Utility is in the Eye of the User: A Critique of NLP Leaderboards

EMNLP 2020

Kawin Ethayarajh

Dan Jurafsky

Stanford | NLP

Benchmark-based leaderboards have driven the creation of more accurate models.

Rank	Name	Model	URL	Score
1	HFL iFLYTEK	MacALBERT + DKM		90.7
2	Alibaba DAMO NLP	StructBERT + TAPT		90.6
3	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6
4	ERNIE Team - Baidu	ERNIE		90.4
5	T5 Team - Google	T5		90.3
14	GLUE Human Baselines	GLUE Human Baselines		87.1

[Wang et al., 2018]

Benchmark-based leaderboards have driven the creation of more accurate models.

[Wang et al., 2018; Wang et al., 2019; McCann et al., 2018; Hu et al., 2020]

But this has been at the expense of other qualities that the NLP community cares about.

size?

inference latency?

fairness?

energy efficiency?

training time?

ease of use?

[Rogers, 2019; Crane, 2018; Linzen, 2020]

How to frame the divergence between what's incentivized by leaderboards and what's valued by practitioners?

- Microeconomics!
- The *utility* of a good is the satisfaction that a *consumer* receives from it.
- Both leaderboards and practitioners are consumers of models.
- Each consumer has a unique *utility function*.

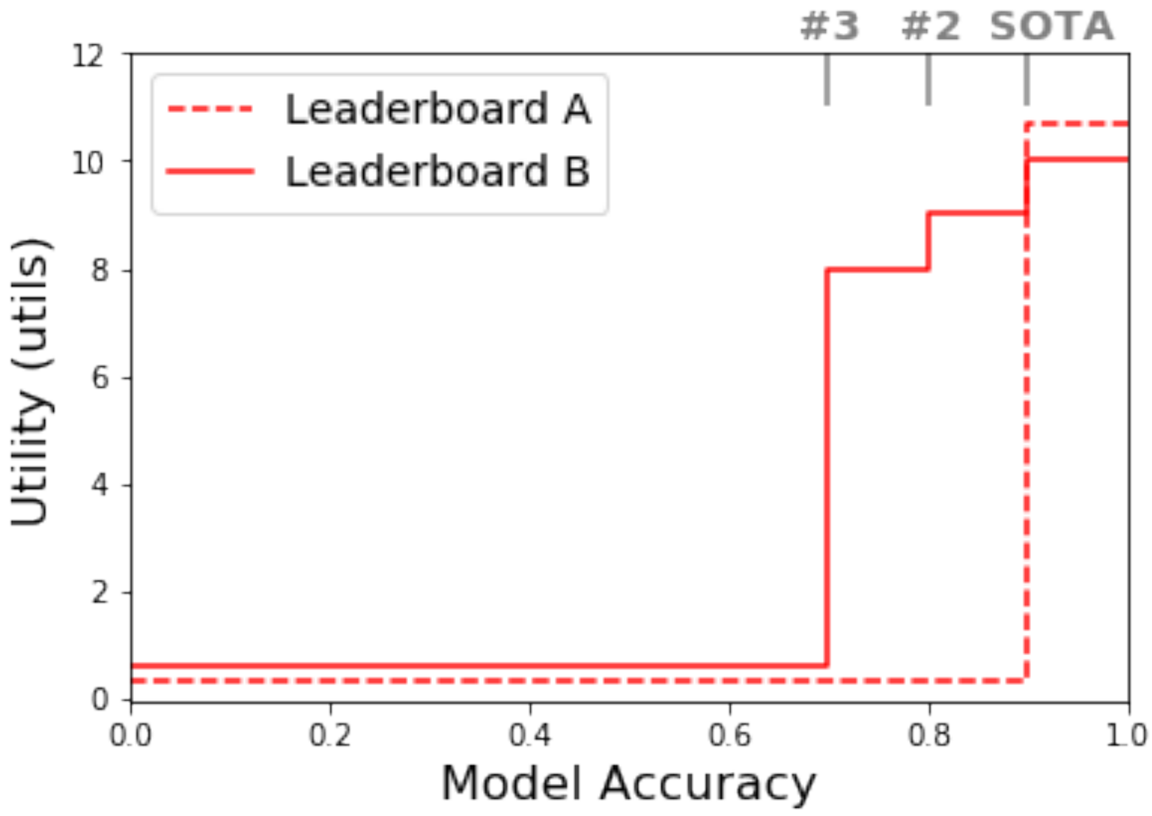
[Mankiw, 2020]

How to frame the divergence between what's incentivized by leaderboards and what's valued by practitioners?

