A Practitioner’s Perspective: AI Assurance in the Current Modeling Paradigm

Ari Chadda — Data Scientist, IQT Labs

IQT is an Independent, Not-for-Profit Connecting Three Different Worlds

GOVERNMENT PARTNERS

VENTURE CAPITAL

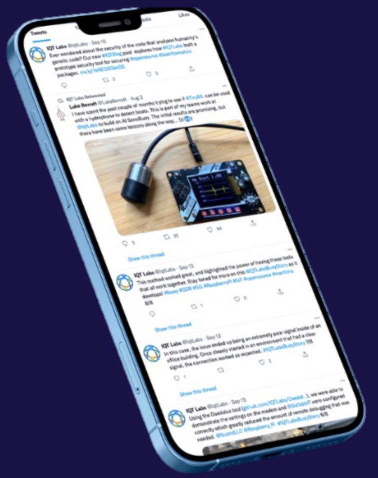
3,000+
co-investors

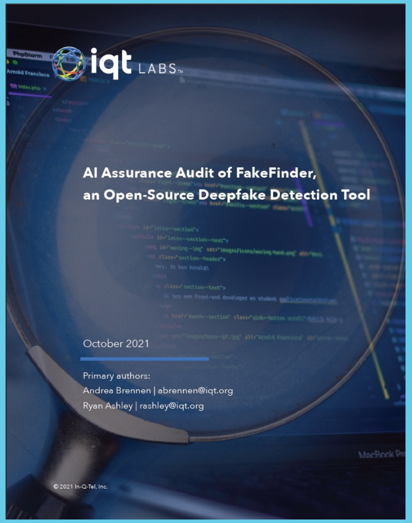
COMMERCIAL STARTUPS

18,000+
company engagements

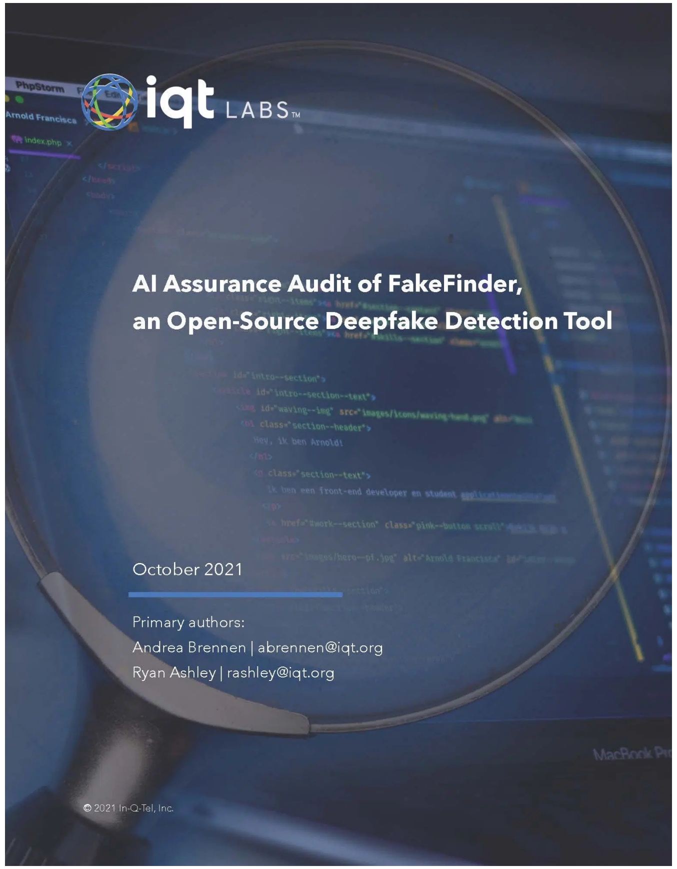
IQT Labs at a Glance

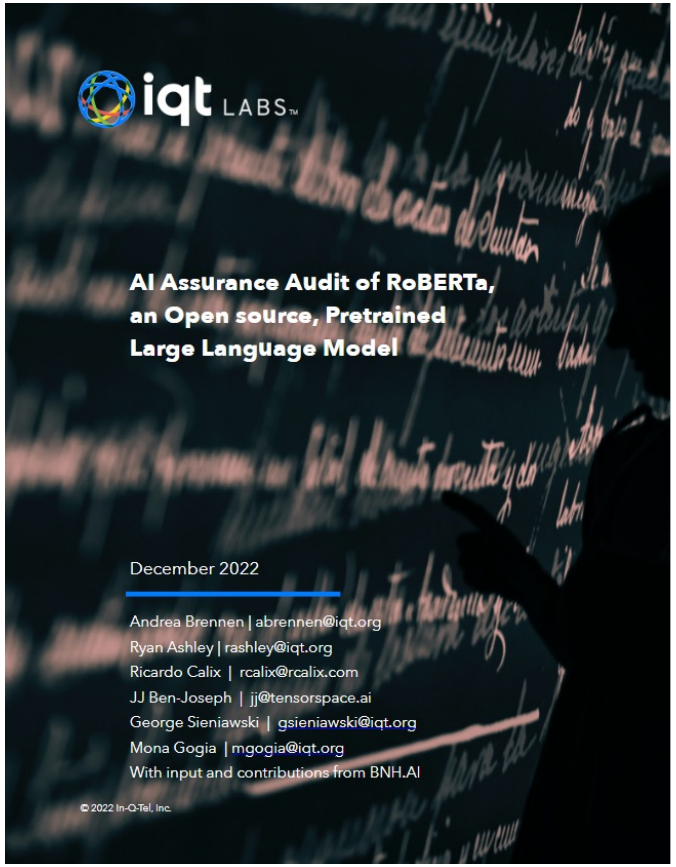
- Open source-driven insights
- Unclassified proxy problems
- Focus on prototyping & proofs-of-concept
- Understand and de-risk emerging technologies

