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!

Tutorial webpage: eacl2024irt.github.io

- Slides
- Jupyter notebooks
- Reading list

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: Evaluation 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/

Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

Part 1. Evaluation for NLP

João Sedoc[1]

[1] New York University

https://joaosedoc.com

What Do We Evaluate in NLP?

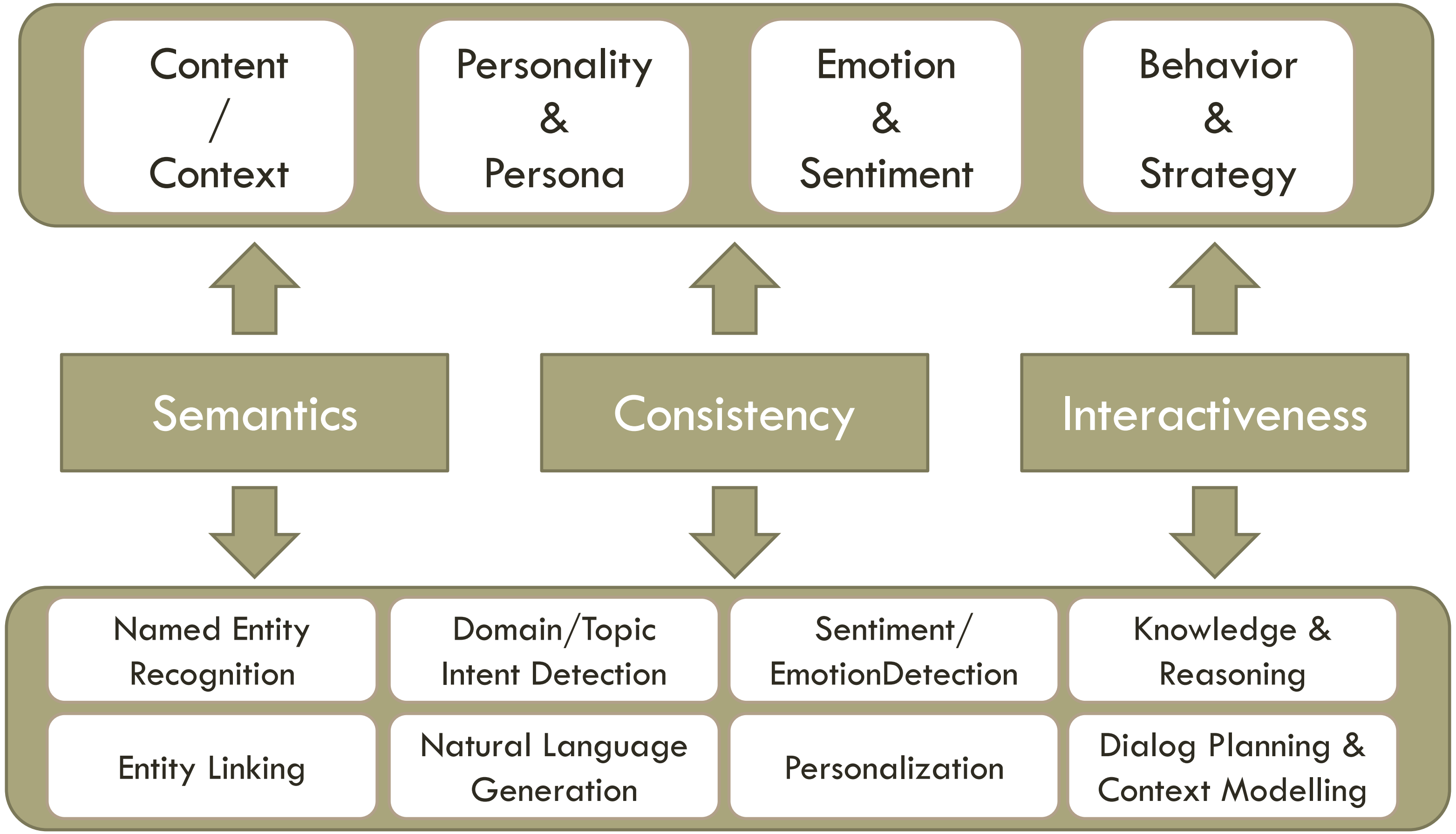
EVALUATIONS ARE AT SEVERAL LEVELS

1) System-level evaluations
   - This is probably the most common evaluation type (MT, Dialog, NLI, etc…)
2) Machine learning method evaluations
   - E.g., LSTM vs Transformer
3) Metrics
   - E.g., BLEU, BERTScore, etc
4) Annotations
   - Annotation error estimates
5) Data
   - Quality, domain similarity, toxicity

SYSTEM EVALUATIONS

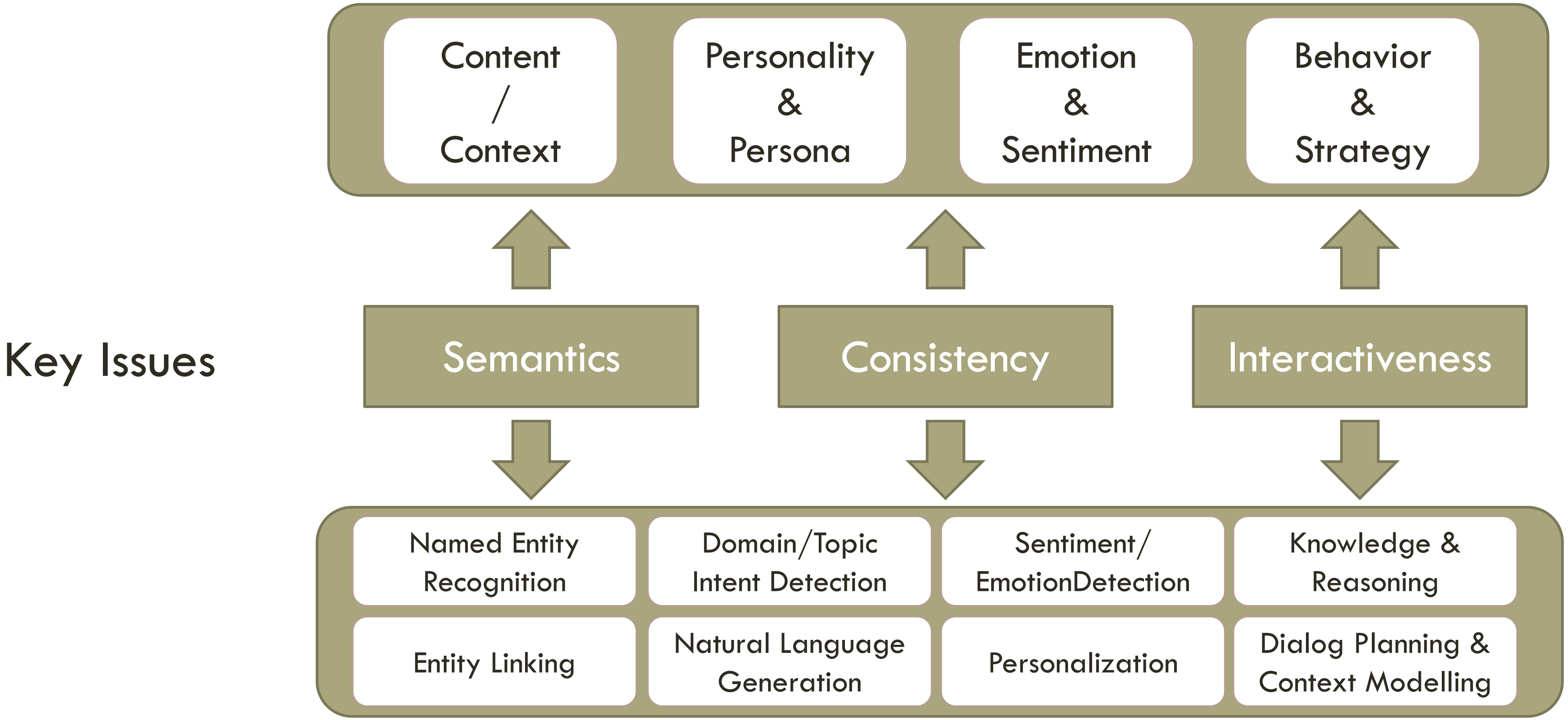
1. Extrinsic task based evaluation
2. Intrinsic evaluation
3. Human evaluation
4. Automatic metric evaluation
5. A/B testing
6. Error analysis

CHALLENGES FOR DIALOG SYSTEMS



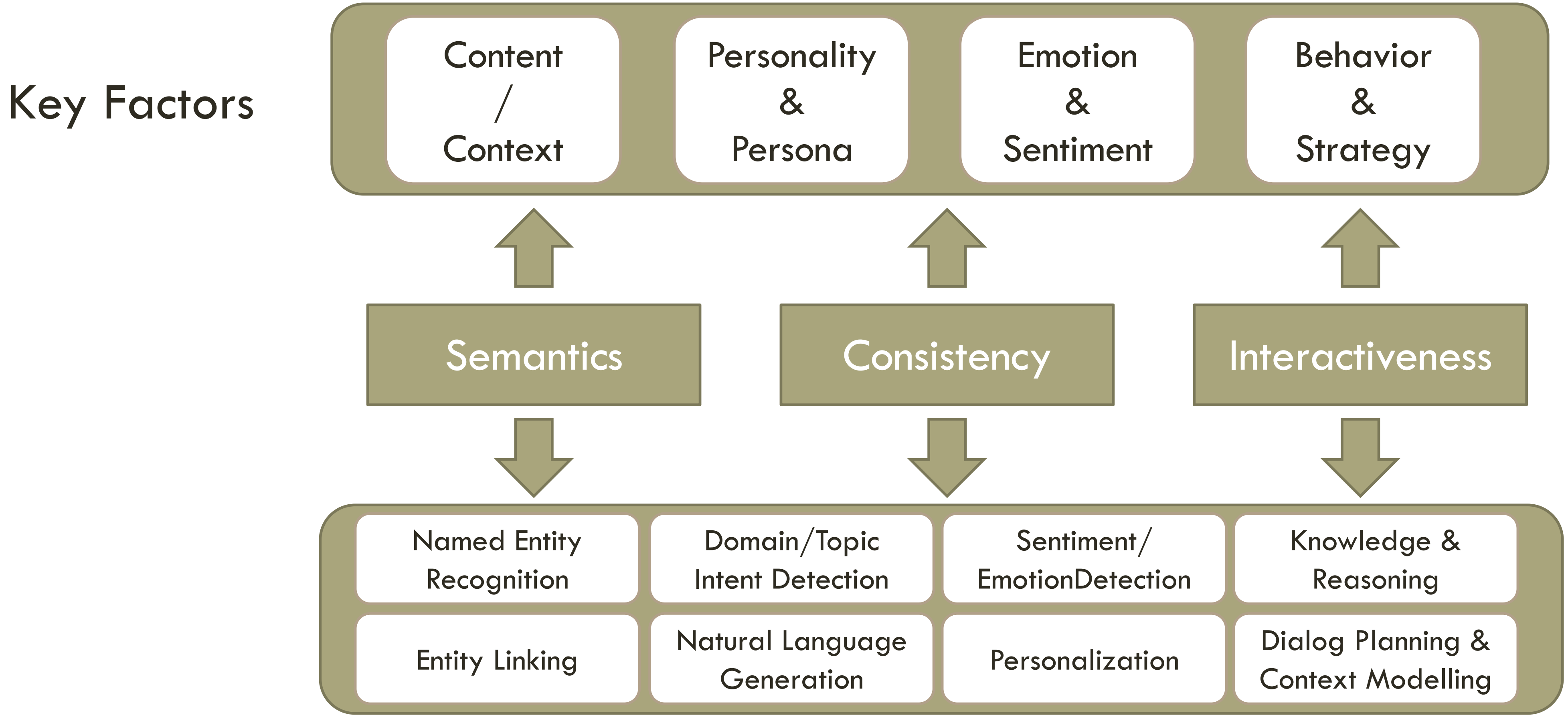
From Huang et al., 2019, “Challenges in Building Intelligent Open-Domain Systems”

CHALLENGES FOR DIALOG SYSTEMS



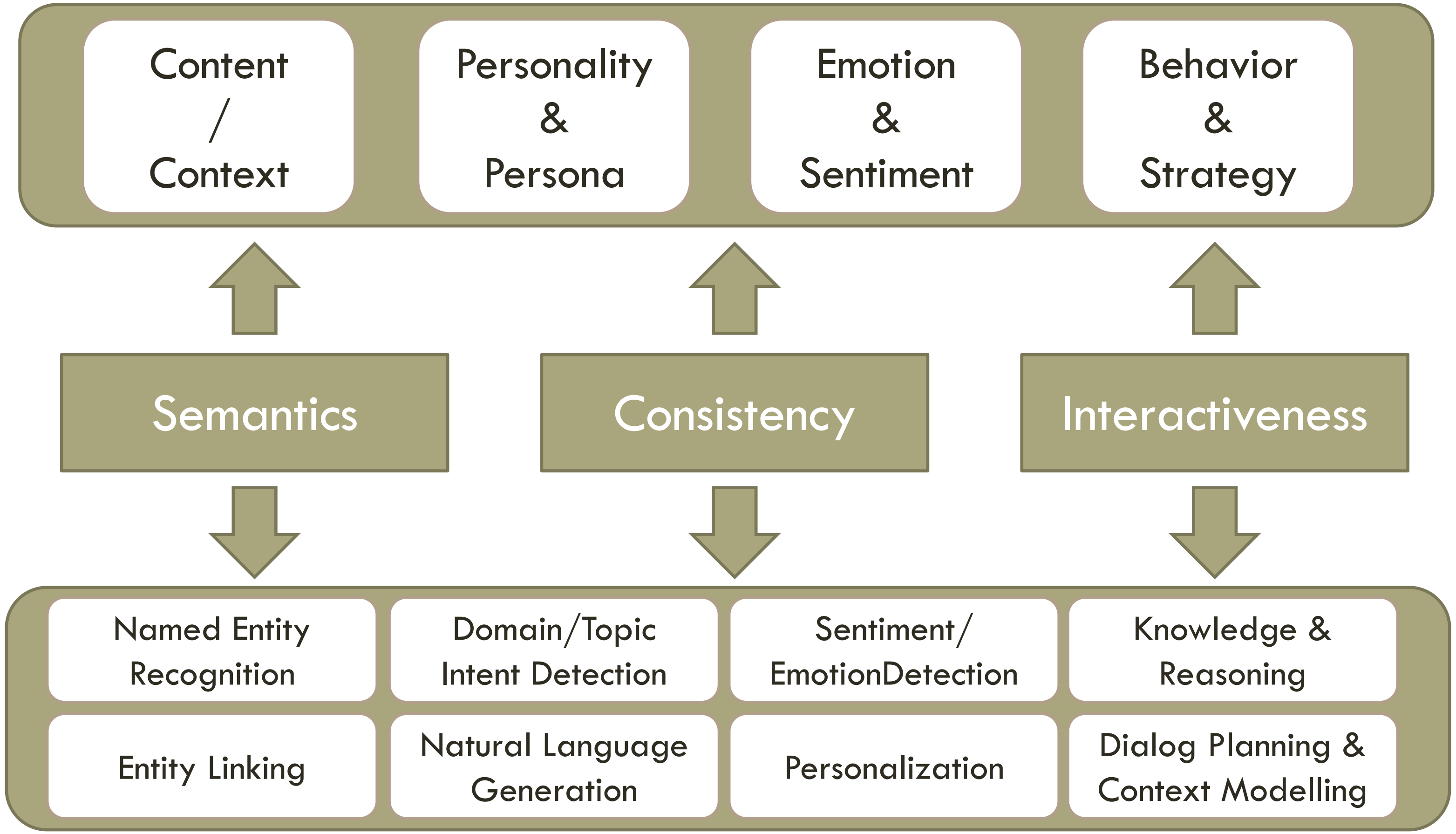
From Huang et al., 2019, “Challenges in Building Intelligent Open-Domain Systems”

CHALLENGES FOR DIALOG SYSTEMS



From Huang et al., 2019, "Challenges in Building Intelligent Open-Domain Systems"

CHALLENGES FOR DIALOG SYSTEMS



Key Technologies

From Huang et al., 2019, "Challenges in Building Intelligent Open-Domain Systems"

COMMON TASK FRAMEWORK & LEADERBOARDS

There is general agreement that these competitive evaluations had a striking and beneficial effect on the performance of various systems tested over the years. However, it is also recognized (albeit less generally) that these evaluation experiments also had the, less beneficial, effect that the participating systems focused increasingly more narrowly on those few parameters that were measured in the evaluation, to the detriment of more general properties.

- Schwitter et al. 2000

Focusing on headline state-of-the-art numbers "provide(s) limited value for scientific progress absent insight into what drives them" and where they fail.