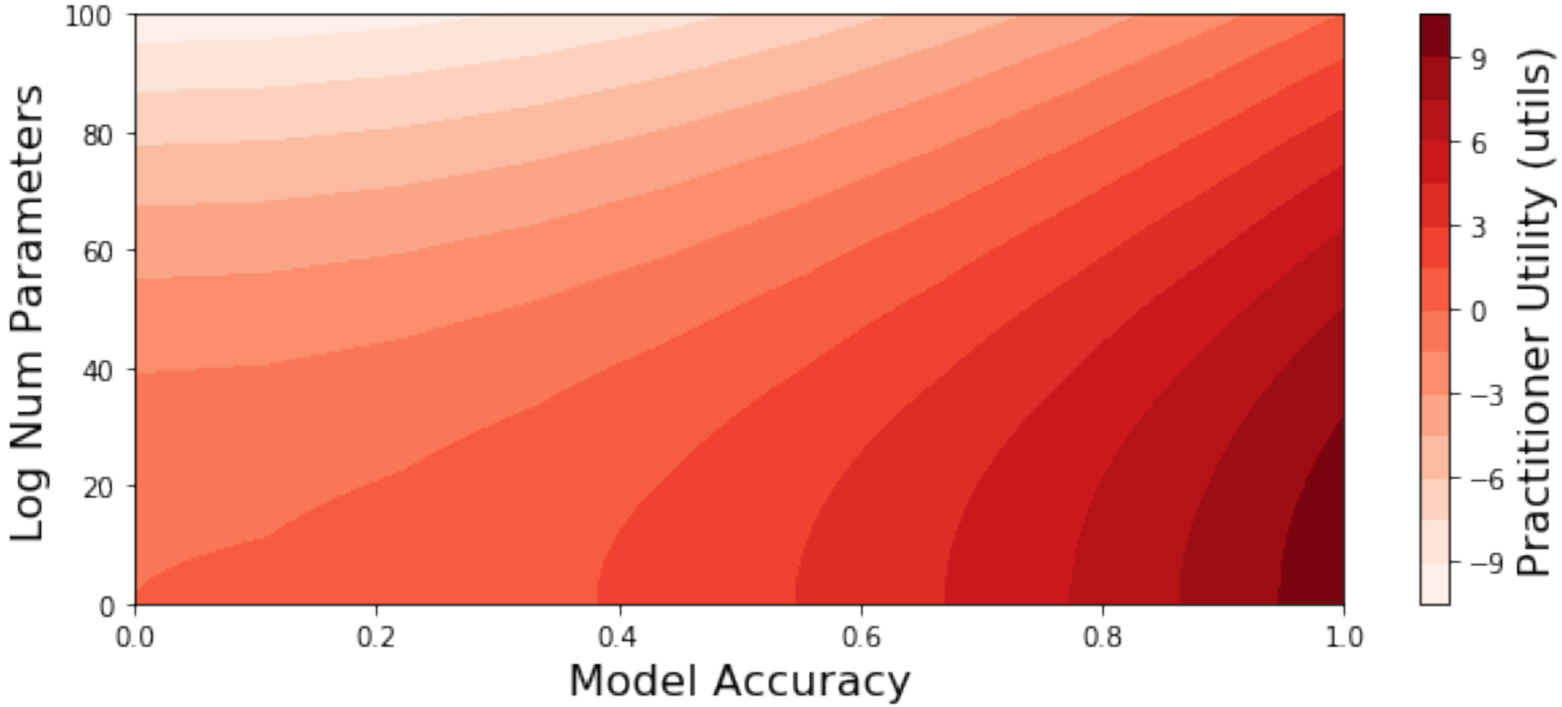
- Microeconomics!
- The *utility* of a good is the satisfaction that a *consumer* receives from it.
- Both leaderboards and practitioners are consumers of models.
- Each consumer has a unique *utility function*.
- **IDEA**: Compare leaderboards and practitioners using their utility functions.

