AI Evaluation Authorities

Ari Chadda[1], Sean McGregor[2], Jesse Hostetler[2], Andrea Brennen[1]

[1] IQT Labs, [2] Digital Safety Research Institute

***Evaluation Authority**: A **programmatic** and **secured** instantiation of one or more tests **maintained by a trusted organization** for the purpose of establishing and iterating safety standards and/or rankings.*

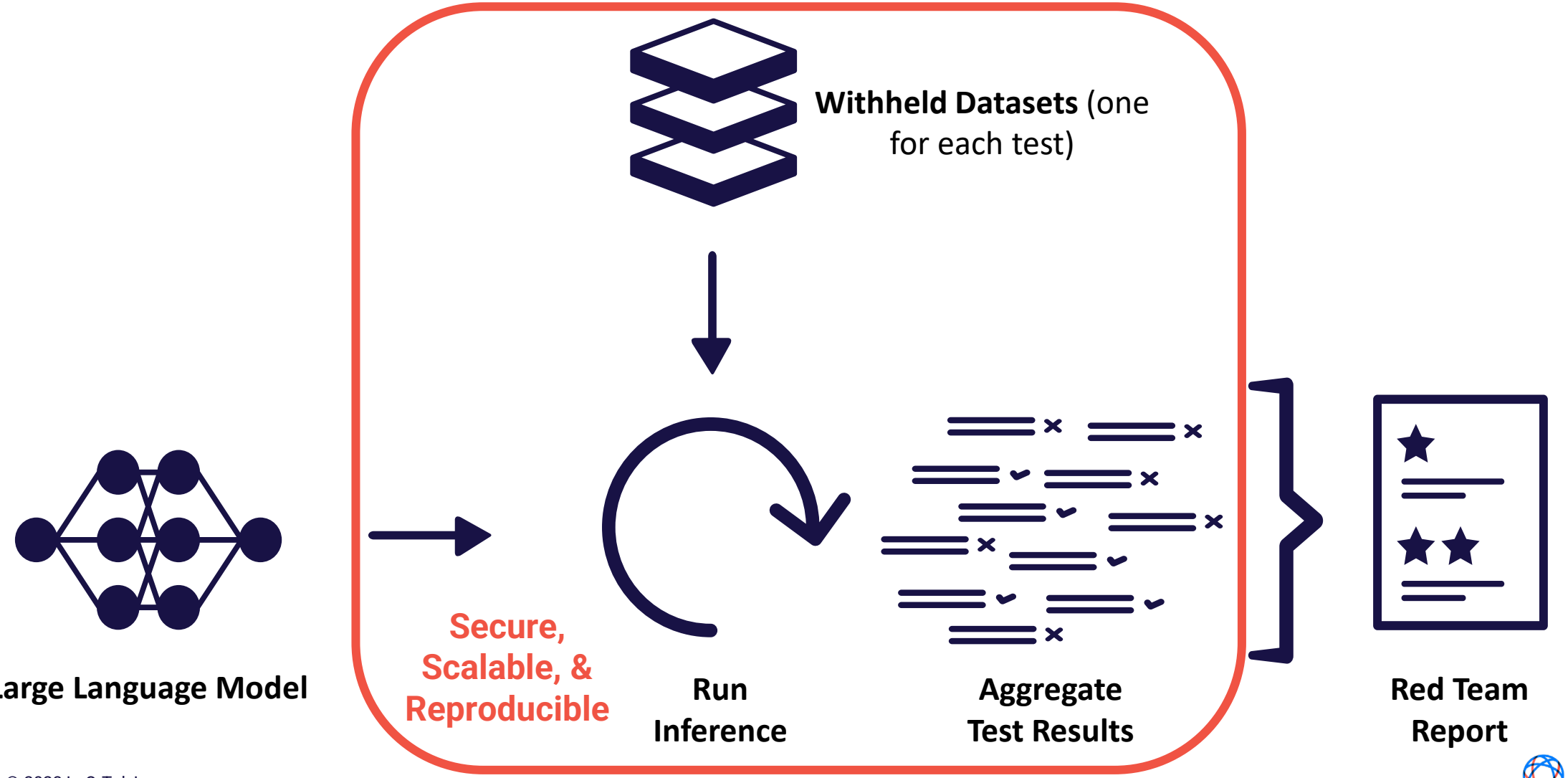
Why do we need Evaluation Authorities?

