Please join the github org:
https://github.com/Robust-AI-Bench/Perpetual-Multiparty-Evaluation

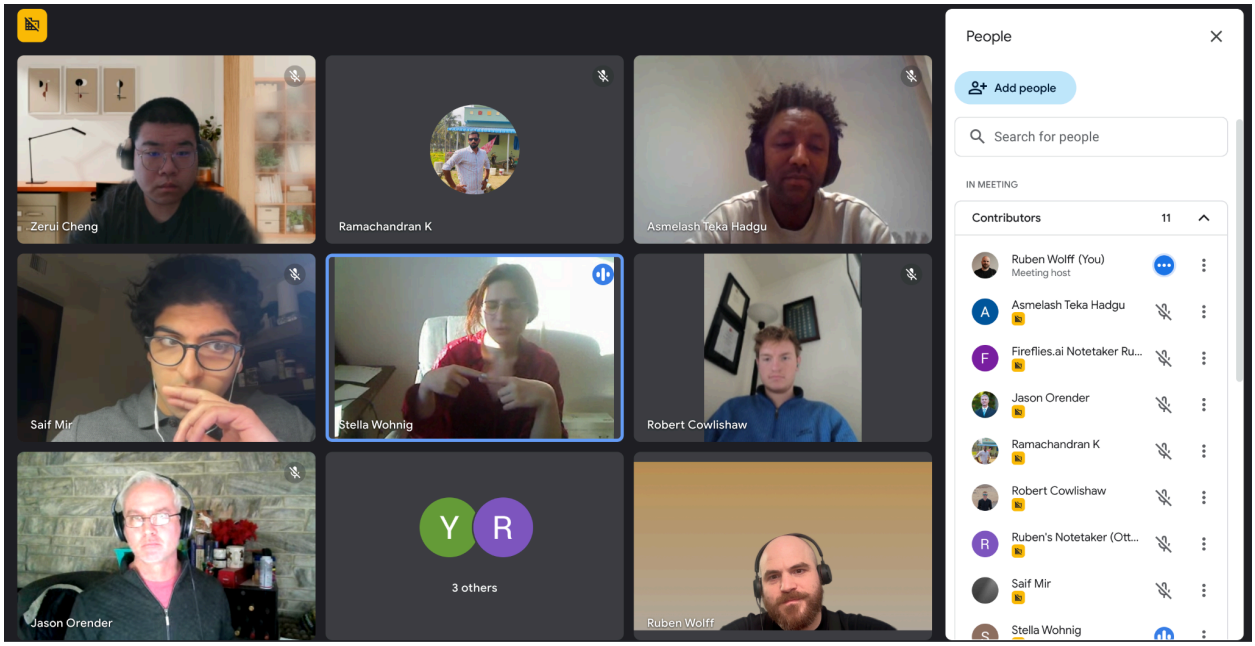
Feb 18, 2025 Meeting

Audio Recording MP3 HERE
Meeting Transcript JSON PDF
Firefly Audio + Transcription synced [requires login]

Agenda

1. Goal of this meeting [5min]
    a. Relatively open
    b. Figure out how to organize ourselves
        i. Make a new foundation
        ii. Make a Working group inside an existing foundation
    c. Who wants to specialize on what
        i. Development / Practice
        ii. Theory / Academic
        iii. Organizational / Political
    d. Start IMportant Decisions
        i. Design questions : how to register, how much decentralization, is it correct to stay away from blockchain
2. Who is involved [10min]
    a. https://www.linkedin.com/in/orenderjw/
    b. https://www.linkedin.com/in/ling-han-brian/
    https://www.linkedin.com/in/saif-mir-94686a314/
    https://www.linkedin.com/in/zerui-cheng/
    c. https://www.linkedin.com/in/asmelashteka
    d. https://www.linkedin.com/in/harshith-reddy-takkala/
    e. https://www.linkedin.com/in/ramakavanan/
    f. https://www.linkedin.com/in/stella-wohnig-600a3b226/
    g. https://www.linkedin.com/in/yufan-jiang-25751214a/