[Mankiw, 2020]

A leaderboard is a consumer whose preferences are perfectly revealed through its rankings: SOTA > #2 > ...



Practitioners derive utility from multiple properties of the model being consumed (e.g., accuracy, efficiency, latency).



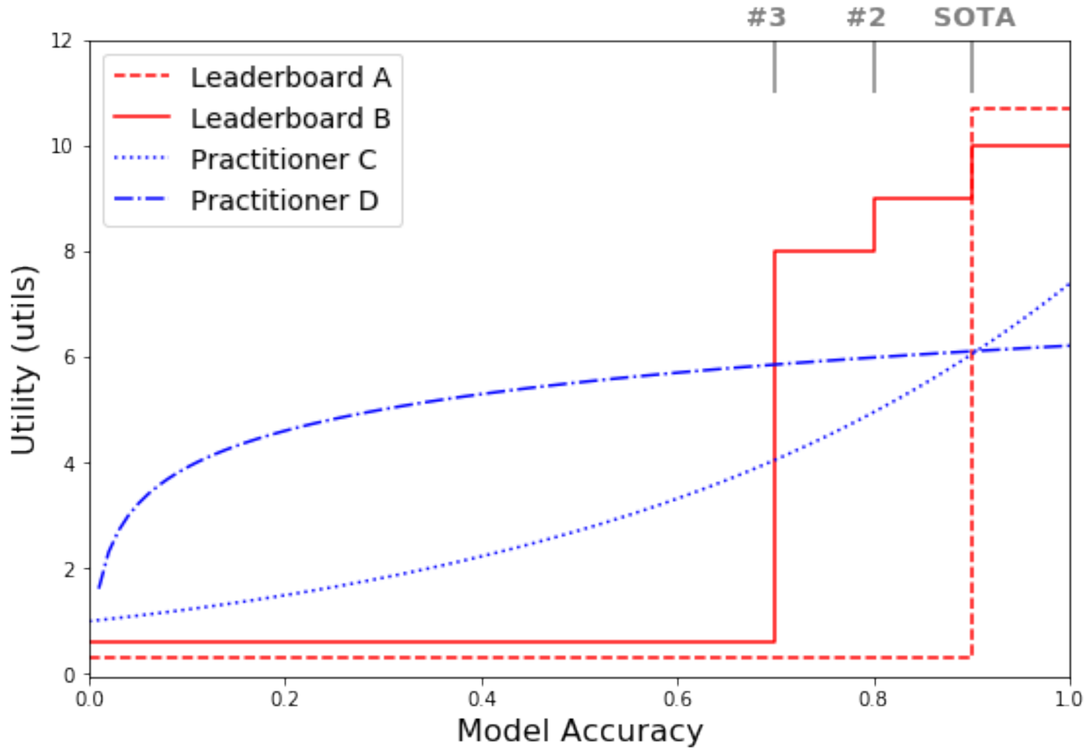
We can formally critique leaderboards by contrasting their utility functions with practitioners'.

- We don't know the exact shapes of utility functions, but we *do* know their properties: monotonicity, (in)sensitivity to certain attributes, etc.
- *Most* critiques apply to *most* leaderboards, but not all: StereoSet ranks by fairness; SNLI reports model size, etc.