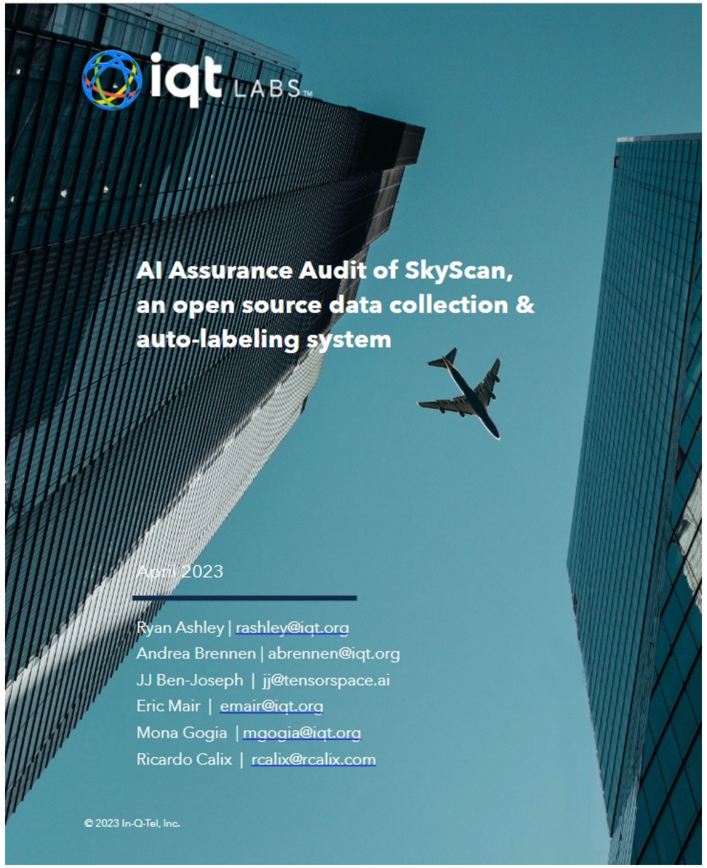




Audits of AI Tools and Systems







assets.iqt.org/pdfs/IQTLabs_SkyScanAudit_May_2023.pdf
assets.iqt.org/pdfs/IQTLabs_RoBERTaAudit_Dec2022_final.pdf
assets.iqt.org/pdfs/IQTLabs_AiA_FakeFinderAudit_DISTRO__1_.pdf

5 Takeaways from Auditing AI

- **Avoid Groupthink:** An audit team needs more than data scientists
- **Audit the tool, (not) just the model:** Risk requires a use case
- **Go beyond accuracy:** Evaluate models based on real world harm
- **Don’t be blind to model blindspots:** Users don’t know what they don’t know
- **Look for vulnerabilities across the ML stack:** Attackers take the easy way in

