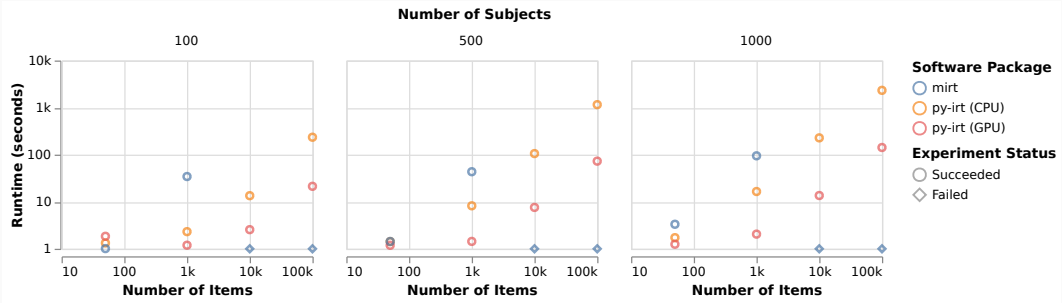
IRT in Python: py-irt

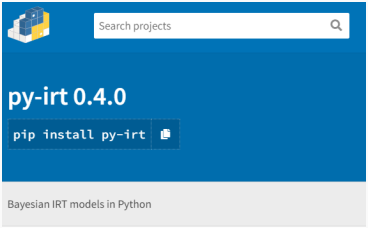
```
{"subject_id": "pedro",    "responses": {"q1": 1, "q2": 0, "q3": 1, "q4": 0}}
{"subject_id": "pinguino", "responses": {"q1": 1, "q2": 1, "q3": 0, "q4": 0}}
{"subject_id": "ken",      "responses": {"q1": 1, "q2": 1, "q3": 1, "q4": 1}}
{"subject_id": "burt",     "responses": {"q1": 0, "q2": 0, "q3": 0, "q4": 0}}
```

```
py-irt train 1pl data/data.jsonlines output/1pl/
```


```
{
  "ability": [
    -1.7251124382019043,
    -0.06789101660251617,
    1.6059941053390503,
    -0.20248053967952728
  ],
  "diff": [
    0.008014608174562454,
    9.654741287231445,
    -5.5452165603637695,
    -0.2792229950428009
  ],
  "irt_model": "1pl",
  "item_ids": {
    "0": "q2",
    "1": "q4",
    "2": "q1",
    "3": "q3"
  },
  "subject_ids": {
    "0": "burt",
    "1": "pinguino",
    "2": "ken",
    "3": "pedro"
  }
}
```

IRT in Python: py-irt





https://github.com/nd-ball/py-irt

Let’s look at the code

Intro to IRT notebook 2 – 2_pyirt_example.ipynb

Break!

- Back in 15 minutes
- Next section: IRT in NLP

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

In this session

IRT Applications

Improving Model Training

Finding Annotation Error

Evaluation Metrics

IRT Applications

IRT for NLP

Overview of IRT Applications:

- Dataset Construction
- Model Training
- Evaluation

Assumptions for IRT + NLP

Basic assumptions of the data and parameterization we have:

- A dataset with items indexed by i.
- A set of subjects indexed by j.
- Responses r_{ij} from graded responses of subjects to each item.
- An IRT parameterization, e.g., one with item difficulty β_i, discriminability γ_i, and skill θ_j might assume:

$$p(r_{ij} = 1 | \beta_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

What IRT Yields

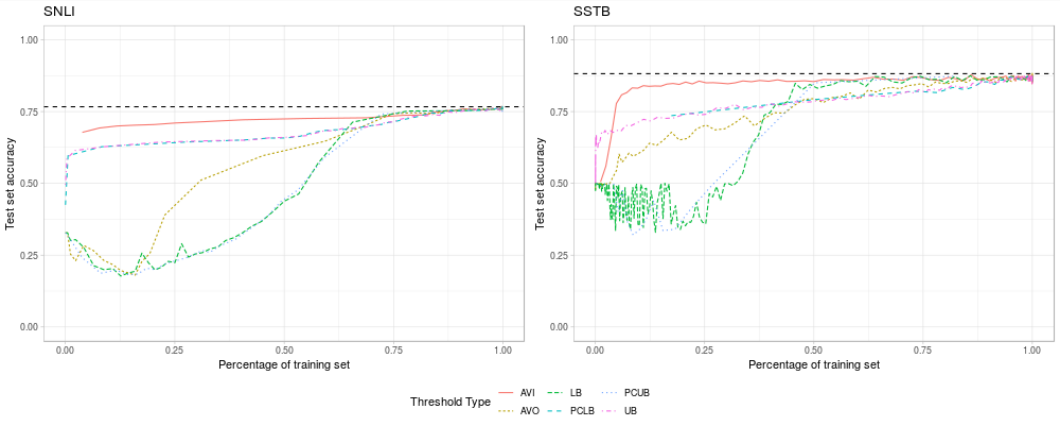
Given the previous information, IRT will yield estimates for chosen parameters, i.e.: item difficulty β_i, discriminability γ_i, and skill θ_j.

Consider two scenarios:

- What if the dataset is the training data?
- What if the dataset is a test set?

Improving Model Training

Data set filtering



- AVI: $|b_i| < \tau$
- UB: $b_i < \tau$
- PCUB: $pc_i < \tau$

- AVO: $|b_i| > \tau$
- LB: $b_i > \tau$
- PCLB: $pc_i > \tau$

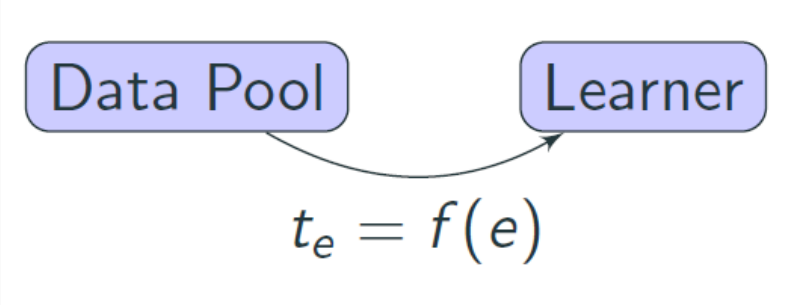
MT-DNN Results

Strategy	% of Training Data		
	0.1%	1%	10%
Random (reported)	82.1	85.2	**88.4**
Random (small batch)	81.79	84.90	88.32
Lower-bound	43.68	41.56	39.89
Upper-bound	81.62	80.46	79.06
AVI	**82.44**	**85.44**	86.73
AVO	43.60	42.05	40.81

Biggest Differences

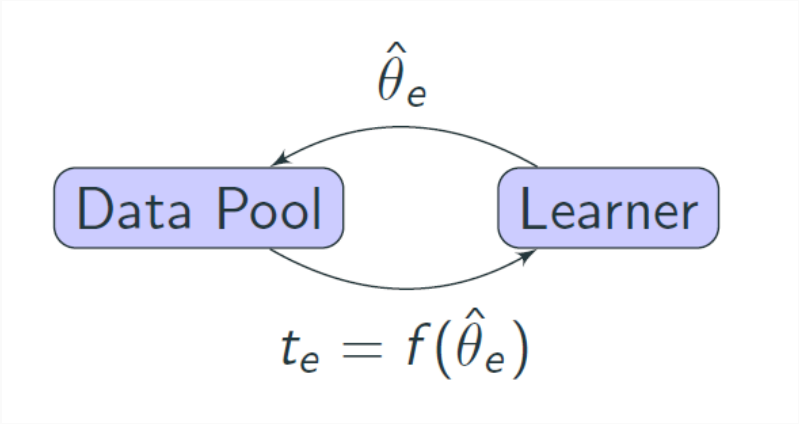
Task	Label	Item Text	Difficulty ranking		
			Humans	LSTM	NSE
SNLI	Con.	*P:* Two dogs playing in snow. *H:* A cat sleeps on floor	168	1	5
	Ent.	*P:* A girl in a newspaper hat with a bow is unwrapping an item. *H:* The girl is going to find out what is under the wrapping paper.	55	172	176
SSTB	Pos.	Only two words will tell you what you know when deciding to see it: Anthony. Hopkins.	9	103	110
	Neg.	...are of course stultifyingly contrived and too stylized by half. Still, it gets the job done–a sleepy afternoon rental.	128	46	41

Traditional Curriculum Learning



- Example difficulty based on heuristics
  - Replace heuristic with IRT difficulty
- Strategy is static
- Competence-based CL: $t_e = f(e, c_0)$ (Platanios et al., 2019)

Dynamic Data Selection



- Example difficulty is learned
- Training set *dynamically selected* as a function of model ability

Estimating θ

Gather responses from model j for items with known difficulties

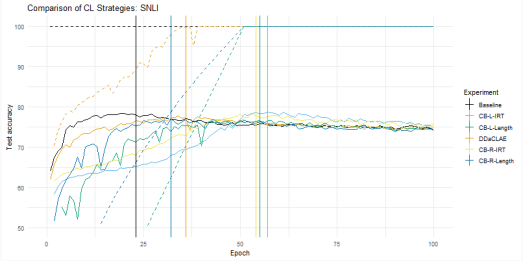
$$
\begin{aligned}
Z_j &= \forall_{y \in Y} \mathbf{I}[y_i = \hat{y}_i] \\
L(\theta_j | Z_j) &= p(Z_j | \theta_j) \\
\hat{\theta}_j &= \arg\max_{\theta_j} \prod_{i=1}^{I} p(z_{ij} = y_{ij} | \theta_j)
\end{aligned}
$$

DDaCLAE

Dynamic Data selection for Curriculum Learning via Ability Estimation

- At each epoch e:
  - Label all data: $\hat{Y}$
  - Estimate $\hat{\theta}_e$: $score(Y, \hat{Y}, B)$
  - Select training data: $b_i \leq \hat{\theta}_e$

Results



(a) SNLI

(b) SSTB

Results

Metric	Experiment	MNIST	CIFAR	SSTB	SNLI
%Δ Train Size	Baseline	0	0	0	**0**
	DDaCLAE	**-9.37**	**-53.71**	**-88.68**	33.51
	CB Lin	-8.22	-21.56	-73.17	38.07
	CB Root	11.29	-22.63	10.23	60.08
%Δ Accuracy	Baseline	**0**	0	0	0
	DDaCLAE	-0.17	**0.66**	**0.45**	-1.08
	CB Lin	-0.01	-0.90	-0.18	**0.69**
	CB Root	-0.06	0.13	-0.38	-0.37

Results

Label	Review	Δ_d
Pos	Heart	67342
Pos	The year's greatest adventure, and Jackson's limited but enthusiastic adaptation has made literature literal without killing its soul – a feat any thinking person is bound to appreciate.	67334
Pos	Hip	67332
Neg	Exit	67346
Neg	There's an admirable rigor to Jimmy's relentless anger, and to the script's refusal of a happy ending, but as those monologues stretch on and on, you realize there's no place for this story to go but down.	67330

Results

Label	Premise	Hypothesis	Δ_d
Con.	Two men in a jogging race on a black top street, one man wearing a black top and pants and the other is dressed as a nun with bright red tennis shoes, while onlookers stand in a grassy area and watch from behind a waist high metal railing.	There is no metal railing.	549179
Ent.	Two dogs in the water.	They are swimming	549180
Neut.	Male musicians are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	Male musicians with long hair are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	549184
Neut.	A dog in a lake.	A dog is swimming.	549183

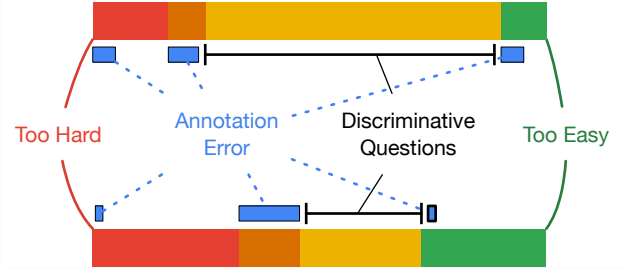
Remarks

- Correlation between parameters between human and machine IRT models
- Downstream effectiveness of difficulty
- Qualitative check of learned parameters
- What about θ?