    h. https://www.linkedin.com/in/daniel-kirste/
    i. https://www.linkedin.com/in/robert-cowlishaw/
3. Problem Discussions [5min]
    a. Ruben
    b. Asme
4. Solution Design [15min]
    a. Existing
        i. Dynabench
        ii. CodaBench
        iii. SEAL
    b. Ours
        i. **Continuously eval, Hidden test data, multiple independent testers, closed source AI allowed**
5. Ruben's list Open questions - Solution Design. [15min]
    a. Validator > Model communication
    b. Validator Registration
    c. Model Registration
6. Org type [5min]
    a. Foundation
    b. Consortium
    c. Do we write a charger on goals
7. How to work together [5min]
    a. Async on github
    b. Async on Google docs
    c. Synchronous meetings
    d. Slack / Discord



Ruben's meeting notes Feb 18, 2025 :

Stella brings up the issue of some companies might not want to give out their models for public testing. < Ruben response, they could choose to only

Jason : maybe having anonymouse testing could be value add for the private models.

Stella: maybe don't put your whole model on-chain.

Zerui : We need to automate the evaluation. IT can't just be humans following

Zerui : there is a lot of high quality data in AI conferences

Jason : Establish credibility with publishing 1 paper first

Asme: He agrees we should gather existing scattered datasets. B/C our benchmarks should be compared to existing test benchmarks

Jason : Maybe academics will want to show the world that they should bring their test datasets to show their test dataset is good. Accurately predict how well it generatlzies

Jason : Smaller test data creation companies could use this to prove to the world that they are producign good test dataset. Maybe much cheaper than what scale.ai does

Stella : there she thinks there is issue

Robert : Go build first don;'t worry about the org.
Asme : Lets just start prototyping , And start writing . Once we have build something.
Saif : I also agree build first. Don't come to orgs without something to show.

Jason: Prefer colab over github. Has assignments included. Create cards.

Name ideas :

- Robust Multi Party Rolling Evaluation (RMRE)
- Perpetual Multiparty Evaluation (PMRE)
- Perpetual Decentralized Evaluation (PDE)
- Robust Generalization Error (RGRE)
- Perpetual Blind Generalization Error (PBGE)

Problem Statement

The problem: Public benchmark test data sets make AI model performance comparable. But this creates an incentivization for closed source models in particular to game the benchmarks by creating heuristics for them or overfitting their training data to include solutions to the known testsets.

Our proposed solution : A network of independent experts evaluating models continuously with newly generated private test data.

Related Projects

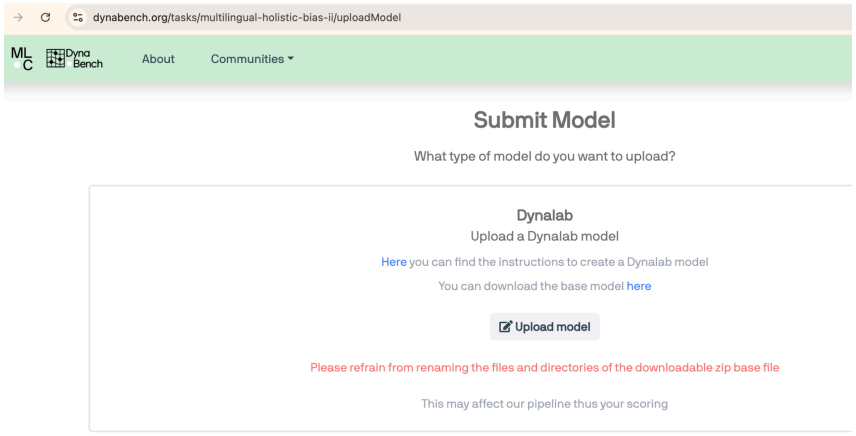
CodaBench - Private datasets

https://www.codabench.org/
https://github.com/codalab/codabench
https://github.com/codalab/codalab-competitions/wiki/Community-Governance