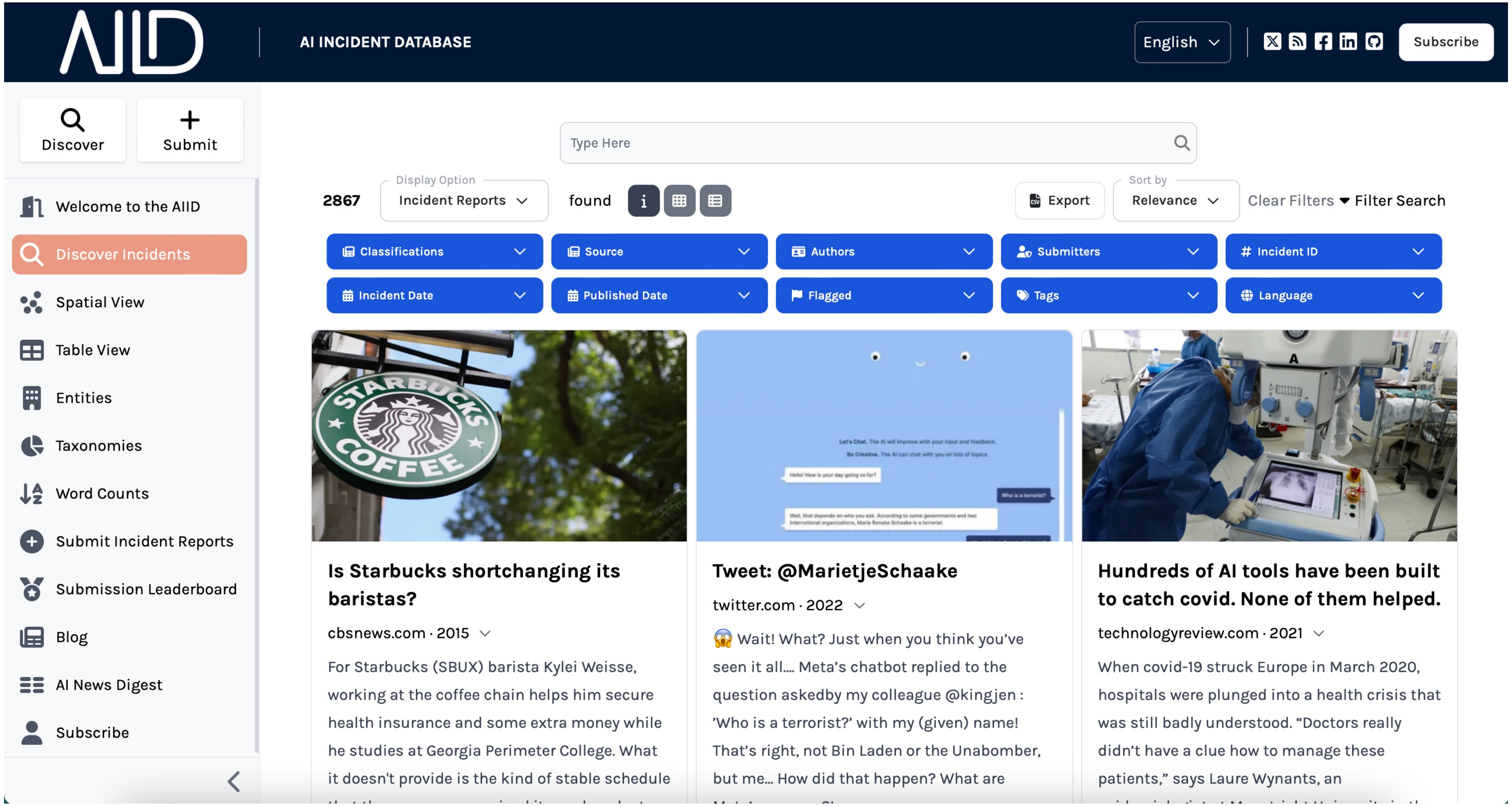
iqt.org/five-key-lessons-for-auditing-ai-from-avoiding-model-groupthink-to-revealing-models-blind-spots/

