NLP SUMMIT HEALTHCARE

Green AI

Roy Schwartz

The Hebrew University of Jerusalem

THE HEBREW UNIVERSITY OF JERUSALEM

A Little about Me

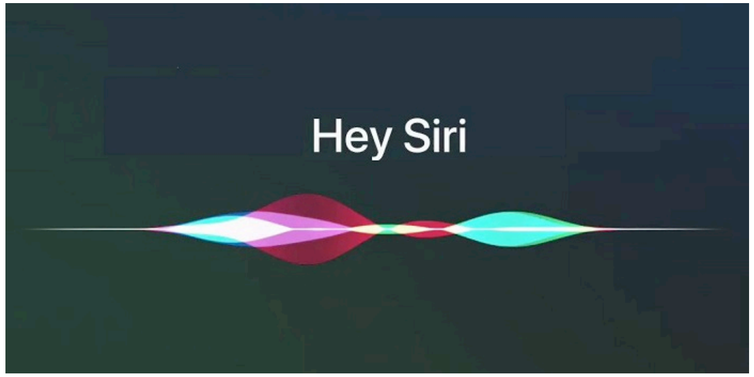
Roy Schwartz

- A senior lecturer (i.e., asst. Prof.) at the School of CS at the Hebrew U. of Jerusalem
  - I was a postdoc and a research scientist at The University of Washington and the Allen Institute for AI (AI2)
- I study Artificial Intelligence (AI) and focus on Natural Language Processing (NLP)
  - **Understanding** AI models
  - Revealing **biases** in datasets
  - Making AI more **sustainable**

AI Today

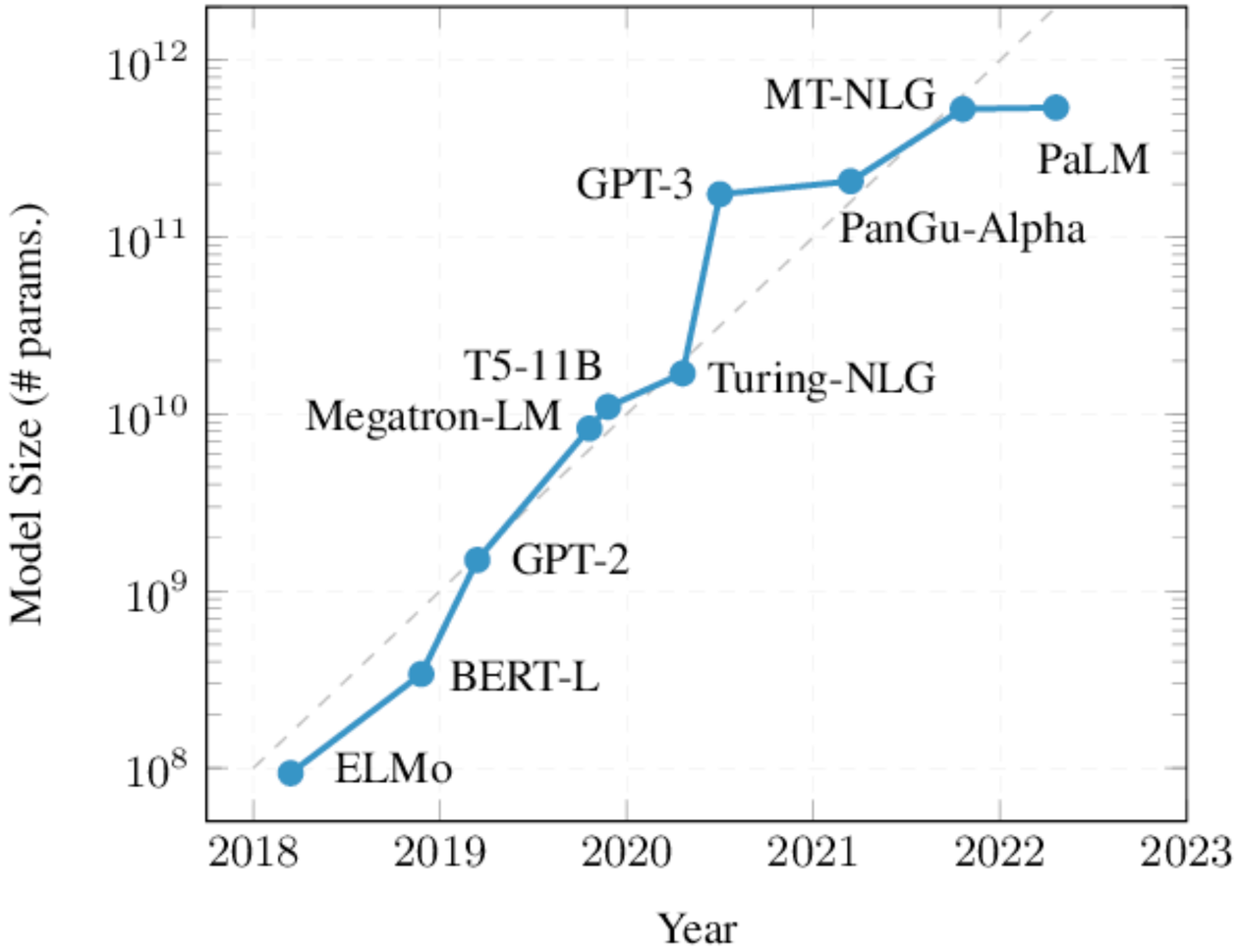
ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms"	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?"	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?"	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021



