Ethics

October 10th

Ethics - Levels of concern

- Laws
  - What should be enacted by the government with regards to AI
- Social Morality
  - “Recognition that nor all of the socially entrenched standards that properly govern our lives are, or should be, legal standards”
- Individual decisions
  - Individuals will still need to exercise their own moral judgement

Today’s topics

Privacy

Bias & Fairness

Environmental impact

Reproducibility

Interpretability

Accountability

Use Cases

Privacy


STORES KNOW YOU'RE PREGNANT
Father Asked Target Why Daughter Got Baby Coupons
Daughter Was Pregnant, Told No One
But Had Been Shopping At Target
TARGET
9
WCPO

EM Talks #8

Please Don't Use GitHub Copilot

Research shows that we can adversarially generate data seen during training

2022 IEEE Symposium on Security and Privacy (SP)

Membership Inference Attacks From First Principles

Nicholas Carlini*[1] Steve Chien[1] Milad Nasr[1,2] Shuang Song[1] Andreas Terzis[1] Florian Tramèr[1]
[1] *Google Research* [2] *University of Massachusetts Amherst*

***Abstract*—A membership inference attack allows an adversary to query a trained machine learning model to predict whether or not a particular example was contained in the model's training dataset. These attacks are currently evaluated using average-case "accuracy" metrics that fail to characterize whether the attack can confidently identify any members of the training set. We argue that attacks should instead be evaluated by computing their true-positive rate at low (e.g., $\leq 0.1\%$) false-positive rates, and find most prior attacks perform poorly when evaluated in this way. To address this we develop a Likelihood Ratio Attack (LiRA) that carefully combines multiple ideas from the literature. Our attack is 10× more powerful at low false-positive rates, and also strictly dominates prior attacks on existing metrics.**

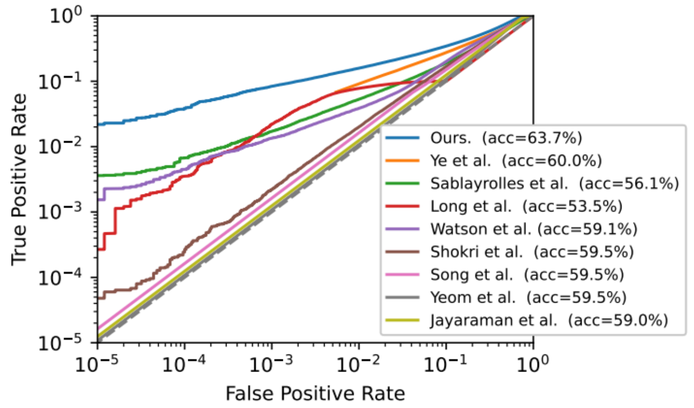


Fig. 1: Comparing the true-positive rate vs. false-positive rate of prior membership inference attacks reveals a wide gap in effectiveness. An attack's average *accuracy* is not indicative of its performance at low FPRs. By extending on the most effective ideas, we improve membership inference attacks by 10×, for a non-overfit CIFAR-10 model (92% test accuracy).

I. Introduction

Neural networks are now trained on increasingly sensitive datasets, and so it is necessary to ensure that trained models are privacy-preserving. In order to empirically verify if a model is in fact private, membership inference attacks [60] have become the de facto standard [42, 63] because of their simplicity. A membership inference attack receives as input a trained model and an example from the data distribution, and predicts if that example was used to train the model.

Unfortunately as noted by recent work [44, 69], many prior membership inference attacks use an incomplete evaluation methodology that considers average-case success metrics (e.g., accuracy or ROC-AUC) that aggregate an attack's accuracy over an entire dataset and over all detection thresholds [6, 18, 26, 33–35, 45, 52, 54, 54–57, 61, 63, 66, 70]. However, privacy is not an average case metric, and should not be evaluated as such [65]. Thus, while existing membership inference attacks do appear effective when evaluated under this average-case