Finding Annotation Error

IRT Applications: Finding Annotation Error

Test examples can be: too hard, discriminative, too easy, or erroneous [1]



How can we use IRT to identify each example type?

[1] Boyd-Graber and Börschinger (2020)

IRT Applications: Finding Annotation Error

What makes examples bad?

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct
- What parameter could identify this?

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct
- What parameter could identify this?
- We can use IRT discriminability γ_i to find bad examples!

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Skill $\sim U(-4, 4)$

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$
- Responses for invalid items: $r_{ij} = u > .5, u \sim U(0, 1)$

Then, train a 3PL IRT model with py-irt

IRT Applications: Setup for Finding Annotation Error

IRT Parameters

- Item Difficulty: $\beta_i \sim$ Normal
- Item Discriminability: $\gamma_i \sim$ LogNormal
- Subject Skill $\theta_j \sim$ Normal

IRT Model

$$p(r_{ij} = 1 | \beta_i, \gamma_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

IRT Applications: Setup for Finding Annotation Error

IRT Parameters

- Item Difficulty: $\beta_i \sim$ Normal
- Item Discriminability: $\gamma_i \sim$ LogNormal
- Subject Skill $\theta_j \sim$ Normal

IRT Model

$$p(r_{ij} = 1 | \beta_i, \gamma_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Note:

- Why $\gamma_i \sim$ LogNormal? Following Vania et al. (2021), forces γ_i to be non-negative.
- Other variables are zero centered.

IRT Applications: Sample Code for Finding Errors

Sample Code

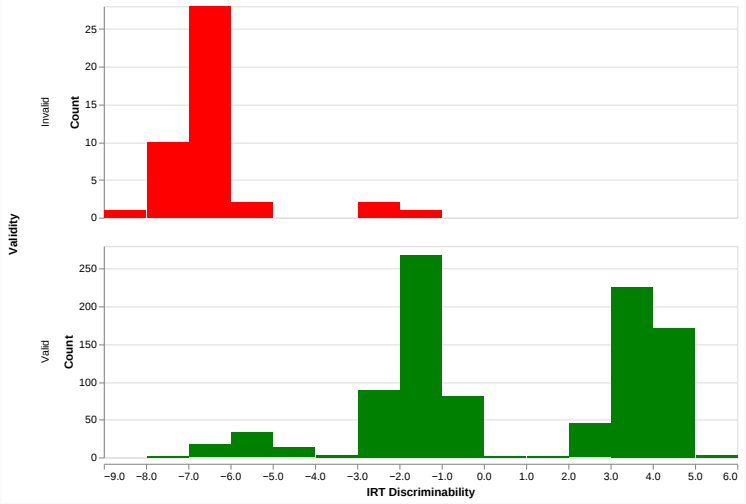
```
dataset = Dataset.from_jsonlines("/tmp/irt_dataset.jsonlines")
config = IrtConfig(
  model_type='tutorial', log_every=500, dropout=.2
)
trainer = IrtModelTrainer(
  config=config, data_path=None, dataset=dataset
)
trainer.train(epochs=5000, device='cuda')
```

IRT Applications: Simulation Results

Can we distinguish valid from invalid items based on discriminability γ_i?

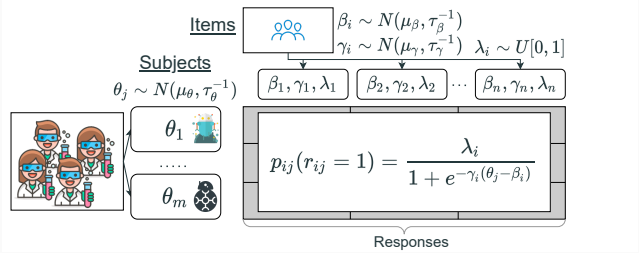
IRT Applications: Simulation Results

Can we distinguish valid from invalid items based on discriminability γ_i?



IRT Applications: Finding Annotation Error

In Rodriguez et al. (2021), we used a slightly different model to do this for SQuAD:

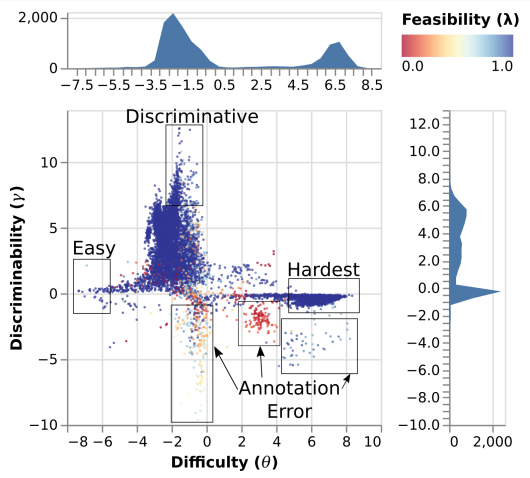


Differences

- Discriminability γ_i could be negative, which is inconvenient
- Feasibility λ_i more difficult to control

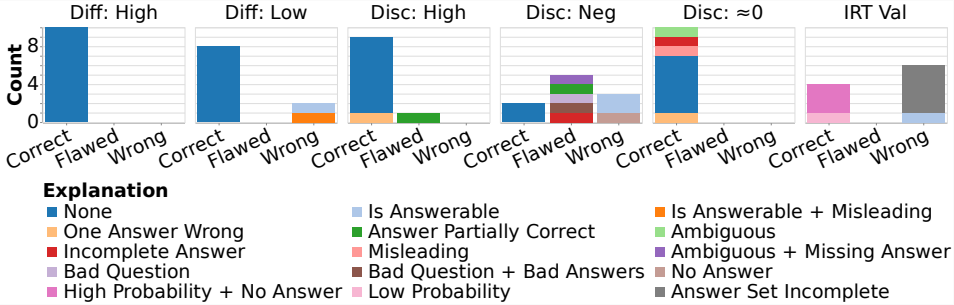
IRT Applications: Finding Annotation Error

Plotting IRT parameters:



IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors



Things to note:

- Difficulty can be high or low, not an issue itself
- Negative discriminability identifies errors

Evaluation Metrics

IRT Applications: Evaluation Metrics

Simple Idea: Instead of accuracy, use subject skill θ_j to rank.

IRT Applications: Evaluation Metrics

Simple Idea: Instead of accuracy, use subject skill θ_j to rank.

What are the tradeoffs?

IRT Applications: Evaluation Metrics Example

Suppose the following:

- As before, 1,000 Test Examples
- A set of 800 easy examples
- A set of 150 moderate examples
- A set of 50 hard examples
- 10 Subjects, similar setup as before

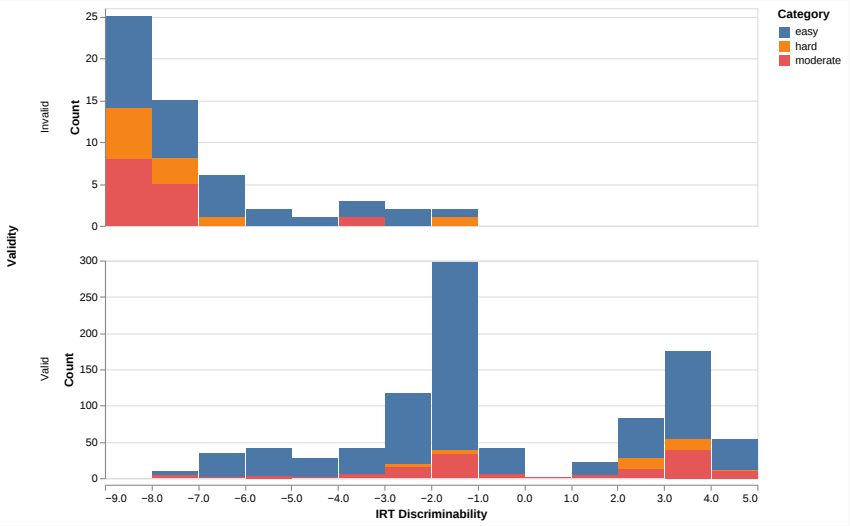
IRT Applications: Evaluation Metrics Example

- Subjects sorted by True skill
- Accuracy gaps vary
- IRT can account for some of this variability

True	IRT	Total	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100
-3.000	-7.61	0.256	0.301	0.066	0.100
-2.645	-4.88	0.325	0.380	0.093	0.140
-1.214	0.348	0.543	0.650	0.113	0.120
-1.156	1.40	0.560	0.667	0.120	0.160
-0.748	2.68	0.602	0.712	0.146	0.200
-0.455	3.36	0.631	0.746	0.193	0.100
0.232	5.76	0.729	0.848	0.293	0.120
2.16	11.1	0.865	0.956	0.586	0.240
2.50	14.2	0.897	0.971	0.686	0.340

