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?

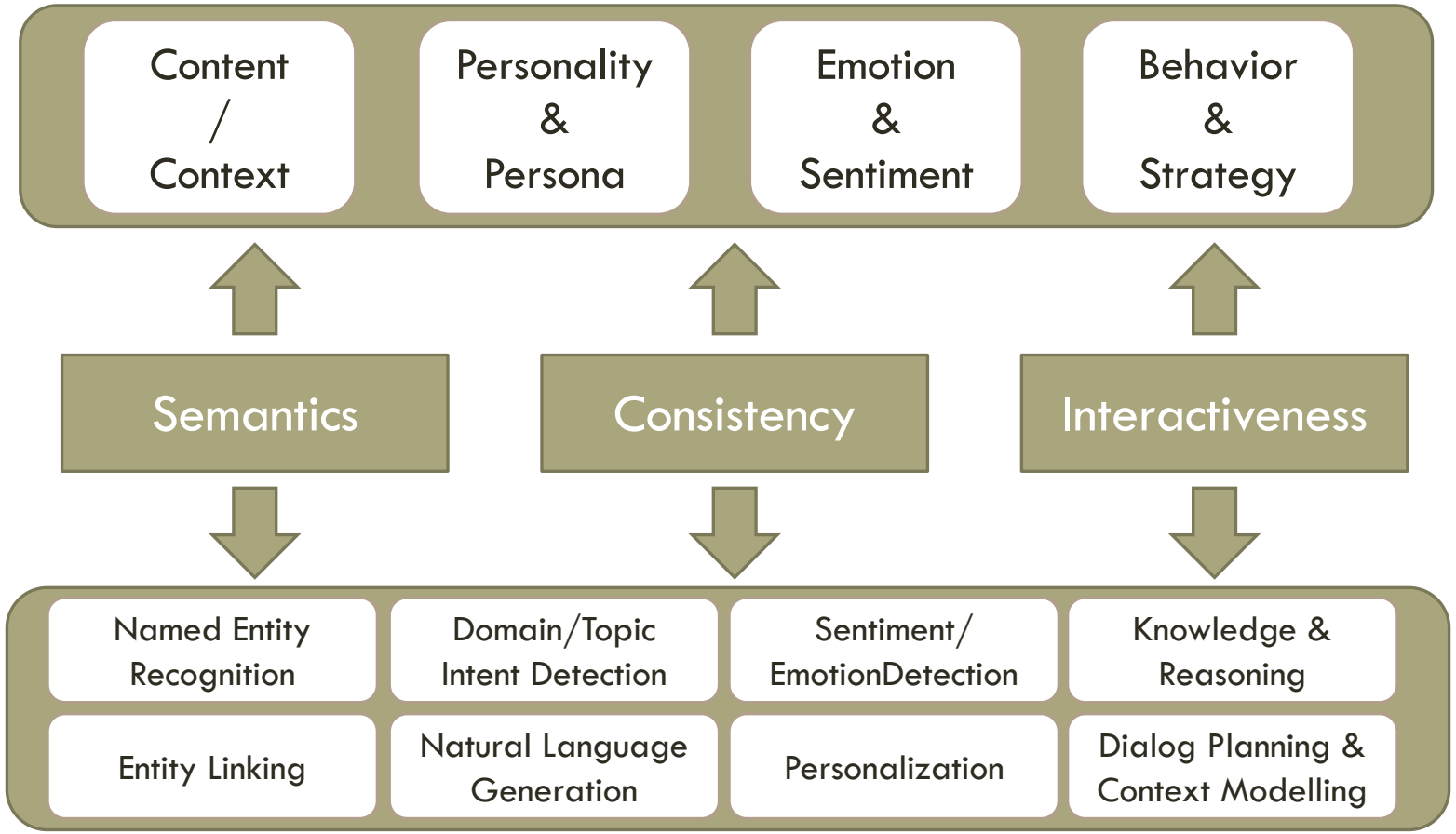
EVALUATIONS ARE AT SEVERAL LEVELS

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

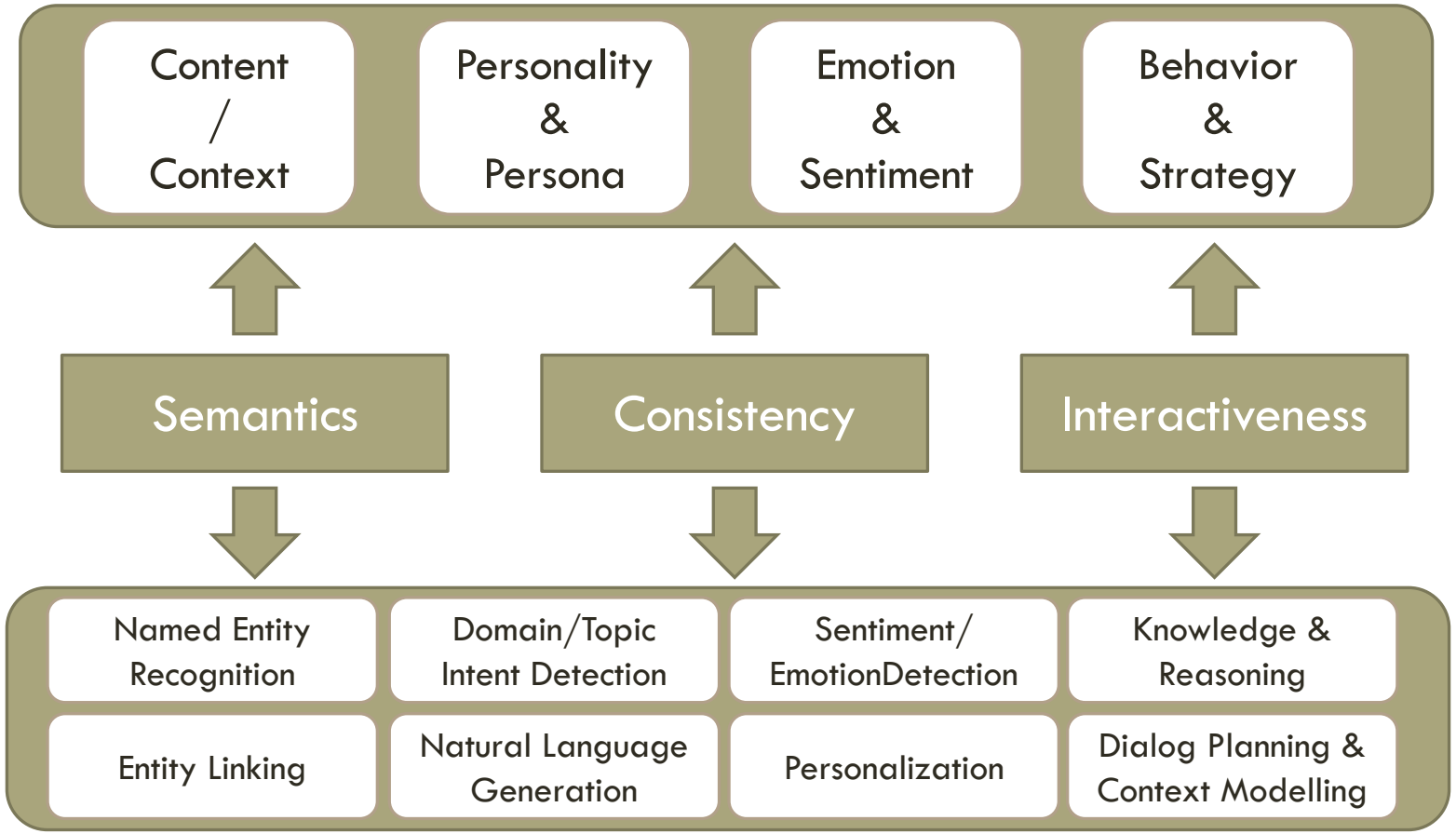
CHALLENGES FOR DIALOG SYSTEMS



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

CHALLENGES FOR DIALOG SYSTEMS

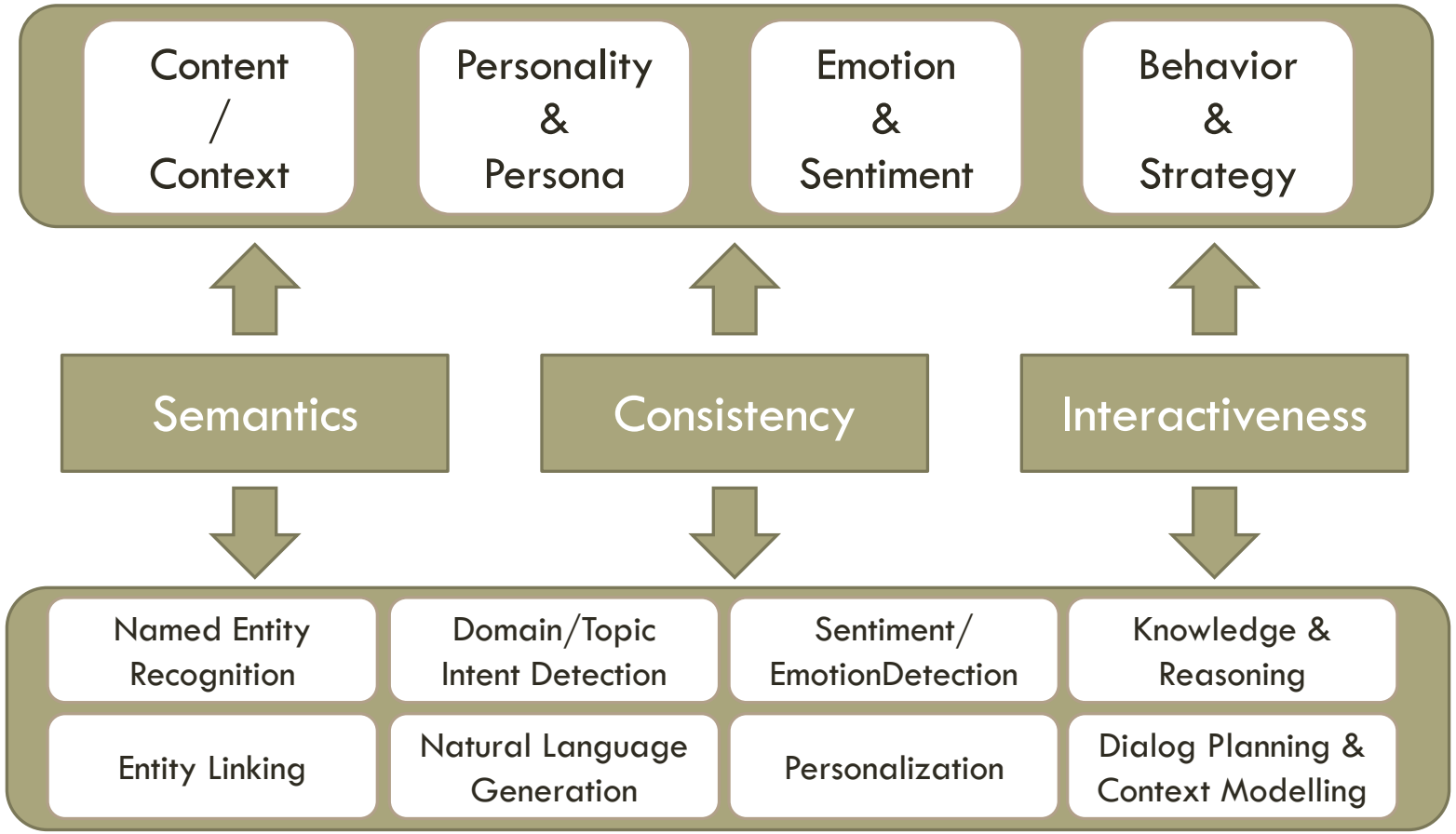
