Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

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

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

- [ ] Show gated/private/deleted models

Model types

- [x] 🟢 pretrained
- [x] 🔶 fine-tuned
- [x] ⭕ instruction-tuned
- [x] 🟦 RL-tuned
- [x] ?

Precision

- [x] float16
- [x] bfloat16
- [x] 8bit
- [x] 4bit
- [x] GPTQ
- [x] ?

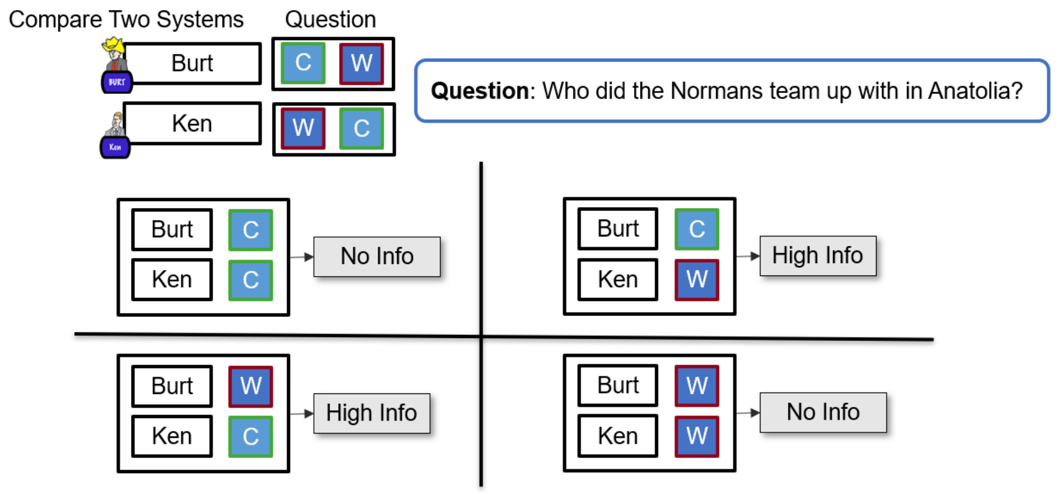
Model sizes (in billions of parameters)

- [x] ?
- [x] ~1.5
- [x] ~3
- [x] ~7
- [x] ~13
- [x] ~35
- [x] ~60
- [x] 70+

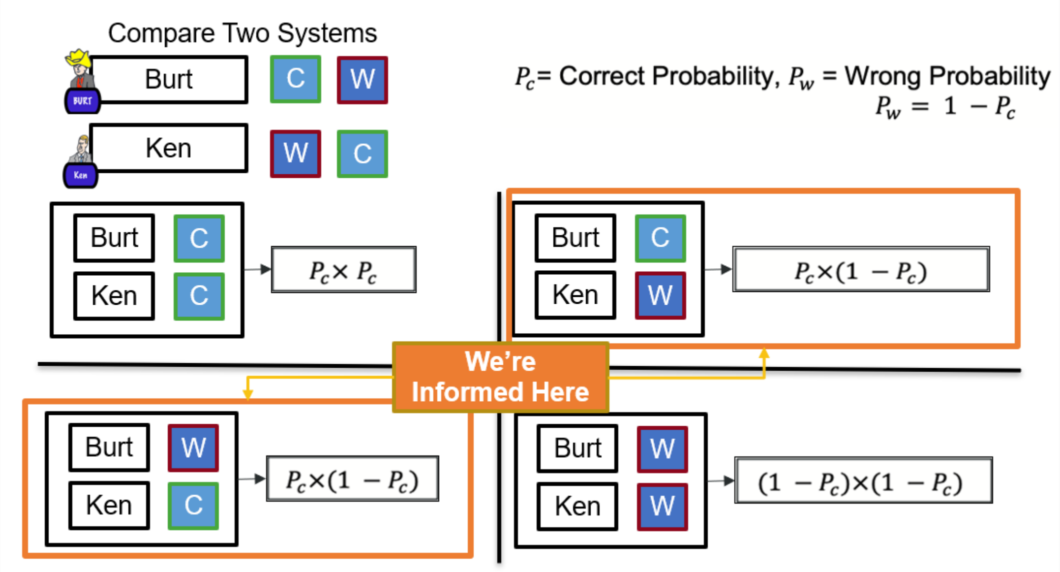
T	Model	Average ⬆️	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	DROP
🔶	TigerResearch/tigerbot-70b-chat-v2	69.76	87.03	82.83	66	75.4	79.16	46.02	51.9
⭕	bhenrym14/platypus-yi-34b	68.96	68.43	85.21	78.13	54.48	84.06	47.84	64.55
🟢	01-ai/Yi-34B	68.68	64.59	85.69	76.35	56.23	83.03	50.64	64.2
🟢	chargoddard/Yi-34B-Llama	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