IRT Applications: Discounting Bad Examples

- Invalid examples sorted down
- Harder examples tend to be more discriminating



IRT Applications: Rank Reliability in Evaluation Metrics

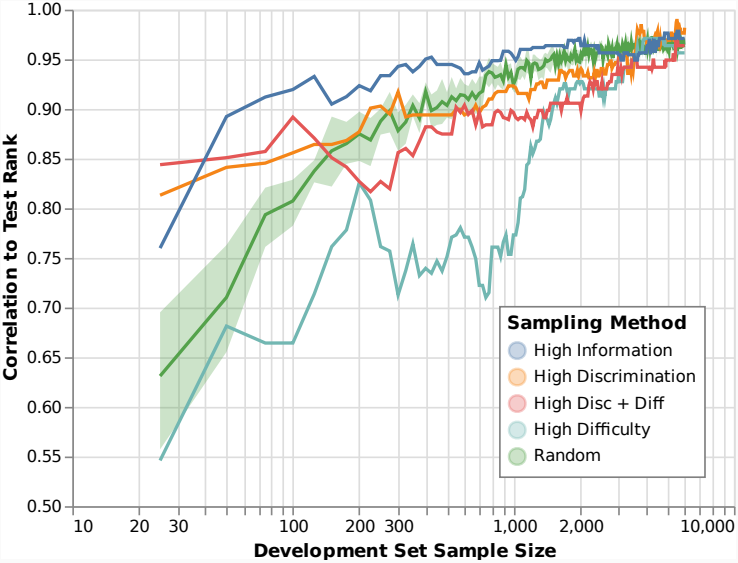
In Rodriguez et al. (2021), we examined a case like the SAT where we have:

- Pre-existing set of annotated responses for subjects/items
- Have a set of subjects (i.e., new models), same items.
- We want to minimize the number of subject responses to annotate, while maximizing the reliability of the resulting ranking.
- Baseline: Random sample
- IRT Methods: Sample based on different parameters

IRT Applications: Rank Reliability in Evaluation Metrics

Overall best method: pick item that maximizes Fisher information content, i.e.,

$$I_i(\theta_j) = \gamma_i^2 p_{ij}(1 - p_{ij})$$
$$Info(i) = \sum_j I_i(\theta_j)$$



Additional Work

- Alternate Evaluation Metrics, e.g., Subject skill θ_j (Lalor et al., 2018)
- Estimate Longevity of Tasks (Vania et al., 2021)
- Efficient Test Set Selection (non-irt) (Vivek et al., 2024)
- Building Tiny Benchmarks (Polo et al., 2024)

Break!

- Back in 15 minutes
- Next section: Advanced Topics

References

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

John P. Lalor and Hong Yu. 2020. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 545–555, Online. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online. Association for Computational Linguistics.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, St. Julian's, Malta. Association for Computational Linguistics.

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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/

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)

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

- 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

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*.

- In regression!
  - The loss function is open (MAE/MSE: [0,∞] loss)
  - Correspond to Gamma or some other distributions.
  - Gamma IRT models with 3 parametres (mapping difficulty, discrimination and ability to the Gamma)

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 on ability is needed to explain system performance.

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

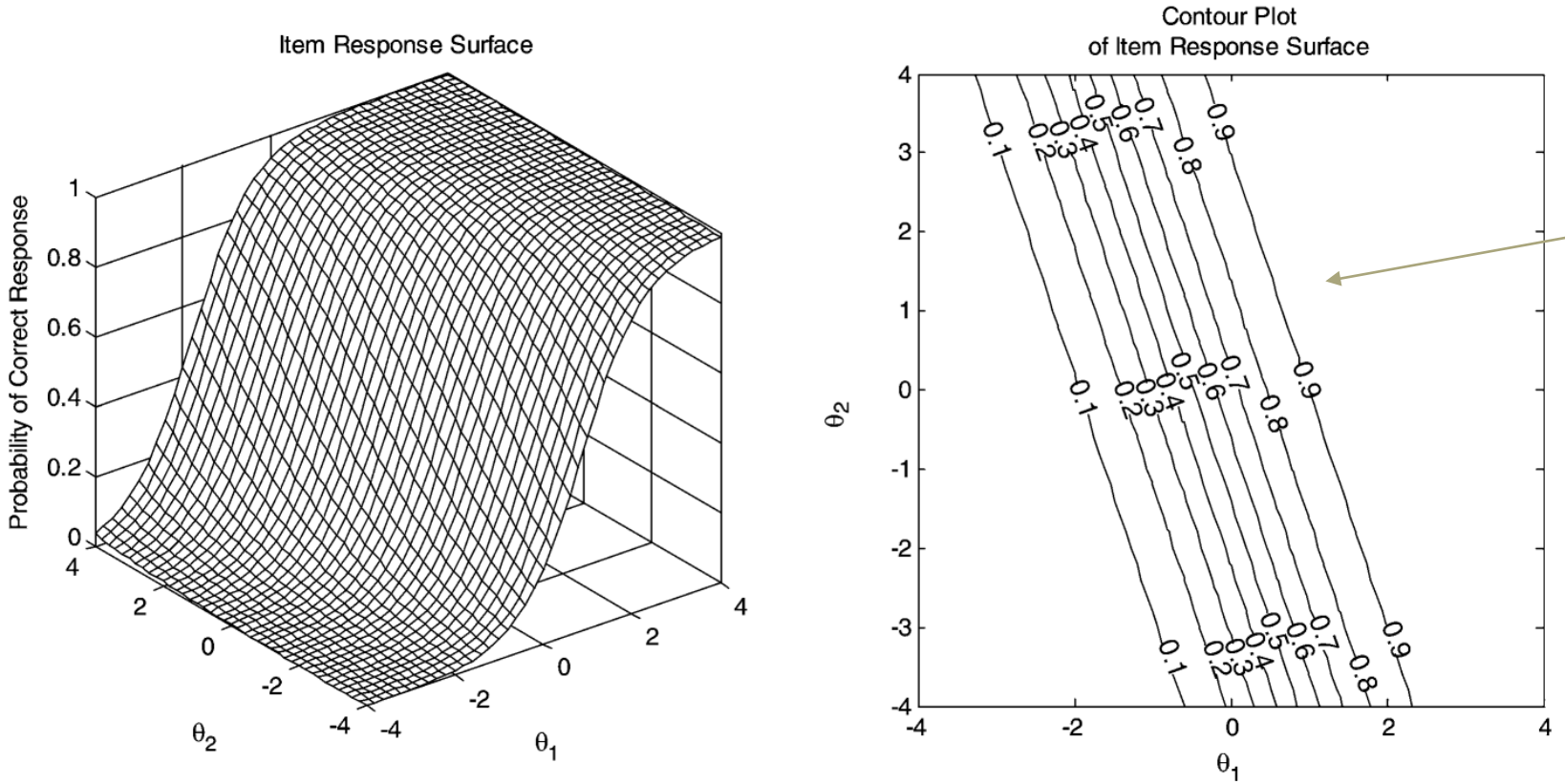
- Ability θ becomes a latent vector and/or difficulty d becomes a latent vector:

$$P(u_i = 1|\boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}}{1 + e^{\mathbf{a}_i'\boldsymbol{\theta}_j + d_i}} \qquad P(u_i = 1|\boldsymbol{\theta}_j) = \frac{e^{\mathbf{a}_i'\boldsymbol{\theta}_j + \mathbf{w}_i'\mathbf{d}}}{1 + e^{\mathbf{a}_i'\boldsymbol{\theta}_j + \mathbf{w}_i\mathbf{d}}}$$

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



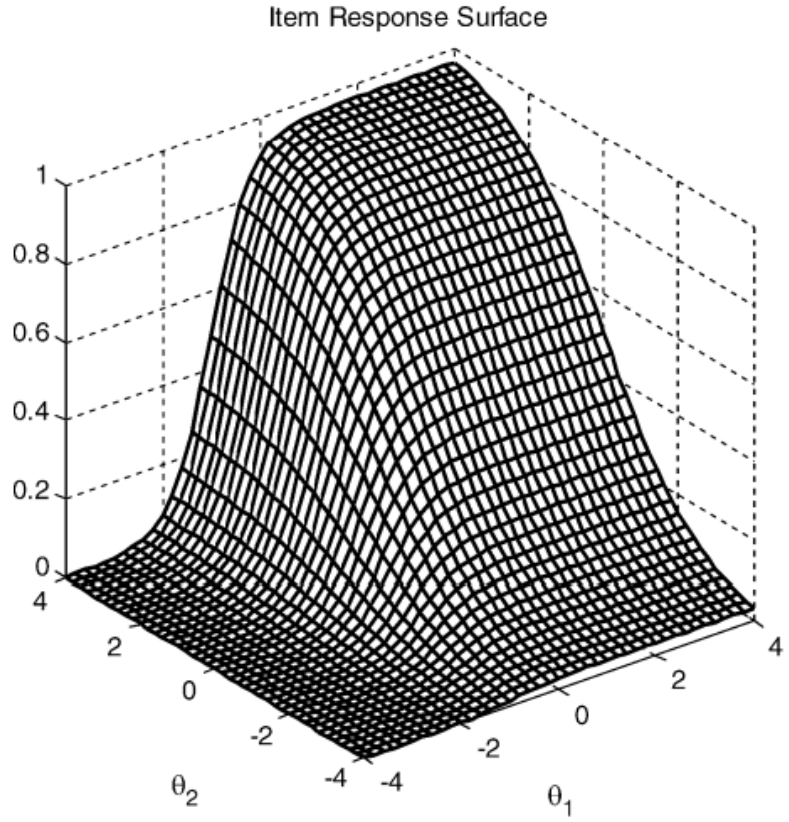
Graphic representations of the compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, and $d_i = .7$.

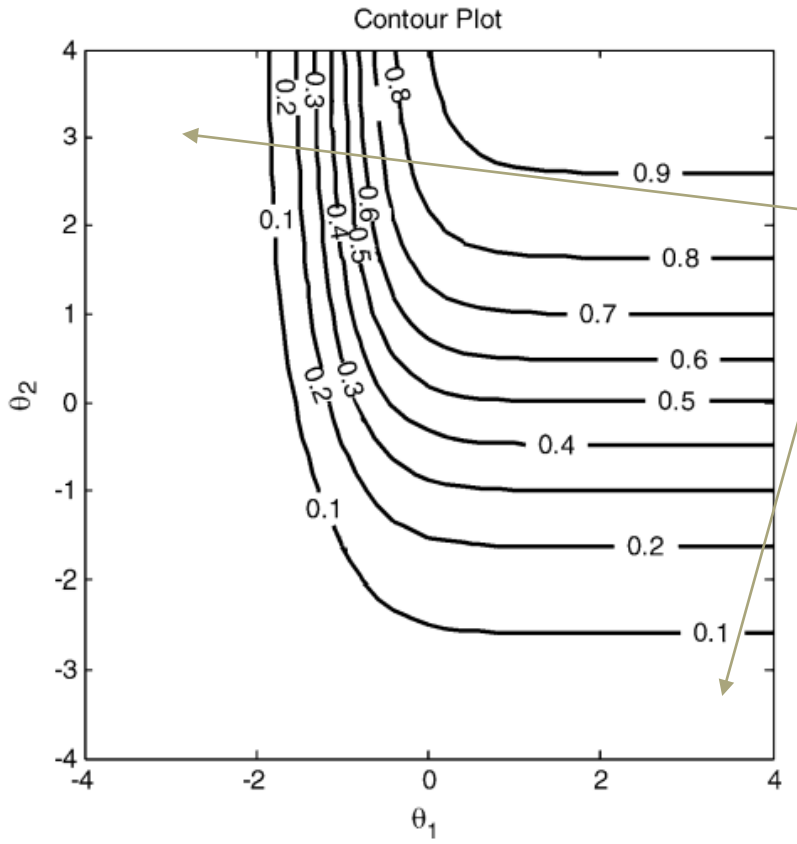
Asymmetric compensation: Given this angle, ability 1 can compensate for ability 2 but not vice versa.<

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

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:

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

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.