[Nadeem et al., 2020; Bowman et al., 2015]

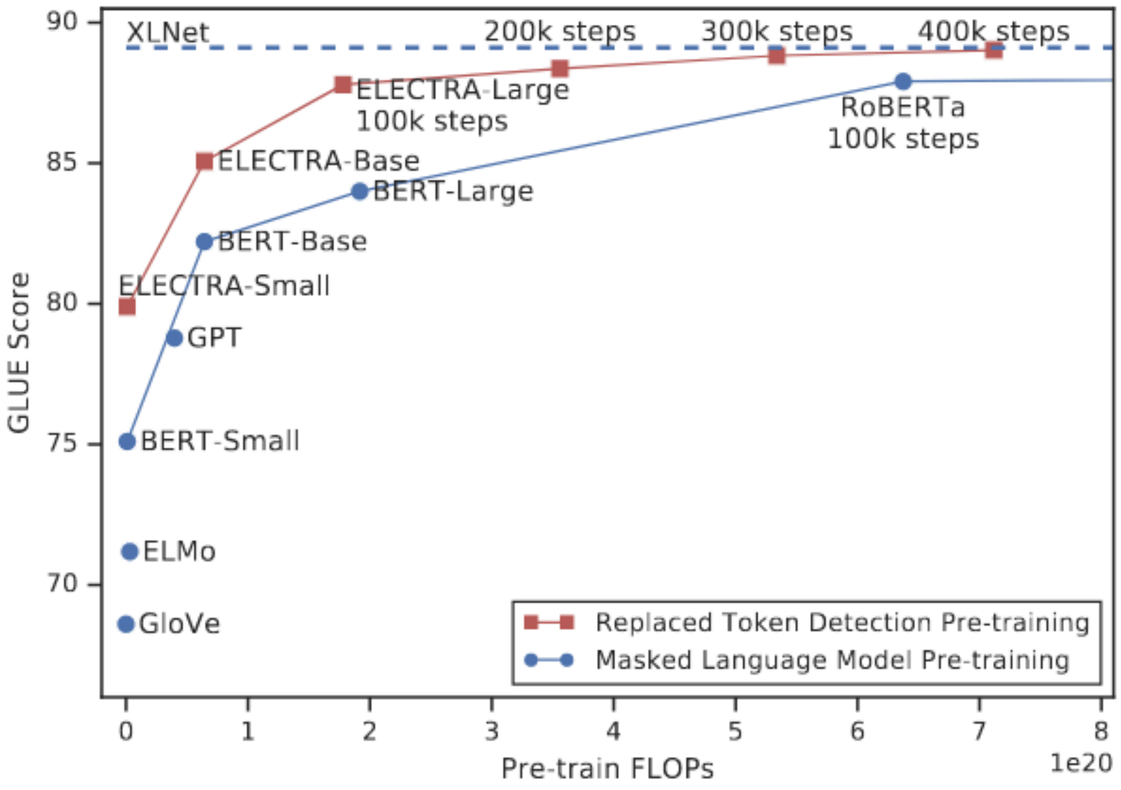
Critique #1: Non-Smoothness of Utility

- Leaderboards only gain utility from increased accuracy when it improves rank.
- The utility of practitioners is smooth with respect to accuracy.