5 Takeaways from Auditing AI

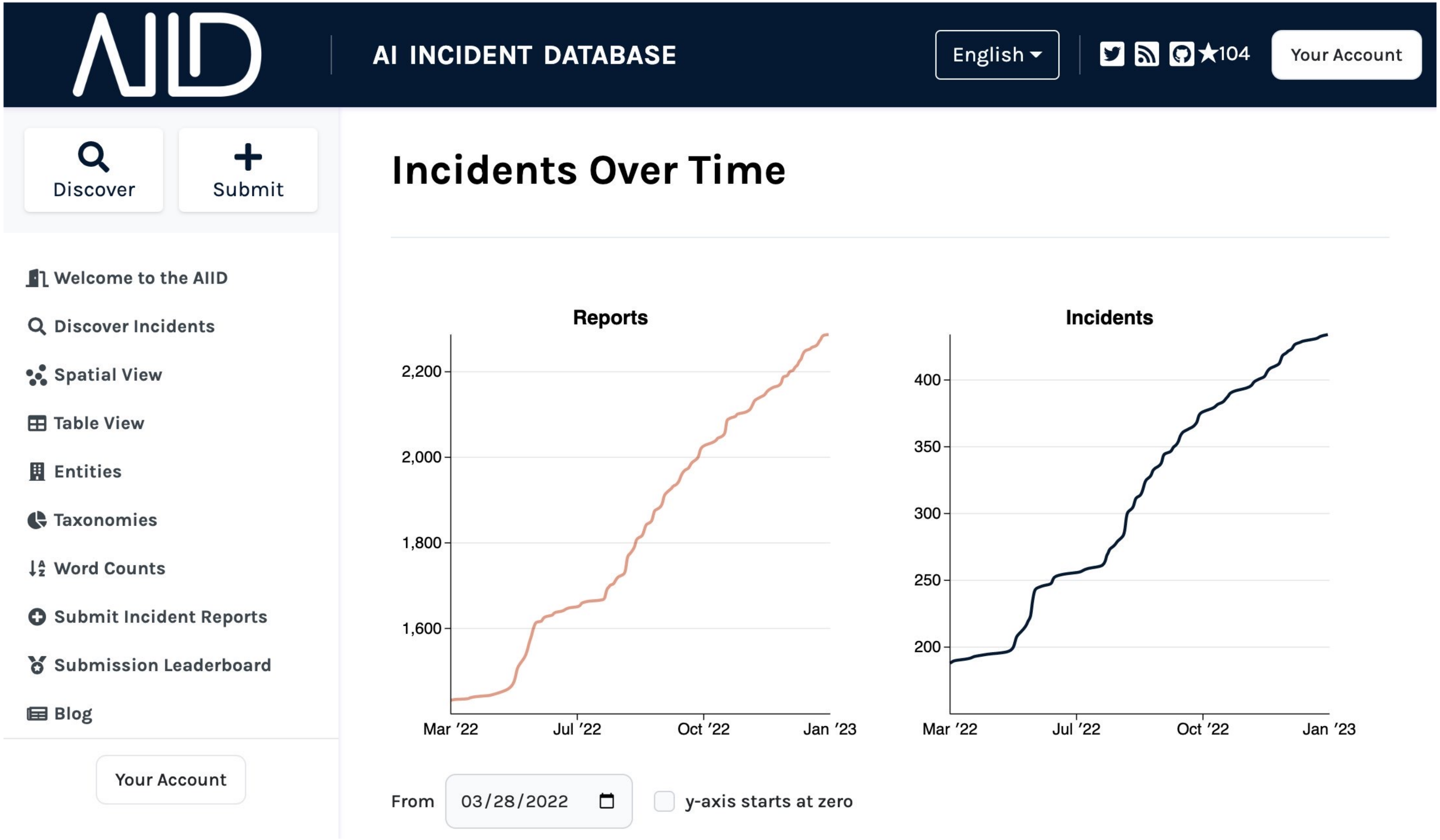
- **Avoid Groupthink:** An audit team needs more than data scientists
- **Audit the tool, (not) just the model:** Risk requires a use case
- **Go beyond accuracy:** Evaluate models based on real world harm
- **Don't be blind to model blindspots:** Users don't know what they don't know
- **Look for vulnerabilities across the ML stack:** Attackers take the easy way in

iqt.org/five-key-lessons-for-auditing-ai-from-avoiding-model-groupthink-to-revealing-models-blind-spots/