Scaling

5,000X in 4 Years



Taken from Lakim et al. (2022)

Green AI

Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Problems: inclusiveness, environment
- Green AI
  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results
  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy

Problems with Scaling

Inclusiveness

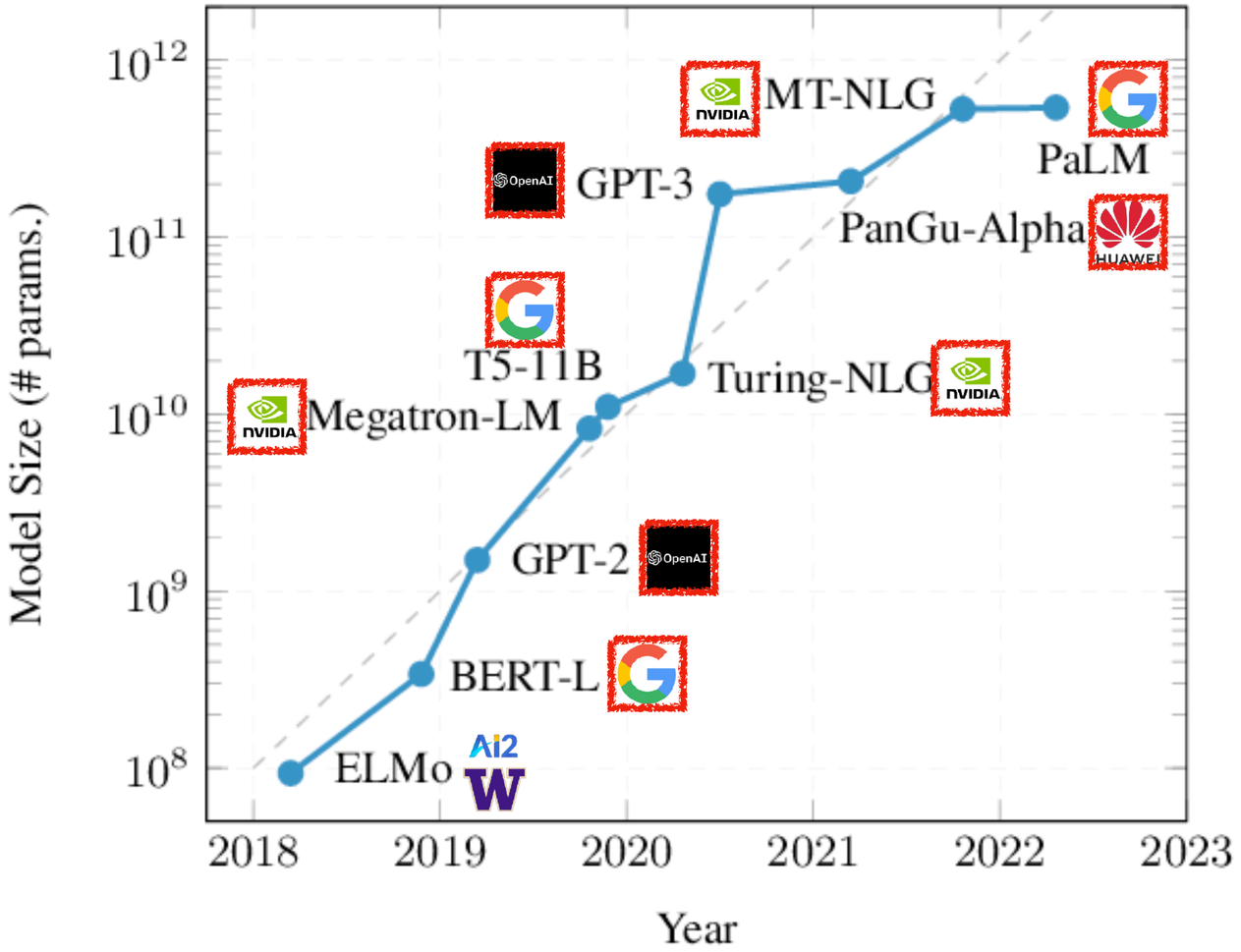
https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Training Costs