- Lipton and Steinhardt, 2019

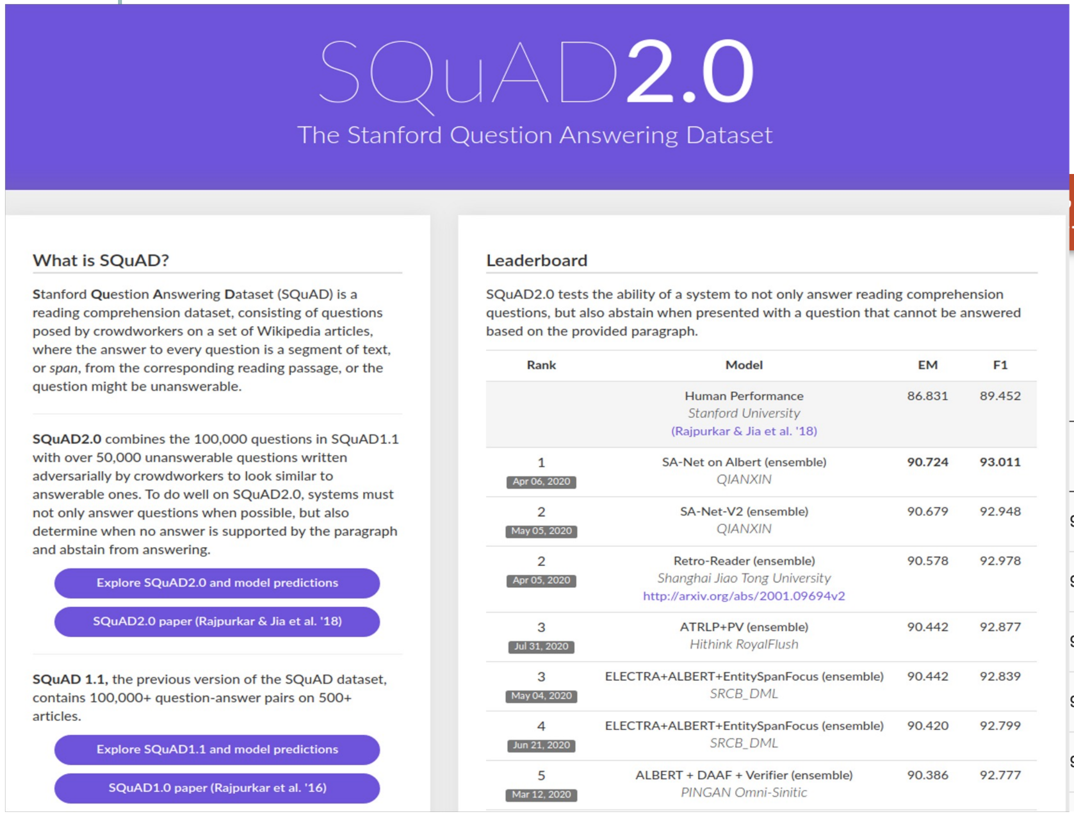
LOTS OF LEADERBOARDS

SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

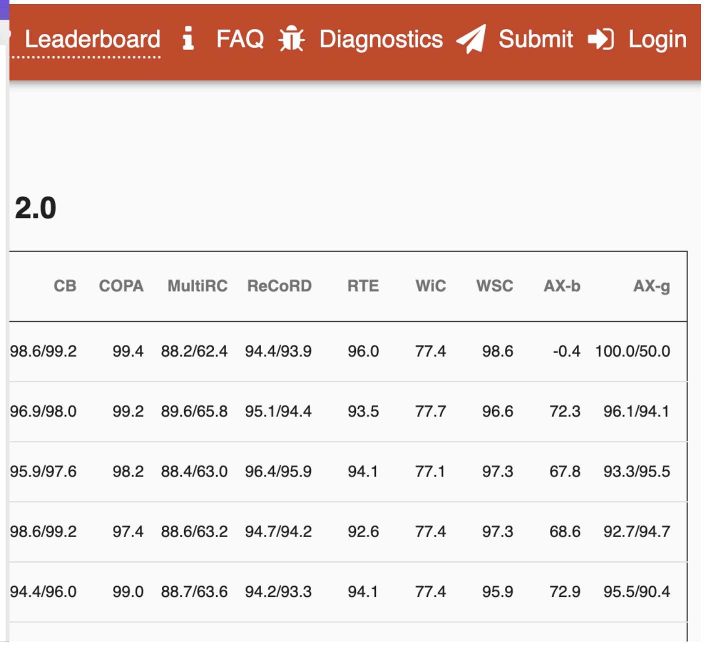
Leaderboard Version: **2.0**

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	JDExplore d-team	Vega v2		91.3	90.5	98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
2	Liam Fedus	ST-MoE-32B		91.2	92.4	96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
3	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
4	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
5	Yi Tay	PaLM 540B		90.4	91.9	94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4

LOTS OF LEADERBOARDS



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694v2	90.578	92.978
3 Jul 31, 2020	ATRLP+PV (ensemble) *Hithink RoyalFlush*	90.442	92.877
3 May 04, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.442	92.839
4 Jun 21, 2020	ELECTRA+ALBERT+EntitySpanFocus (ensemble) *SRCB_DML*	90.420	92.799
5 Mar 12, 2020	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.386	92.777



CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
98.6/99.2	99.4	88.2/62.4	94.4/93.9	96.0	77.4	98.6	-0.4	100.0/50.0
96.9/98.0	99.2	89.6/65.8	95.1/94.4	93.5	77.7	96.6	72.3	96.1/94.1
95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
94.4/96.0	99.0	88.7/63.6	94.2/93.3	94.1	77.4	95.9	72.9	95.5/90.4

LOTS OF LEADERBOARDS

SQu

The Stanford Qu

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Spaces: mteb/**leaderboard** like 2 Running on CPU UPGRADE

App Files and versions Community 2 Linked Models Linked Datasets

Massive Text Embedding Benchmark (MTEB) Leaderboard. To submit, refer to the MTEB GitHub repository 🤗

- **Total Datasets:** 56
- **Total Languages:** 112
- **Total Scores:** >2380
- **Total Models:** 34

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Retrieval | Reranking | STS | Summarization

Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English, refer to task tabs for others

Rank	Model	Embedding Dimensions	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)	STS Average (10 datasets)
1	sentence-t5-xxl	768	59.51	73.42	43.72	85.06	56.42	42.24	82.63
2	gtr-t5-xxl	768	58.97	67.41	42.42	86.12	56.66	48.48	78.38
3	SGPT-5.8B-weightedmean-msmarco-specb-bitfit	4096	58.81	68.13	40.34	82	56.56	50.25	78.1

LOTS OF LEADERBOARDS

SQ

The Stanfo

What is SQuAD?

Stanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000+ question-answer pairs on 500+ articles.

Explore SQuAD1.1 and model predictions

SQuAD1.0 paper (Rajpurkar et al. '16)

Spaces: mteb/**leaderboard** like 2 Running on CPU UPGRADE

App | Files and versions | Community 2 | Linked Models | Linked Datasets

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **400,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **73**. Total #votes: **408144**. Last updated: March 13, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-1106-preview	1251	+5/-4	48226	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1249	+5/-6	22282	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1247	+6/-6	14854	Anthropic	Proprietary	2023/8
4	Bard (Gemini Pro)	1202	+6/-7	12623	Google	Proprietary	Online
4	Claude 3 Sonnet	1190	+6/-6	14845	Anthropic	Proprietary	2023/8
5	GPT-4-0314	1185	+4/-6	27245	OpenAI	Proprietary	2021/9
7	GPT-4-0613	1159	+4/-5	43783	OpenAI	Proprietary	2021/9

SHARED TASKS

English→Czech

Range	Ave.	Ave. z	System
1	91.2	0.335	HUMAN–C
2	90.9	0.279	Online-W
3	88.6	0.158	JDExploreAcad.
4-6	85.3	0.045	Online-B
4-6	87.1	0.041	Lan-Bridge
4-6	85.1	0.029	HUMAN-B
7-10	84.2	−0.059	CUNI-Bergamot
7-10	83.7	−0.074	CUNI-DocTransf.
7-10	84.0	−0.087	Online-A
7-10	83.2	−0.128	CUNI-Transf.
11-12	83.3	−0.258	Online-G
11-12	80.8	−0.310	Online-Y

SHARED TASKS

English→Czech

Range	Ave.	Ave. z	System
1	91.2	0.335	HUMAN–C
2	90.9		
3	88.6		
4-6	85.3		
4-6	87.1		
4-6	85.1		
7-10	84.2	–	
7-10	83.7	–	
7-10	84.0	–	
7-10	83.2	–	
11-12	83.3	–	
11-12	80.8	–	

CodaLab   Search Competitions   My Competitions   Help   Sign

Max submissions total: 999

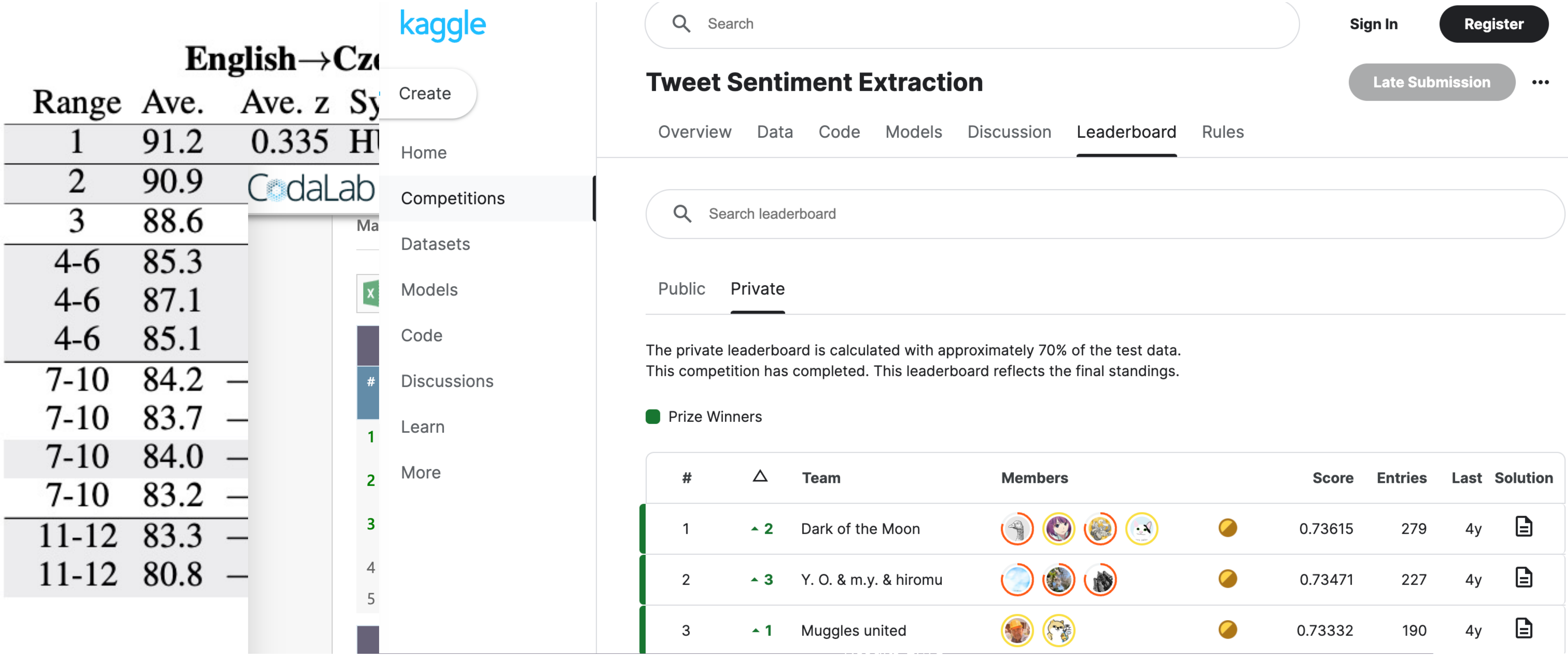
Download CSV

Results EMP

#	User	Entries	Date of Last Entry	Team Name	Averaged Pearson Correlations ▲	Empathy Pearson Correlation ▲	Distress Pearson Correlation ▲
1	jaymundra	18	02/18/21	IITK@WASSA	0.533 (3)	0.558 (1)	0.507 (3)
2	justglowing	12	02/13/21	CompNA	0.554 (2)	0.554 (2)	0.554 (2)
3	atharvakulkarni	4	02/16/21	PVG@WASSA2021	0.557 (1)	0.517 (3)	0.597 (1)
4	vinid	8	02/17/21	MilaNLP	- (4)	- (4)	- (4)
5	kanishksin	21	02/22/21	Phoenix	- (4)	- (4)	- (4)

Results EMO

SHARED TASKS



LEADERBOARDS CAN IMPROVE

1. Questions with the Right Difficulty
2. Discriminative Questions
3. Minimize Ambiguity, Maximize Fairness
4. Don’t be Overly Definitive
5. Be Flexible and Introspective

METHODS FOR RANKING

1. Average score
2. Z-scored ratings
3. Preference ranking
   - Bradley-Terry-Leech
   - Elo rating system
   - Trueskill
   - Item Response Theory

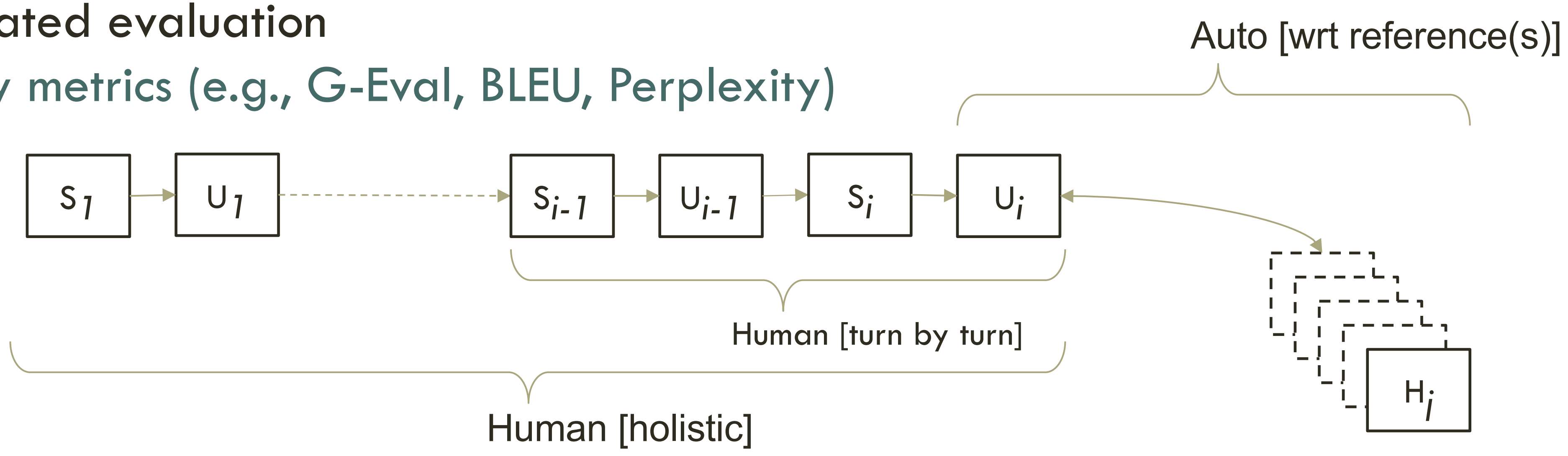
HUMAN / AUTOMATIC METRICS

Human evaluation

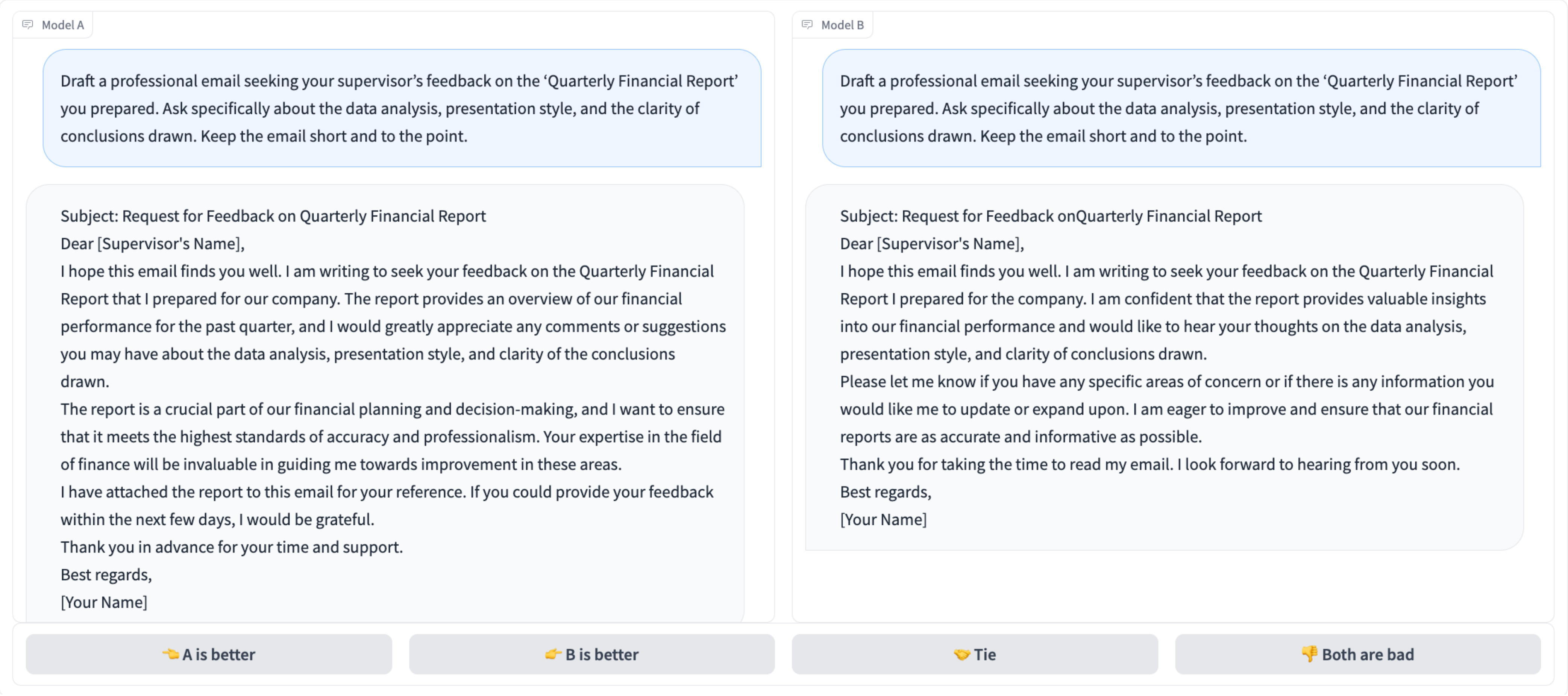
- Expert judges (WOCHAT, Alexa)
- Crowd-sourced (non-expert) judgments (DBDC)

Automated evaluation

- Proxy metrics (e.g., G-Eval, BLEU, Perplexity)



A / B TESTING


Model A
Draft a professional email seeking your supervisor’s feedback on the ‘Quarterly Financial Report’ you prepared. Ask specifically about the data analysis, presentation style, and the clarity of conclusions drawn. Keep the email short and to the point.
Subject: Request for Feedback on Quarterly Financial Report
Dear [Supervisor's Name],
I hope this email finds you well. I am writing to seek your feedback on the Quarterly Financial Report that I prepared for our company. The report provides an overview of our financial performance for the past quarter, and I would greatly appreciate any comments or suggestions you may have about the data analysis, presentation style, and clarity of the conclusions drawn.
The report is a crucial part of our financial planning and decision-making, and I want to ensure that it meets the highest standards of accuracy and professionalism. Your expertise in the field of finance will be invaluable in guiding me towards improvement in these areas.
I have attached the report to this email for your reference. If you could provide your feedback within the next few days, I would be grateful.
Thank you in advance for your time and support.
Best regards,
[Your Name]
Model B
Draft a professional email seeking your supervisor’s feedback on the ‘Quarterly Financial Report’ you prepared. Ask specifically about the data analysis, presentation style, and the clarity of conclusions drawn. Keep the email short and to the point.
Subject: Request for Feedback onQuarterly Financial Report
Dear [Supervisor's Name],
I hope this email finds you well. I am writing to seek your feedback on the Quarterly Financial Report I prepared for the company. I am confident that the report provides valuable insights into our financial performance and would like to hear your thoughts on the data analysis, presentation style, and clarity of conclusions drawn.
Please let me know if you have any specific areas of concern or if there is any information you would like me to update or expand upon. I am eager to improve and ensure that our financial reports are as accurate and informative as possible.
Thank you for taking the time to read my email. I look forward to hearing from you soon.
Best regards,
[Your Name]
👈 A is better
👉 B is better
🤝 Tie
👎 Both are bad

ERROR ANALYSIS

1. Categorize error types
2. Investigate sources
3. Identify possible explanations

Annotations

EVALUATION OF ANNOTATIONS

1. Inter-annotator agreement (IAA)
   - Cohen’s Kappa
   - Krippendorff’s alpha
   - Fleiss’ Kappa

2. Accuracy, Precision/Recall/F-score
3. Consistency checks
4. Error Analysis

Data

UNDERLYING DATA ANALYSIS

1. Quality of the examples
2. Difficulty of data
3. Usefulness for evaluation
4. Error Analysis

THANK YOU!