Qualitative versus Quantitative Risk



incidentdatabase.ai/apps/discover

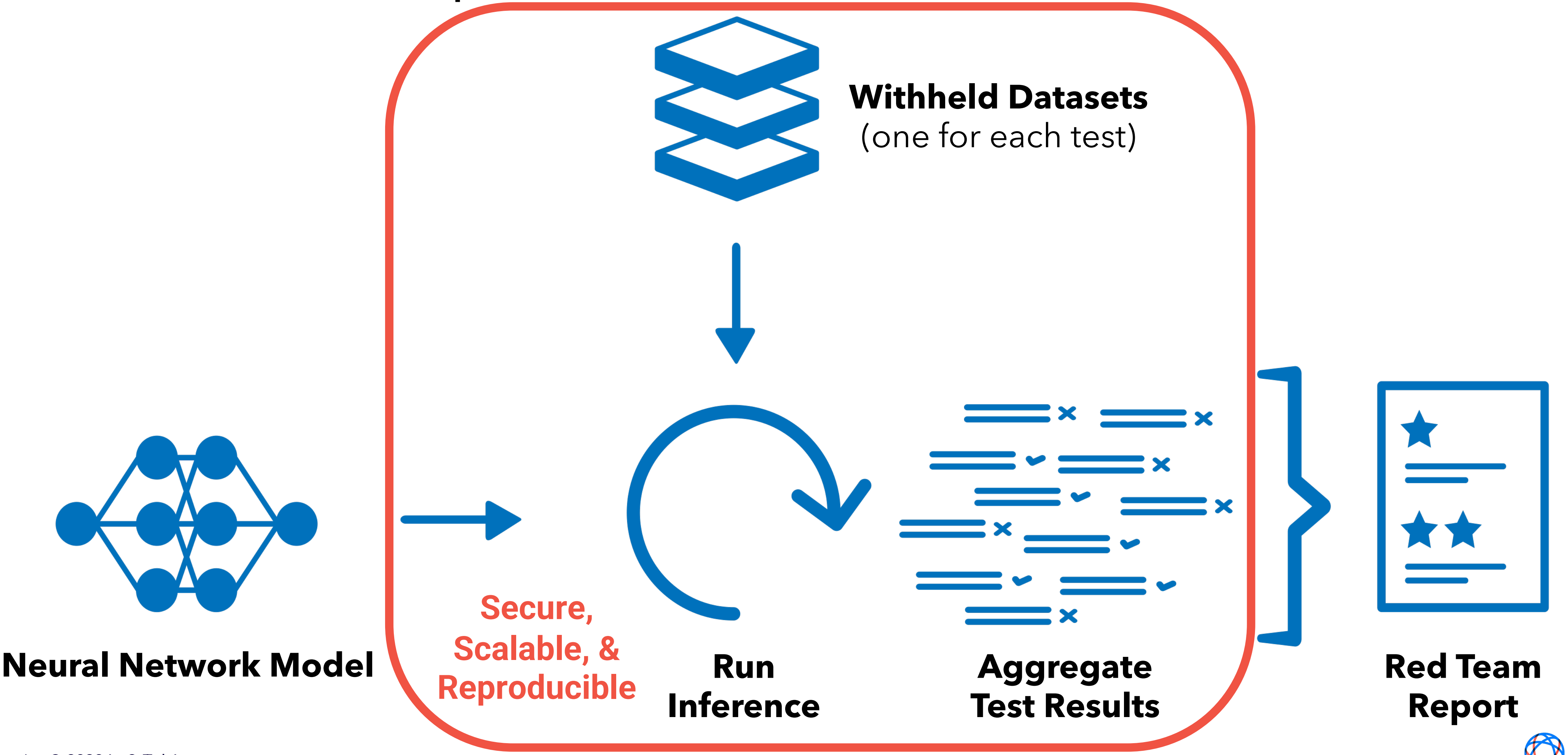


twitter.com/IncidentsDB/status/1613417066962505733

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.*

Evaluation Authority

"Consumer Reports" for AI Models



Proof of Concept Red Team Report

localhost:8050