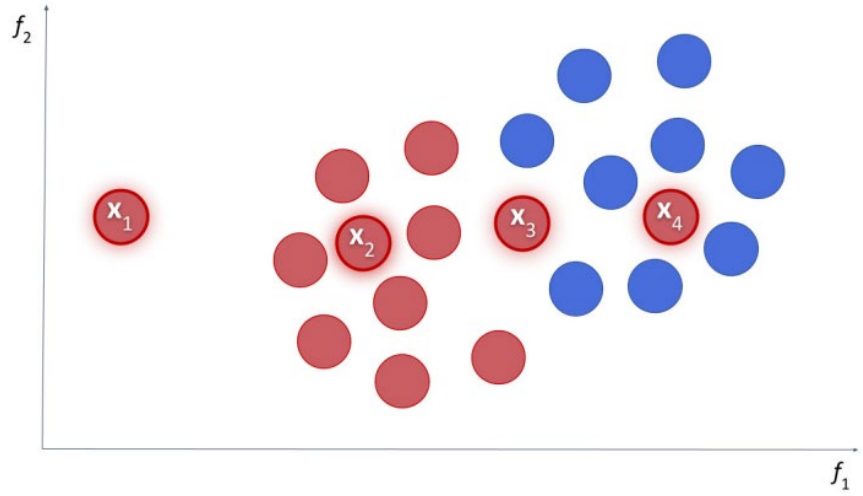
- Values can be > 1

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 (it 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.

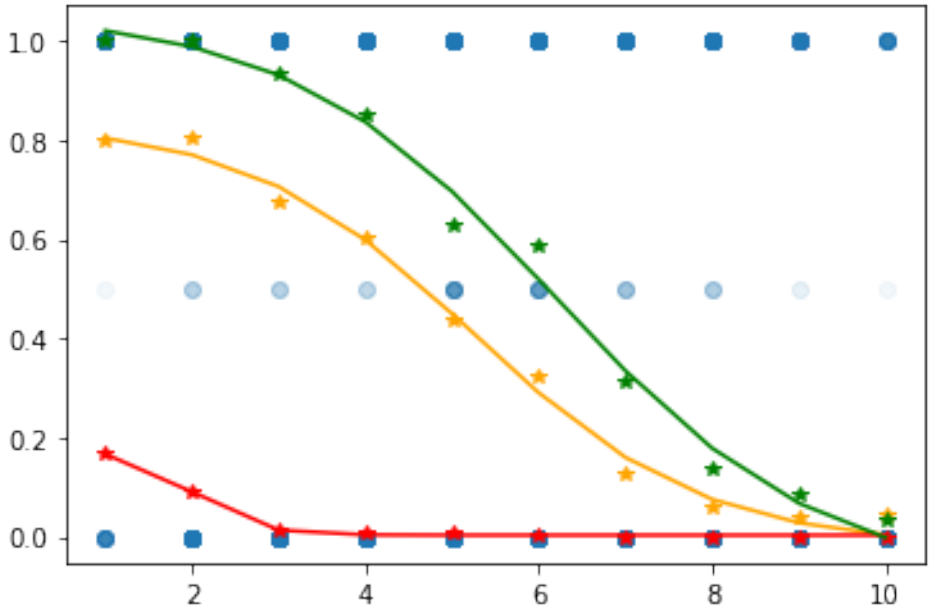


X1: medium
X2: easy
X3: hard
X4: very hard.

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

- Syntantic 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 ultracrepidarian 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

Yael Moros-Daval "Automated Annotation of Meta-Features for Predicting Language Model Performance in Natural Language Processing Tasks", 2023

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*.

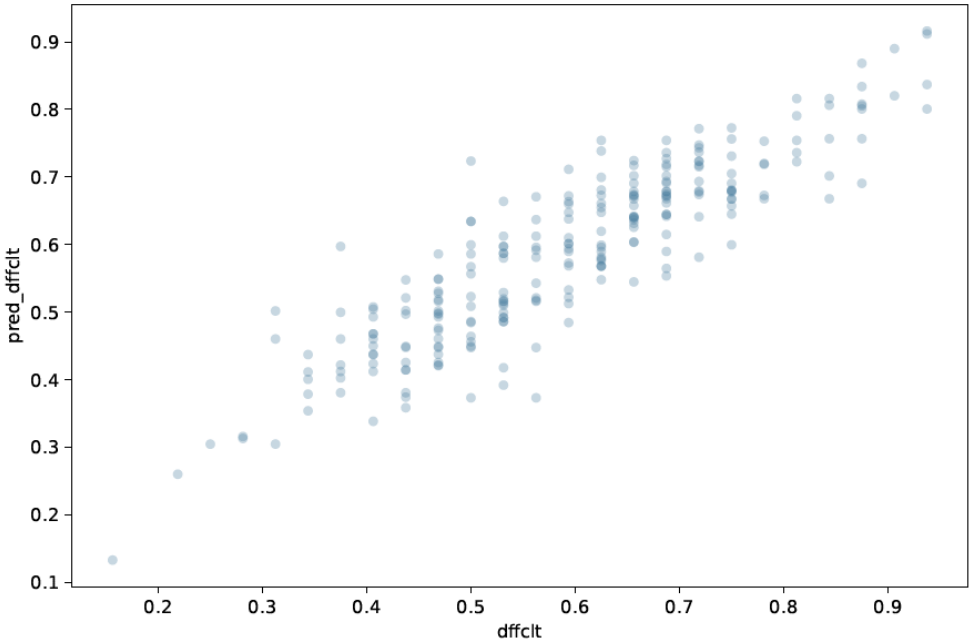
Creator	Model	Number of Parameters
AI21 Labs	J1-Jumbo v1	178B
AI21 Labs	J1-Large v1	7.5B
AI21 Labs	J1-Grande v1	17B
AI21 Labs	J1-Grande v2 beta	17B
Aleph Alpha	Luminous Base	13B
Aleph Alpha	Luminous Extended	30B
Aleph Alpha	Luminous Supreme	70B
Anthropic	Anthropic-LM v4-s3	52B
BigScience	BLOOM	176B
BigScience	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	
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:

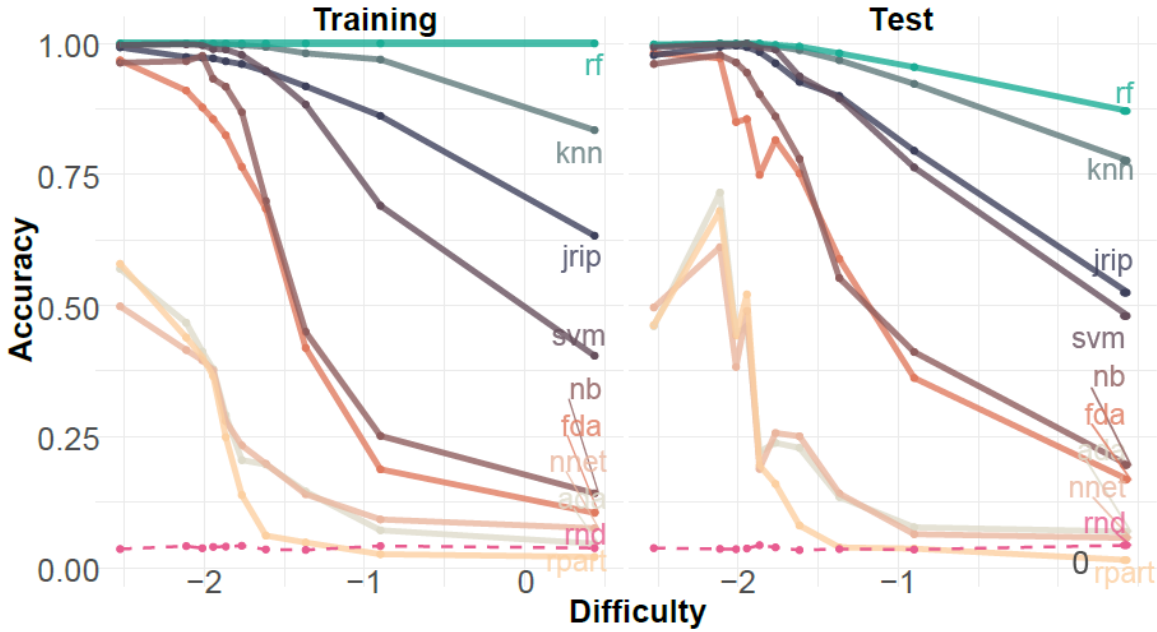


Figure 5: (Left) SCC obtained with the 70% of the letter benchmark and the observed difficulties $\hbar$. (Right) SCC obtained with the test set (30%), using estimated difficulties $\hat{\hbar}$.

Martínez-Plumed, F., Castellano, D., Monserrat-Aranda, C., & Hernández-Orallo, J. (2022, June). When ai difficulty is easy: The explanatory power of predicting IRT difficulty. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 7, pp. 7719-7727).

Predicting Performance Directly: Assessors

JH Orallo, W Schellaert, FM Plumed
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:

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

π	μ	r
Resnet, θ_1, θ_2, …	Image3, χ_1, χ_2, …	1
Resnet, θ_1, θ_2, …	Image23, χ_1, χ_2, …	0
…	…	…
Inception, θ_1, θ_2, …	Image3, χ_1, χ_2, …	1
Inception, θ_1, θ_2, …	Image78, χ_1, χ_2, …	1
…	…	…

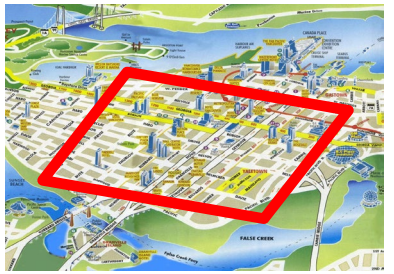
It is trained (and evaluated) on test data:

- 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.