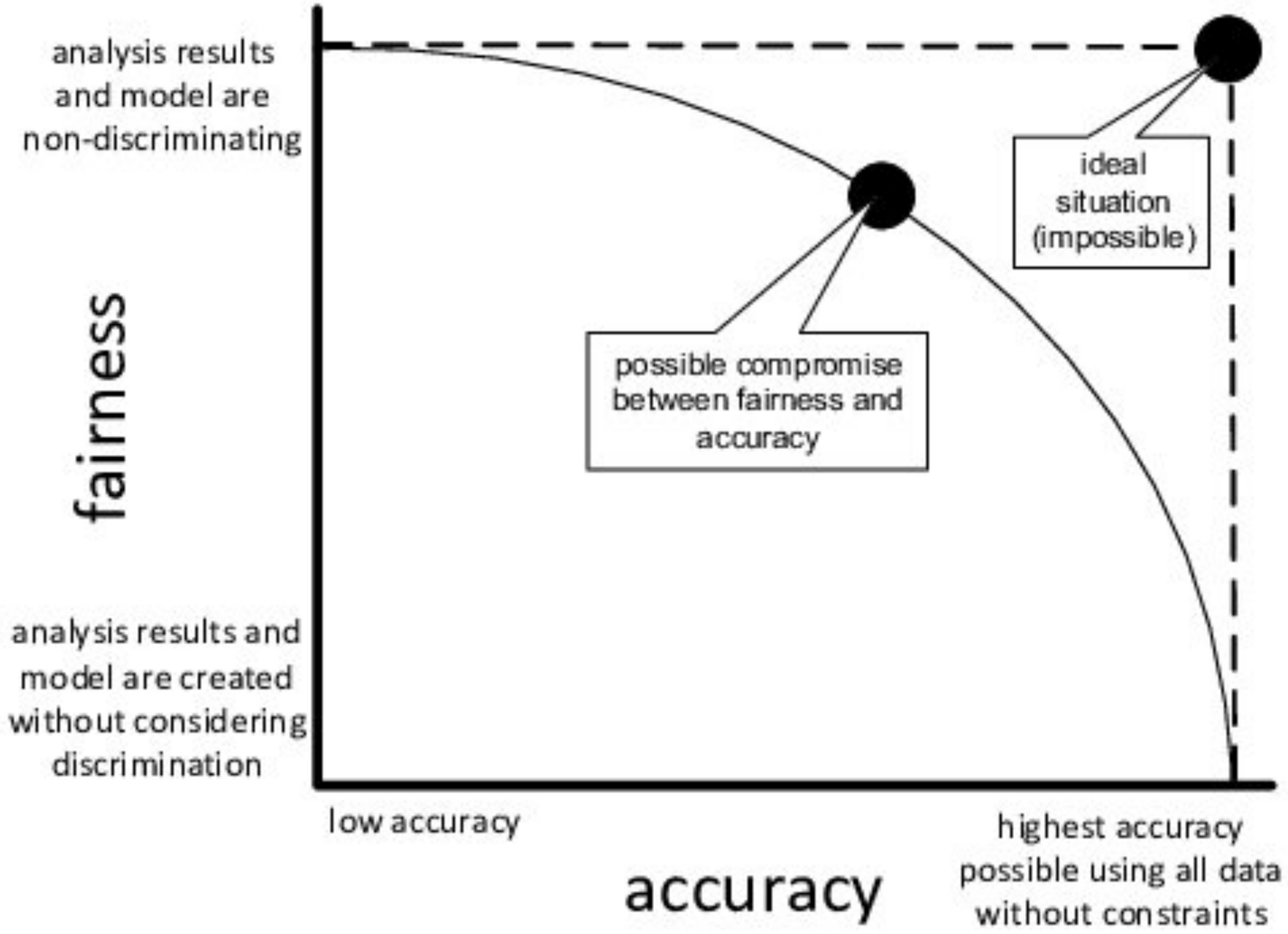
metrics (e.g., AUC) are often uncorrelated with low FP success rates. For example the attack of Yeom et al. [70] has a high accuracy (59.5%) yet fails completely at low FPRs, and the attack of Long et al. [36] has a much lower accuracy (53.5%) but achieves higher success rates at low FPRs.

We develop a Likelihood Ratio Attack (LiRA) that succeeds 10× more often than prior work at low FPRs—but still strictly dominates prior attacks on aggregate metrics introduced previously. Our attack combines per-example difficulty

Privacy in NLP

- How much risk should we be willing to assume?
- Are there any situations in which we shouldn’t use models to avoid privacy leaks?
- Informed Consent - when is it ethical to collect and use someone’s data for training?