iqt 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

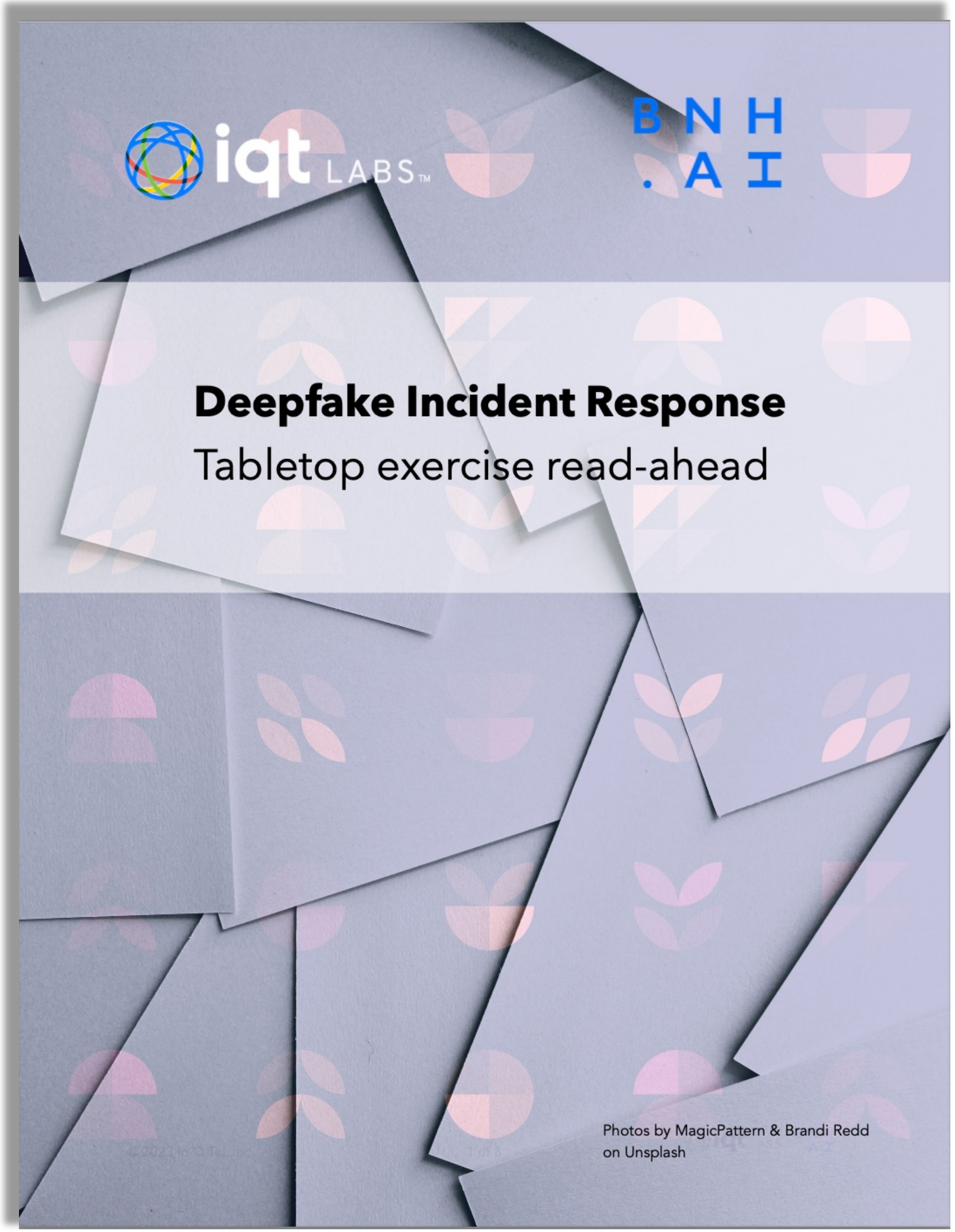
Language	Precision	Recall	F1 Score
Amis	0.73	0.77	0.75
Chinese	0.74	0.79	0.76
English	0.77	0.83	0.8
Finnish	0.8	0.86	0.82
Greek	0.79	0.84	0.82
Hebrew	0.78	0.83	0.8
Icelandic	0.73	0.79	0.76
Korean	0.8	0.85	0.82
Saisiyat	0.25	0.05	0.09