- BERT (Devlin et al, 2019) was trained on **16** Cloud TPUs for **4** days
- RoBERTa (Liu et al., 2019) was trained on **1024** V100 GPUs for approximately **1** day
- PaLM (Chowdhery et al., 2022) was trained on **6144** TPU v4 chips for **50** days and **3072** TPU v4 chips for **15** days

We need better **reporting**!

It's a Rich Man's World



Problems with Scaling

Environment

Consumption	CO_2e (lbs)
Air travel, 1 person, 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 & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Strubell et al. (2019)

Is AI creating an environmental problem?

Google's Answer: No!

BLOG ›

Good News About the Carbon Footprint of Machine Learning Training

TUESDAY, FEBRUARY 15, 2022

Posted by David Patterson, Distinguished Engineer, Google Research, Brain Team

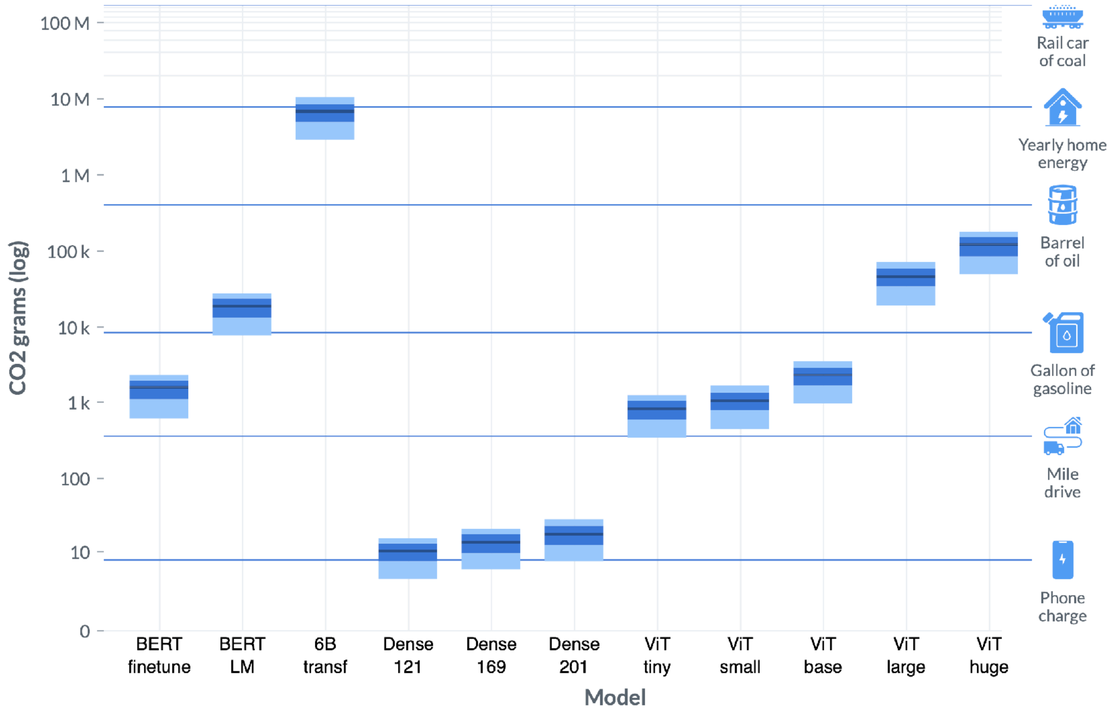
Strubell et al.'s energy estimate for NAS ended up ***18.7X*** *too high for the average organization (…) and* ***88X*** *off in emissions for energy-efficient organizations like Google*