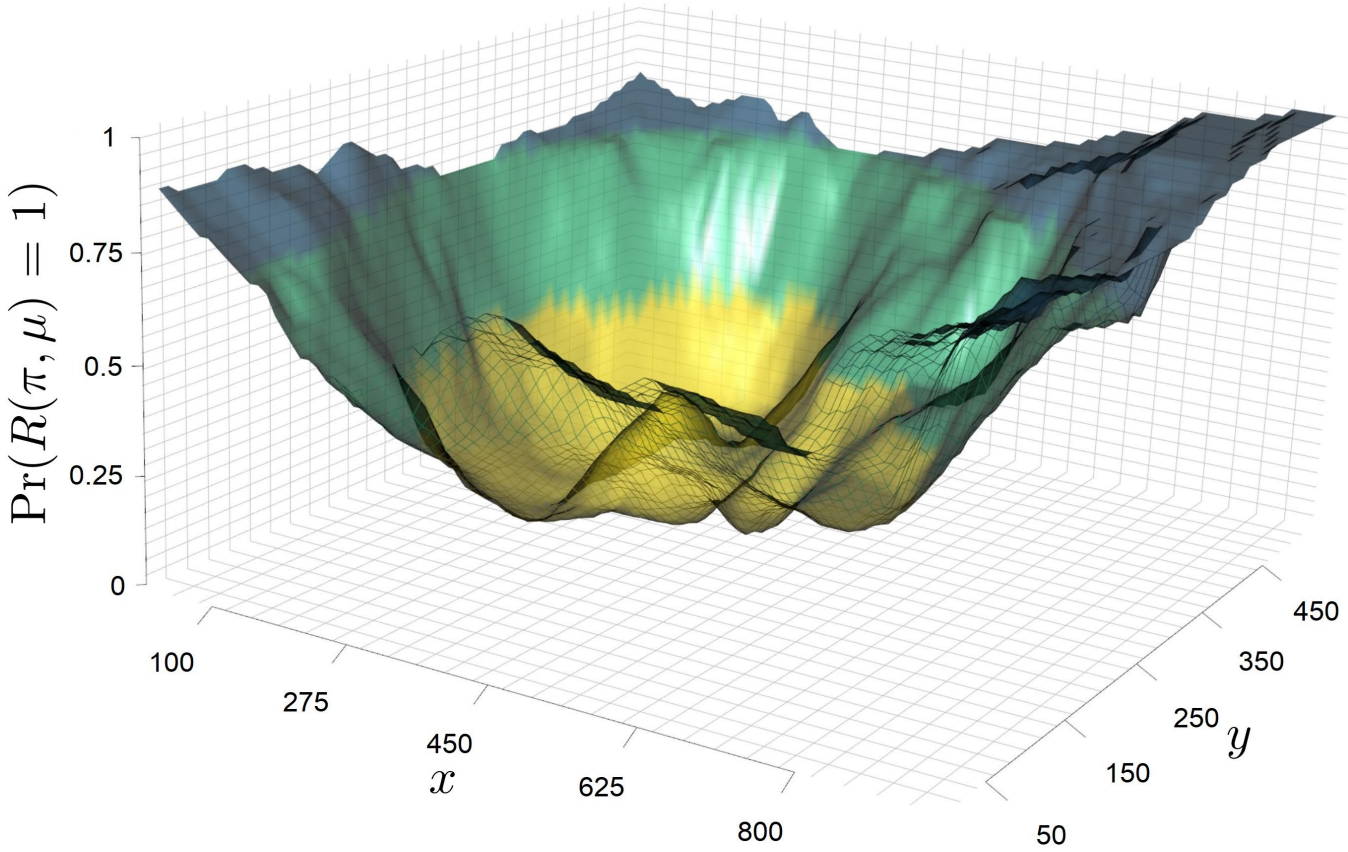
PROBLEM SPACE

Downtown Vancouver



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

- 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)



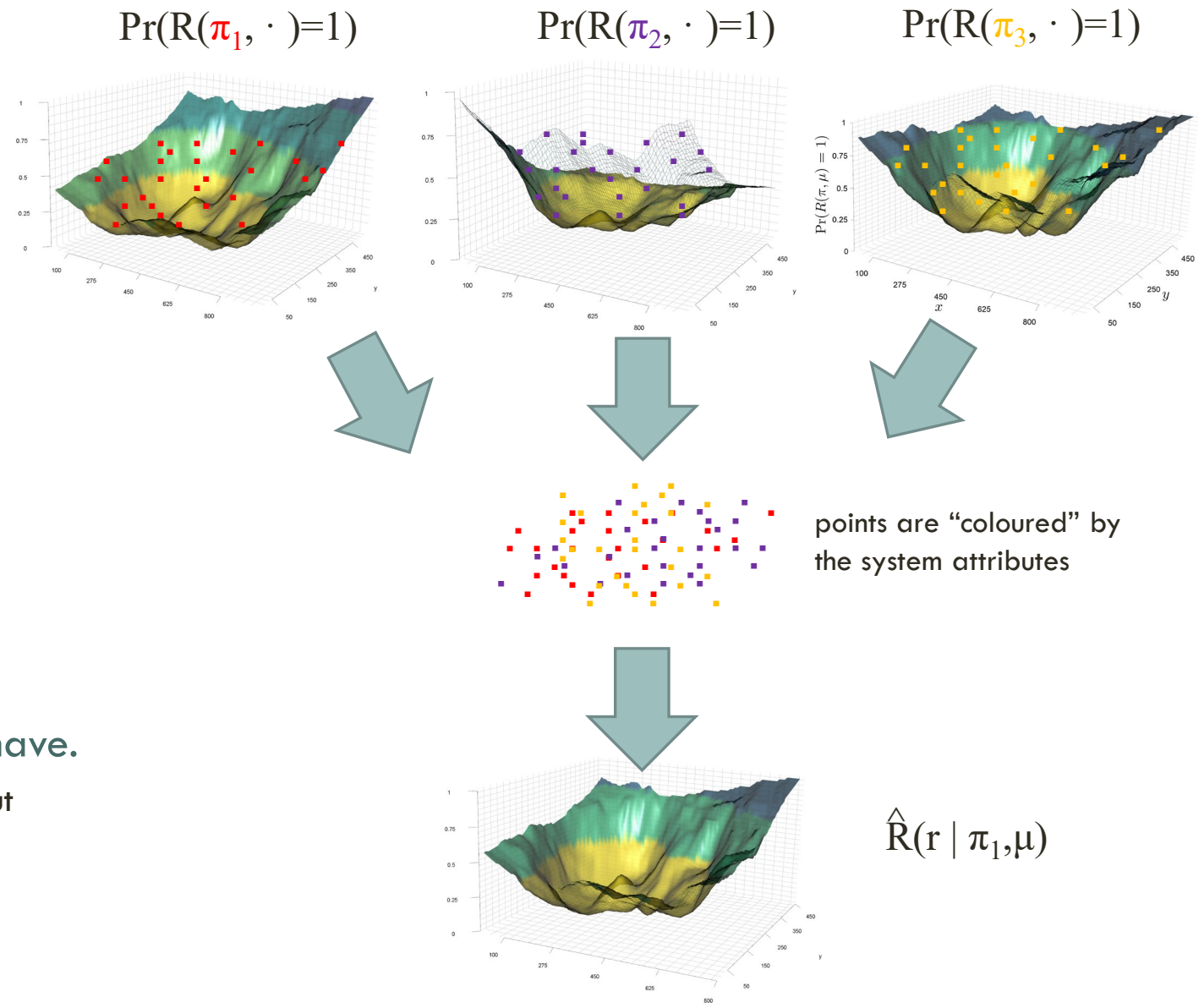
SYSTEM SPACE

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

- 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!