Key Issues



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

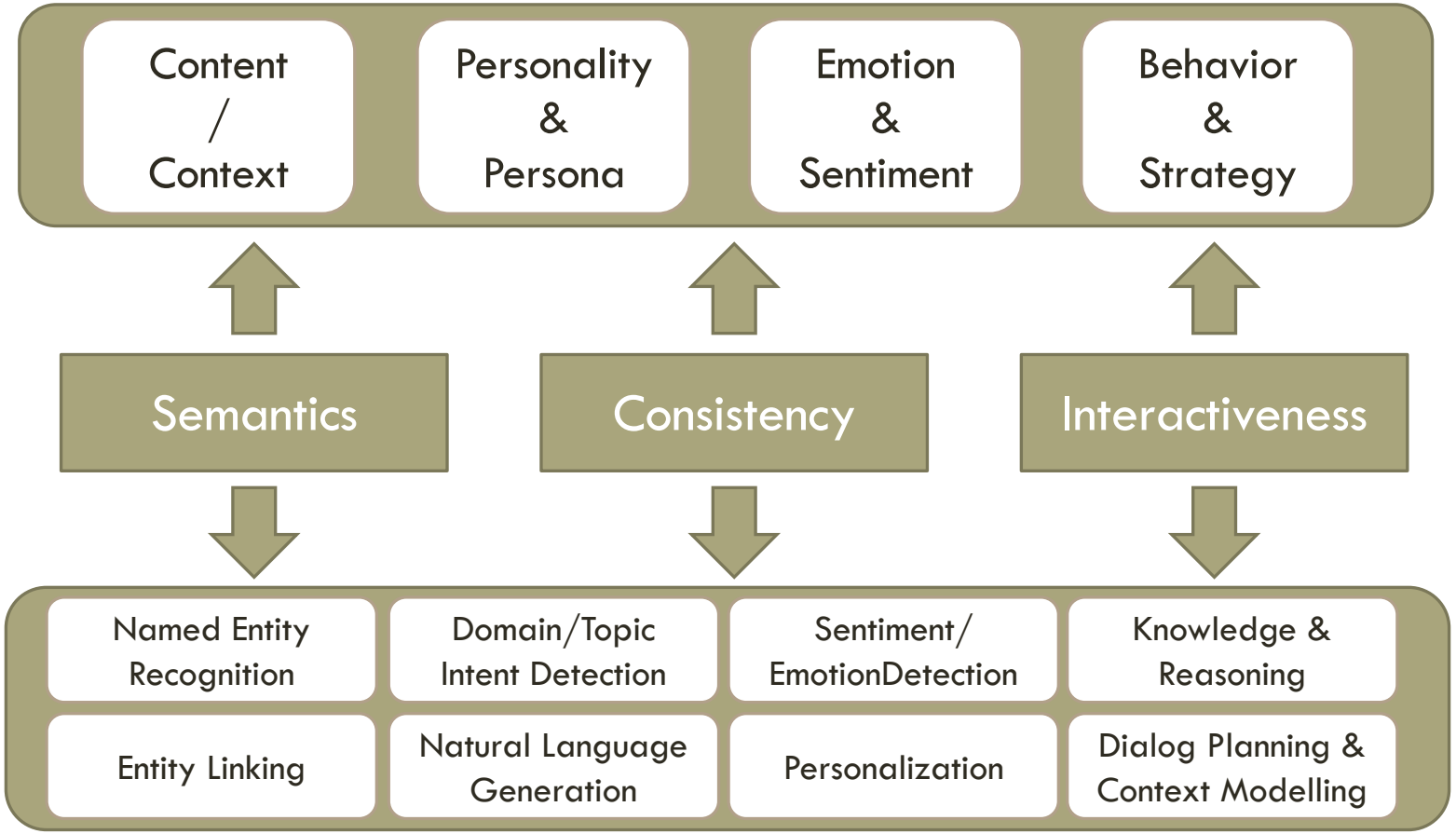
CHALLENGES FOR DIALOG SYSTEMS

Key Factors



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

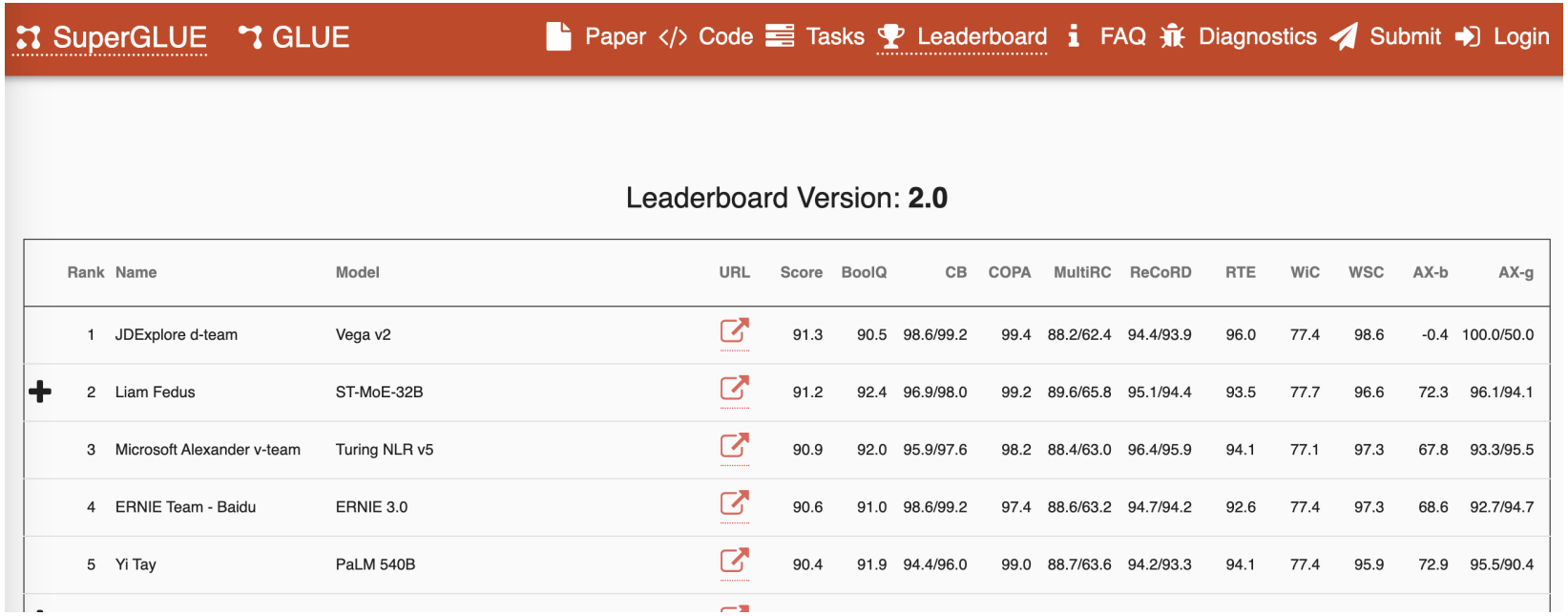
There is general agreement that these competitive evaluations had a striking and beneficial effect