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

iqt

As you’re **deploying AI** (it’s inevitable), think of **quantitative measures** for more than just task-based **performance and thresholds that are acceptable** to your organization/use case


iqt LABS
BNH.AI
Deepfake Incident Response
Tabletop exercise read-ahead
Photos by MagicPattern & Brandi Redd on Unsplash


1/20
SKYNET NEEDS DRIVER'S ED
Your company's autonomous car system kills a pedestrian in self-driving mode. Discuss your Response Plan and then draw 4 injects.
Real-World Harm: Serious Injury (Physical Trauma, Grievous Bodily Harm, Organ Failure, and Loss of Life)
2/40
OOPS MY BAD!
Your intern plugged in a USB drive with ransomware and now your workstations are frozen until your organization pays up! Update your Response Plan and then roll D10.
HAIWIRE
INJECT
STAY AHEAD OF THE GAME WITH iqt LABS

 |

Contact Me

Website: iqtlabs.org

Email: achadda@iqt.org

Phone: 571.481.7257

AI Evaluation Authorities: A Case Study Mapping Model Audits to Persistent Standards

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

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

achadda@iqt.org, sean.mcgregor@ul.org, jesse.hostetler@ul.org, abrennen@iqt.orgt

Abstract

Intelligent system impact assessments are organizationally consuming assurance activities that are typically performed once and discarded along with the opportunity to programmatically test all similar products for the market. This study illustrates how several incidents (i.e., harms) involving Named Entity Recognition (NER) can be prevented by scaling up a previously-performed audit of NER systems. The audit instrument's diagnostic capacity is maintained through a security model that protects the underlying data (i.e., it addresses Goodhart's Law). An open-source evaluation infrastructure is released along with an example derived from a real-world audit on top of the dataset security model that reports high-level findings without exposing the underlying data.

Introduction

Many real-world applications of knowledge discovery, knowledge extraction, search, and computer network security involve a Named Entity Recognition (NER) step. NER is the task of recognizing a variety of "entities" within text. For example, the text "2012's [DATE] AlexNet [PRODUCT] is named for Alex Krizhevsky [PERSON]," has three entity types for dates, products, and persons.

Examples of failed NER appear frequently in the AI Incident Database (AIID) of (McGregor 2021), which catalogs examples of AI harms produced in the real world. Though not referenced explicitly in the incident reports, NER is a foundational Natural Language Processing (NLP) task undergirding a great many products. Most incident reports related to NER center on the user-facing issues of the technologies, including Incidents 317 (Bug in Facebook's Anti-Spam Filter Allegedly Blocked Legitimate Posts about

wrapped by a variety of user interfaces to facilitate the system's overall use case. This complexity obfuscates the explicit role that the NER model plays in any incidents produced by the system, but it is standard practice when it comes to moving from Research and Development (R&D) to end-user-facing production. Still, the frequency of incidents that likely link to NER models underscores the importance of the NER task and the need to mitigate incidents arising from it. Any multi-component system can fail if an individual component produces erroneous or harmful outputs.

While the specific issues leading to these incidents cannot be localized without proprietary knowledge of Meta's implementation, the incidents in Table 1, are similar and involve probable NER failures on Meta's Facebook platform.

IQT Labs[1] has conducted several audits where we assessed the safety and fairness properties of AI tools and systems (Brennen and Ashley 2021; Brennen et al. 2022; Ashley et al. 2023). The current paper focuses on our audit of the RoBERTa model (Liu et al. 2019) and variants thereof (Conneau et al. 2019), which are pre-trained LLM architectures we audited over several months. The variants audited included RoBERTa-base, RoBERTa-large, XLM-RoBERTa-base, and XLM-RoBERTa-large, which collectively were downloaded about 27.2M times on HuggingFace between July 15th and August 15th (HuggingFace 2023a,b,c,d). As part of that audit, we developed a multilingual NER programmatic assessment that exposed model limitations and highlighted vulnerabilities in the system attack surface (Calix et al. 2022).

Here we extend our prior work on NER auditing by adding reproducibility and scalability to the programmatic assessment—a nontrivial exercise in developing and implementing an applied framework for reproducible model as-

Accepted IAAI, 2024