JOAO SEDOC
http://joaosedoc.com/
jsedoc@nyu.edu

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

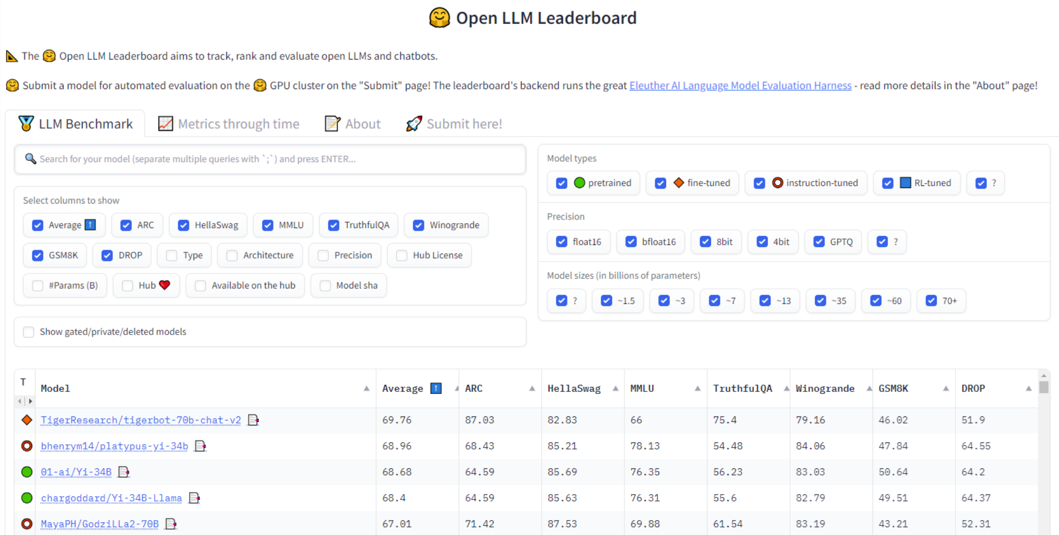
Natural language inference (NLI)

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



🤗 Open LLM Leaderboard

📐 The 🤗 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤗 Submit a model for automated evaluation on the 🤗 GPU cluster on the "Submit" page! The leaderboard's backend runs the great Eleuther AI Language Model Evaluation Harness - read more details in the "About" page!

🏅 LLM Benchmark | 📈 Metrics through time | 📝 About | 🚀 Submit here!

🔍 Search for your model (separate multiple queries with `;`) and press ENTER...

Select columns to show

☑ Average ⬆️ ☑ ARC ☑ HellaSwag ☑ MMLU ☑ TruthfulQA ☑ Winogrande ☑ GSM8K ☑ DROP ☐ Type ☐ Architecture ☐ Precision ☐ Hub License ☐ #Params (B) ☐ Hub ❤️ ☐ Available on the hub ☐ Model sha

☐ Show gated/private/deleted models

Model types

☑ 🟢 pretrained ☑ 🔶 fine-tuned ☑ ⭕ instruction-tuned ☑ 🟦 RL-tuned ☑ ?

Precision

☑ float16 ☑ bfloat16 ☑ 8bit ☑ 4bit ☑ GPTQ ☑ ?

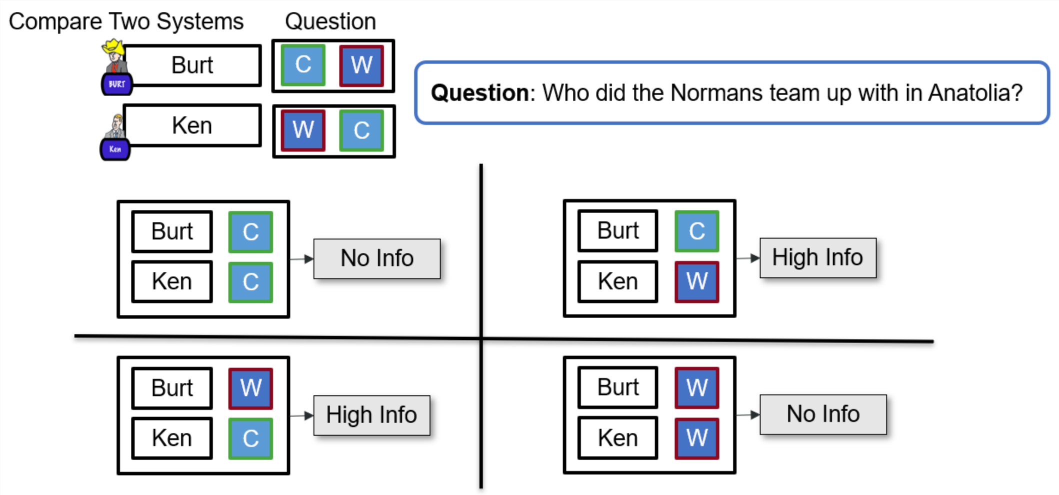
Model sizes (in billions of parameters)

☑ ? ☑ ~1.5 ☑ ~3 ☑ ~7 ☑ ~13 ☑ ~35 ☑ ~60 ☑ 70+

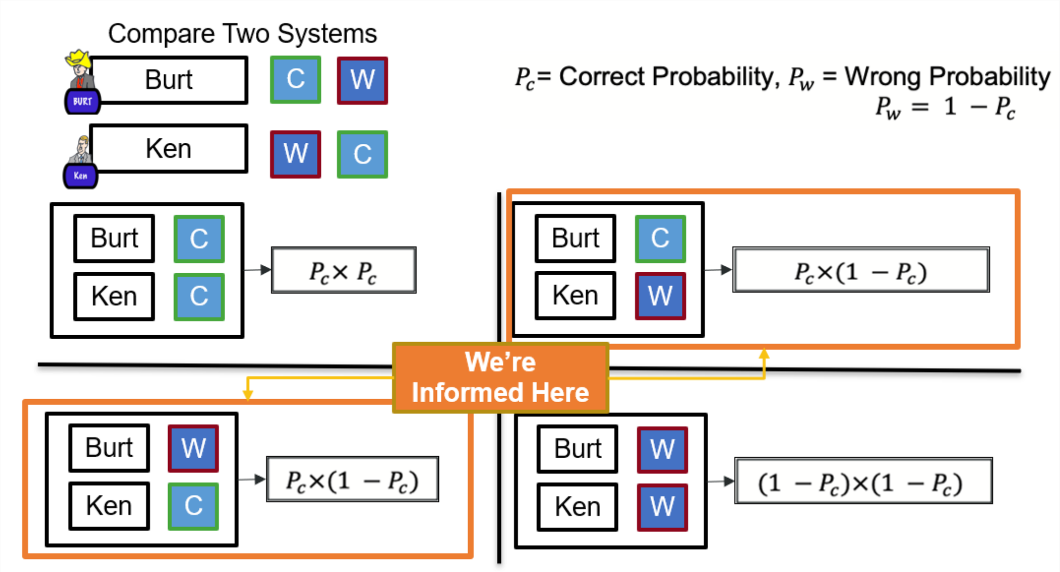
T	Model	Average ⬆️	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	DROP
🔶	TigerResearch/tigerbot-70b-chat-v2	69.76	87.03	82.83	66	75.4	79.16	46.02	51.9
⭕	bhenrym14/platypus-yi-34b	68.96	68.43	85.21	78.13	54.48	84.06	47.84	64.55
🟢	01-ai/Yi-34B	68.68	64.59	85.69	76.35	56.23	83.03	50.64	64.2
🟢	chargoddard/Yi-34B-Llama	68.4	64.59	85.63	76.31	55.6	82.79	49.51	64.37
⭕	MayaPH/GodziLLa2-70B	67.01	71.42	87.53	69.88	61.54	83.19	43.21	52.31

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

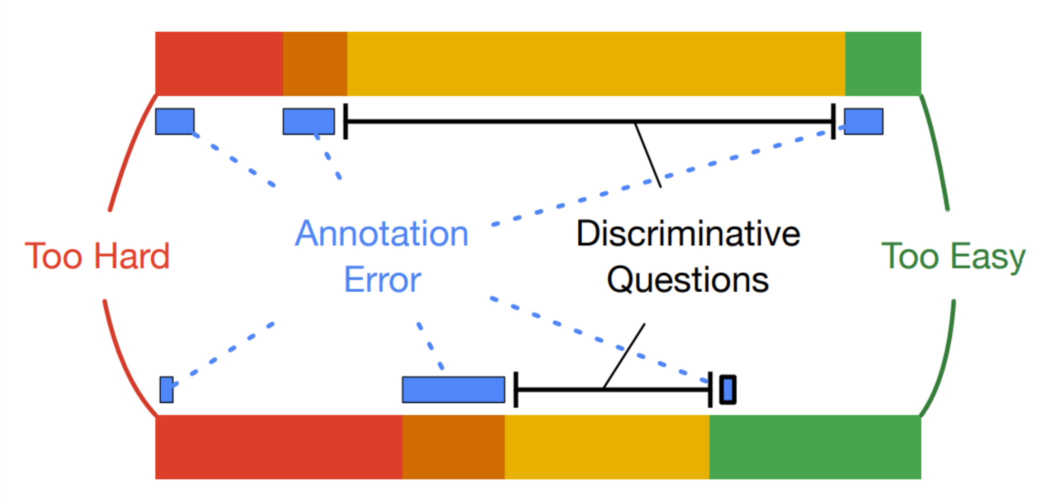
Differences in Questions



Differences in Questions



Differences in Questions



Source: Boyd-Graber and Börschinger (2020)

Introducing IRT

Psychometrics

Psychometrics: study of quantitative measurement practices

- Building instruments for measurement (standardized tests)
- Development of theoretical approaches to measurement

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

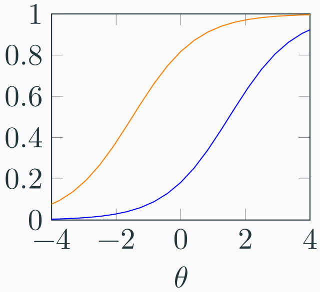
IRT: 1 Parameter Logistic Model (1PL)

Also known as *Rasch model*

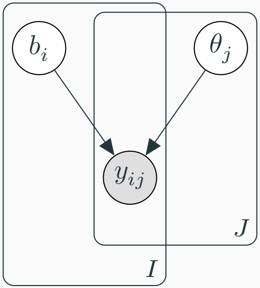
$$p(y_{ij} = 1 | b_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$$

θ_j: latent ability

b_i: difficulty



1PL Plate Notation



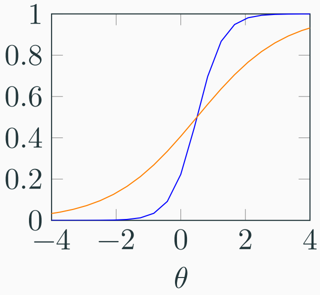
IRT: Other Examples (2PL)

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

θ_j: latent ability

b_i: difficulty

a_i: discriminability



IRT: Other Examples (3PL)

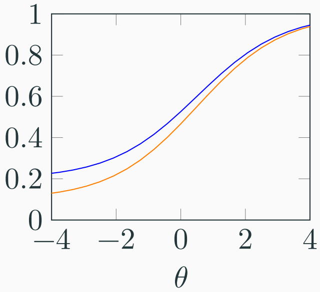
$$p(y_{ij} = 1 | a_i, b_i, c_i, \theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability

b_i: difficulty

a_i: discriminability

c_i: guessing



IRT: Other Examples (Feasibility)

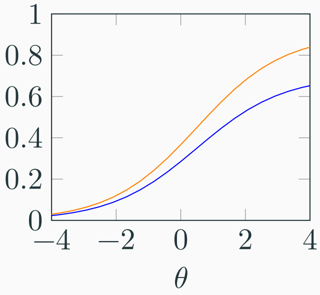
$$p(y_{ij} = 1 | a_i, b_i, c_i, \theta_j) = \frac{\gamma_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability

b_i: difficulty

a_i: discriminability

γ_i: feasibility



Parameter Estimation

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

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

- $p(Y | B, \Theta)$ – model
- $q(\Theta, B)$ – guide (variational distribution)

Natesan et al. (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%

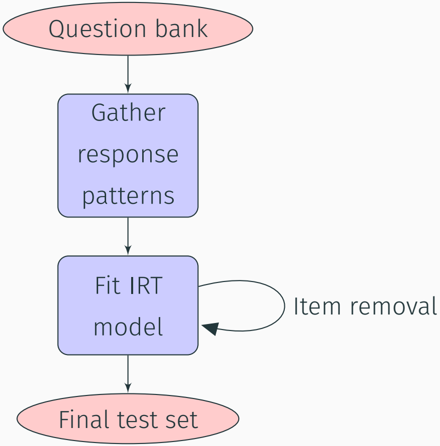
Source: Lalor et al. (2016)

Human Bottleneck

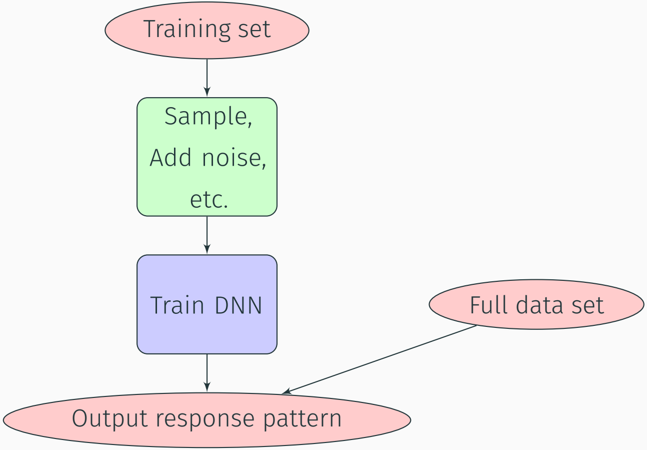
- Gathering human response patterns is expensive
- Can we use ensembles of models to gather response patterns?
- Even if we can, should we?