on the performance of various systems tested over the years. However, it is also recognized (albeit less generally) that these evaluation experiments also had the, less beneficial, effect that the participating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



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

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **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)

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

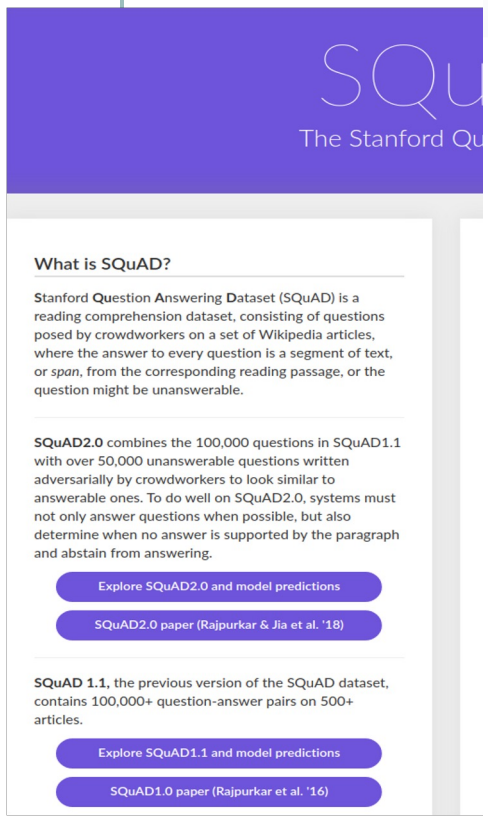
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	**90.724**	**93.011**
2 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
2 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* 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

