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?

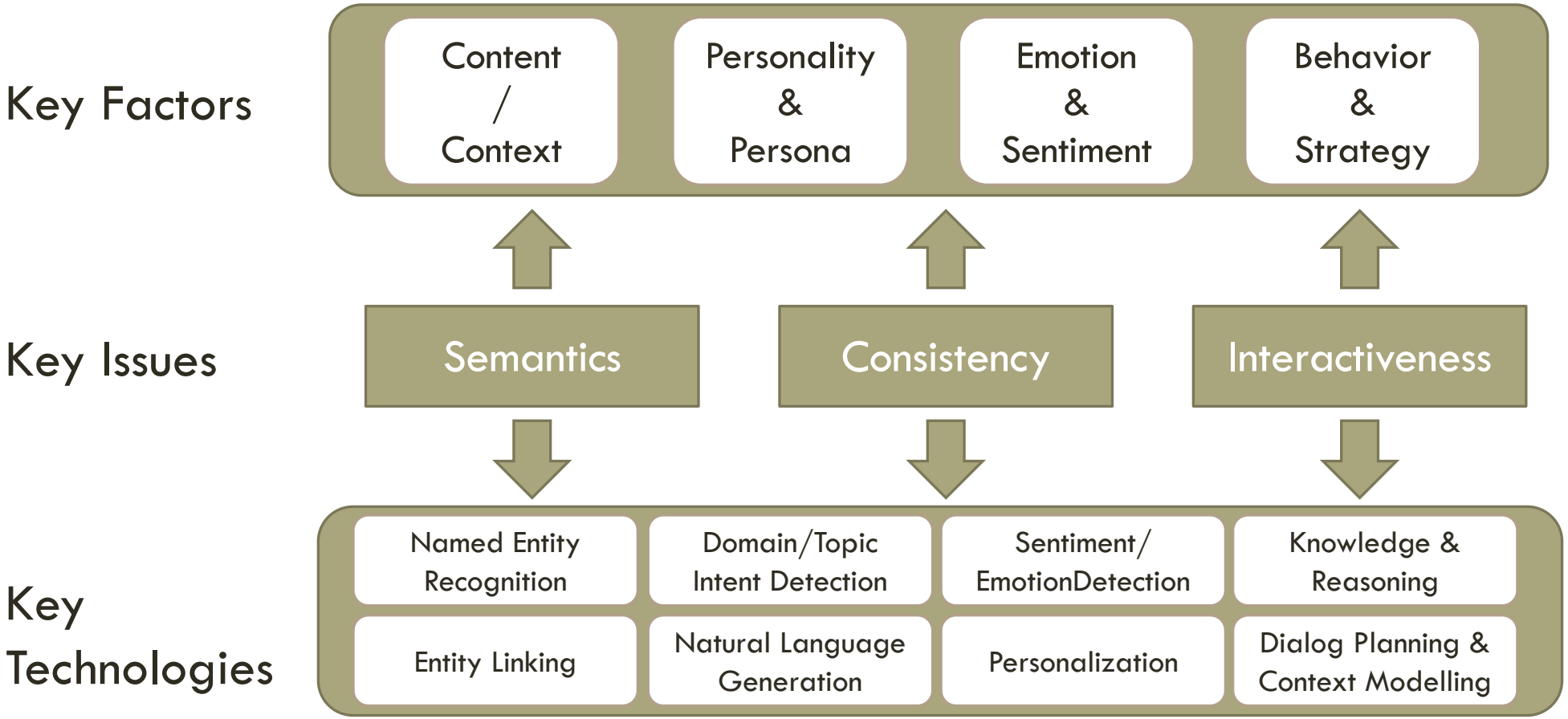
EVALUATIONS ARE AT SEVERAL LEVELS

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"

COMMON TASK FRAMEWORK & LEADERBOARDS

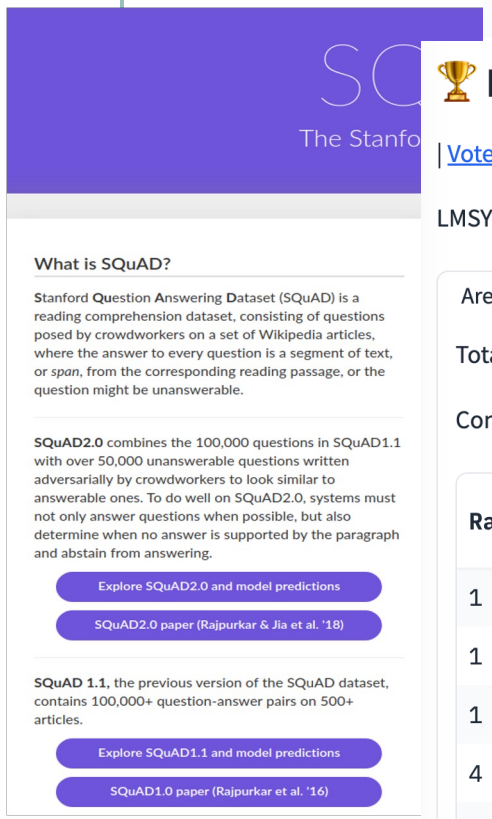
There is general agreement that these competitive evaluations had a striking and beneficial effect on the performance of various systems tested over the years. However, it is also recognized (albeit less generally) that these evaluation experiments also had the, less beneficial, effect that the participating 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



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.