With LLMs or LLM-enabled applications:

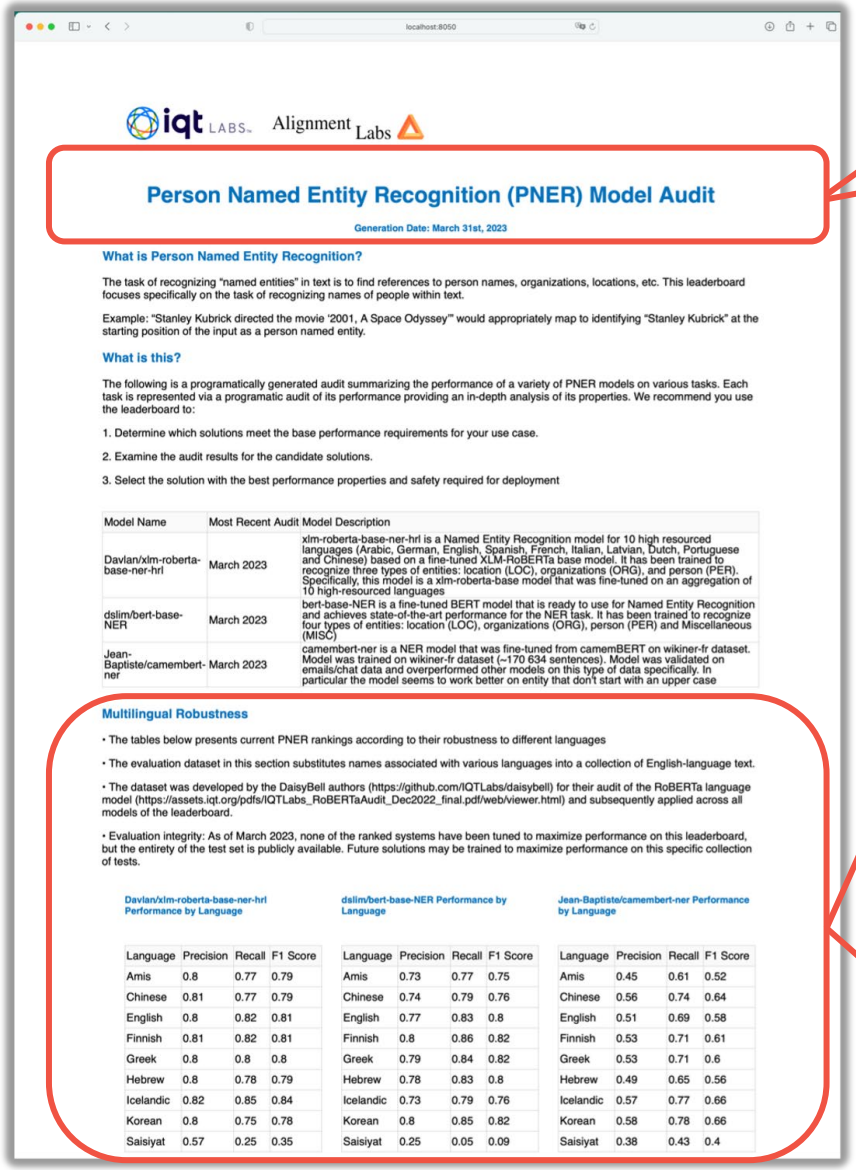
- How do you know which one to use or purchase?
- How do you know if it performs safely?
- What are its limitations?
- Is it robust to adversarial attack?
- How is it biased?
- What ethical considerations with this tool?

Evaluation Authority

”Consumer Reports” for AI Models



Proof of Concept Red Team Report



iqt LABS. Alignment Labs

Person Named Entity Recognition (PNER) Model Audit

Generation Date: March 31st, 2023

What is Person Named Entity Recognition?

The task of recognizing "named entities" in text is to find references to person names, organizations, locations, etc. This leaderboard focuses specifically on the task of recognizing names of people within text.

Example: "Stanley Kubrick directed the movie '2001, A Space Odyssey'" would appropriately map to identifying "Stanley Kubrick" at the starting position of the input as a person named entity.

What is this?

The following is a programatically generated audit summarizing the performance of a variety of PNER models on various tasks. Each task is represented via a programatic audit of its performance providing an in-depth analysis of its properties. We recommend you use the leaderboard to:

1. Determine which solutions meet the base performance requirements for your use case.
2. Examine the audit results for the candidate solutions.
3. Select the solution with the best performance properties and safety required for deployment

Model Name	Most Recent Audit	Model Description
Davlan/xlm-roberta-base-ner-hrl	March 2023	xlm-roberta-base-ner-hrl is a Named Entity Recognition model for 10 high resourced languages (Arabic, German, English, Spanish, French, Italian, Latvian, Dutch, Portuguese and Chinese) based on a fine-tuned XLM-RoBERTa base model. It has been trained to recognize three types of entities: location (LOC), organizations (ORG), and person (PER). Specifically, this model is a xlm-roberta-base model that was fine-tuned on an aggregation of 10 high-resourced languages
dslim/bert-base-NER	March 2023	bert-base-NER is a fine-tuned BERT model that is ready to use for Named Entity Recognition and achieves state-of-the-art performance for the NER task. It has been trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC)
Jean-Baptiste/camembert-ner	March 2023	camembert-ner is a NER model that was fine-tuned from camemBERT on wikiner-fr dataset. Model was trained on wikiner-fr dataset (~170 634 sentences). Model was validated on emails/chat data and overperformed other models on this type of data specifically. In particular the model seems to work better on entity that don't start with an upper case