IRT Models with Artificial Crowds

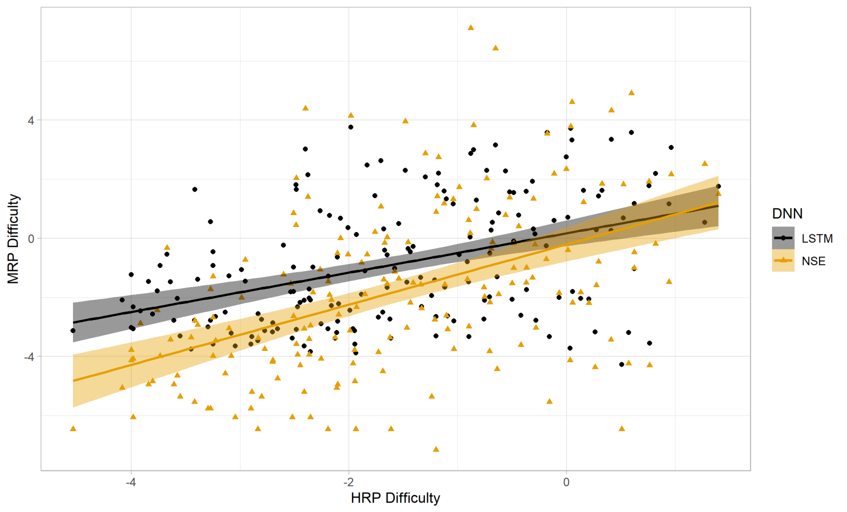
Building IRT Test Sets



Artificial Crowd Construction

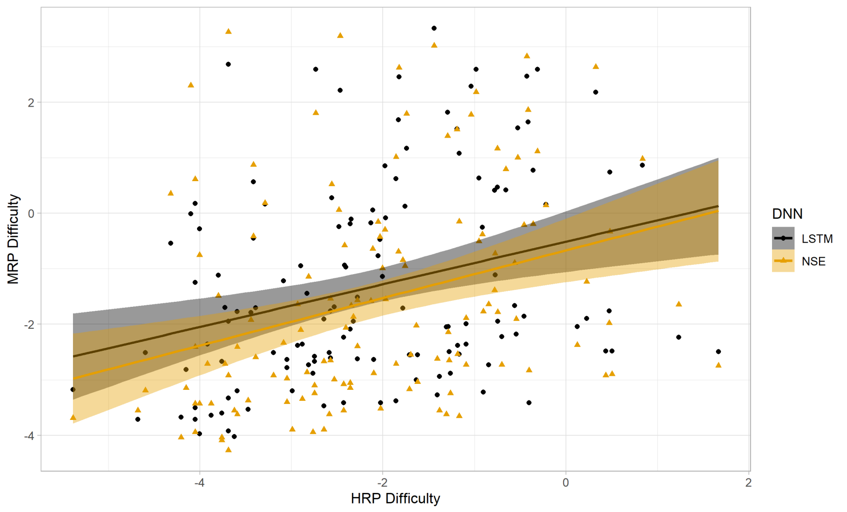


Human-Machine Correlation



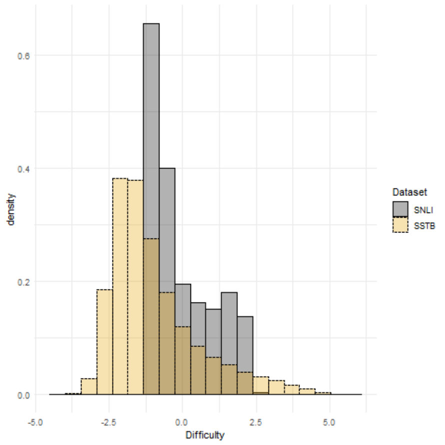
- Spearman ρ (NLI): 0.409 (LSTM) and 0.496 (NSE) (Lalor et al., 2019)

Human-Machine Correlation



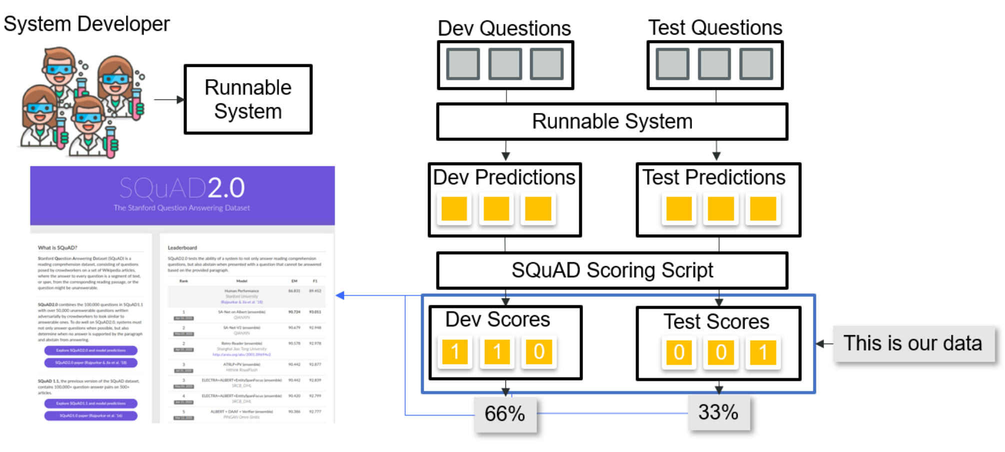
- Spearman ρ (SA): 0.332 (LSTM) and 0.392 (NSE) (Lalor et al., 2019)

Difficulty Distribution



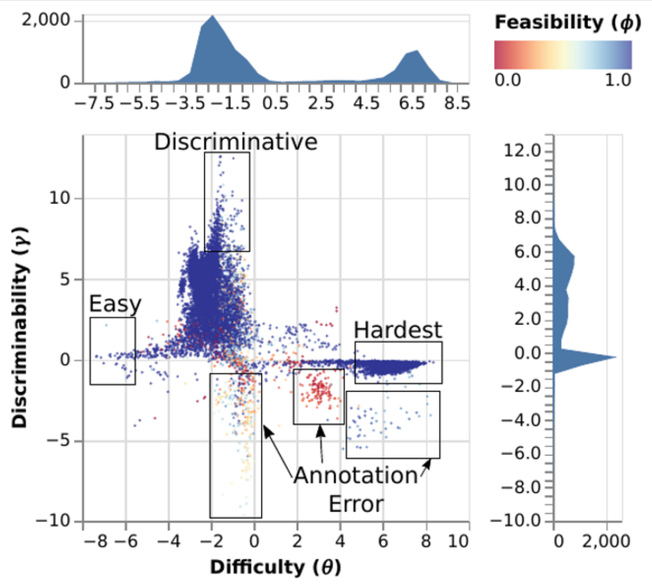
Source: Lalor et al. (2019)

IRT for Leaderboards (SQuAD)

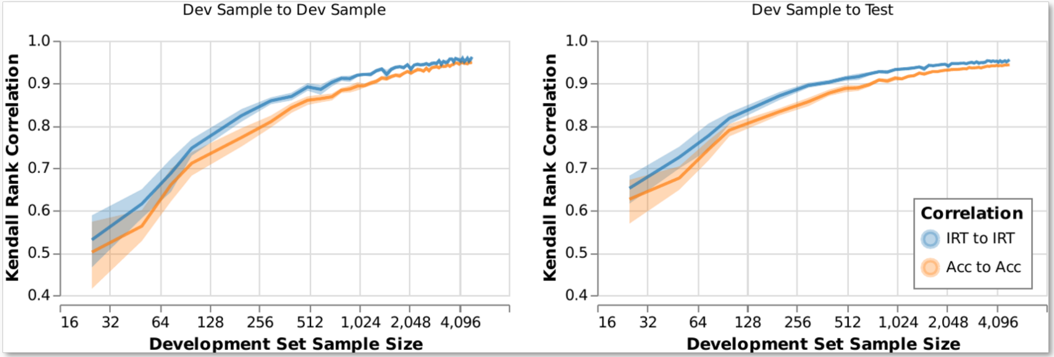


- 1.9 million subject-item pairs

IRT for SQuAD



Ranking Performance



Source: Rodriguez et al. (2021)

The py-irt Package

IRT in Python: py-irt

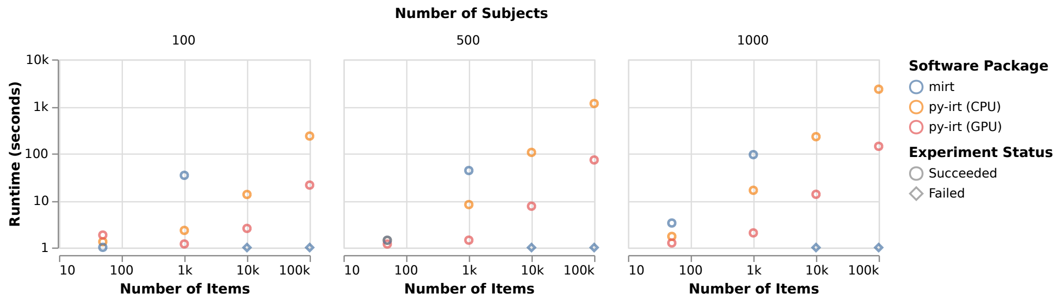

```
{"subject_id": "pedro",    "responses": {"q1": 1, "q2": 0, "q3": 1, "q4": 0}}
{"subject_id": "pinguino", "responses": {"q1": 1, "q2": 1, "q3": 0, "q4": 0}}
{"subject_id": "ken",      "responses": {"q1": 1, "q2": 1, "q3": 1, "q4": 1}}
{"subject_id": "burt",     "responses": {"q1": 0, "q2": 0, "q3":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

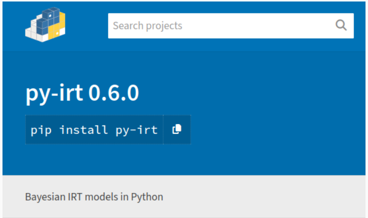

```
py-irt train 1pl data/data.jsonlines output/1pl/
```

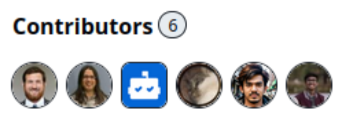

```
{
  "ability": [
    -1.7251124382019043,
    -0.06789101660251617,
    1.6059941053390503,
    -0.20248053967952728
  ],
  "diff": [
    0.008014608174562454,
    9.654741287231445,
    -5.5452165603637695,
    -0.2792229950428009
  ],
```

```
  "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

Lalor and Rodriguez (2022)

Let’s look at the code

Intro to IRT notebook 2 – 2_pyirt_example.ipynb

References

Frank B Baker. 2001. *The basics of item response theory.* ERIC.

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 and Pedro Rodriguez. 2022. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. 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.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

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.

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

Introduction

Improving Model Training

Finding Annotation Error

Evaluation Metrics

Introduction

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 ability θ_j might assume:

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

IRT Applications: Example of Model Behavior

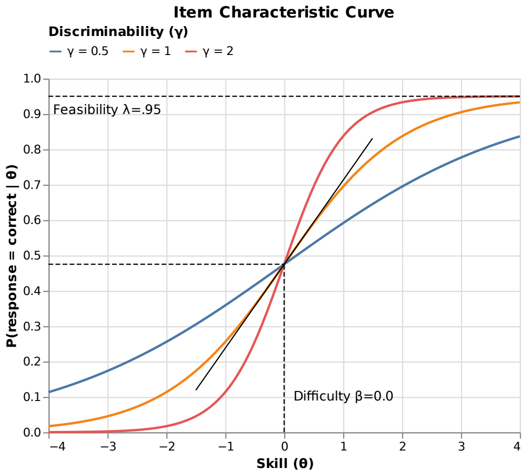
Likelihood of correct answer for subject j on item i.

$$\overbrace{p(y_{ij} = 1 \mid \gamma_i, \beta_i, \lambda_i, \theta_j)}^{\text{Likelihood of correct answer for subject } j \text{ on item } i} = \frac{\lambda_i}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Discriminability of item i

Ability of subject j

Difficulty of item i



What IRT Yields

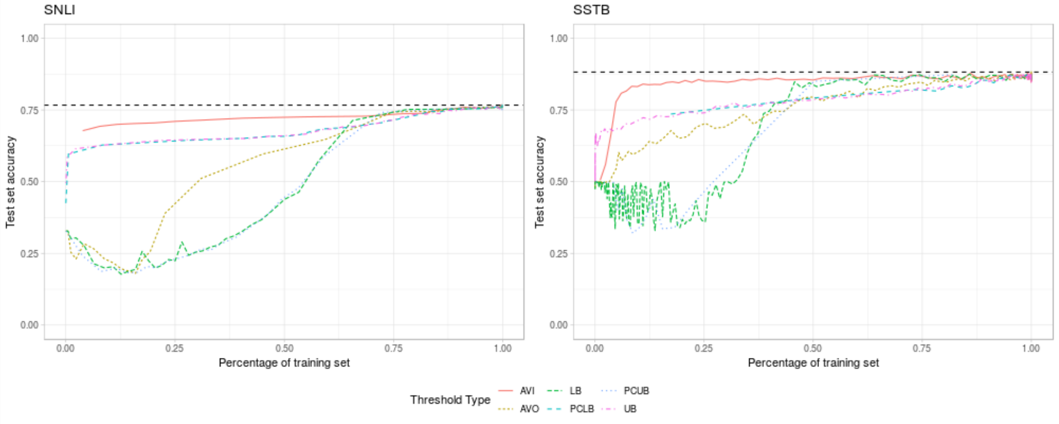
Given the previous information, IRT will yield estimates for chosen parameters, i.e.: item difficulty β_i, discriminability γ_i, and ability θ_j.

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: $pc_i > \tau$

Source: Lalor et al. (2019)

Biggest Differences

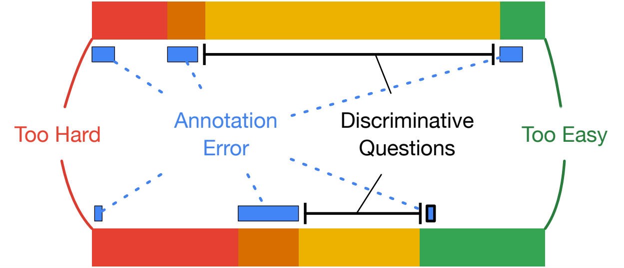
Task	Label	Item Text	Difficulty ranking		
			Humans	LSTM	NSE
SNLI	Con.	*P:* Two dogs playing in snow. *H:* A cat sleeps on floor	168	1	5
	Ent.	*P:* A girl in a newspaper hat with a bow is unwrapping an item. *H:* The girl is going to find out what is under the wrapping paper.	55	172	176
SSTB	Pos.	Only 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

Source: Lalor et al. (2019)

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:

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/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: $r_{ij} = u > .5, u \sim U(0, 1)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/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: $r_{ij} = u > .5, u \sim U(0, 1)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/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

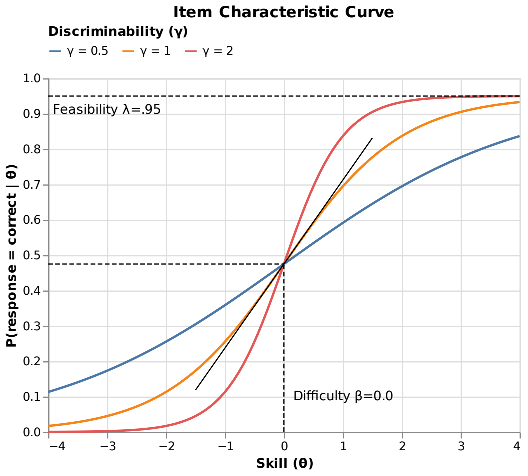
IRT Applications: 3PL Model

$$\overbrace{p(y_{ij} = 1 \mid \gamma_i, \beta_i, \lambda_i, \theta_j)}^{\text{Likelihood of correct answer for subject } j \text{ on item } i.} = \frac{\lambda_i}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Discriminability of item i

Ability of subject j

Difficulty of item i



IRT Applications: Setup for Finding Annotation Error

IRT Parameters

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

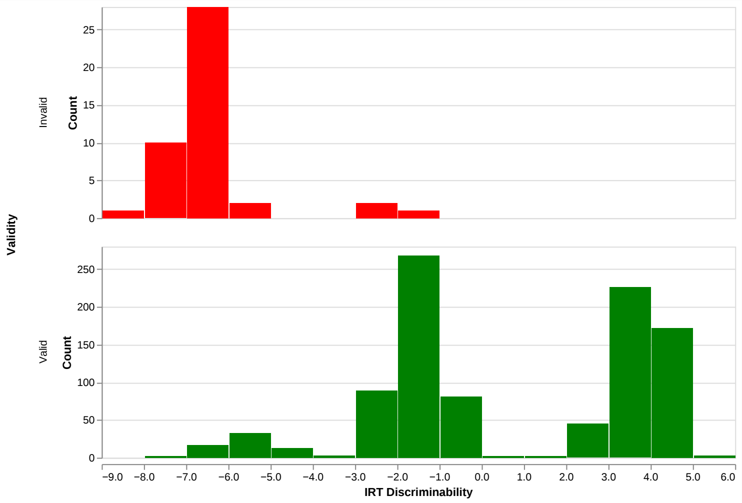
```
dataset = Dataset.from_jsonlines("/tmp/irt_dataset.jsonlines")
config = IrtConfig(
  model_type='tutorial', log_every=500, dropout=.2
)
trainer = IrtModelTrainer(
  config=config, data_path=None, dataset=dataset
)
trainer.train(epochs=5000, device='cuda')
```

IRT Applications: Simulation Results

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

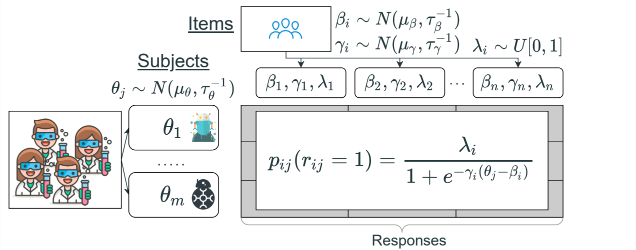
IRT Applications: Simulation Results

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:

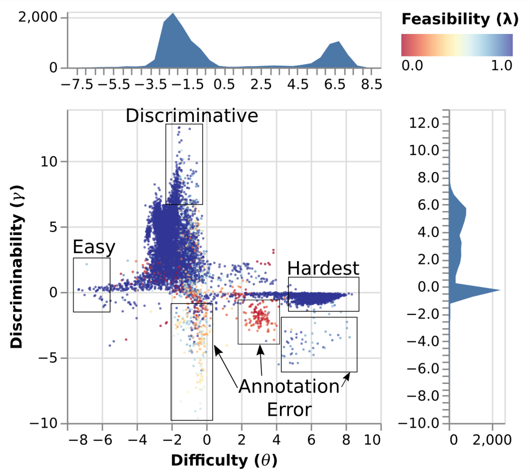


Differences

- Discriminability γ_i could be negative, which is inconvenient.
- Feasibility λ_i.