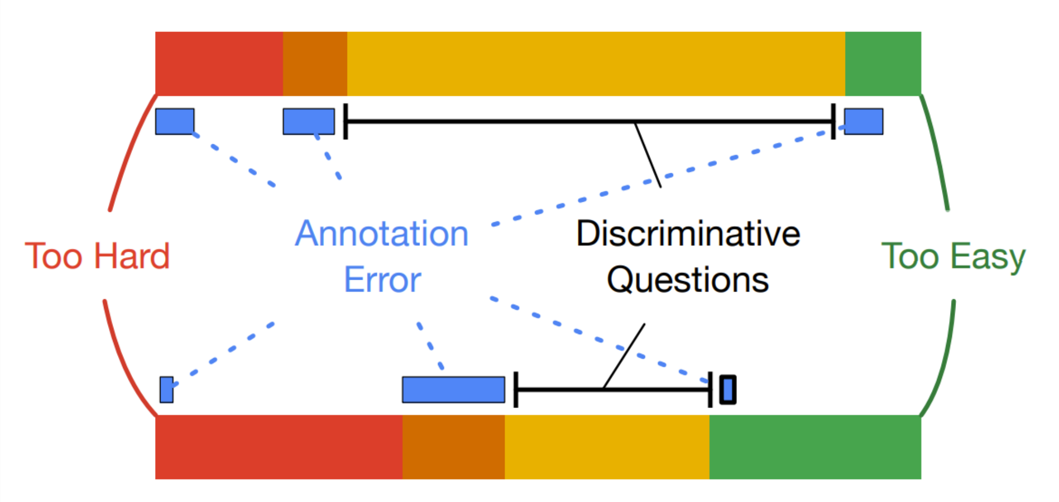
Differences in Questions



Differences in Questions



Differences in Questions



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

Psychometrics

Psychometrics: study of quantitative measurement practices

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

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

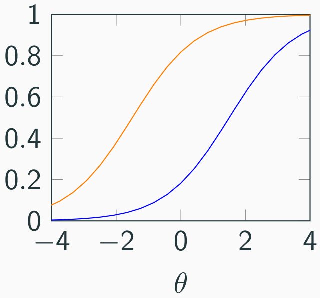
IRT: 1 Parameter Logistic Model (1PL)

Also known as *Rasch model*

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

θ_j: latent ability

b_i: difficulty



Parameter Estimation

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

$$p(y_{ij} = 0 | b_i, \theta_j) = 1 - p(y_{ij} = 1 | b_i, \theta_j)$$

$$L = \prod_{j=1}^{J} \prod_{i=1}^{I} p(Y_{ij} = y_{ij} | b_i, \theta_j)$$

$$q(\Theta, B) = \prod_{j} \pi_j^{\theta}(\theta_j) \prod_{i} \pi_i^{b}(b_i)$$

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

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

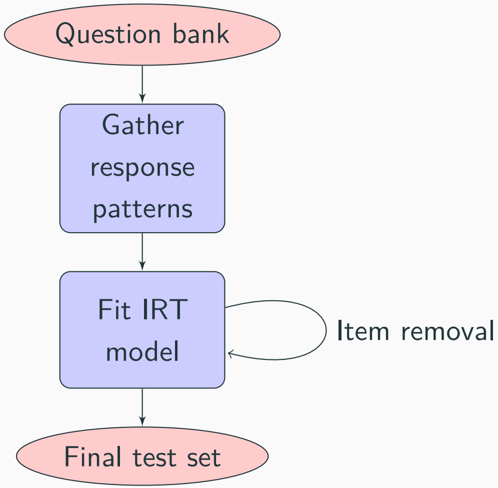
Item Set	Theta Score	Percentile	Test Acc.
5GS			
Entailment	-0.133	44.83%	96.5%
Contradiction	1.539	93.82%	87.9%
Neutral	0.423	66.28%	88%
4GS			
Contradiction	1.777	96.25%	78.9%
Neutral	0.441	67%	83%