Critique #1: Non-Smoothness of Utility

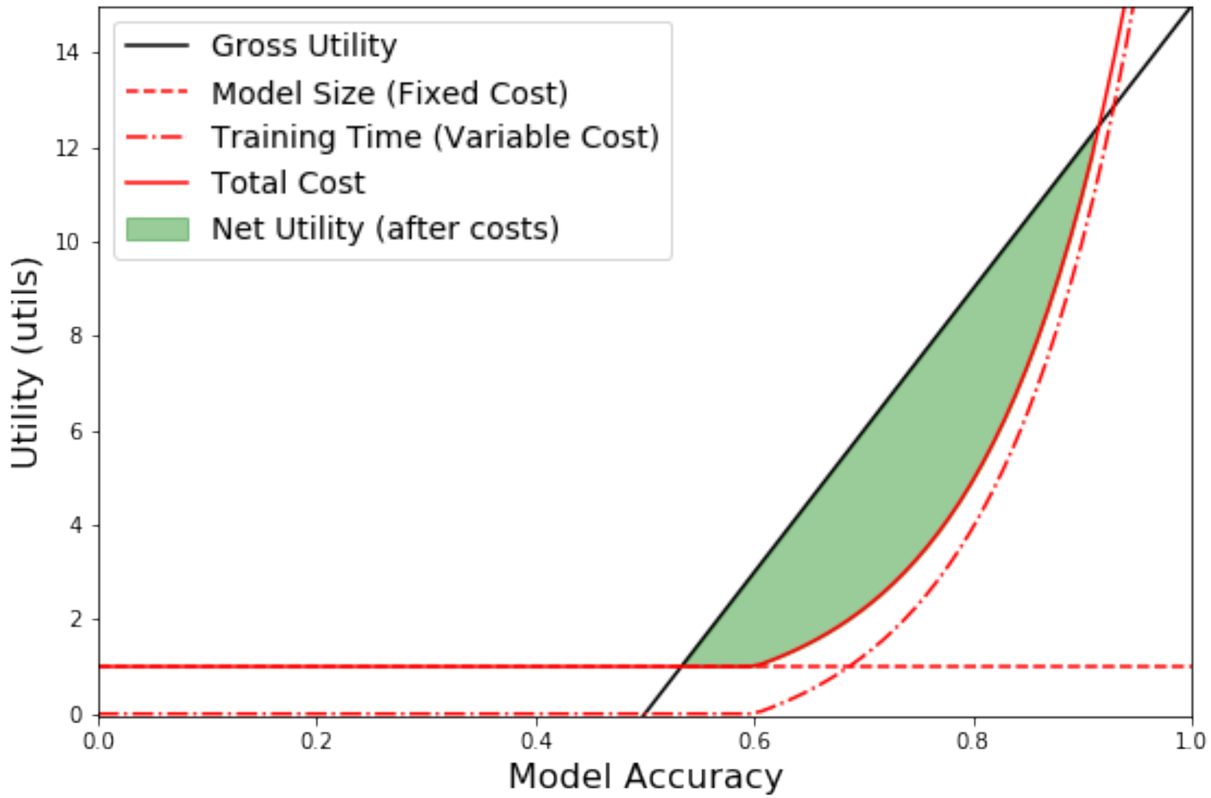
- Practitioners who are content with less-than-SOTA — e.g., for low latency or Green AI — are under-served; those who want competitive-with-SOTA are over-served.



[Clark et al., 2020; Schwartz et al., 2019; Strubell et al., 2019]

Critique #2: Cost-Ignorance

- Leaderboards rank by prediction *value*: accuracy, F1 score, exact match rate, etc.
- They ignore prediction *costs*: size, latency, energy efficiency, training time, etc.



Critique #2: Cost-Ignorance

- Practitioners can't afford to be cost-ignorant (especially the poorly-resourced)!
- Cost-sensitive rankings would
  - incentivize the creation of low-cost models like ELECTRA
  - allow practitioners to better estimate model utility

[Clark et al., 2020]

Critique #3: Robustness

- Over-fitting via resubmission is possible, even on private test sets.
- Most practitioners — but not most leaderboards — would gain utility from
  - robustness to adversarial examples
  - generalization to OOD data
  - Rawlsian fairness

[Linzen, 2020; Rawls, 2001; Hashimoto et al., 2018; Jia & Liang, 2017; Hardt, 2017]

Critique #3: Robustness

- This problem is being actively tackled:

Winogender Schemas

Winogender Schemas (inspired by Winograd Sche
pronoun in the sentence, designed to test for the p
sentence template has three mentions: an `OCCUPA`
either `OCCUPATION` or `PRONOUN`). Here are two ex
"patient."

1. **The nurse** notified the patient that...
   i. **her** shift would be ending in an hour.
   ii. **his** shift would be ending in an hour.
   iii. **their** shift would be ending in an hour.