IRT Applications: Finding Annotation Error

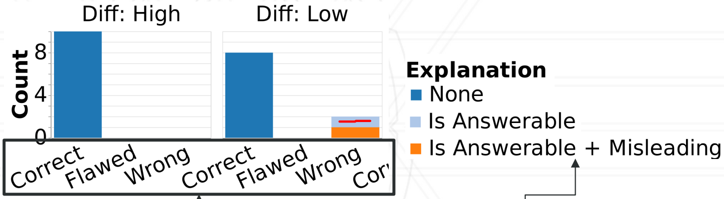
Plotting IRT parameters:



IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors

Was the example correct?
- Question makes sense
- Answer is correct
- No ambiguity
- …

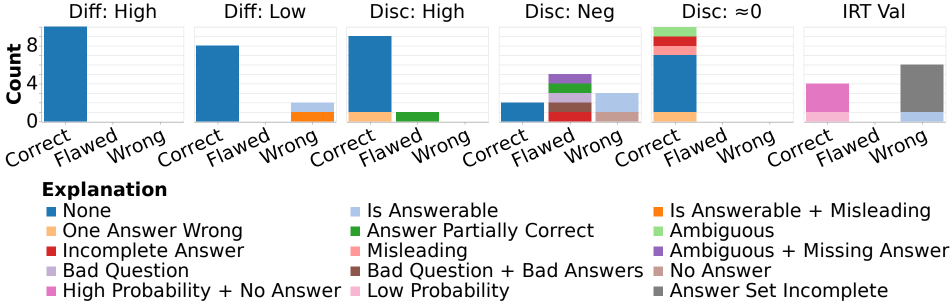


If the example is wrong, then why?
- It is *"Wrong/Flawed"* because it *"Explanation"*

Example:
One low difficulty questionwas wrong, because although the label says it is not answerable, it is answerable

IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors



Things to note:

- Negative discriminability identifies errors

IRT Applications: Finding Annotation Error

Example of bad example identified by IRT

discriminability: -9.63 **Difficulty**: -0.479 **Feasibility**: 0.614 **Mean Exact Match**: 0.472
Wikipedia Page: Economic inequality **Question ID:** 572a1c943f37b319004786e3
Question: Why did the demand for rentals decrease?
Official Answer: demand for higher quality housing
Context: A number of researchers (David Rodda, Jacob Vigdor, and Janna Matlack), argue that a shortage of affordable housing – at least in the US – is caused in part by income inequality. David Rodda noted that from 1984 and 1991, the number of quality rental units decreased as the demand for higher quality housing increased (Rhoda 1994:148). Through gentrification of older neighbourhoods, for example, in East New York, rental prices increased rapidly as landlords found new residents willing to pay higher market rate for housing and left lower income families without rental units. The ad valorem property tax policy combined with rising prices made it difficult or impossible for low income residents to keep pace.

Evaluation Metrics

IRT Applications: Evaluation Metrics

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

IRT Applications: Evaluation Metrics

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

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%
- A set of 150 moderate examples $\sim U(0, 3)$, Validity Rate 90%

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%
- A set of 150 moderate examples $\sim U(0, 3)$, Validity Rate 90%
- A set of 50 hard examples $\sim U(3, 4)$, Validity Rate 80%

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset
  - Moderate subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset
  - Moderate subset
  - Hard subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset
  - Moderate subset
  - Hard subset
- What does the data show?

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset → Asking the right question matters!

True	IRT	Overall	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: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset → Asking the right question matters!
- Fewer hard examples → noisier subset.

True	IRT	Overall	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: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset → Asking the right question matters!
- Fewer hard examples → noisier subset.
- Accuracy difference between best two subjects is not large.

True	IRT	Overall	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: Evaluation Metrics Example

The data shows:

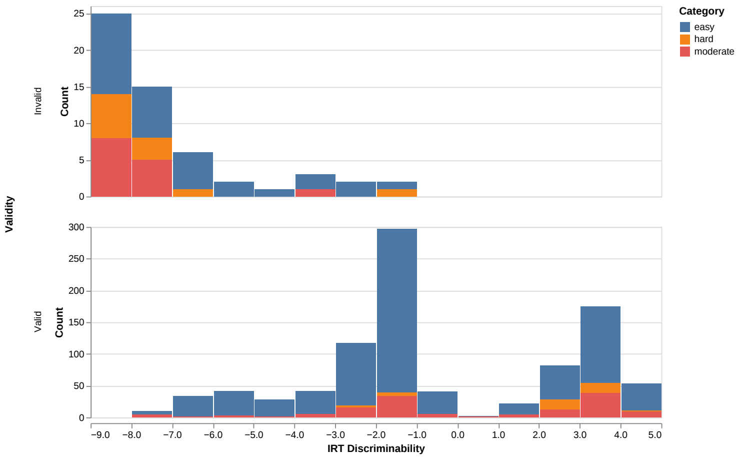
- Variation in true/inferred ability and accuracy by subset $\rightarrow$ Asking the right question matters!
- Fewer hard examples $\rightarrow$ noisier subset.
- Accuracy difference between best two subjects is not large.
- IRT is well suited to this type of data.

True	IRT	Overall	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

What do we see?

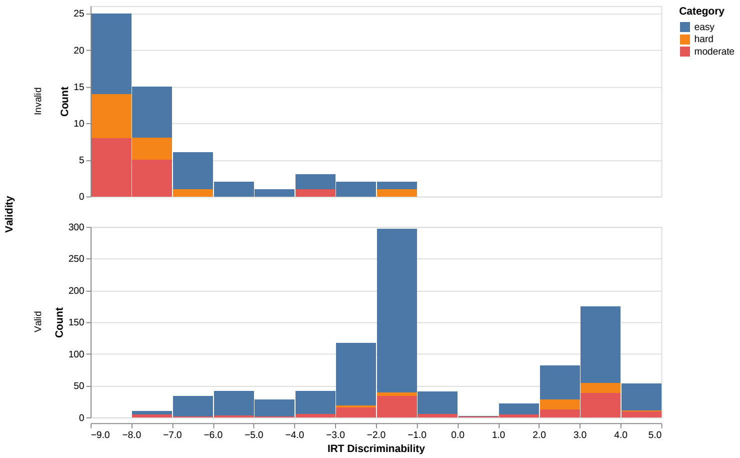
- Invalid examples sorted down



IRT Applications: Discounting Bad Examples

What do we see?

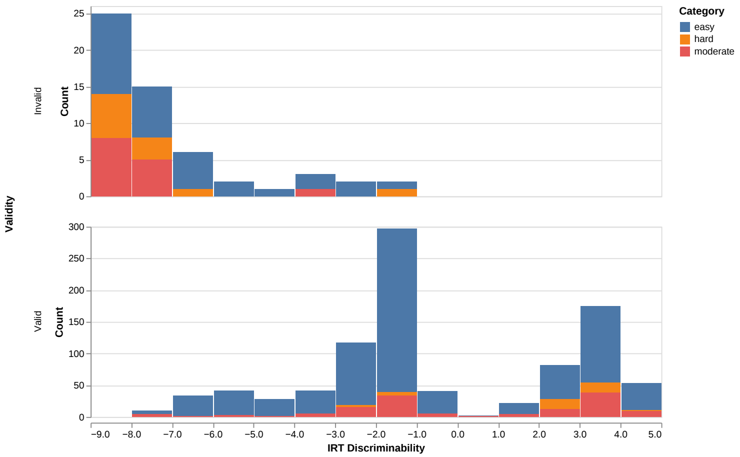
- Invalid examples sorted down
- Proportion of invalid examples represented



IRT Applications: Discounting Bad Examples

What do we see?

- Invalid examples sorted down
- Proportion of invalid examples represented
- Valid Hard examples are more discriminating



IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics

IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics
- Noise metrics $\rightarrow$ noisy rankings

IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics
- Noise metrics $\rightarrow$ noisy rankings
- IRT is one way to mitigate the effect of noisy examples by directly modeling them!

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
    - Minimize annotation cost

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
  - Minimize annotation cost
  - Maximize correlation to ranking if fully annotate

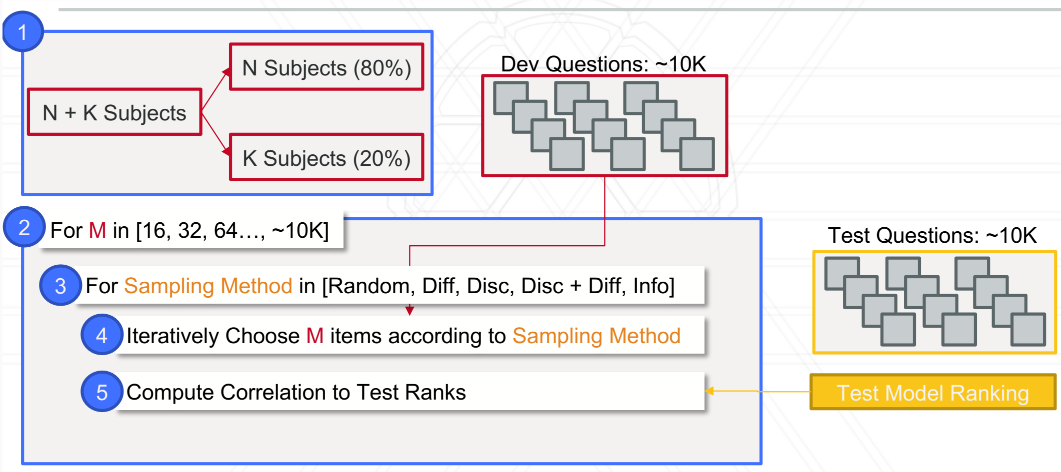
IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

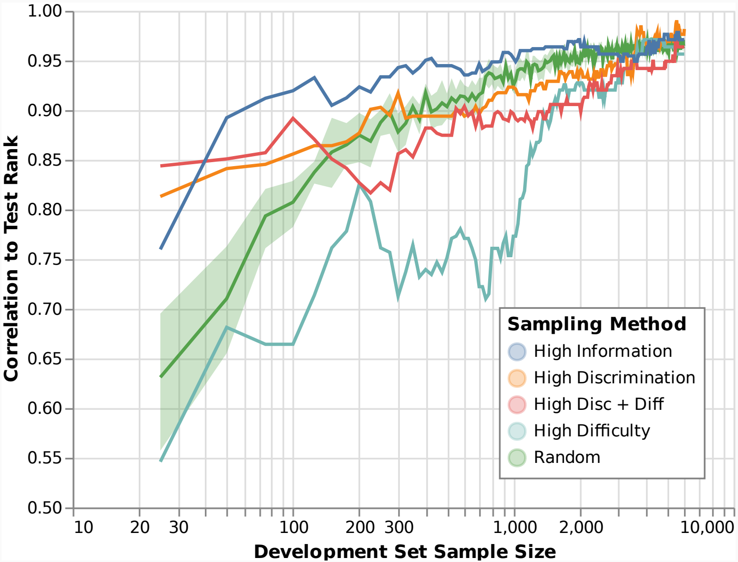
- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
  - Minimize annotation cost
  - Maximize correlation to ranking if fully annotate
- Experiment: What method for selecting subset to annotate is best?

IRT Applications: Rank Reliability in Evaluation Metrics

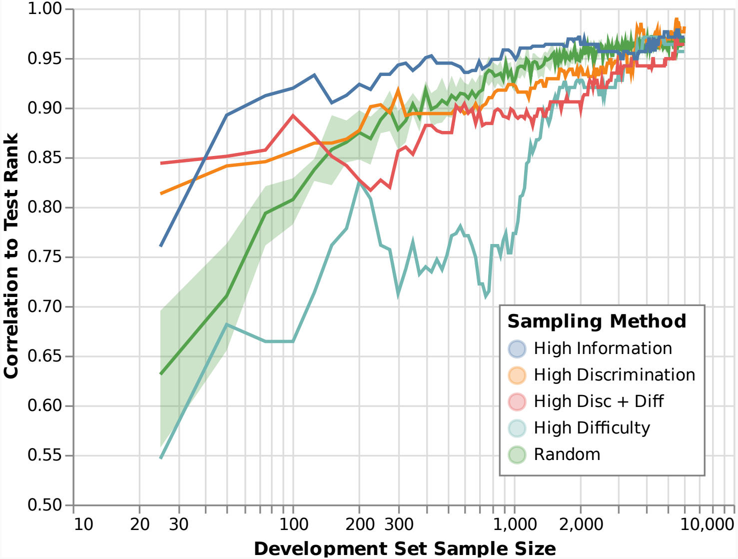
We test this setup with SQuAD leaderboard data:



IRT Applications: Rank Reliability in Evaluation Metrics



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

- Adaptive Language-based Mental Health Assessment with Item-Response Theory (Varadarajan et al., 2023)
- Alternate Evaluation Metrics, e.g., Subject ability θ_j (Lalor et al., 2018)
- Anchor Points: Benchmarking Models with Much Fewer Examples (Vivek et al., 2024)
- tinyBenchmarks: evaluating LLMs with fewer examples (Polo et al., 2024)
- Comparing Test Sets with Item Response Theory (Vania et al., 2021)
- IRT for Efficient Human Evaluation of Chatbots (Sedoc and Ungar, 2020)

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.

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.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, 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.

Vasudha Varadarajan, Sverker Sikström, Oscar NE Kjell, and H Andrew Schwartz. 2023. Adaptive language-based mental health assessment with item-response theory. *arXiv preprint arXiv:2311.06467*.

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). β3-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). β4-IRT: A New β3-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). Item Response Theory for Evaluating Regr 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

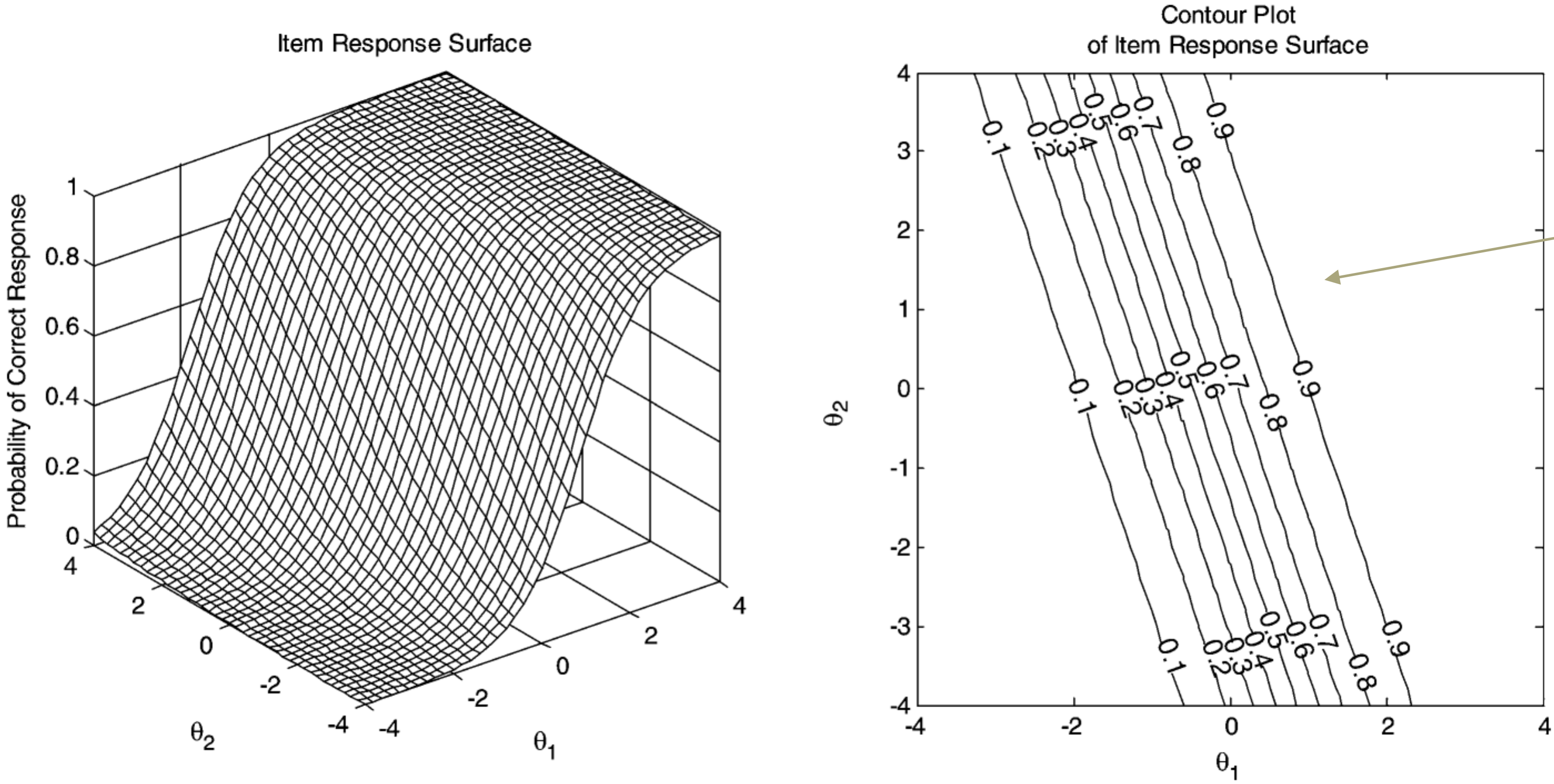
- Ability θ becomes a latent vector and/or difficulty d becomes a latent vector:

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

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



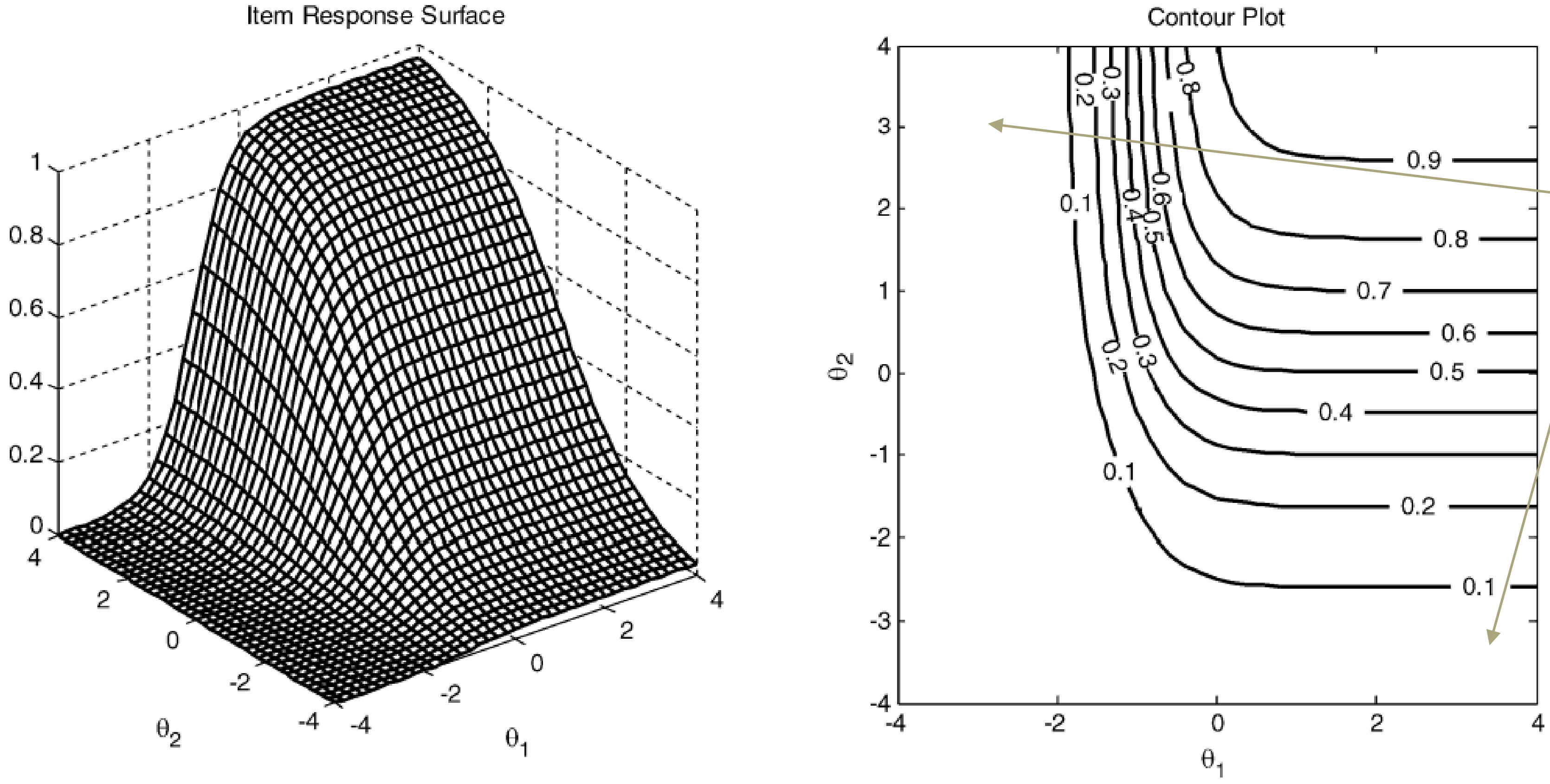
Graphic representations of the compensatory model – item response surface and equiprobable contours for an item with $a_{i1} = 1.5$, $a_{i2} = .5$, and $d_i = .7$.

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

INTRINSIC (OBSERVABLE) DIFFICULTIES

- Frequently, we have intuitions of what makes an instance difficult.
  - “What’s 31+26?” -> very easy
  - “What’s 39+96?” -> easy
  - “What’s 316184915+269435716?” -> hard
  - “What’s 111111111+333333333?” -> easy

q_1= #digits,
q_2= carrying
q_3= digit diversity

- Can we use these K=3 “features” or “characteristics” (q_1, q_2, q_3) as a proxy for difficulty?
  - Do we know how much each of them contributes to difficulty?

LINEAR LOGISTIC TEST MODELS (LLTM)

- 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.

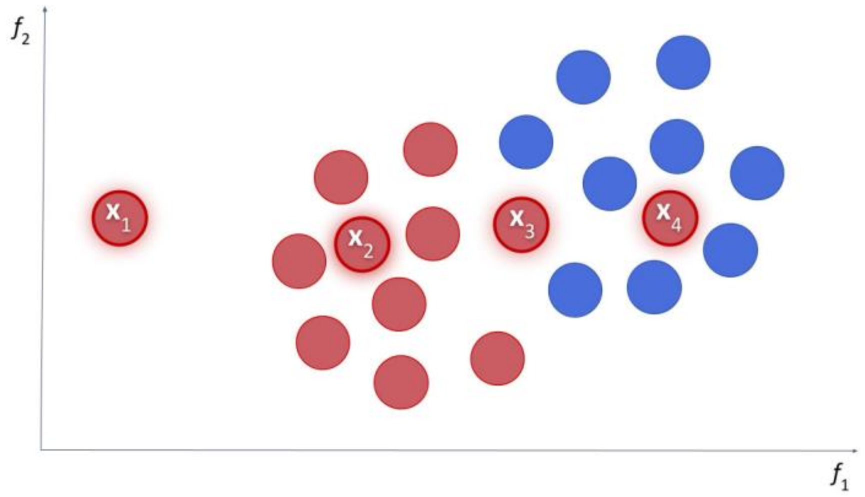
- Values can be > 1

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 is **extrinsic**: 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.

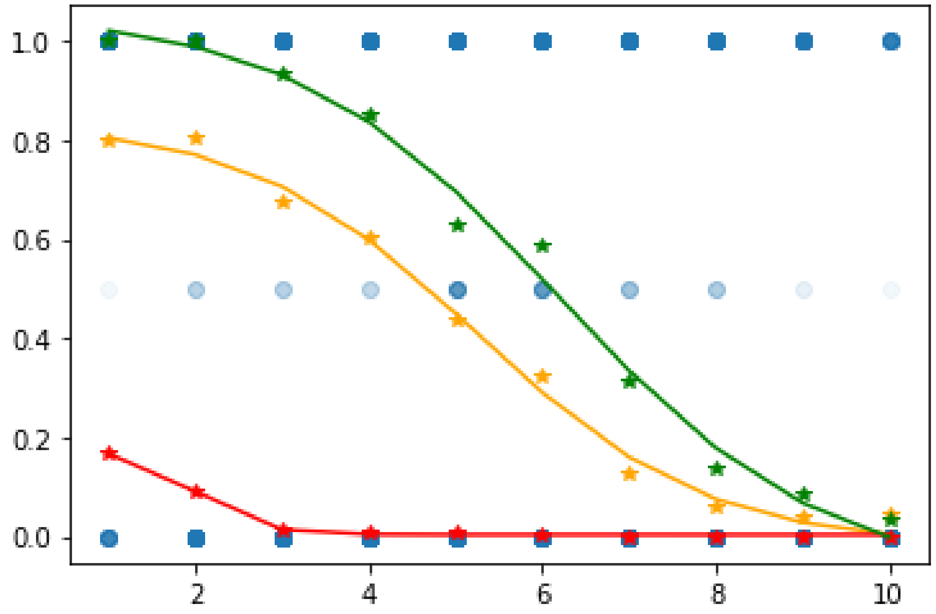


X1: medium
X2: easy
X3: hard
X4: very hard.

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

- Use “topic modelling” to extract the demands?
- 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):

```
You must help me annotate the level of {META-FEATURE} of
some text. Note that {META-FEATURE DEFINITION}. I will
first give you a few examples to illustrate it. Then you
will have to determine the level of {META-FEATURE} for the
text on a scale from {META-FEATURE SCALE}.
{META-FEATURE EXAMPLES}
Sentence: {INSTANCE} Level of {META-FEATURE}:"
```

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	0.83
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

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

Task	Linguistic Meta-features	Traditional Metrics
Abstract Narrative Understanding	0.06	-0.01
BBQ	$\boxed{0.62}$	$\boxed{0.5}$
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	0.83
OpenbookQA	-0.04	0.01
TruthfulQA	$\boxed{0.59}$	$\boxed{0.56}$

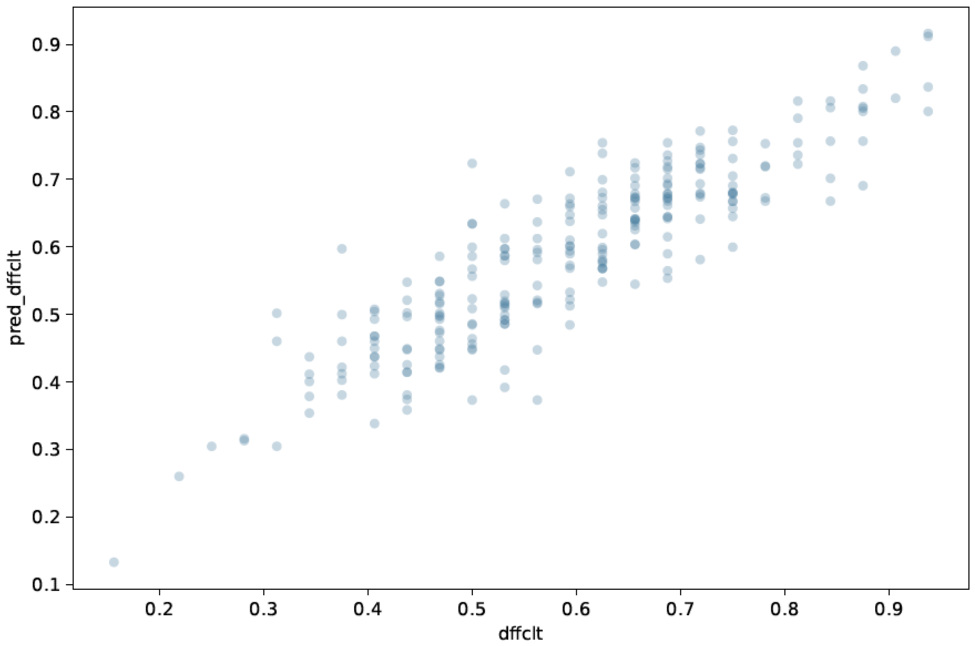
Table 5.1: R^2 obtained in the test split when predicting difficulty with linguistic meta-features and lexical and readability metrics

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	$\boxed{0.8}$	
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 examples?

YES, WE CAN

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

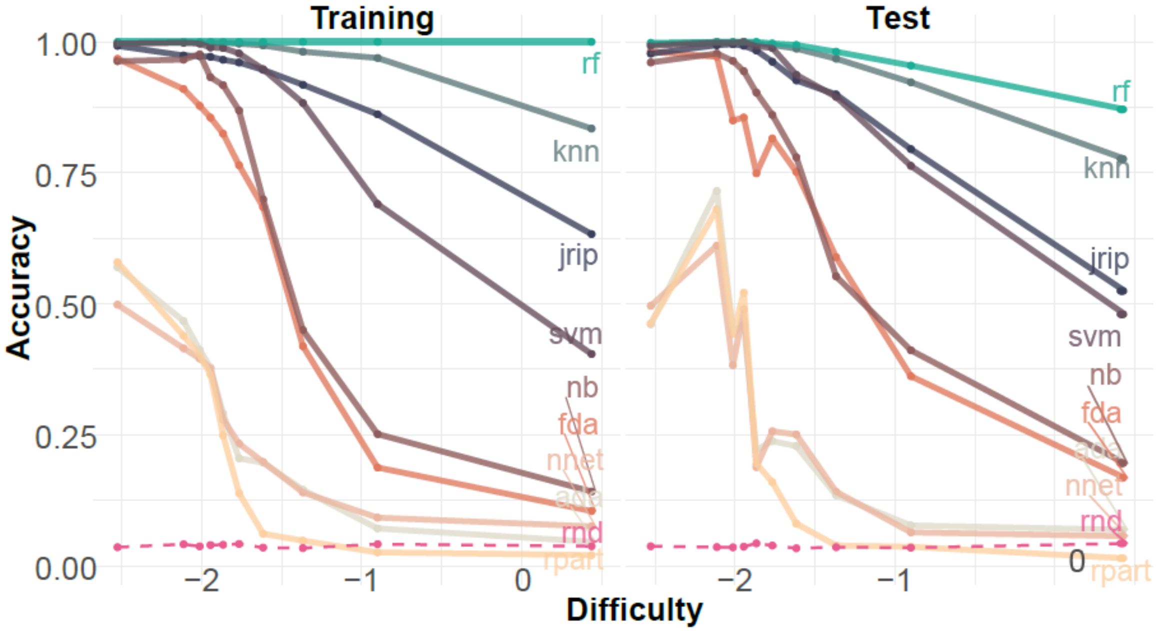


Figure 5: (Left) SCC obtained with the 70% of the letter benchmark and the observed difficulties $\hbar$. (Right) SCC obtained with the test set (30%), using estimated difficulties $\hat{\hbar}$.

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:

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

It is trained (and evaluated) on test data:

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

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
…	…	…

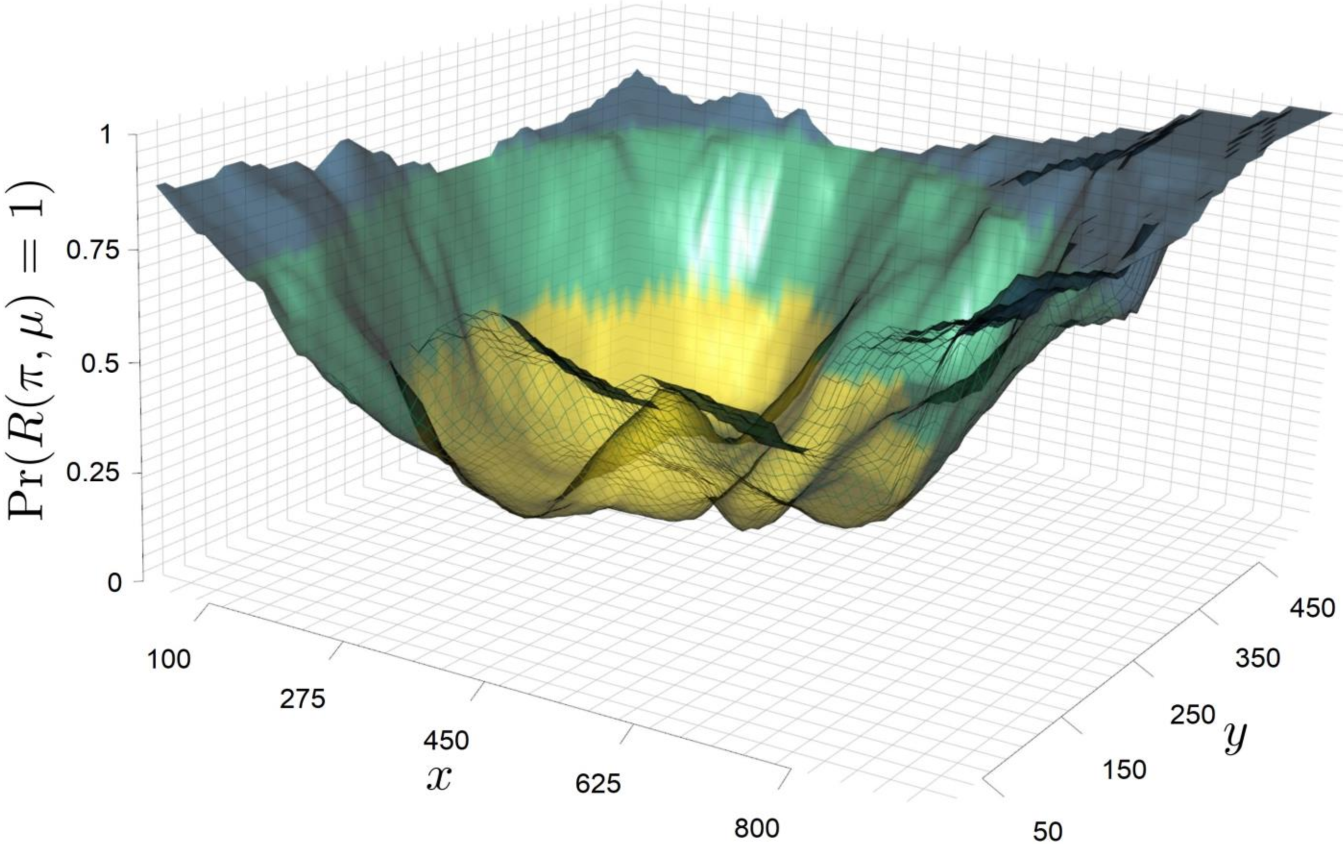
It is applied during deployment, before π does any inference or even starts.