Human Bottleneck

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

Building IRT Models with Artificial Crowds

Building IRT Test Sets

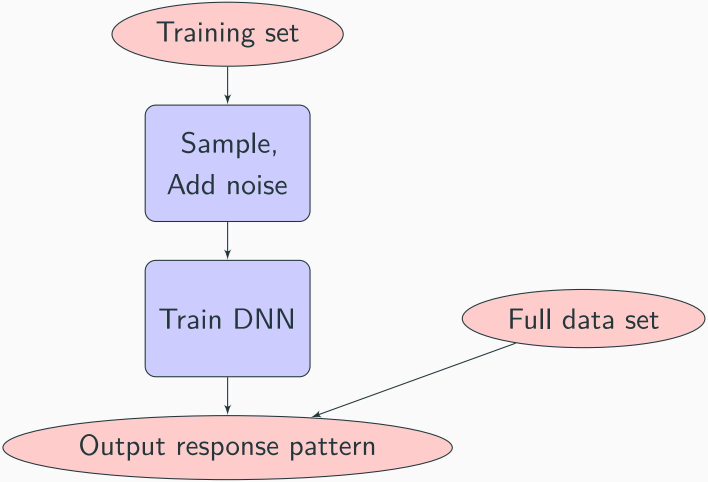


IRT for NLP: Human Annotations

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

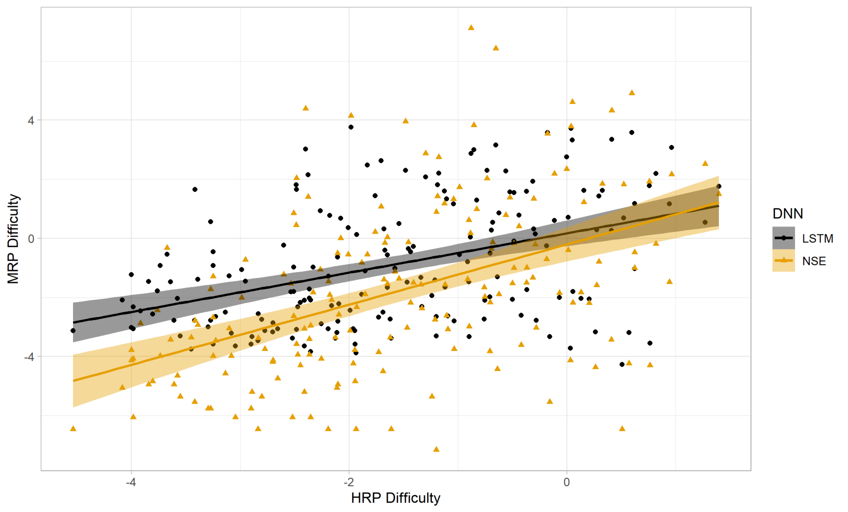
Artificial Crowd Construction



Experiments

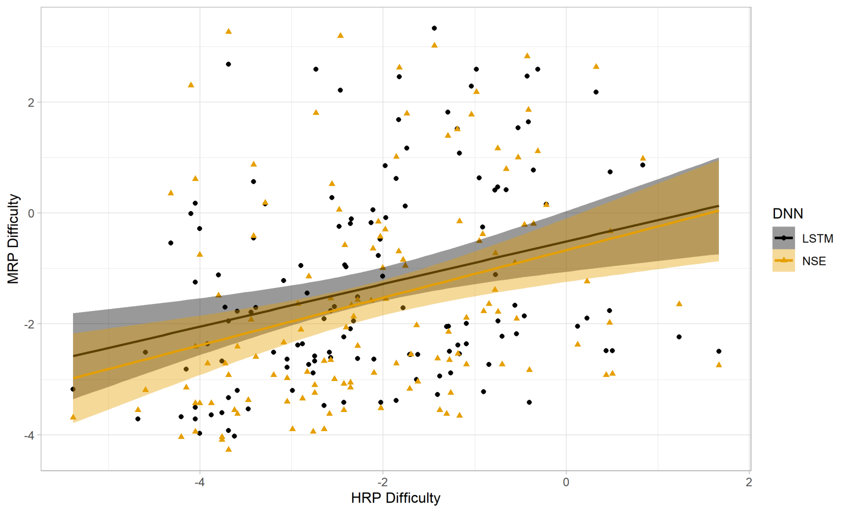
- Parameter comparison between models fit with human and machine response patterns
- Downstream use-case: training set filtering
- Qualitative evaluation: how do they look?