SQuAD2.0
The Stanford Question Answering Dataset

Dyna Bench

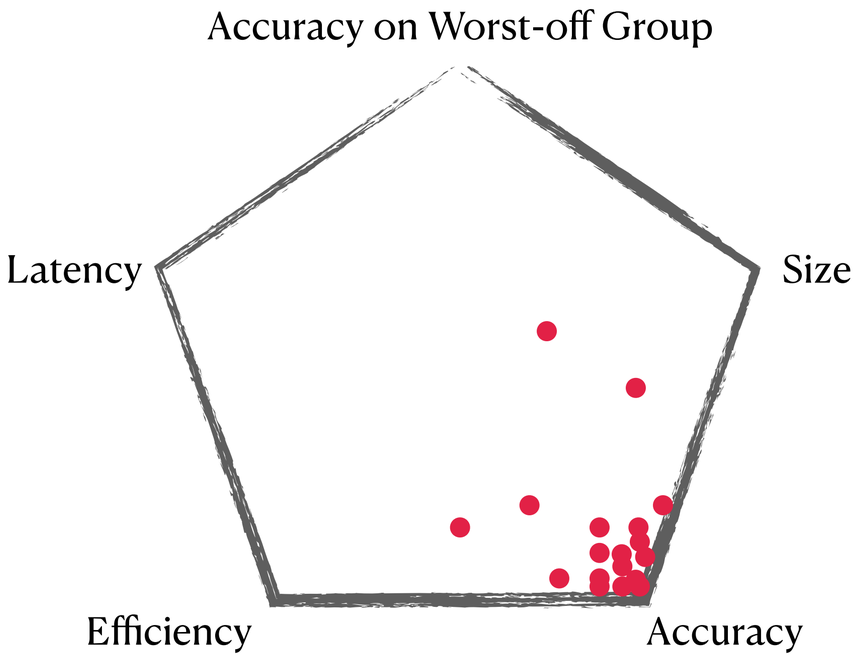
[Rudinger et al., 2018; Rajpurkar et al., 2018; Nie et al., 2019]

The Future of Leaderboards: One for Every User

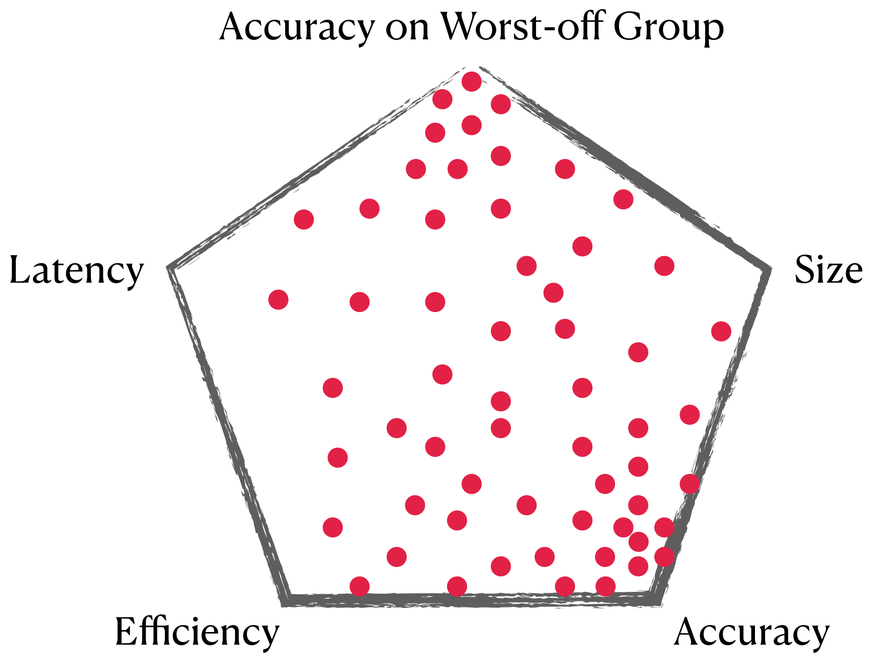
- Every practitioner has a unique utility function — no one-size-fits-all leaderboard.
- Leaderboards should demand transparency: require the reporting of metrics that are of practical concern (e.g., training time, model size, etc).
- Allow users to dynamically re-rank models based on their priorities over these statistics (i.e., align leaderboard's utility with their own).

Diverse Preferences, Diverse Models

2020



A More Enlightened Age



Thank you!