Bias and Fairness


analysis results and model are non-discriminating
fairness
analysis results and model are created without considering discrimination
ideal situation (impossible)
possible compromise between fairness and accuracy
low accuracy
accuracy
highest accuracy possible using all data without constraints

[Source]

Research shows that models amplify bias

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

Jieyu Zhao[§] **Tianlu Wang**[§] **Mark Yatskar**[‡]
Vicente Ordonez[§] **Kai-Wei Chang**[§]
[§]University of Virginia
{jz4fu, tw8cb, vicente, kc2wc}@virginia.edu
[‡]University of Washington
my89@cs.washington.edu

Abstract

Language is increasingly being used to define rich visual recognition problems with supporting image collections sourced from the web. Structured prediction models are used in these tasks to take advantage of correlations between co-occurring labels and visual input but risk inadvertently encoding social biases found in web corpora. In this work, we study data and models associated with multilabel object classification and visual semantic role labeling. We find that (a) datasets for these tasks contain significant gender bias and (b) models trained on these datasets further amplify existing bias. For example, the activity `cooking` is over 33% more likely to involve females than males in a training set, and a trained model further amplifies the disparity to 68% at test time. We

tics from images and require large quantities of labeled data, predominantly retrieved from the web. Methods often combine structured prediction and deep learning to model correlations between labels and images to make judgments that otherwise would have weak visual support. For example, in the first image of Figure 1, it is possible to predict a `spatula` by considering that it is a common tool used for the activity `cooking`. Yet such methods run the risk of discovering and exploiting societal biases present in the underlying web corpora. Without properly quantifying and reducing the reliance on such correlations, broad adoption of these models can have the inadvertent effect of magnifying stereotypes.

In this paper, we develop a general framework for quantifying bias and study two concrete tasks, visual semantic role labeling (vSRL) and multilabel object classification (MLC). In vSRL, we use the imSitu formalism (Yatskar et al., 2016, 2017), where the goal is to predict activities, objects and

What is the 80% Rule?

The 80% rule was created to help companies determine if they have been unwittingly discriminatory in their hiring process. The rule states that companies should be hiring protected groups at a rate that is at least 80% of that of white men. For example, if a firm has hired 100 white men in their last hiring cycle but only hired 50 women, then the company can be found in violation of the 80% rule. The rule itself has no real effect other than to call into question a company's hiring ethics. Those that are found in violation are only asked to provide a legitimate reason as to why they are hiring protected groups at such a lower rate.

[Source]

Bias and Fairness

- Should these models correct existing biases or replicate them?
- What are the potential harms of deploying a biased model?
- What guard rails should we put in place to ensure fairness in models?
- How should we think about the balance between fairness and accuracy?
- How does privacy conflict with fairness?

Accountability

GDPR

General Data Protection Regulation

- Data privacy regulation in the EU
- Articles 13–15: individuals have the right to 'meaningful information about the logic involved' in automated decisions
- Article 17: individuals have the right to have personal data erased

NEWS & COMMENTARY

Holding Facebook Accountable for Digital Redlining

Online ad-targeting practices often reflect and replicate existing disparities, effectively locking out marginalized groups from housing, job, and credit opportunities.

Linda Morris,
Staff Attorney,
ACLU Women's Rights Project

In today's digital world, people rely on online advertising platforms for critical information such as job opportunities or available housing. But unfortunately, thanks to practices known as "digital redlining" — the use of technology to perpetuate

[Source]

Accountability

- Who is responsible for bad predictions?
  - The company?
  - The engineers?
  - The user?
- Should there be regulations on AI systems? If so, what would they look like?

Environmental Impact

Energy and Policy Considerations for Deep Learning in NLP

[Source]

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Environmental Impact

- As a field, should we be concerned about the environmental impact of our models?
- Larger models are more accurate. How do we consider the trade off of performance vs environmental impact?
- Should we focus more on more efficient algorithms to limit our environmental impact?

Reproducibility

Reproducibility Checklists

2019

Show Your Work: Improved Reporting of Experimental Results

Jesse Dodge[♣] **Suchin Gururangan**[◇] **Dallas Card**[♡] **Roy Schwartz**[♠◇] **Noah A. Smith**[♠◇]
[♣]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[◇]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[♡]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{jessed,dcard}@cs.cmu.edu {suching,roys,noah}@allenai.org