Human-Machine Correlation



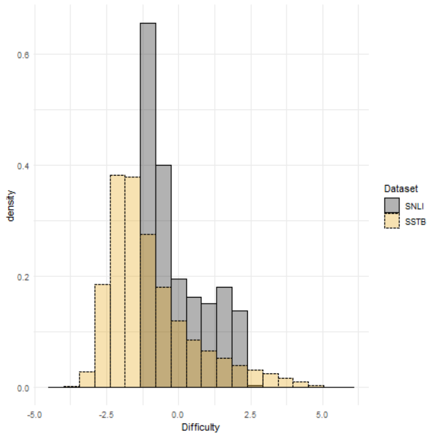
- Spearman ρ (NLI): 0.409 (LSTM) and 0.496 (NSE).

Human-Machine Correlation

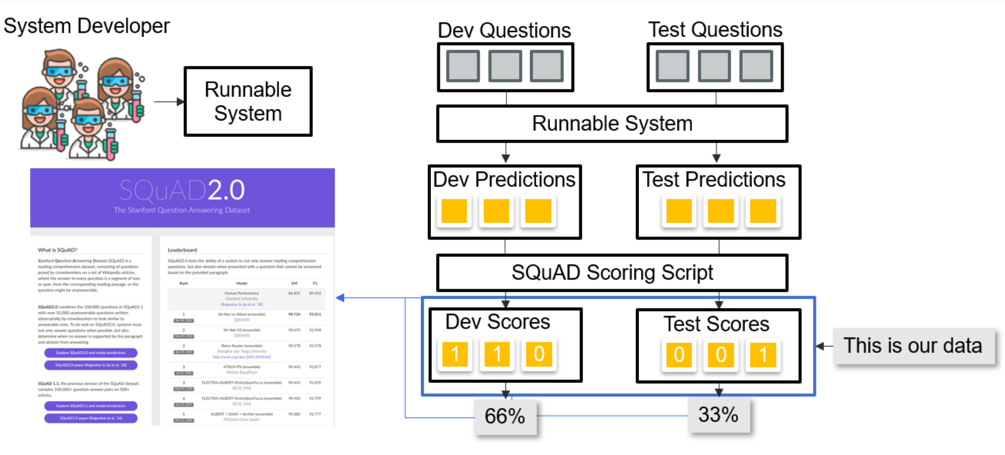


- Spearman ρ (SA): 0.332 (LSTM) and 0.392 (NSE).

Difficulty Distribution

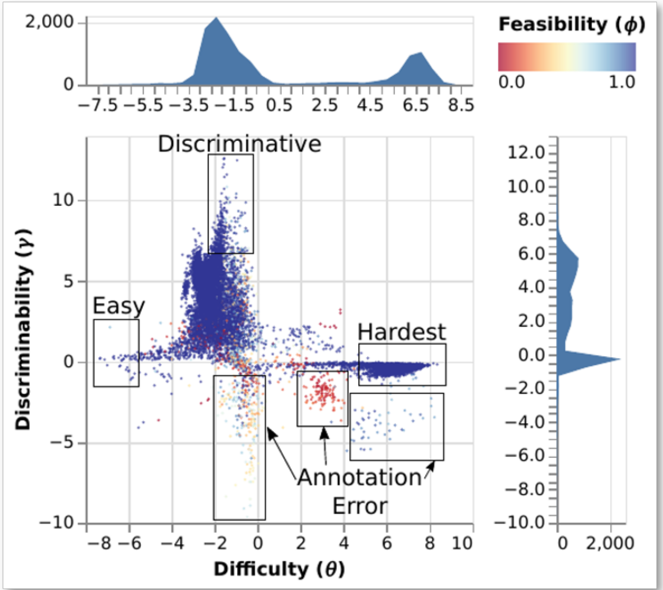


IRT for Leaderboards (SQuAD)