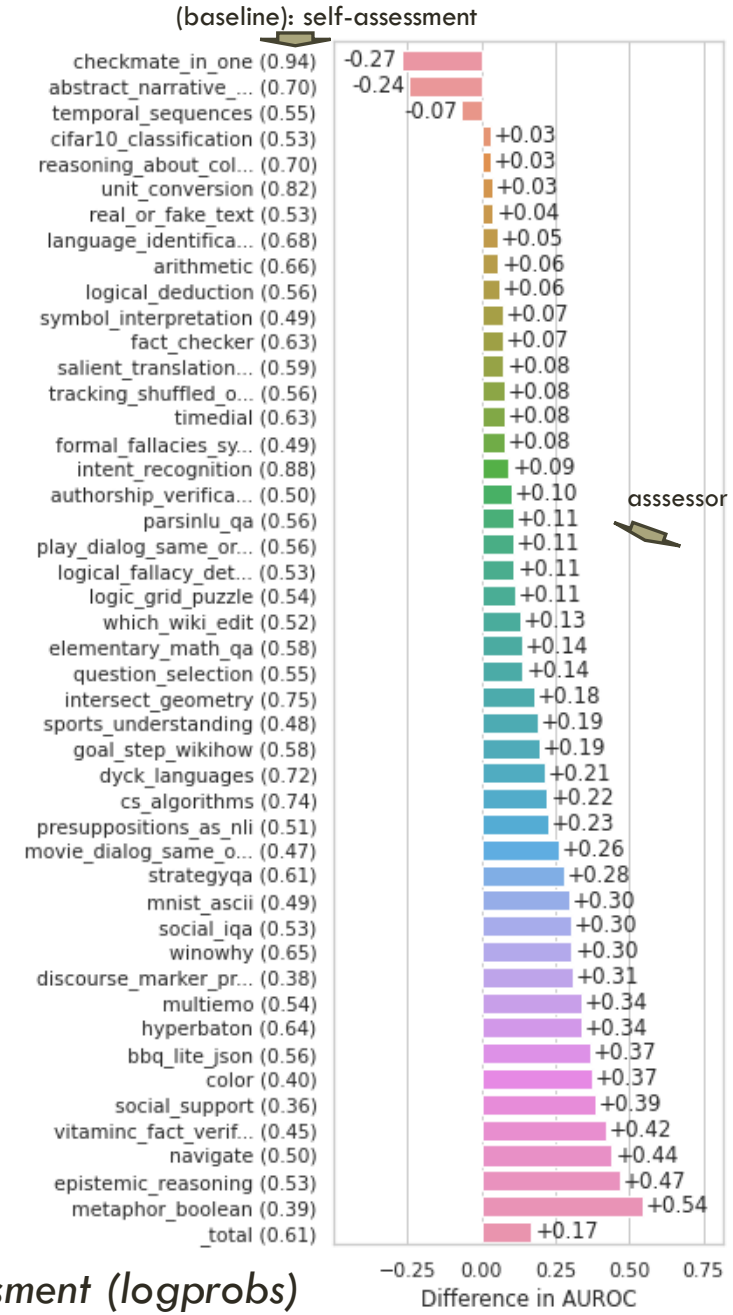
LMs PREDICT LMs

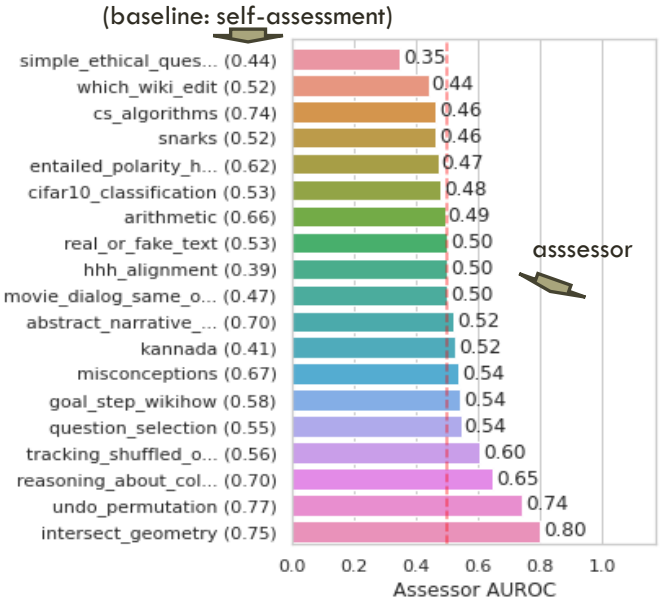
Setup:

- 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)

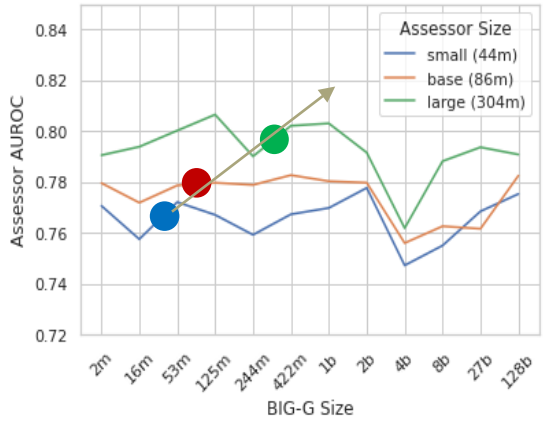
In distribution:

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





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



Bigger assessor = better
Bigger subject = neutral

Schellaert et al. "Validity Predictability Factors in Language Models" (forthcoming)

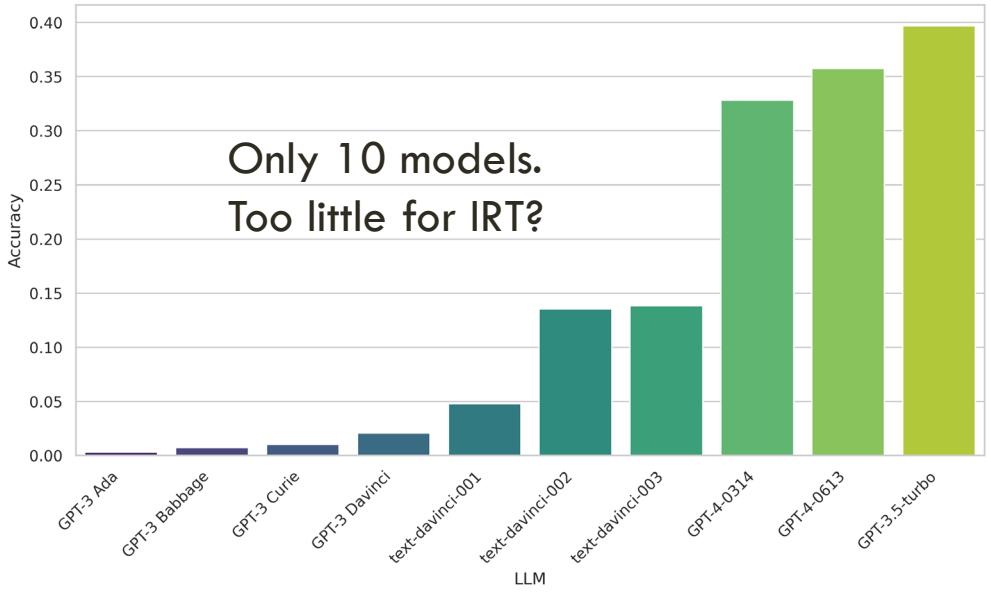
Measurement Layouts

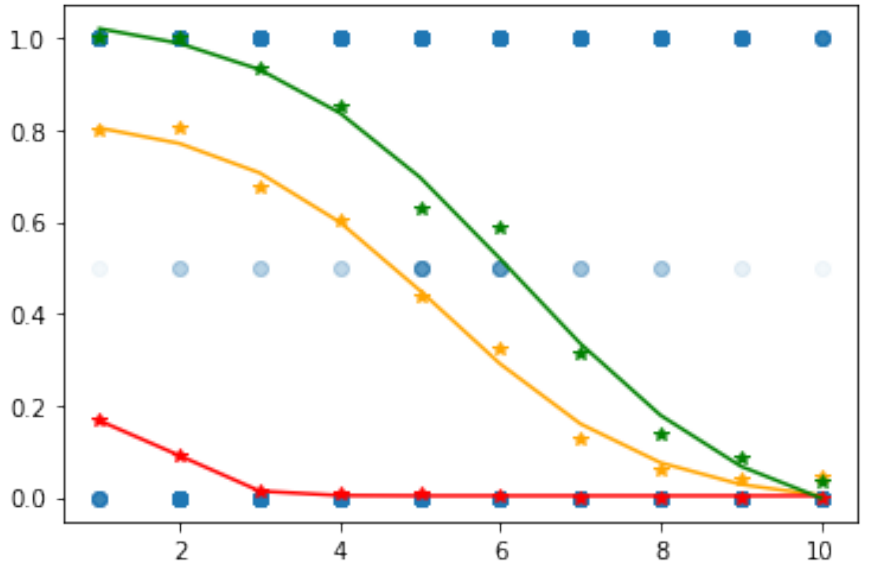
AAAI2024 Tutorial
"Measurement Layouts for Capability-Oriented AI Evaluation",
J. Burden, L. Cheke, J. Hernández-Orallo, M. Tešić, K. Voudouris
https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial

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:





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.