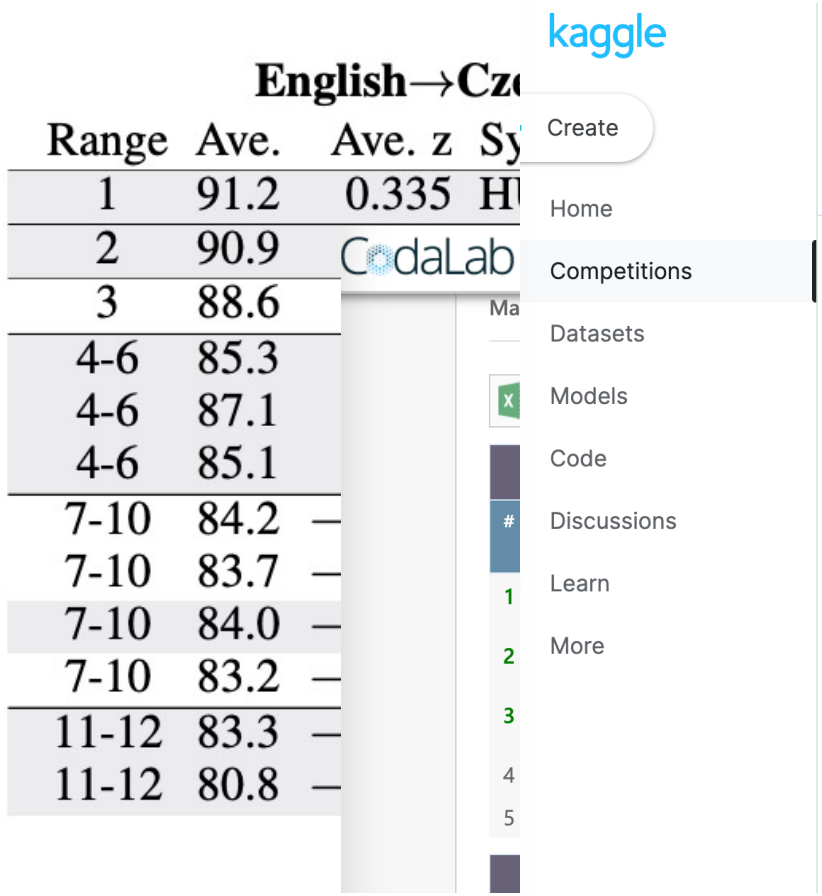
Arena Elo | Full Leaderboard

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→Cz

Range	Ave.	Ave. z	Sy
1	91.2	0.335	H
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