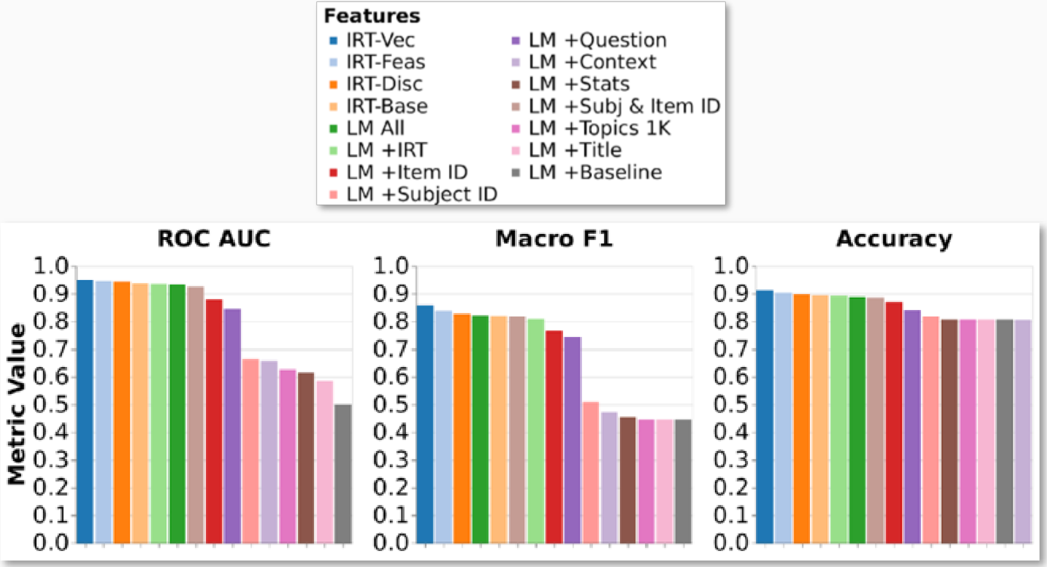


- 1.9 million subject-item pairs

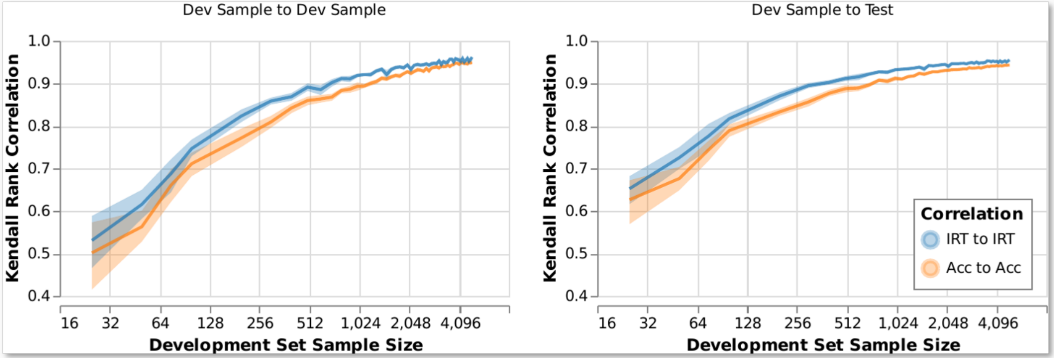
IRT for SQuAD



Predicting Correct Responses



Ranking Performance



IRT in Python: py-irt

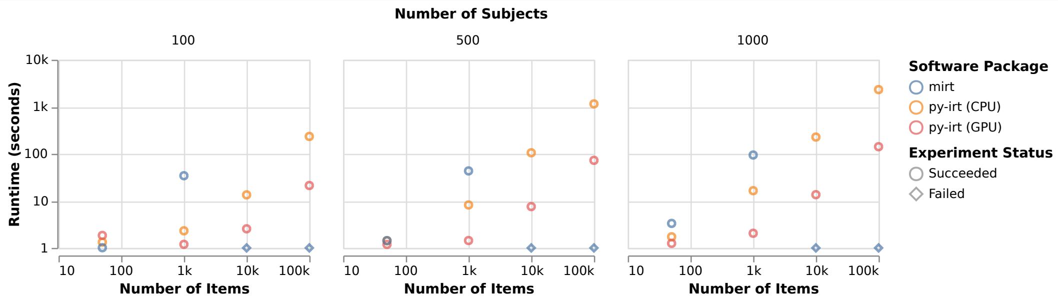

```
{"subject_id": "pedro",    "responses": {"q1": 1, "q2": 0, "q3": 1, "q4": 0}}
{"subject_id": "pinguino", "responses": {"q1": 1, "q2": 1, "q3": 0, "q4": 0}}
{"subject_id": "ken",      "responses": {"q1": 1, "q2": 1, "q3": 1, "q4": 1}}
{"subject_id": "burt",     "responses": {"q1": 0, "q2": 0, "q3":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

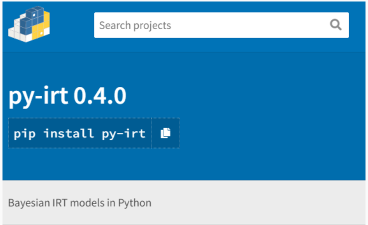

```
py-irt train 1pl data/data.jsonlines output/1pl/
```


```
{
  "ability": [
    -1.725112438201904З,
    -0.06789101660251617,
    1.605994105339050З,
    -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