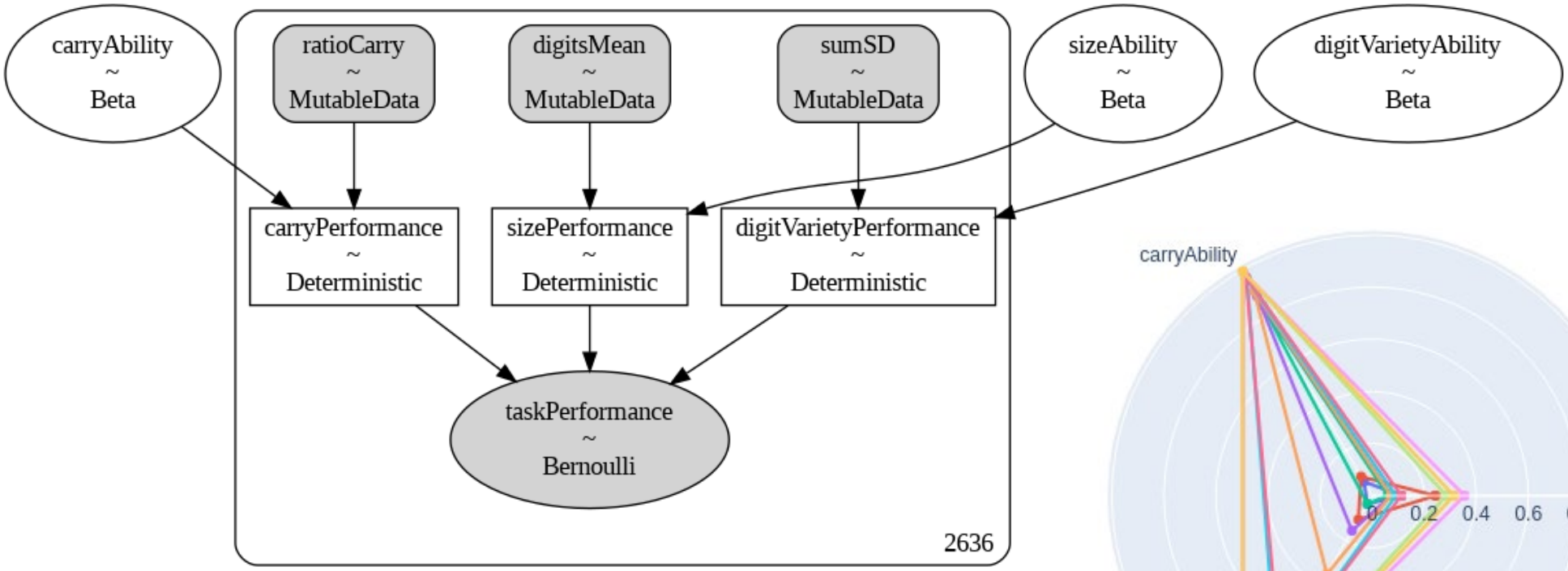
MORE SOPHISTICATED DEMANDS

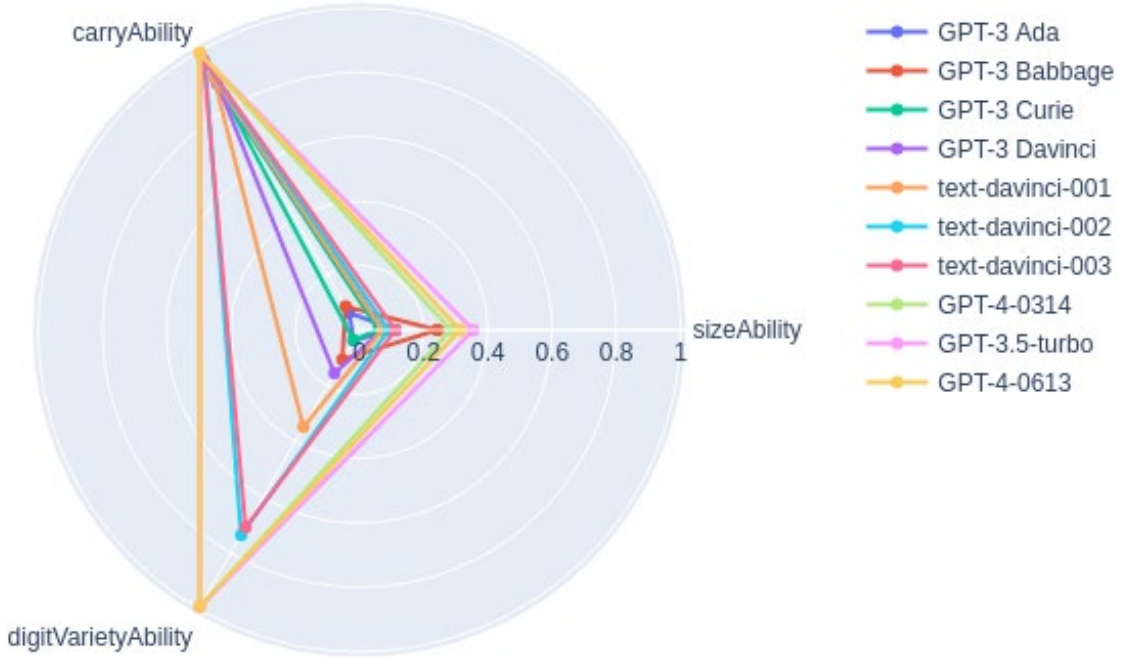
- `digits1`: The number of digits in the first summand.
- `digits2`: The number of digits in the second summand.
- `min_digits`: $min(digits_1, digits_2)$, i.e., the number of digits in the smaller summand.
- `harm_mean`: $2/(1/digits_1 + 1/digits_2)$, i.e., the harmonic mean of the number of digits 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'?