https://github.com/mlcommons/dynabench/

Dynabench: Rethinking Benchmarking in NLP

https://github.com/mlcommons/dynabench
Douwe Kiela, Adina Williams https://arxiv.org/abs/2104.14337
https://dynabench.org/tasks/multilingual-holistic-bias-ii
They talk about some of the problems we want to address and solve them by letting people make new data. They are going the approach of having times by which one must submit their models and times by which one must submit data. I THINK this is not for closed source



https://www.linkedin.com/company/mlcommons/people/ 71 people working in ML commons on this but they have more than dynabench

https://github.com/mlcommons/dynabench/graphs/contributors anyone contributing to this could be interesting to talk to

Dynabench: Rethinking Benchmarking in NLP https://arxiv.org/abs/2104.14337
Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, Adina Williams
We introduce Dynabench, an open-source platform for dynamic dataset creation and model benchmarking. Dynabench runs in a web browser and supports human-and-model-in-the-loop dataset creation

EvalAI: Towards Better Evaluation Systems for AI Agents

https://eval.ai/?t
http://learningsys.org/sosp19/assets/papers/23_CameraReadySubmission_EvalAI_SOSP_2019%20(8)%20(1).pdf
Deshraj Yadav1 Rishabh Jain1 Harsh Agrawal1 Prithvijit Chattopadhyay1 Taranjeet Singh2 Akash Jain3 Shiv Baran Singh4 Stefan Lee1 Dhruv Batra1 1Georgia Institute of Technology 2Paytm 3Zomato 4Cyware

Deshraj Yadav[1] **Rishabh Jain**[1] **Harsh Agrawal**[1] **Prithvijit Chattopadhyay**[1]
Taranjeet Singh[2] **Akash Jain**[3] **Shiv Baran Singh**[4] **Stefan Lee**[1] **Dhruv Batra**[1]
[1]Georgia Institute of Technology [2]Paytm [3]Zomato [4]Cyware

Features	OpenML	Topcoder	Kaggle	AICrowd	ParlAI	CodaLab	EvalAI
AI Challenge Hosting	✗	✓	✓	✓	✗	✓	✓
Custom metrics	✗	✗	✗	✓	✓	✓	✓
Multiple phases/splits	✗	✗	✗	✓	✗	✓	✓
Open Source	✓	✗	✗	✓	✓	✓	✓
Remote Evaluation	✗	✗	✗	✗	✓	✓	✓
Human Evaluation	✗	✗	✗	✗	✓	✗	✓
Environments	✗	✗	✗	✓	✗	✗	✓

Table 1. Head-to-head comparison of capabilities between existing platforms and EvalAI

Looks like they are trying to create a sort of trusted compute environment. I guess this could work for some closed source things but not sure if that black box image could not leak the test data

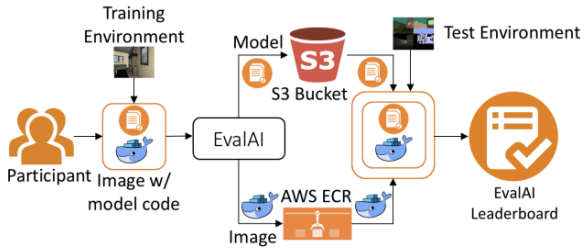


Figure 2. System architecture for code upload challenges

The Benchmark Lottery https://arxiv.org/abs/2107.07002
Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, Oriol Vinyals

https://arxiv.org/abs/2104.02145
What Will it Take to Fix Benchmarking in Natural Language Understanding?
Samuel R. Bowman, George E. Dahl
Evaluation for many natural language understanding (NLU) tasks is broken: Unreliable and biased systems score so highly on standard benchmarks

ForecastBench - future events only

https://www.forecastbench.org/ https://openreview.net/forum?id=IfPkGWXLLf&t

This drastically reduces scope but its cool

SEAL - Centralized Private eval never public

https://scale.com/leaderboard/spanish?utm_campaign=The%20Batch&utm_source=hs_email&utm_medium=email
Developed by Scale's Safety, Evaluations, and Alignment Lab (SEAL)
Developed by Scale's Safety, Evaluations, and Alignment Lab (SEAL)
https://scale.com/leaderboard