PROBLEM SPACE

Downtown Vancouver

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

- 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 $\Pr(x,y)$



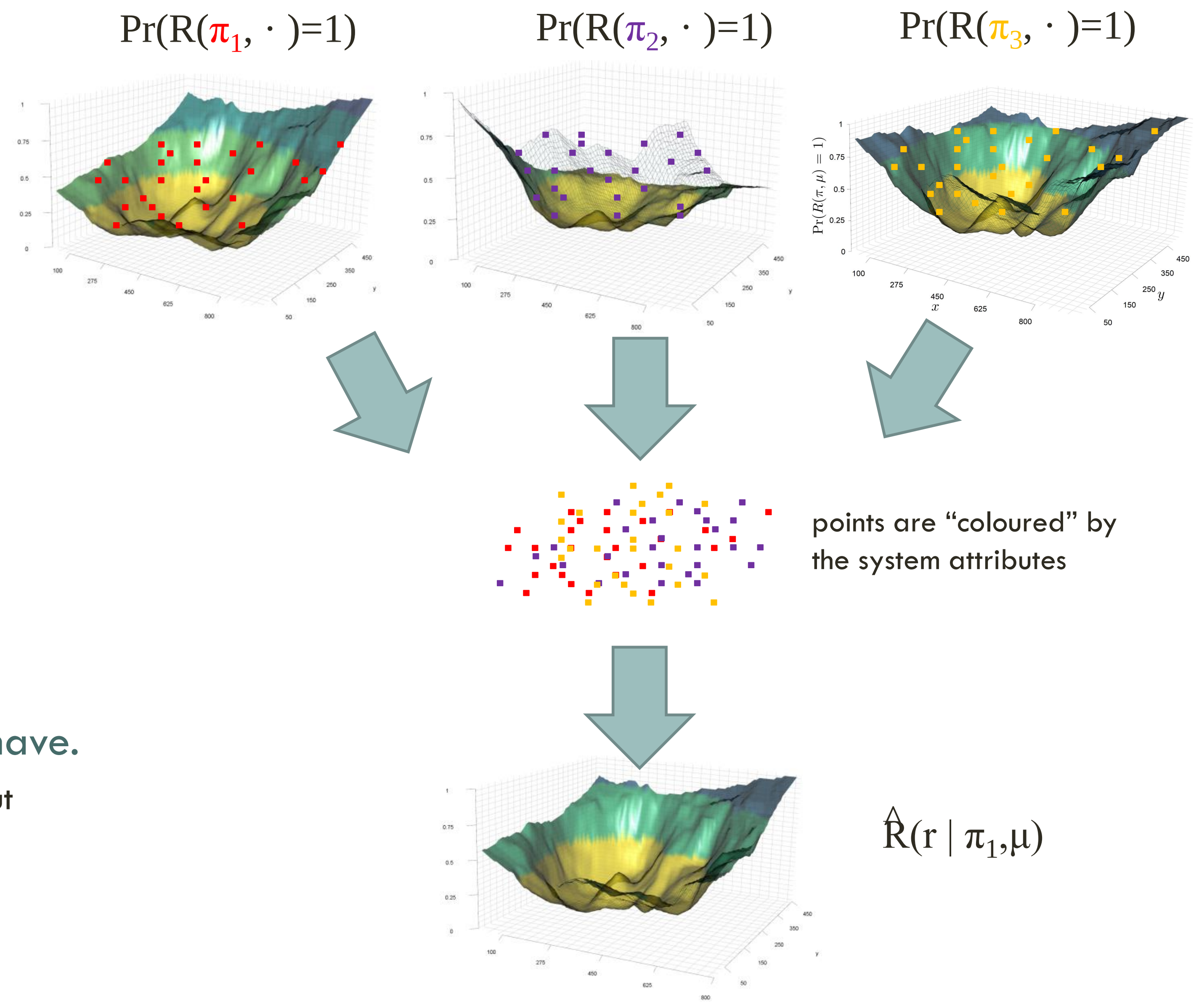
SYSTEM SPACE

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!



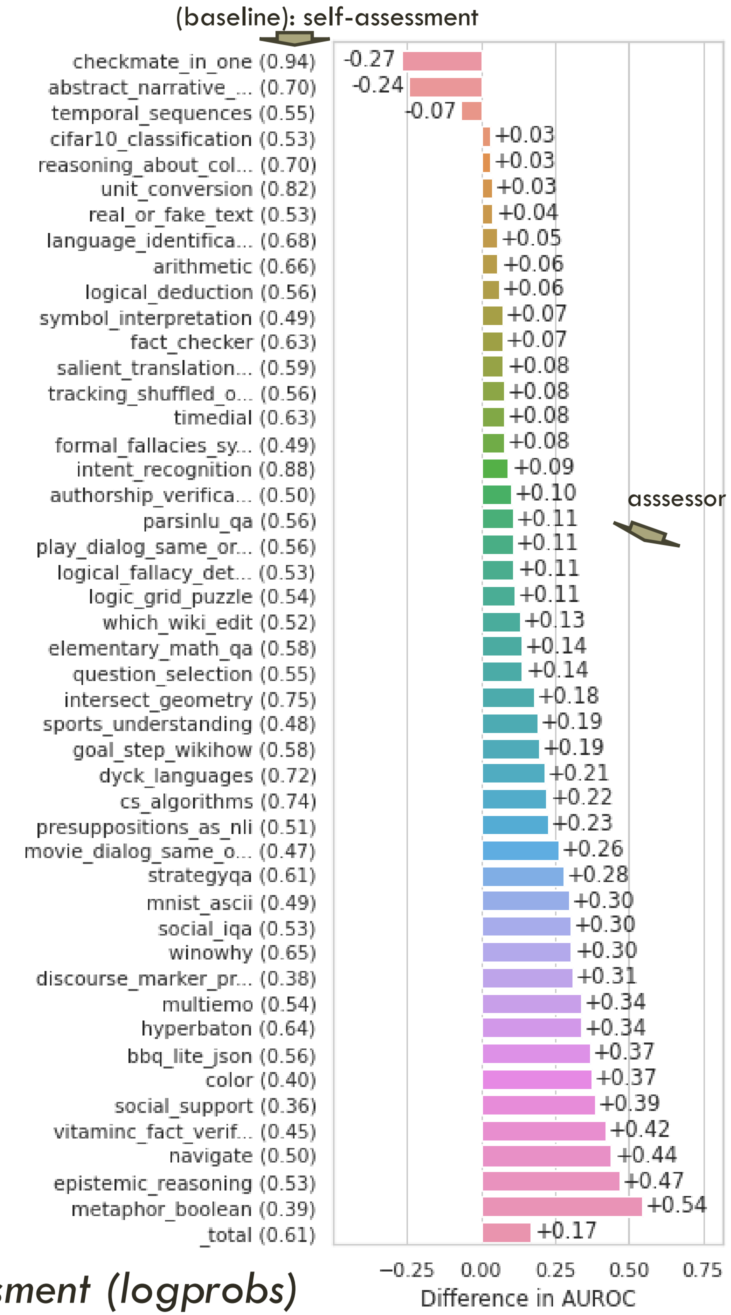
LMs PREDICT LMs

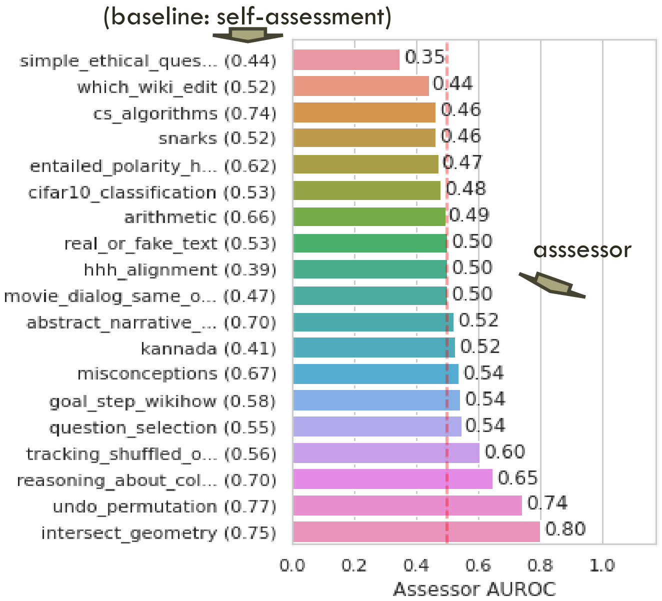
Setup:

- 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)

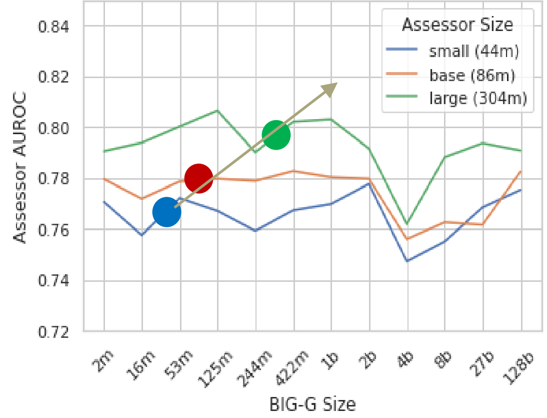
In distribution:

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





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



Bigger assessor = better
Bigger subject = neutral

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

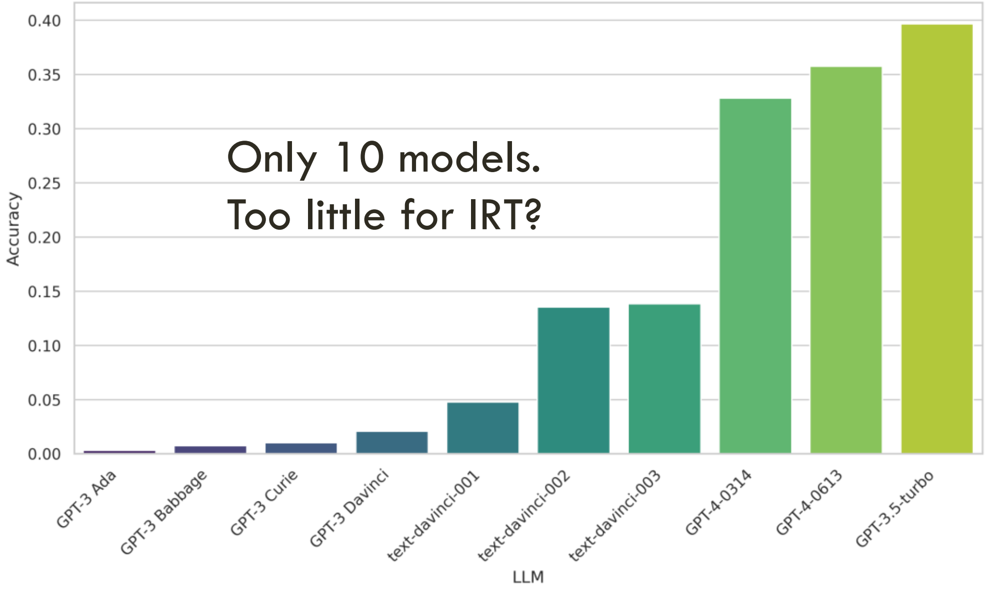
Measurement Layouts

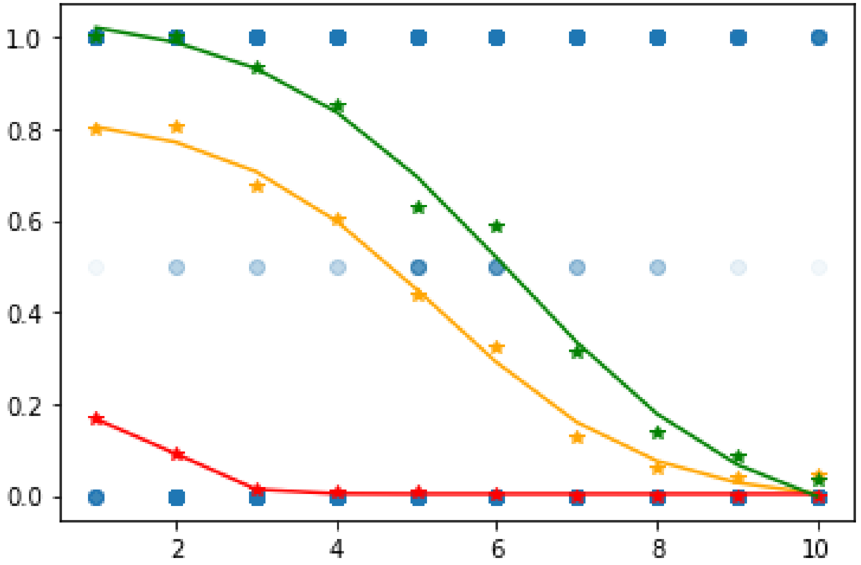
AAAI2024 Tutorial
"Measurement Layouts for Capability-Oriented AI Evaluation",
J. Burden, L. Cheke, J. Hernández-Orallo, M. Tešić, K. Voudouris
https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial

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.

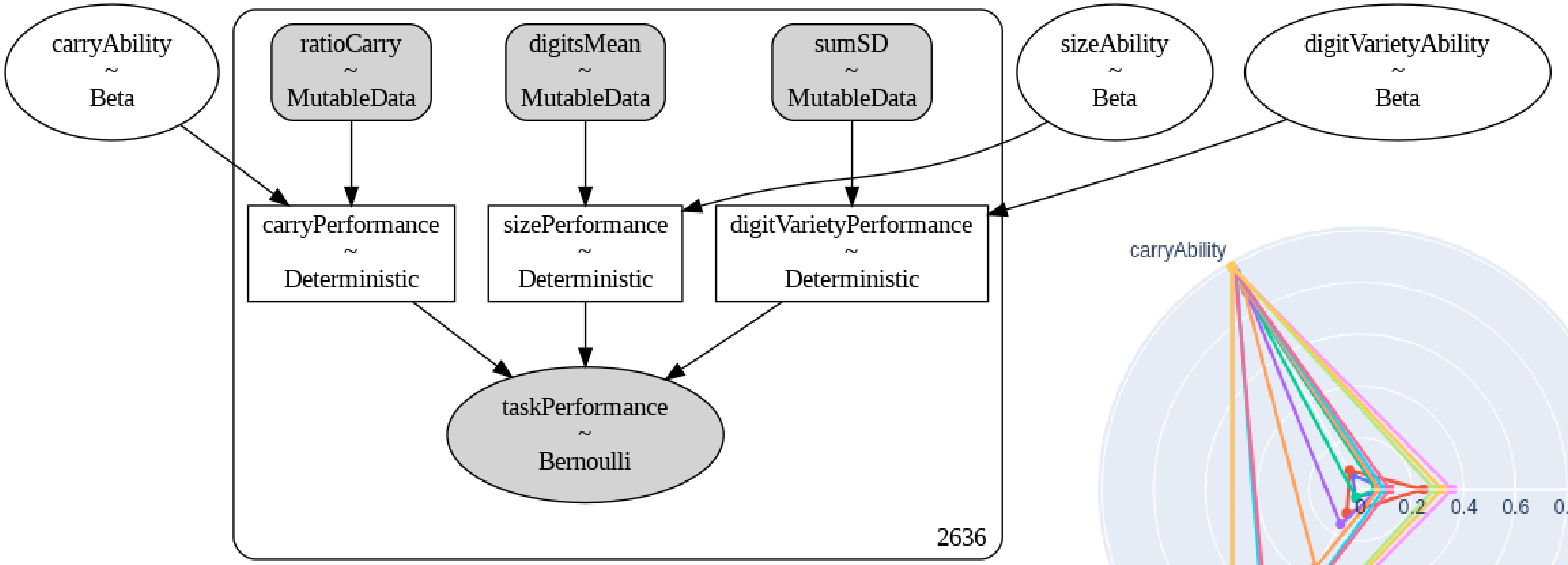
MORE SOPHISTICATED DEMANDS

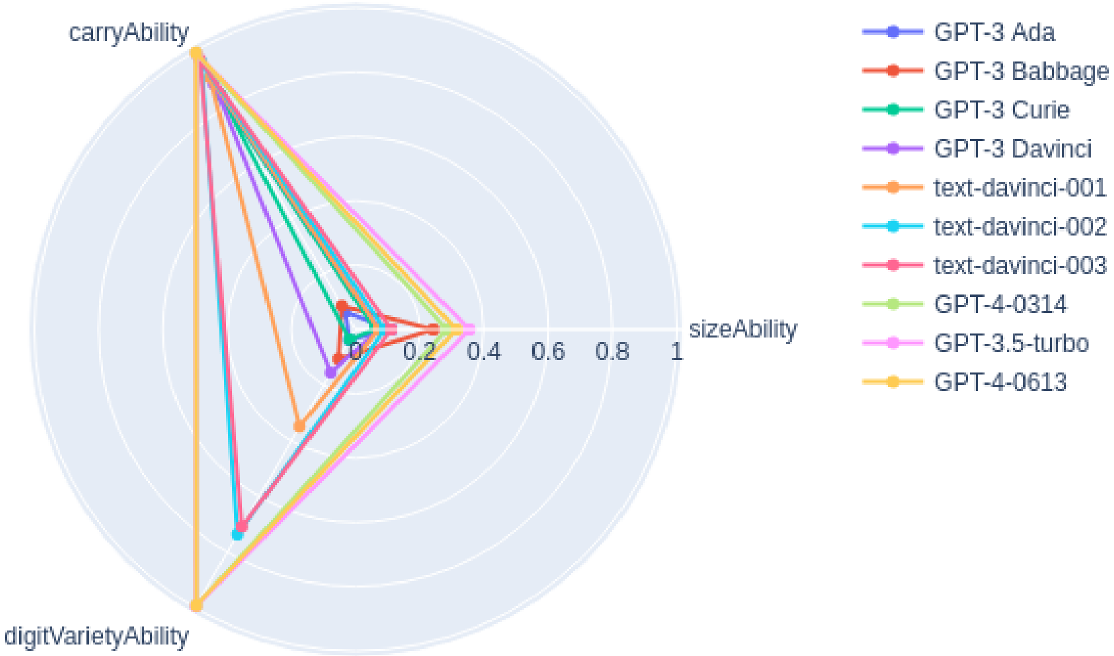
- `digits1`: The number of digits in the first summand.
- `digits2`: The number of digits in the second summand.
- `min_digits`: $min(digits_1, digits_2)$, i.e., the number of digits in the smaller summand.
- `harm_mean`: $2/(1/digits_1 + 1/digits_2)$, i.e., the harmonic mean of the number of digits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