Multilingual Robustness

- The tables below presents current PNER rankings according to their robustness to different languages
- The evaluation dataset in this section substitutes names associated with various languages into a collection of English-language text.
- The dataset was developed by the DaisyBell authors (https://github.com/IQTLabs/daisybell) for their audit of the RoBERTa language model (https://assets.iqt.org/pdfs/IQTLabs_RoBERTaAudit_Dec2022_final.pdf/web/viewer.html) and subsequently applied across all models of the leaderboard.
- Evaluation integrity: As of March 2023, none of the ranked systems have been tuned to maximize performance on this leaderboard, but the entirety of the test set is publicly available. Future solutions may be trained to maximize performance on this specific collection of tests.

Davlan/xlm-roberta-base-ner-hrl Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.8	0.77	0.79
Chinese	0.81	0.77	0.79
English	0.8	0.82	0.81
Finnish	0.81	0.82	0.81
Greek	0.8	0.8	0.8
Hebrew	0.8	0.78	0.79
Icelandic	0.82	0.85	0.84
Korean	0.8	0.75	0.78
Saisiyat	0.57	0.25	0.35

dslim/bert-base-NER Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.73	0.77	0.75
Chinese	0.74	0.79	0.76
English	0.77	0.83	0.8
Finnish	0.8	0.86	0.82
Greek	0.79	0.84	0.82
Hebrew	0.78	0.83	0.8
Icelandic	0.73	0.79	0.76
Korean	0.8	0.85	0.82
Saisiyat	0.25	0.05	0.09

Jean-Baptiste/camembert-ner Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.45	0.61	0.52
Chinese	0.56	0.74	0.64
English	0.51	0.69	0.58
Finnish	0.53	0.71	0.61
Greek	0.53	0.71	0.6
Hebrew	0.49	0.65	0.56
Icelandic	0.57	0.77	0.66
Korean	0.58	0.78	0.66
Saisiyat	0.38	0.43	0.4

Person Named Entity Recognition (PNER) Model Audit

Generation Date: March 31st, 2023

Multilingual Robustness

- The tables below presents current PNER rankings according to their robustness to different languages
- The evaluation dataset in this section substitutes names associated with various languages into a collection of English-language text.
- The dataset was developed by the DaisyBell authors (https://github.com/IQTLabs/daisybell) for their audit of the RoBERTa language model (https://assets.iqt.org/pdfs/IQTLabs_RoBERTaAudit_Dec2022_final.pdf/web/viewer.html) and subsequently applied across all models of the leaderboard.
- Evaluation integrity: As of March 2023, none of the ranked systems have been tuned to maximize performance on this leaderboard, but the entirety of the test set is publicly available. Future solutions may be trained to maximize performance on this specific collection of tests.

Davlan/xlm-roberta-base-ner-hrl Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.8	0.77	0.79
Chinese	0.81	0.77	0.79
English	0.8	0.82	0.81
Finnish	0.81	0.82	0.81
Greek	0.8	0.8	0.8
Hebrew	0.8	0.78	0.79
Icelandic	0.82	0.85	0.84
Korean	0.8	0.75	0.78
Saisiyat	0.57	0.25	0.35

dslim/bert-base-NER Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.73	0.77	0.75
Chinese	0.74	0.79	0.76
English	0.77	0.83	0.8
Finnish	0.8	0.86	0.82
Greek	0.79	0.84	0.82
Hebrew	0.78	0.83	0.8
Icelandic	0.73	0.79	0.76
Korean	0.8	0.85	0.82
Saisiyat	0.25	0.05	0.09

Jean-Baptiste/camembert-ner Performance by Language

Language	Precision	Recall	F1 Score
Amis	0.45	0.61	0.52
Chinese	0.56	0.74	0.64
English	0.51	0.69	0.58
Finnish	0.53	0.71	0.61
Greek	0.53	0.71	0.6
Hebrew	0.49	0.65	0.56
Icelandic	0.57	0.77	0.66
Korean	0.58	0.78	0.66
Saisiyat	0.38	0.43	0.4

Looking Ahead

- We are in the process of scaling now:
  - Assessments (subject matter experts to write evaluations)
  - Models (proprietary models for assessments)
  - Use cases (tasks of interests for testing)
- Watch for the open-source release alongside our publication at IAAI
- Come talk to me (email: achadda@iqt.org)!