Leaderboard | FAQ | Diagnostics | Submit | Login

2.0

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



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

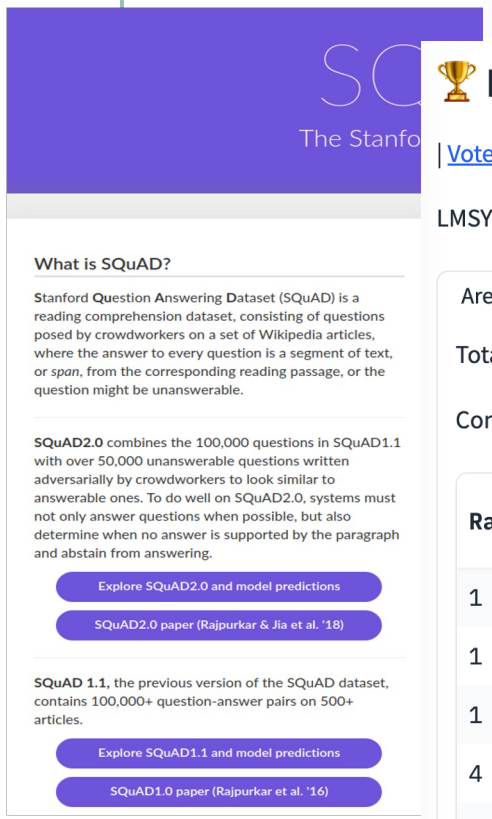
Overall | Bitext Mining | Classification | Clustering | Pair Classification | Retrieval | Reranking | STS | Summarization