Abstract

Research in natural language processing proceeds, in part, by demonstrating that new models achieve superior performance (e.g., accuracy) on held-out test data, compared to previous results. In this paper, we demonstrate that test-set performance scores alone are insufficient for drawing accurate conclusions about which model performs best. We argue for reporting additional details, especially performance on validation data obtained during model development. We present a novel technique for doing so: *expected validation performance* of the best-found model as a function of computation budget (i.e., the number of hyperparameter search trials or the overall training time). Using our approach, we find multiple recent model comparisons where authors would have reached a different conclusion if they had used more (or less) computation. Our approach also allows us to estimate the amount of computation required to obtain a given accuracy; applying it to several recently published results yields massive variation across papers, from hours to weeks. We conclude with a set of best practices for reporting experimental results which allow for robust future comparison, and provide code to allow researchers to use our technique.[1]

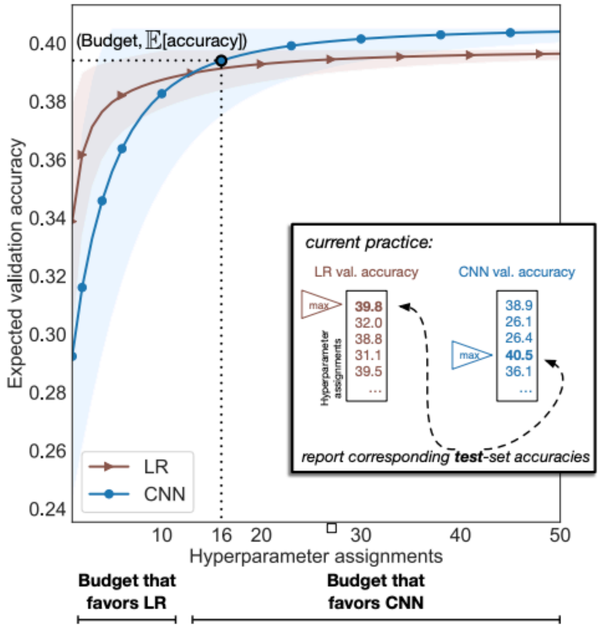


Figure 1: Current practice when compraing NLP models is to train multiple instantiations of each, choose the best model of each type based on validation performance, and compare their performance on test data (inner box). Under this setup, (assuming test-set results are similar to validation), one would conclude from the results above (hyperparameter search for two models on the 5-way SST classification task) that the CNN outperforms Logistic Regression (LR). In our

CALL FOR

Main Conference Papers
Paper Submission FAQ
Call For System Demonstrations

On this page

Reproducibility Checklist

Reproducibility Checklist

For all reported experimental results:

- A clear description of the mathematical setting, algorithm, and/or model
- A link to a downloadable source code, with specification of all dependencies, including external libraries (recommended for camera ready)
- A description of computing infrastructure used
- The average runtime for each model or algorithm, or estimated energy cost
- The number of parameters in each model
- Corresponding validation performance for each reported test result
- A clear definition of the specific evaluation measure or statistics used to report results.