Other related link

Maybe related

- Fraunhofer's AI-DAPT https://www.fokus.fraunhofer.de/en/DPS/projects/AI_DAPT?t
- Anthropic " A new initiative for developing third-party model evaluations" https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations?t
- NIST idk https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv?t
- LNE AI Evaluation Laboratory ; French National Laboratory of Metrology and Testing (LNE) https://www.lne.fr/en/testing/evaluation-artificial-intelligence-systems?t
- Toloka's Deep Evaluation Platform https://toloka.ai/evaluation?t
- [No Author Name….] Private Benchmarks for Fairer Tests https://www.deeplearning.ai/the-batch/issue-254/?t

---

Recent collusion events (discussion section):

https://analyticsindiamag.com/ai-news-updates/openai-just-pulled-a-theranos-with-o3/
OpenAI Just Pulled a Theranos With o3

Movement / Community / Organizational considerations

? What decisions should we make so that **academics** want to participate in test data making and scoring?

- We could bring out quarterly publications where each person that contributed test data gets authorship.

? What decisions should we make so we can attract **test data companies** like 'scale' https://scale.com/leaderboard ?

- TODO go find people that build SEAL inside SCALE
- Maybe scale does not want people auditing their scores ?

?What decisions should we make so that we attract people that are already in the process of releasing a test dataset such that they first gradually release it not all at once?

? AI Startups ?

- Privacy of b2b models
- Asme: AI startups feele like they are not being treated fairly b/c testdatasets are gamed. If you create something that Can not be gamed and you know you have the best models then they will jump

---

Framework Design Questions

1. Validators announcing they are open to validating
    a.

    2. [could be registered on blockchain easily]...
    3. None blockchain
        a. Centralized server controlled by an NGO
    4. Blockchain Near
        a. Insert into TableLand table (blockchain adjacent, use eth key identity)
        b. Insert into Ceramic table (blockchain adjacent, use eth key identity)
    5. DID registry
        a. I have used ION and DID-DHT they come and go few are stable DHT probably the most stable since its using mainline BitTorrent DHT

    b.
6. Validators communicating with Providers
    a.

    7. REST API
        a. ^Probably preferred. If we assume all providers are companies this should not be much of a problem. We are not dealing with things that can get censored here,.
    8. P2P
        a. IROH

    b.
9.
10. Providers giving Validators credentials to query them for free ?
    a. Robert : thinks …
11. Can the providers choose which validators get to query them ? < yes
12. Validators locally have a way to insert test data that it will keep and slowly send out to providers.
13. Validators locally collect AI responses
14. Validators locally grade the quality of results [how this works is task specific and needs to have a separate frontend for each]
15. ?Should we by default copy over all the prompts and answers 1 validator got to all other validators so that they can also score the results of this prompt?
16. The work of generating test data and answering it has already been done
17. This way we immediately get a validator cross agreement metric
18. Downside : more work for validators that might not be required

19. Provider signatures on responses should include the prompt.
    a. (Can one simply extract the signature from REST HTTPS certificate)? I think yes IF the response includes a copy of the prompt
        i. ?How can we prove that the clock on the HTTPS repose signature was not wrong. AKA the provider answered late.
            1. ^Firstly the validator would not accept it
        ii. ?How can we prevent a validator from censoring responses ?
20. ?Do validators need to pre-register publicly hashes of the prompts that they will send. The idea would be that they then can not censor prompt/answer pairs where their favorite providers get worse scores than others. ?
21. ?Should we by default provide an anonymization layer for the API calls to come from a proxy of ours. So its harder for providers to see if a prompt is coming from one of the validators.
22. Sub Tool- Super Broadcast: Announce something to the world on as many channels as possible. So that you can be sure that you will be heard. So that you are never dependent on one system.
    a. [Inspired by IROH discovery service. Maybe we could use this Rust module https://www.iroh.computer/docs/concepts/discovery