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



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

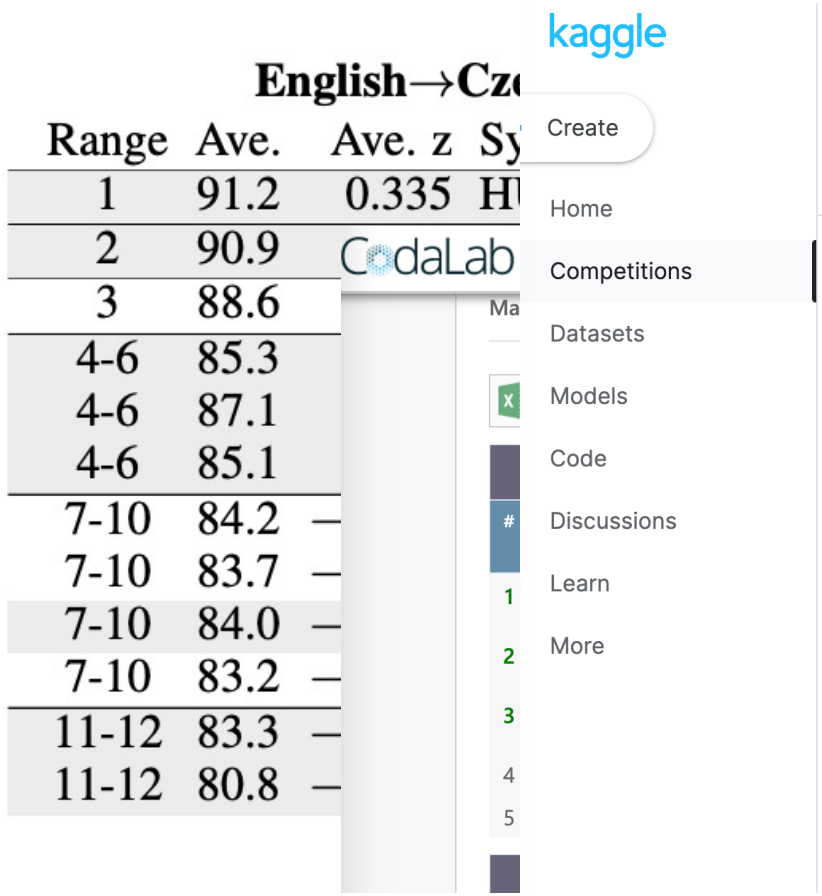
Download CSV

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