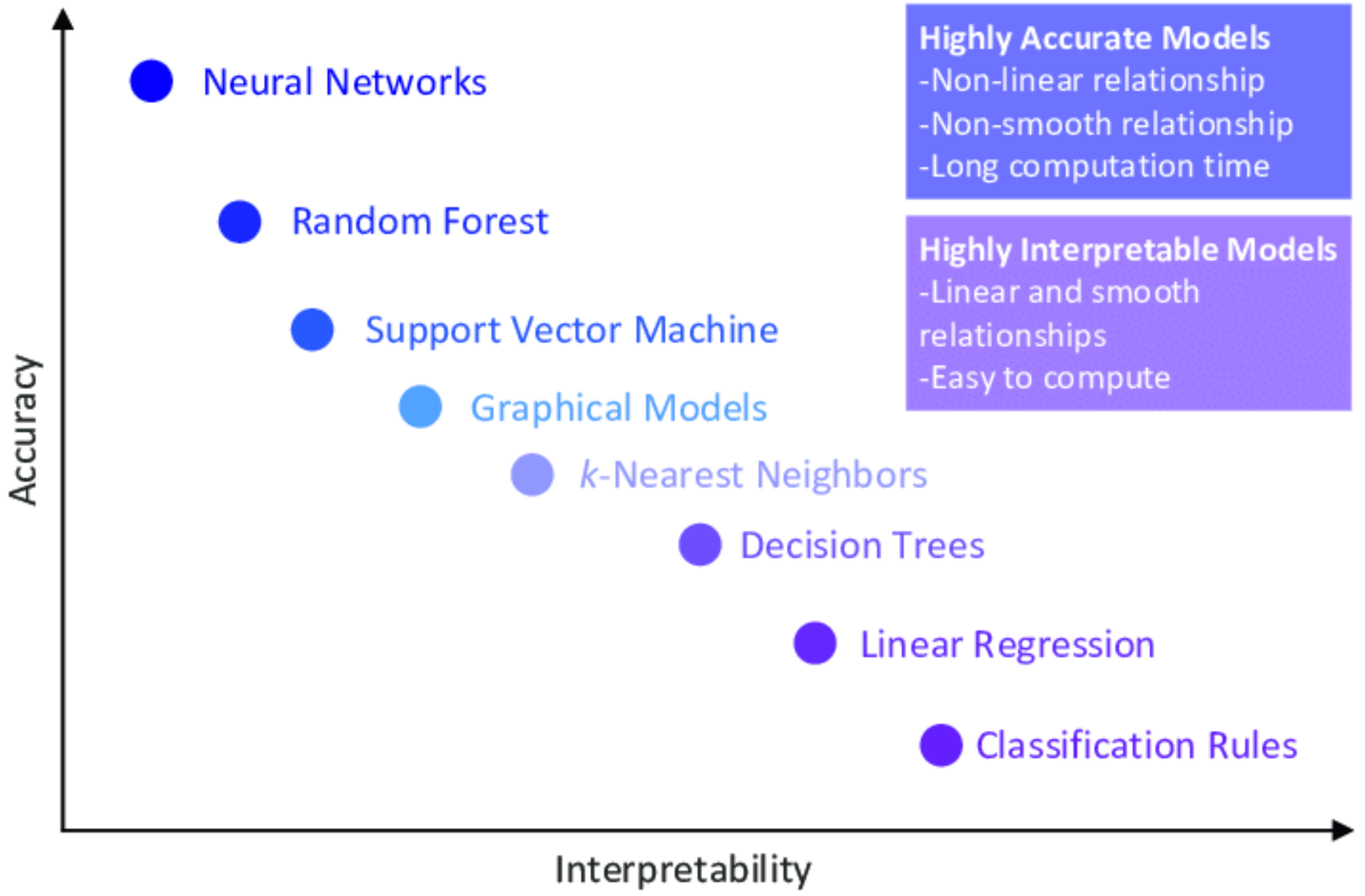
For all results involving multiple experiments, such as hyperparameter search:

[Source]