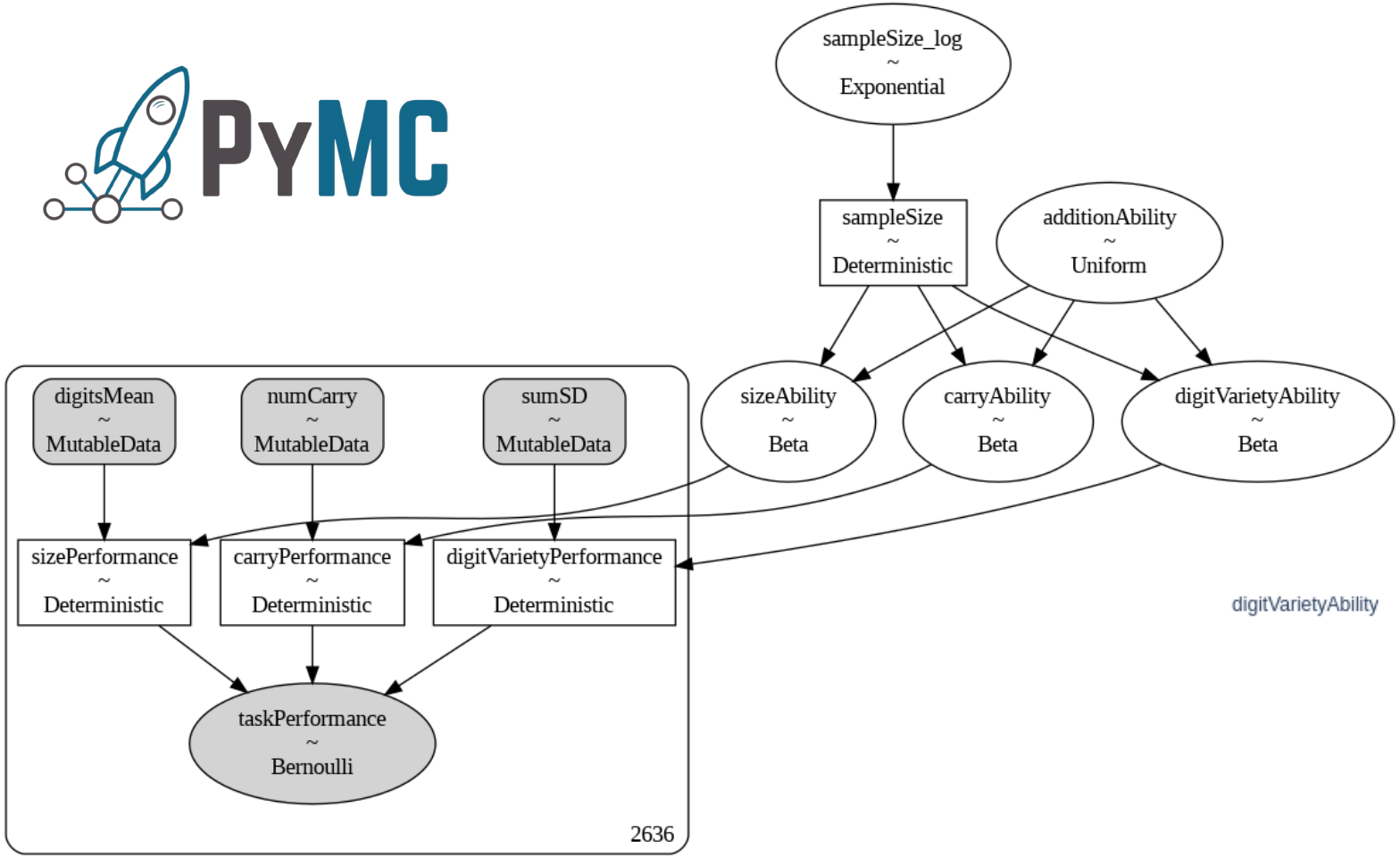
SIMPLE MEASUREMENT LAYOUT

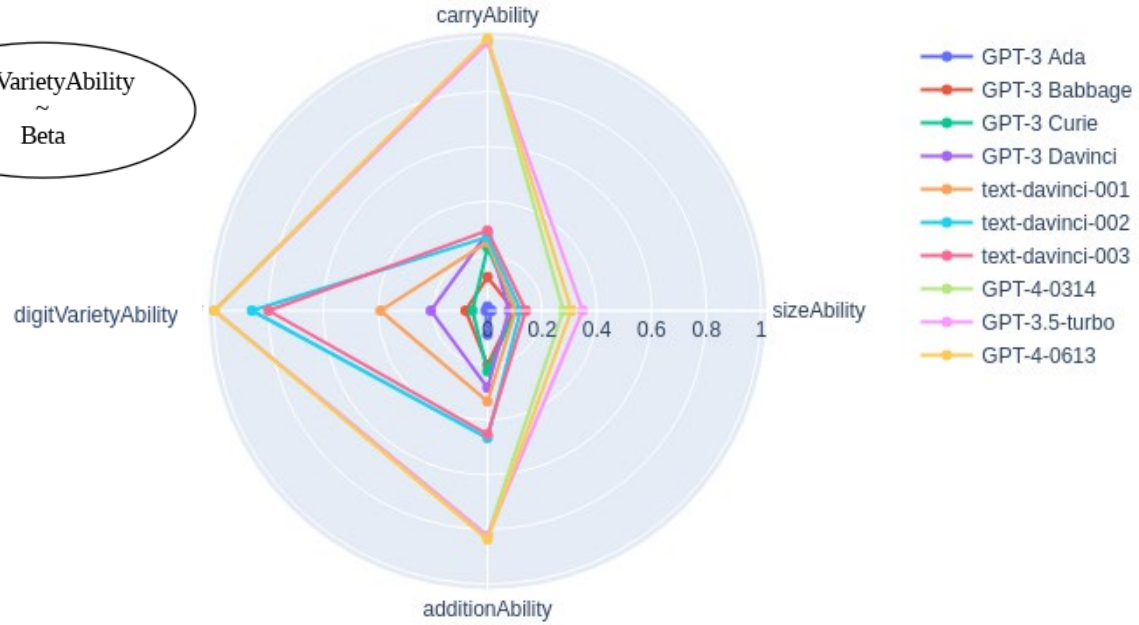




PyMC

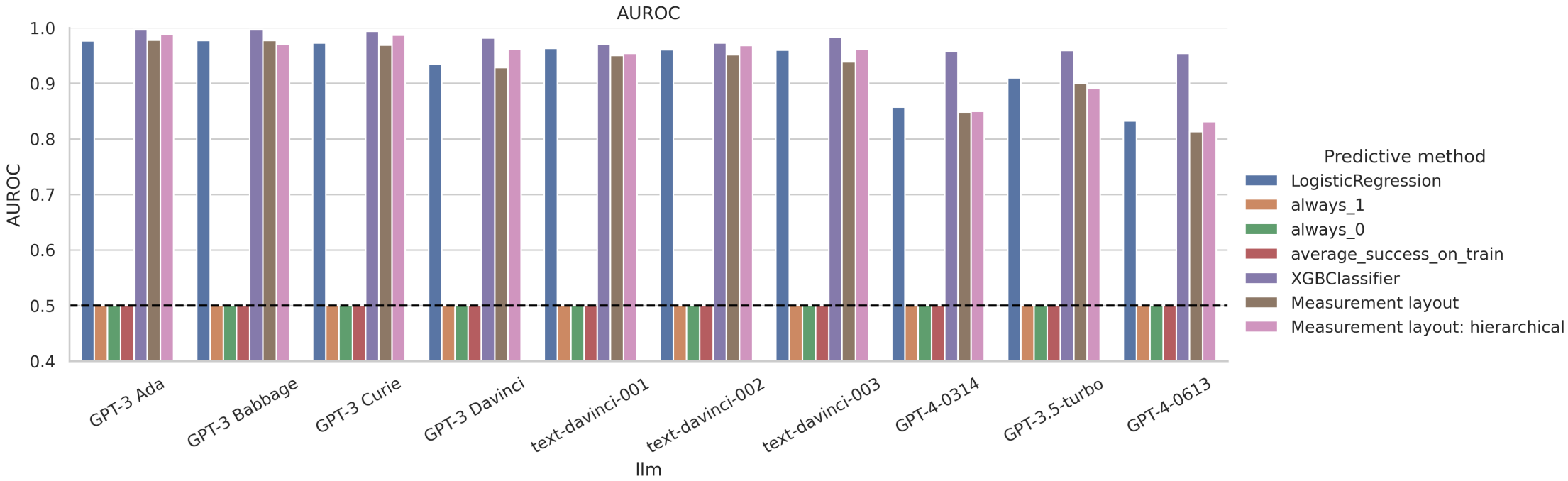
HIERARCHICAL MEASUREMENT LAYOUT





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

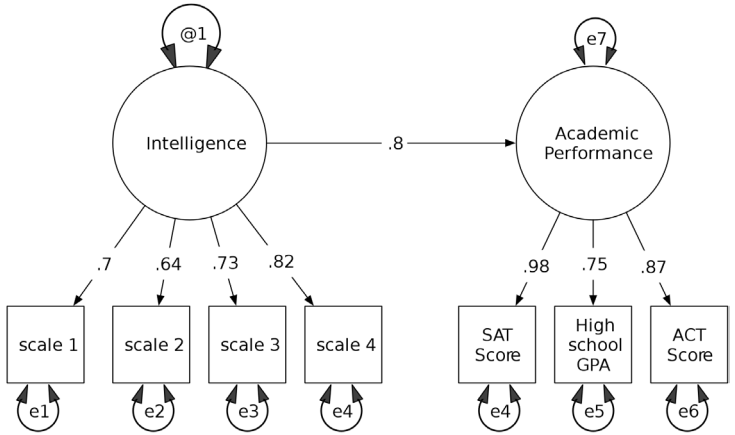
FACTOR ANALYSIS

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

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

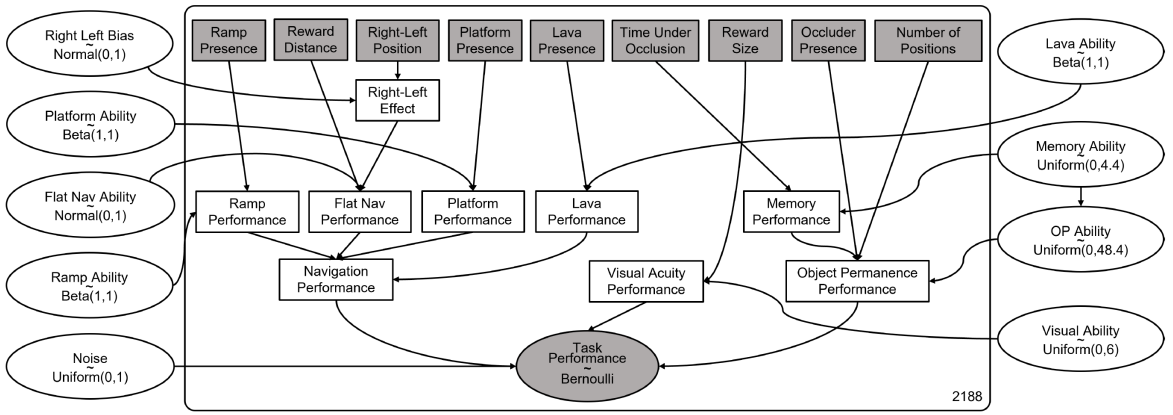
POPULATIONAL? INSTANCE-LEVEL?

- 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

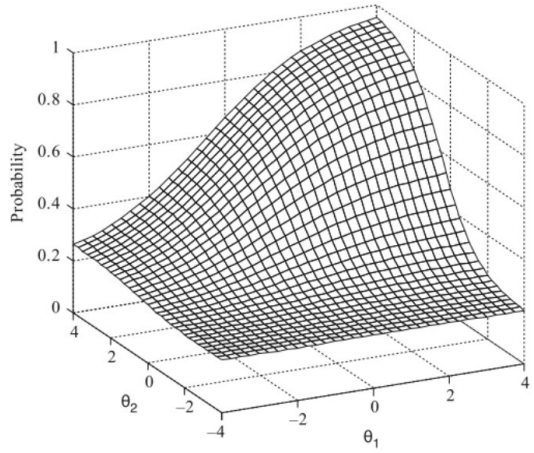
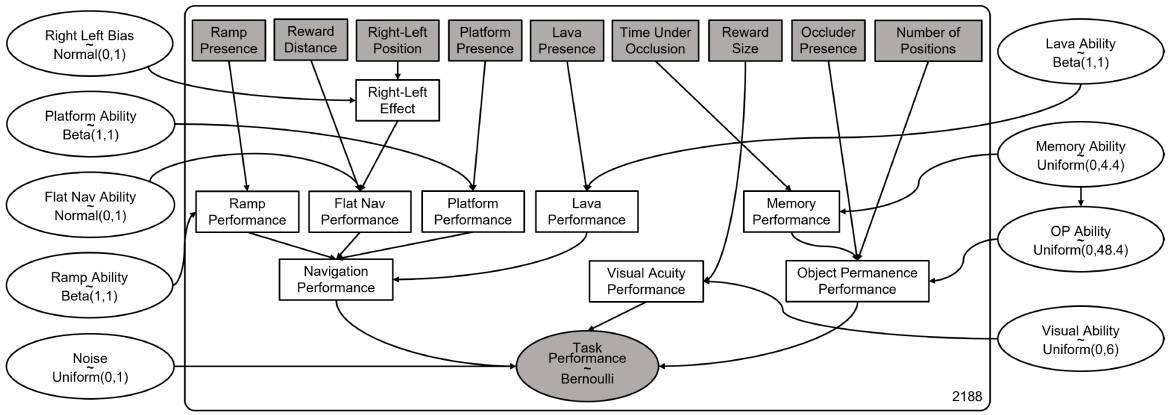


Fig. 4.9 Item response surface for the partially compensatory model when $a_1 = .7$, $a_2 = 1.1$, $b_1 = -.5$, $b_2 = .5$, and $c = .2$

"Multidimensional Item Response Theory" (V. Duran's slides)

- 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 (≥1 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 Burnell[1], Wout Schellaert[2], John Burden[1,3], Tomer D. Ullman[4], Fernando Martinez-Plumed[2], Joshua B. Tenenbaum[5], Danaja Rutar[1], Lucy G. Cheke[1,6], Jascha Sohl-Dickstein[7], Melanie Mitchell[8], Douwe Kiela[9], Murray Shanahan[10,11], Ellen M. Voorhees[12], Anthony G. Cohn[13,14,15,16], Joel Z. Leibo[10], 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 policymakers 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-

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

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

VRAIN

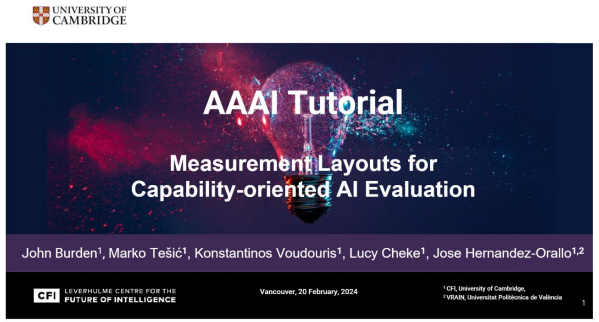
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

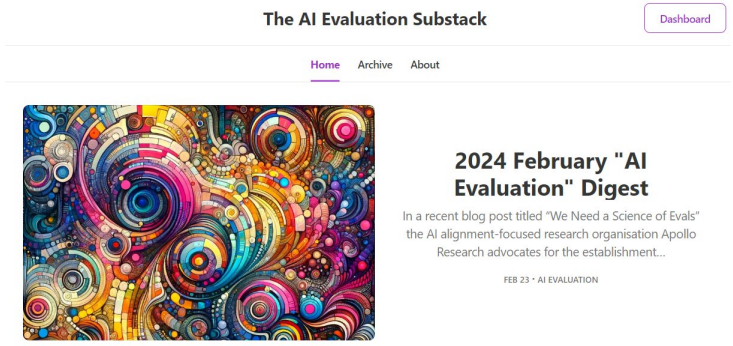
POINTERS

- 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/



Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

Conclusion, Recent Work, and Future Directions

Concluding Remarks and Summary

1. Learned about IRT models
2. How to implement IRT models and/or use py-irt
3. Showed ways to apply IRT to specific NLP problems
    3.1 Annotation Error
    3.2 Evaluation
    3.3 Training
4. Classical IRT is a starting point, but the range of IRT methods is much larger

Future Directions

1. Classical IRT is a starting point, but the range of IRT methods is much larger
2. Future Directions
    - 2.1 LLMs?
    - 2.2 Multidimensional IRT and Big Benchmarks?
    - 2.3 Predictability?

Google Group

Join our IRT in NLP Google group! TODO-link

Recent Work

Do great minds think alike? Investigating Human-AI Complementarity for Question Answering

- Skill/difficulty should be multidimensional, but making it work is difficult (Rodriguez et al., 2022)
- Idea: use BERT-informed embeddings to inform multidim difficulty, etc.
- Compare different proficiencies of humans versus models
- Gor et al. (2024) made it work!

Do great minds think alike?
Investigating Human-AI Complementarity for Question Answering

Maharshi Gor[1] **Tianyi Zhou**[1] **Hal Daumé III**[1,2] **Jordan Boyd-Graber**[1]

[1]University of Maryland [2]Microsoft Research
mgor@cs.umd.edu

Abstract

This study examines question-answering (QA) abilities across human and AI agents. Our framework CAIMIRA addresses limitations in traditional item response theory, by incorporating multidimensional analysis, identifiability, and content awareness, enabling nuanced comparison of QA agents. Analyzing responses from ~ 30 AI systems and 155 humans over thousands of questions, we identify distinct knowledge domains and reasoning skills where these agents demonstrate differential proficiencies. Humans outperform AI systems in scientific reasoning and understanding nuanced language, while large-scale LLMs like GPT-4 and LLAMA-2-70B excel in retrieving specific factual information. The study identifies key areas for future QA tasks and model development, emphasizing the importance of semantic understanding and scientific reasoning in creating more effective and discriminating benchmarks.

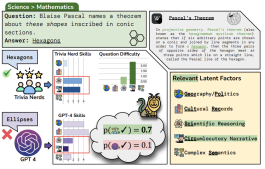


Figure 1: Response Correctness prediction using Agent skills and Question difficulty over relevant latent factors. We list the five latent factors that CAIMIRA discovers, and highlight the relevant ones (green), which contribute to estimating whether an agent will respond to the example question correctly. The agent skills over these relevant factors are highlighted in red boxes.

tion answering, particularly with the new panoply

Related Work

1. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information (Ethayarajh et al., 2022)
2. IRT in Recommender System Benchmarking (Liu et al., 2023)

References

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Maharshi Gor, Tianyi Zhou, III Daumé, Hal, and Jordan Boyd-Graber. 2024. Do great minds think alike? investigating human-ai complementarity for question answering.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2023. What we evaluate when we evaluate recommender systems: Understanding recommender systems' performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 658–670, New York, NY, USA. Association for Computing Machinery.

Pedro Rodriguez, Phu Mon Htut, John Lalor, and João Sedoc. 2022. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112, Dublin, Ireland. Association for Computational Linguistics.

Acknowledgements

Headshots of people we want to thank: Jordan, Hong, Phu, etc.