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

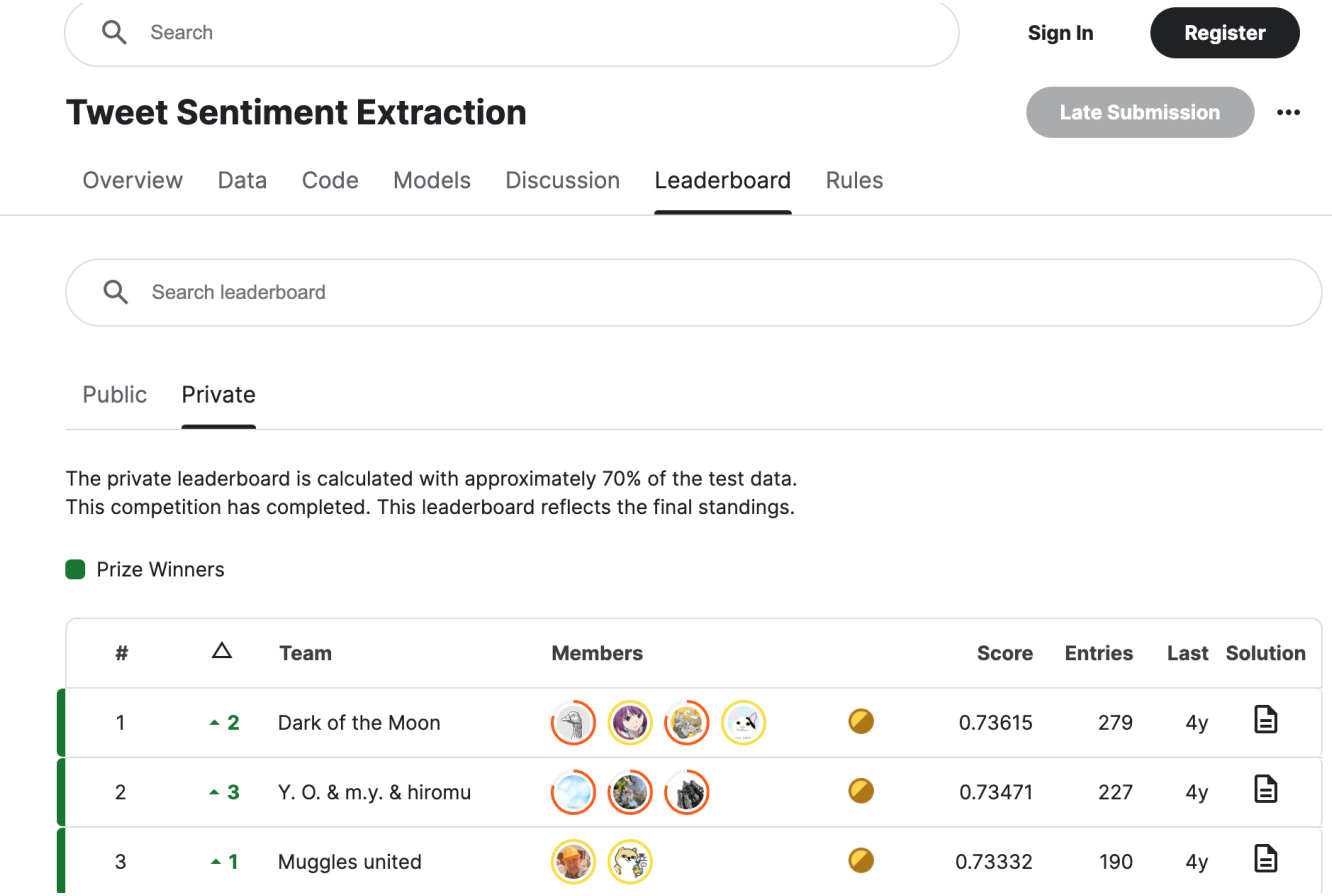
Models

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

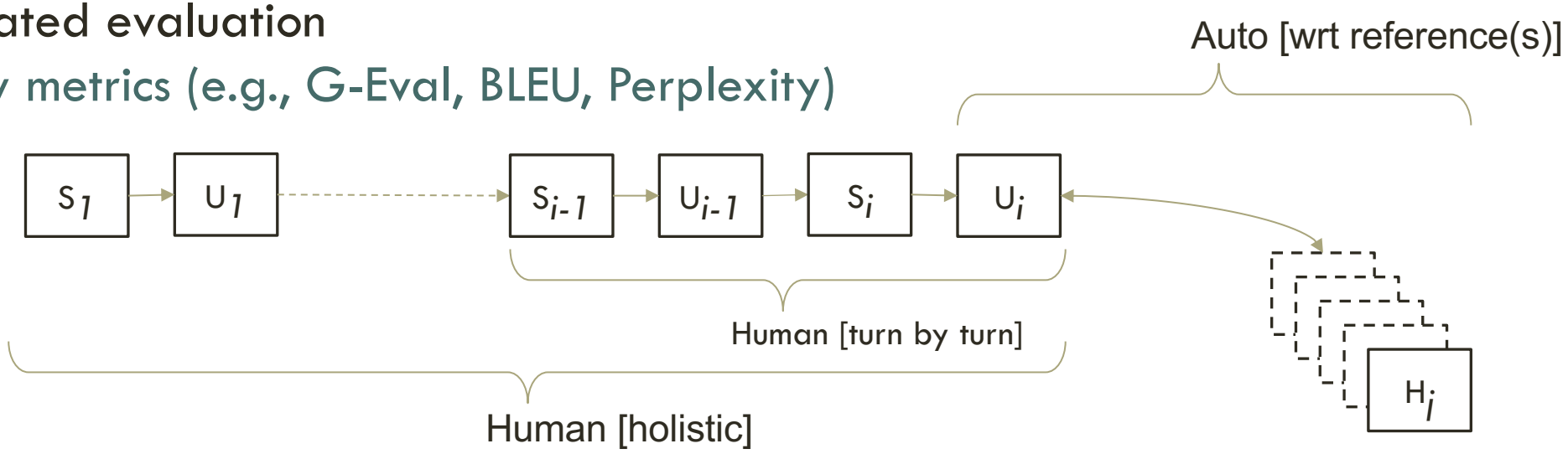