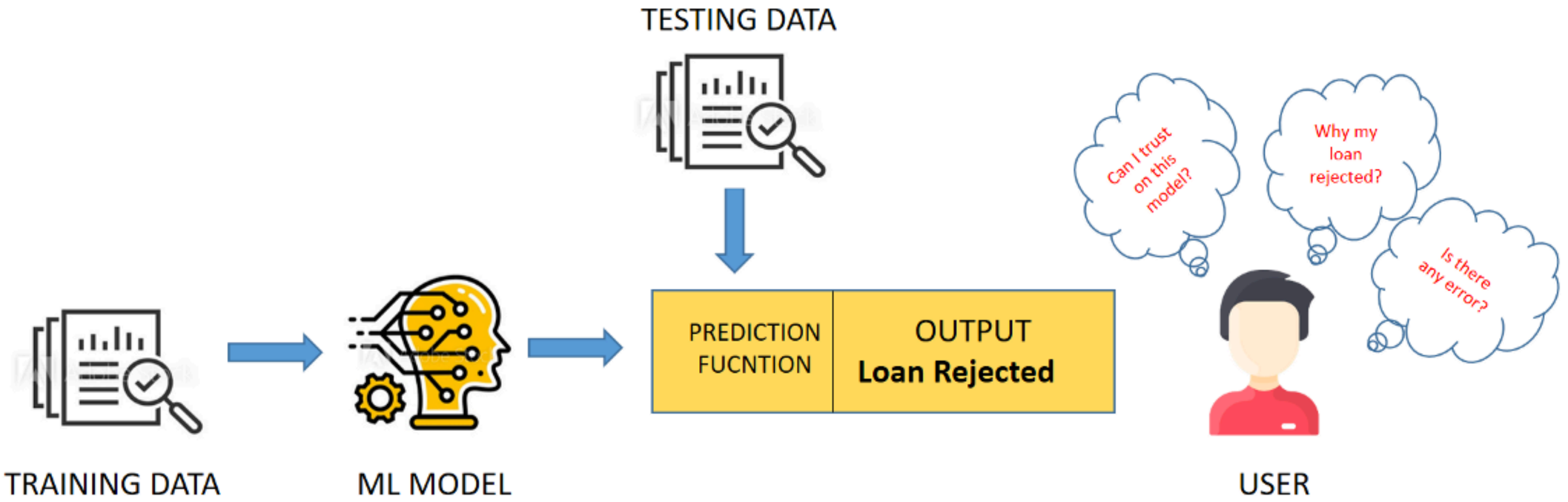
Reproducibility

- How should we consider models that are not reproducible? (Too expensive)
- How much effort should ML researchers put into making their work reproducible?
- Is a checklist the best way to ensure reproducibility?

Interpretability



[Image Credit]



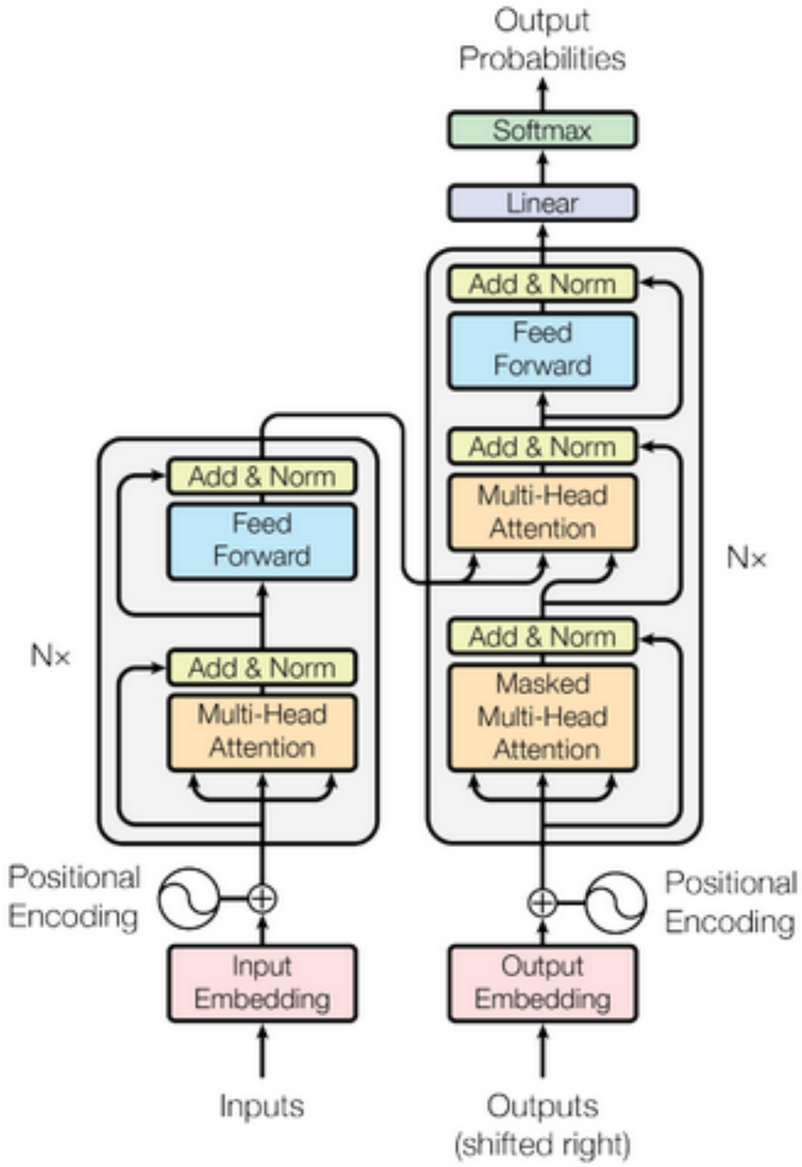
[Image Credit]

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\textbf{where } \text{head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

$$lrate = d_{\text{model}}^{-0.5} \cdot \min(step\_num^{-0.5}, step\_num \cdot warmup\_steps^{-1.5})$$



Interpretability

- Are there cases in which we should prioritize interpretability over accuracy?
- Autonomy: Should the user have full insight into the decision making process?
- Responsibility: Does the ability to understand the predictions affect who is responsible?

Use Cases

Use Cases

Are there any use cases that should be off-limits?

- Crime prediction
- Medicine management
- Demographic prediction
- Essay writing
- Facial recognition