We need better **reporting**!

Our Answer: Maybe?

Measuring the Carbon Intensity of AI in Cloud Instances

JESSE DODGE, Allen Institute for AI, USA
TAYLOR PREWITT, University of Washington, USA
REMI TACHET DES COMBES, Microsoft Research Montreal, USA
ERIKA ODMARK, Microsoft, USA
ROY SCHWARTZ, Hebrew University of Jerusalem, Israel
EMMA STRUBELL, Carnegie Mellon University, USA
ALEXANDRA SASHA LUCCIONI, Hugging Face, USA
NOAH A. SMITH, Allen Institute for AI and University of Washington, USA
NICOLE DECARIO, Allen Institute for AI, USA
WILL BUCHANAN, Microsoft, USA



CO2 Relative Size Comparison

Green Software Foundation

Microsoft Azure

AI and the Environment

- Evidence around the **most expensive experiments**
  - More recent models consume 2-3 orders of magnitude more CO_2 (Luccioni et al., 2022)
  - But these are typically run very few times
- What about “normal” experiments?
  - Much cheaper, but run hundreds / thousands of times a day?
- What about **inference** operations?
  - Very cheap (though increasingly more expensive)
  - Run billions of times a day?
  - 80-90% of AI computation is spent on inference

We need better reporting!

Green AI

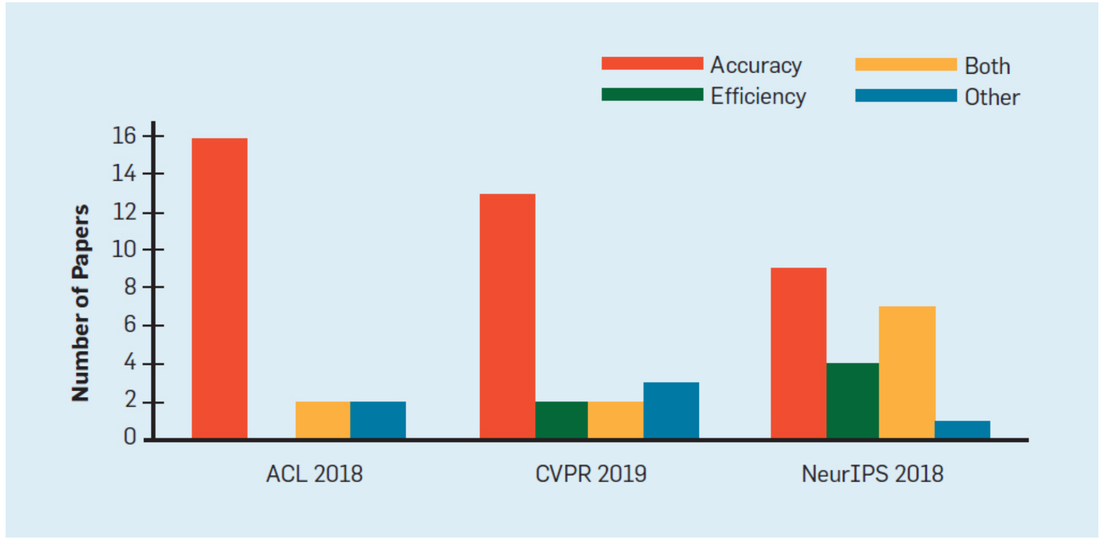
Schwartz*, Dodge*, Smith & Etzioni, CACM 2020

- Red AI
  - Problems: inclusiveness, environment
- Green AI

  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results

  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy

Accuracy or Efficiency?