HUMAN / AUTOMATIC METRICS

Human evaluation

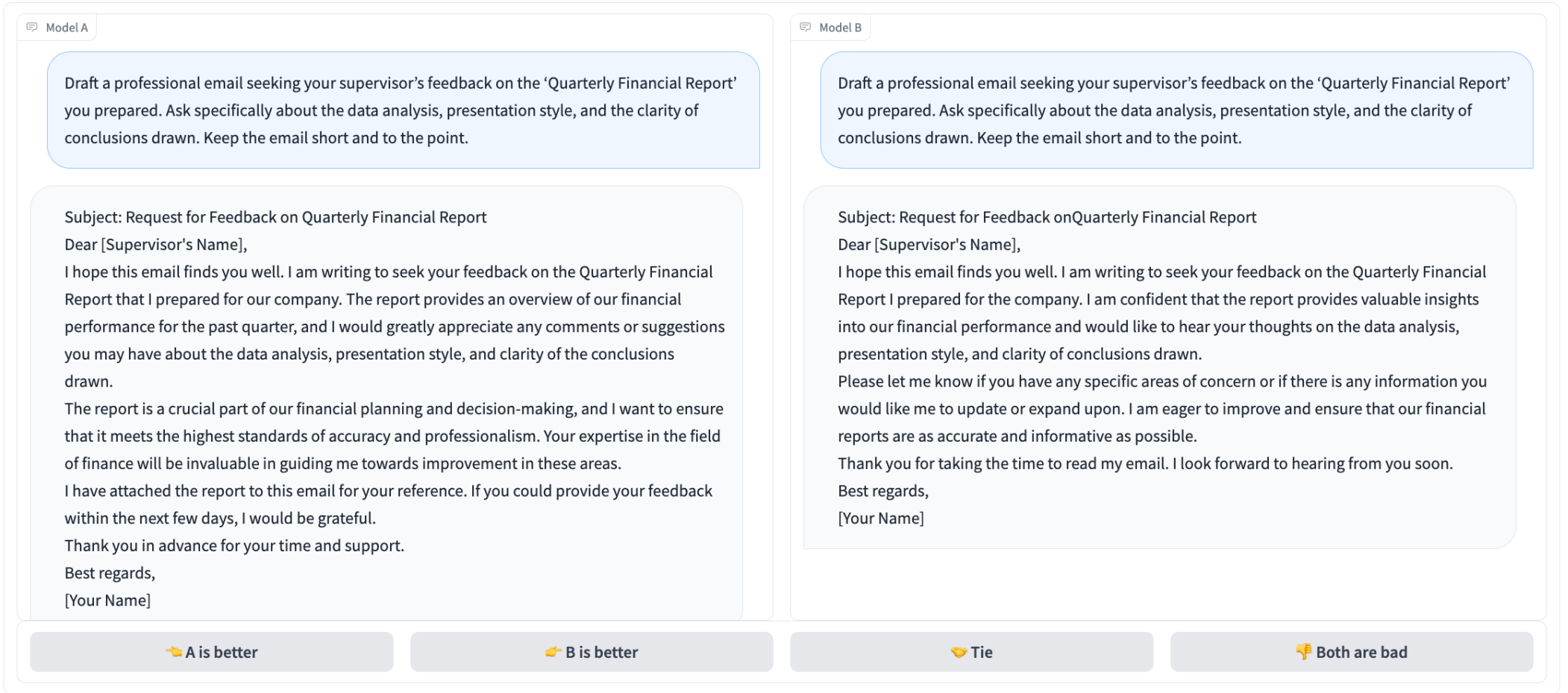
- Expert judges (WOCHAT, Alexa)
- Crowd-sourced (non-expert) judgments (DBDC)

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