Ma

kaggle

Create

Home

Competitions

Datasets

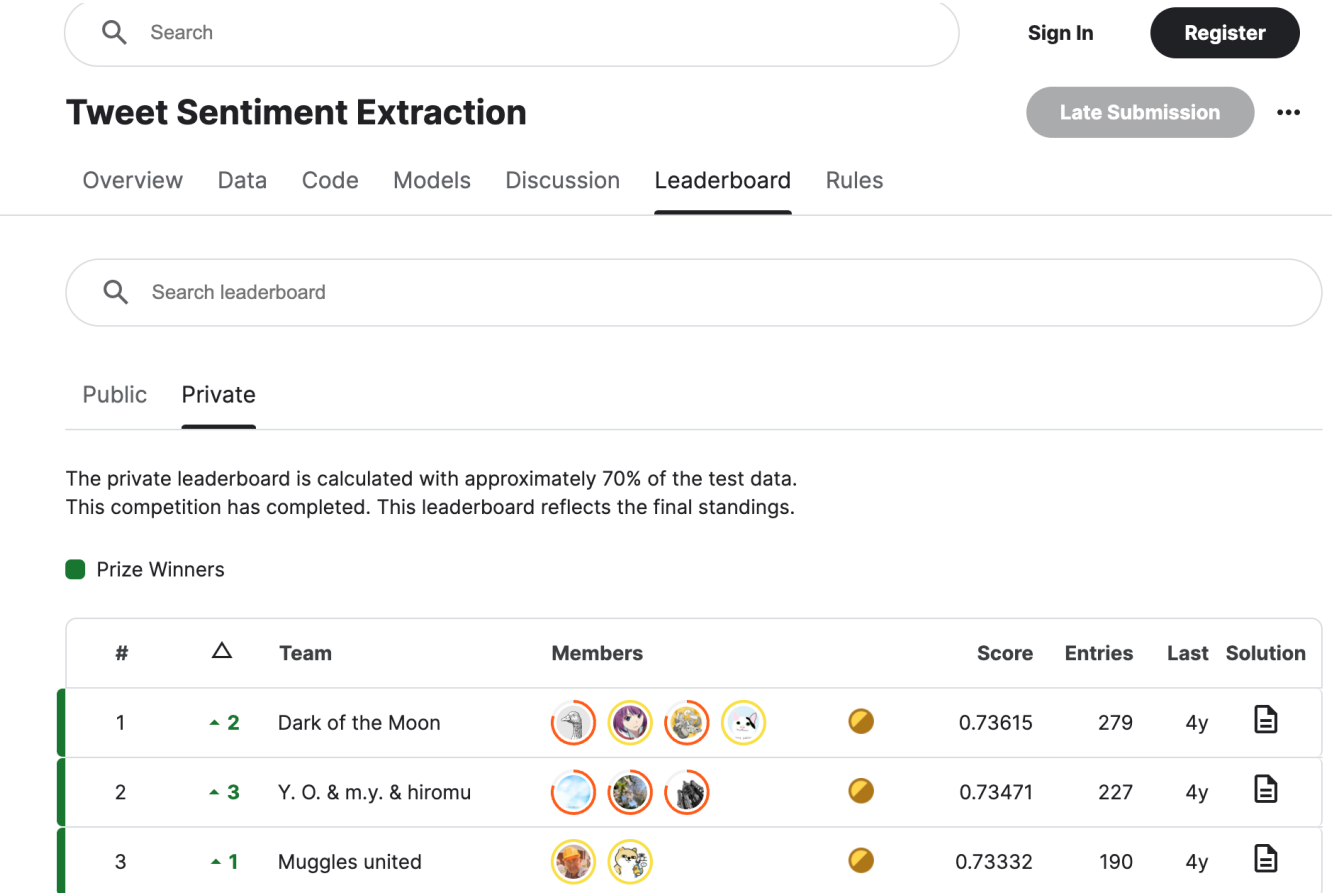
Models

Code

Discussions

Learn

More



Search

Sign In

Register

Tweet Sentiment Extraction

Late Submission

Overview Data Code Models Discussion Leaderboard Rules

Search leaderboard

Public Private

The private leaderboard is calculated with approximately 70% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Prize Winners

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 2	Dark of the Moon		0.73615	279	4y	
2	▲ 3	Y. O. & m.y. & hiromu		0.73471	227	4y	
3	▲ 1	Muggles united		0.73332	190	4y	

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

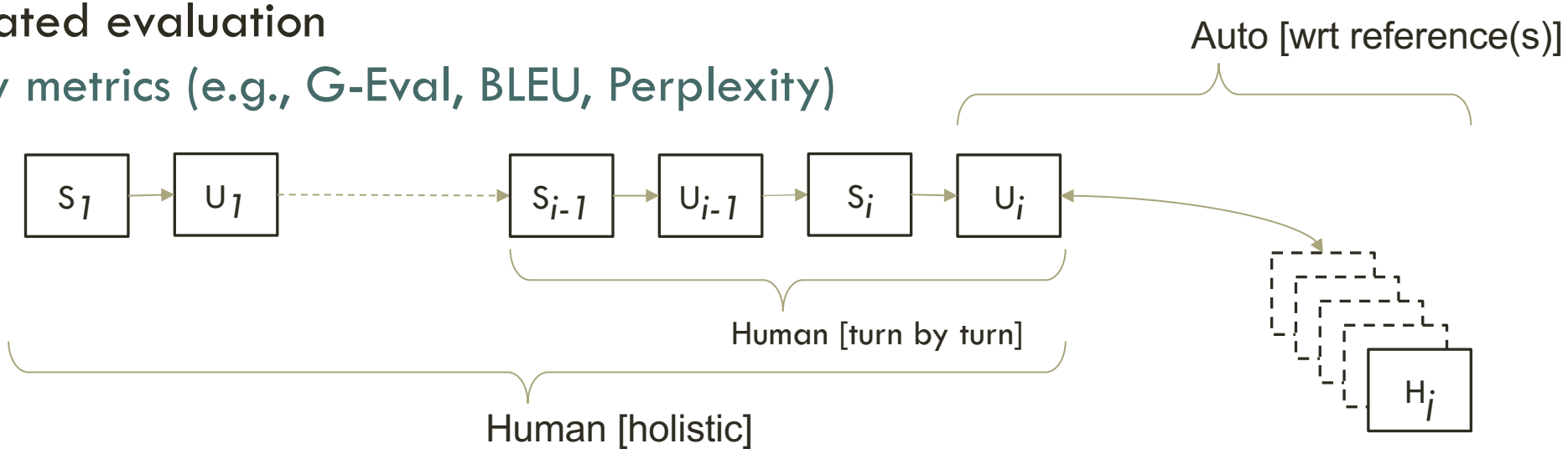
HUMAN / AUTOMATIC METRICS

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