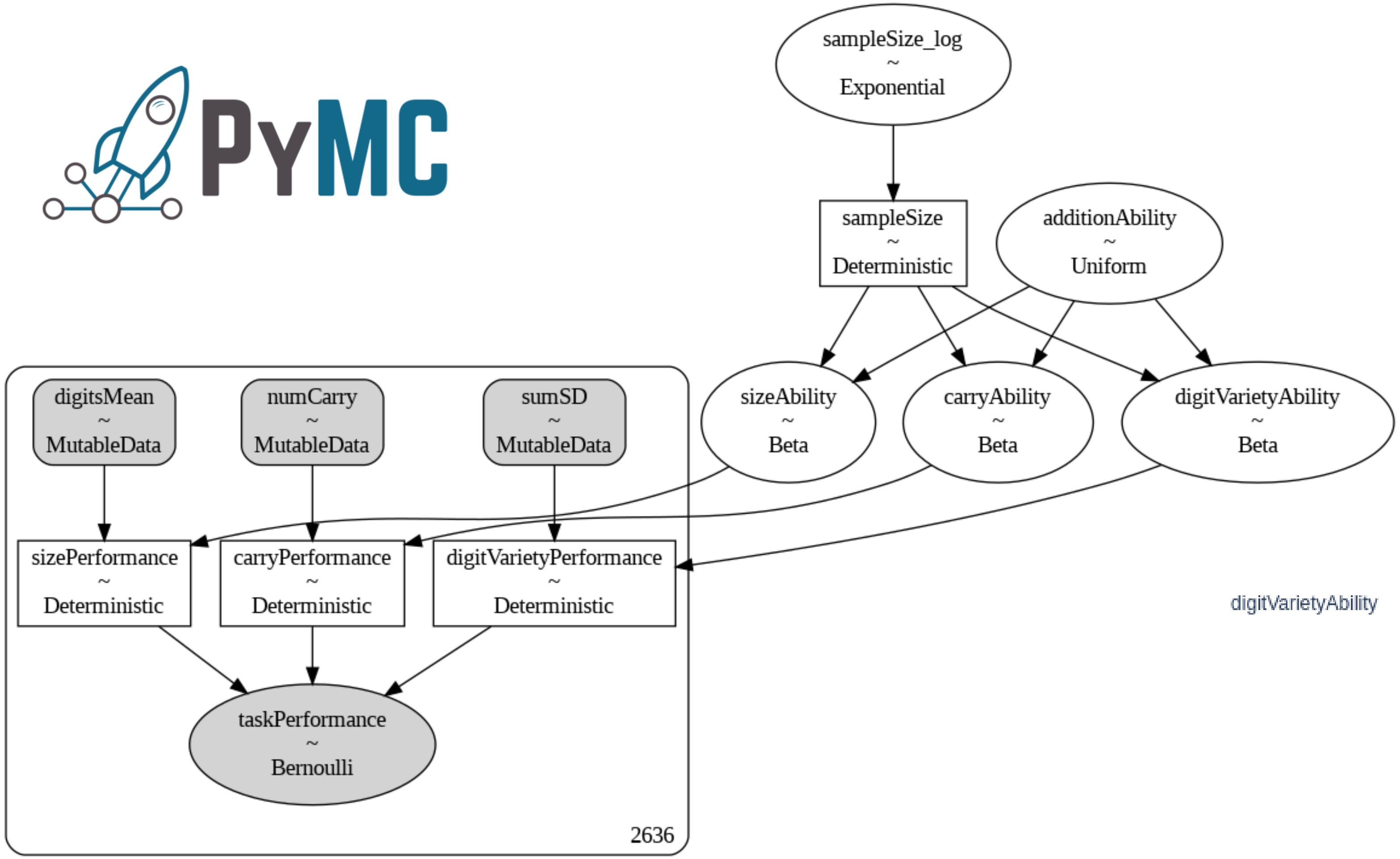
SIMPLE MEASUREMENT LAYOUT

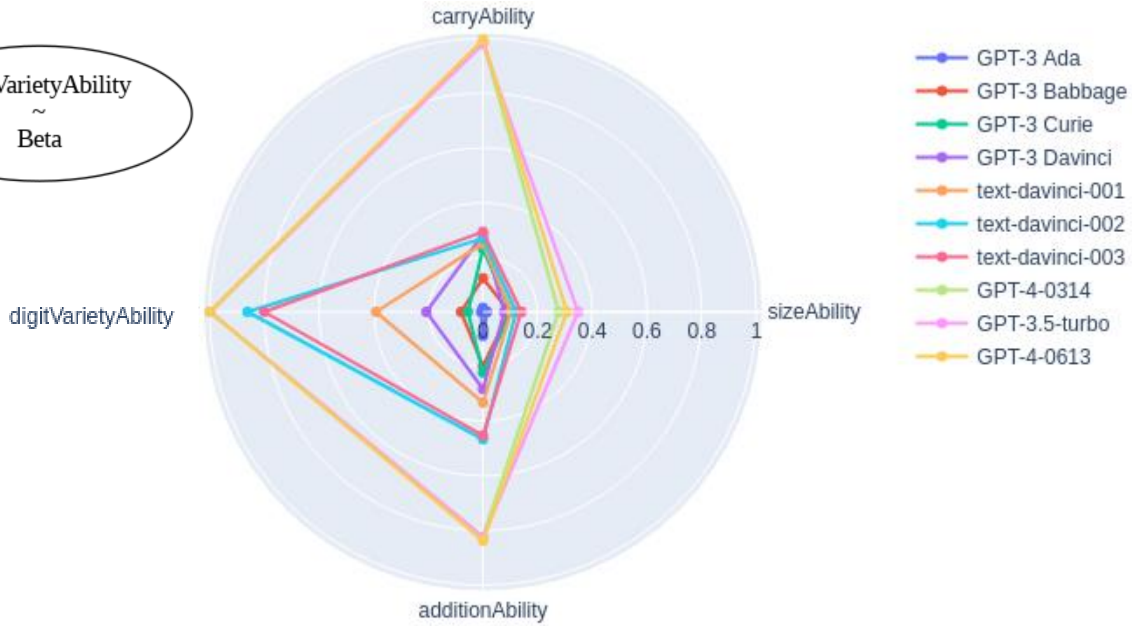




PyMC

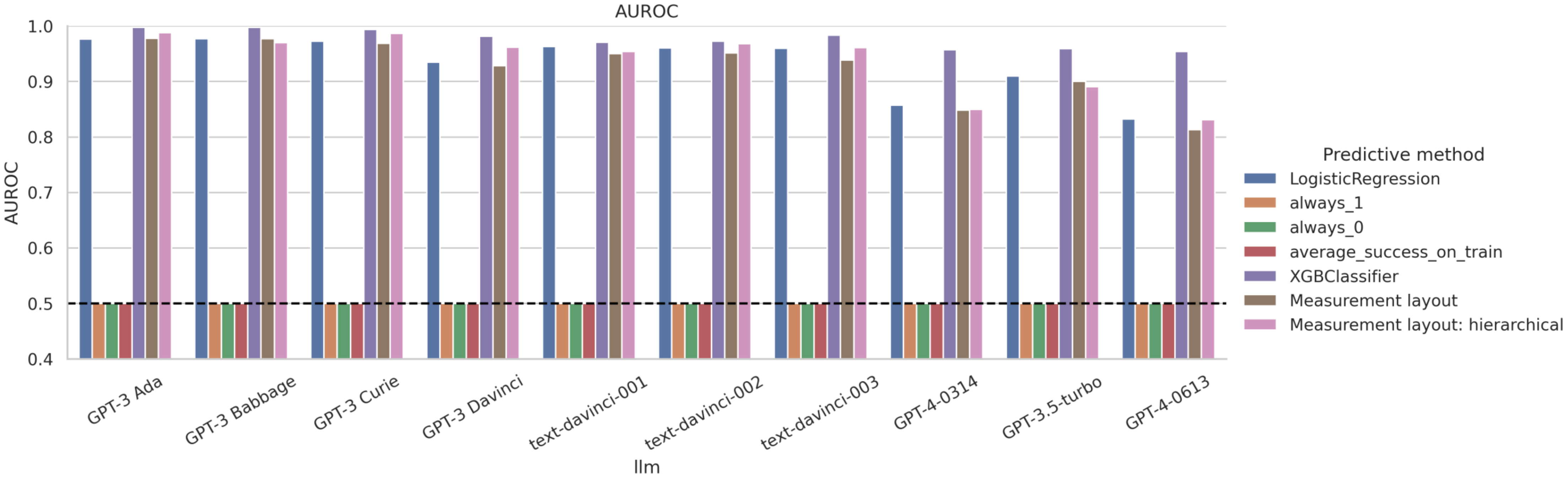
HIERARCHICAL MEASUREMENT LAYOUT





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 PERFORMANCE

From Games and AI:

- Elo-Ranking, TrueSkill (Microsoft)

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

From AI:

- Scaling laws

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

From Psychometrics:

- 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

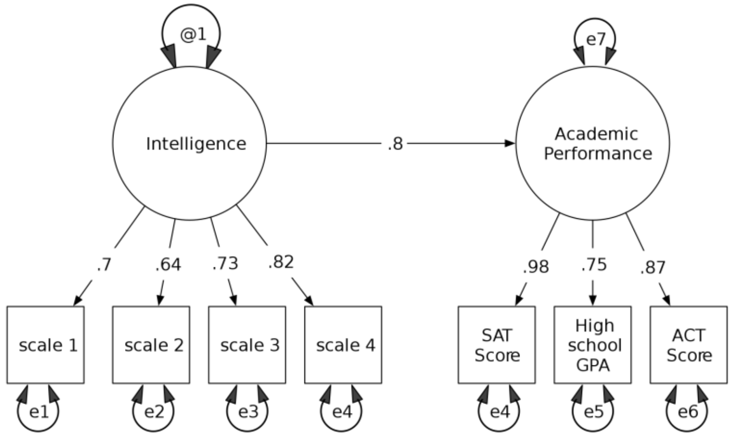
FACTOR ANALYSIS

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

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

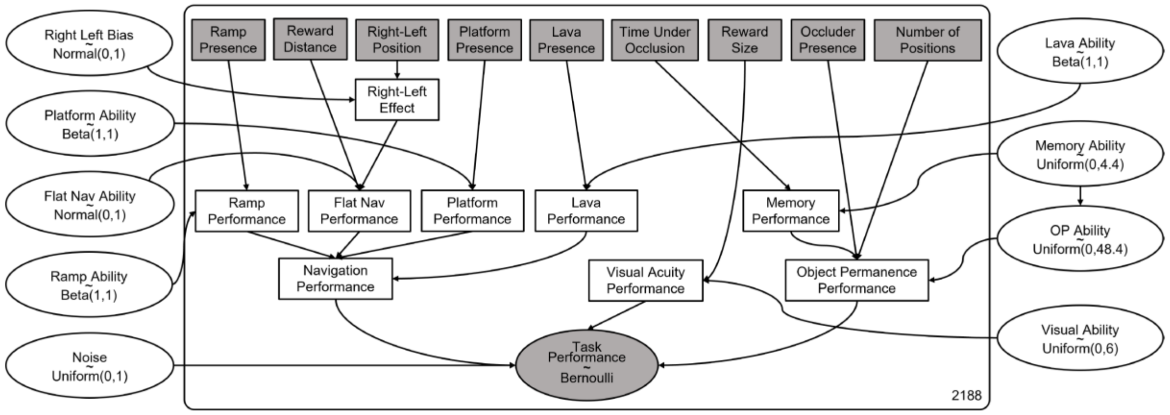
POPULATIONAL? INSTANCE-LEVEL?

- 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

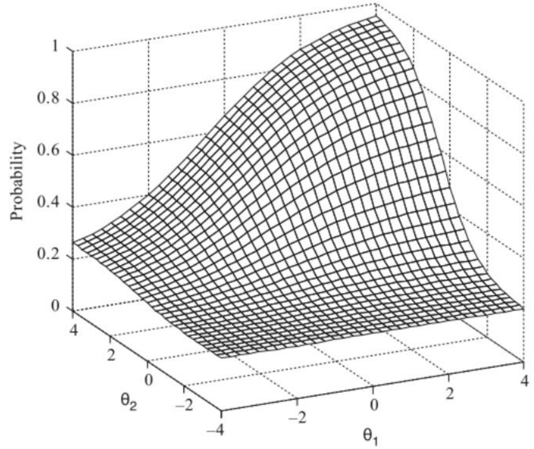
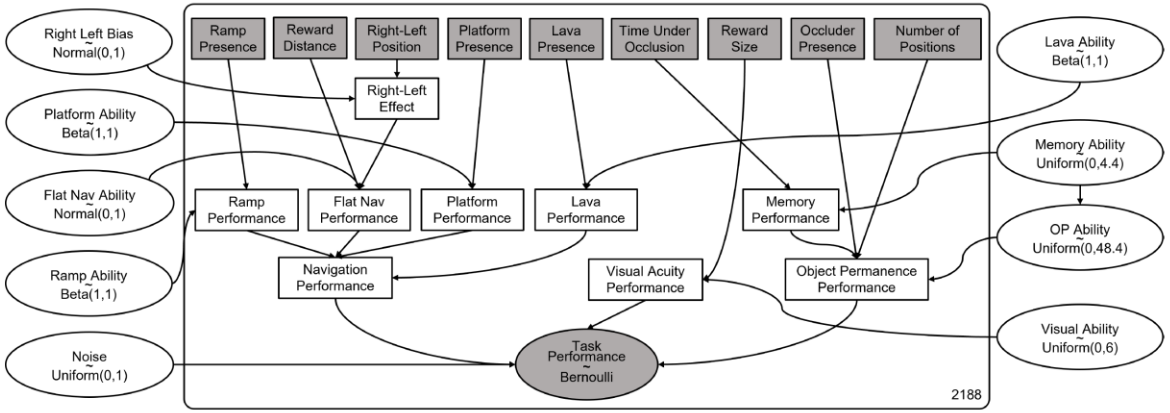


Fig. 4.9 Item response surface for the partially compensatory model when $a_1 = .7$, $a_2 = 1.1$, $b_1 = -.5$, $b_2 = .5$, and $c = .2$

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

SUMMARY OF APPROACHES

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

SUMMARY OF APPROACHES

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

SUMMARY OF APPROACHES

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

SUMMARY OF APPROACHES

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

SUMMARY OF APPROACHES

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

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 policy-makers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-by-instance evaluation results that could be used to unpack these aggregate metrics are rarely made available (*1*). Here, we propose a path forward in which results are presented in more nuanced ways and instance-by-instance evaluation results are made publicly available.

Across most areas of AI, system evaluations follow a similar structure. A system is first built or trained to perform a particular set of functions. Then, the performance of the system is tested on a set of tasks relevant to the desired functionality of the system. In many areas of AI, evaluations use standardized sets of tasks known as "benchmarks." For each task, the system will be tested on a number of example "instances" of the task. The system would then be given a score for each instance based on its performance, e.g., 1 if it classified an image correctly, or 0 if it was incorrect. For other systems, the score for each instance might be based on how quickly the system completed its task, the quality of its outputs, or the total reward it obtained. Finally, performance across the various instances and tasks is usually aggregated to a small number of metrics that summarize how well the system performed, such as percentage accuracy.

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

AGGREGATE METRICS

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

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

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

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

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

[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

136 14 APRIL 2023 • VOL 380 ISSUE 6641 science.org SCIENCE

TAKE-AWAYS

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

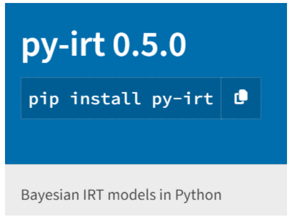
> Instead of aggregating performance, the key idea is to estimate a model of the AI system (e.g., capabilities) so that we can explain/predict performance at the instance level!

THANK YOU!

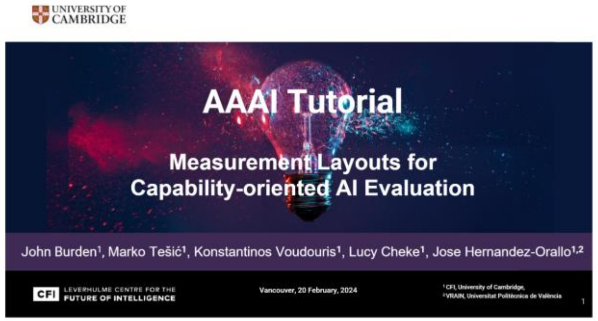
JOSE H. ORALLO
http://josephorallo.webs.upv.es/
jorallo@upv.es

POINTERS

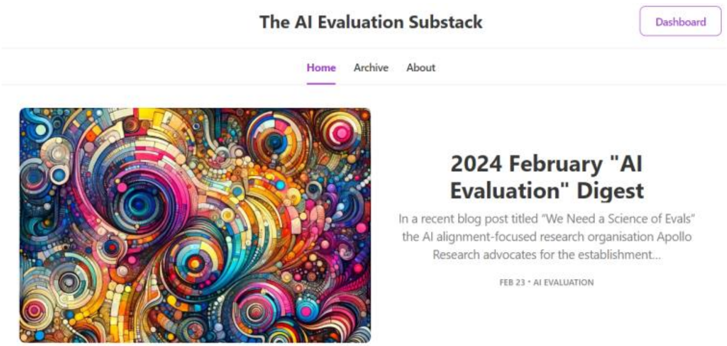
- References: You've been given a reference list…
- Libraries:
  - PY-IRT: https://github.com/nd-ball/py-irt/
  - flexMIRT, MIRT, Stan, JAGS, Mplus, SPSS



- AAAI2024 Tutorial on Measurement Layouts:
  - https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial



- AI Evaluation Digest (monthly)
  - https://aievaluation.substack.com/



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?

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.

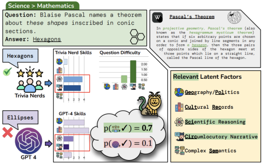


Figure 1: Response Correctness prediction using Agent skills and Question difficulty over relevant latent factors. We list the five latent factors that CAIMIRA discovers, and highlight the relevant ones (green), which contribute to estimating whether an agent will respond to the example question correctly.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

Structured references on the website!

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.

Interested in continuing the conversation?

https://forms.gle/rwAhu6ufgcYgioKm6

Acknowledgements

Thank you!

Web page: http://eacl2024irt.github.io