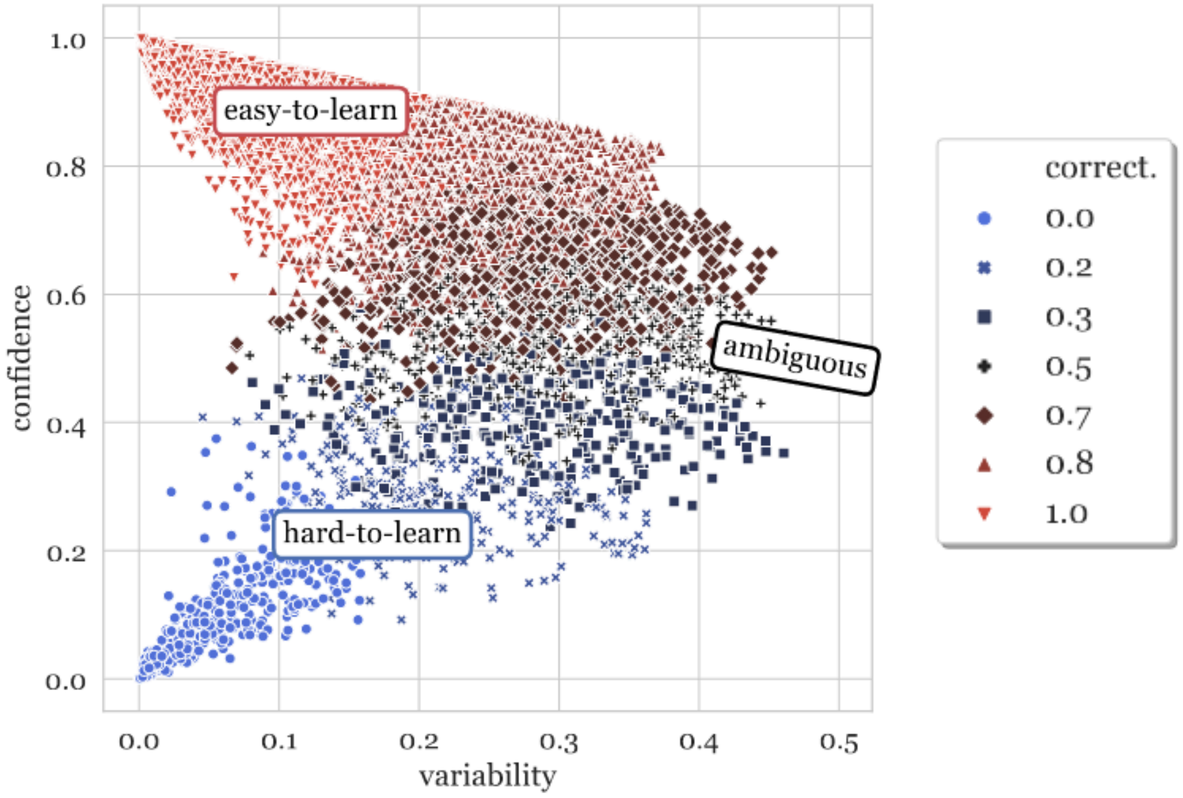
S. et al. (2020)

Smart Filtering For Faster Training

Swayamdipta, S. et al., EMNLP 2020

- Not all training instances contribute the same to learning
  - Some are “*easy-to-learn*”, others are more challenging

Dataset Maps



Experiments

WinoGrande, RoBERTa-Large

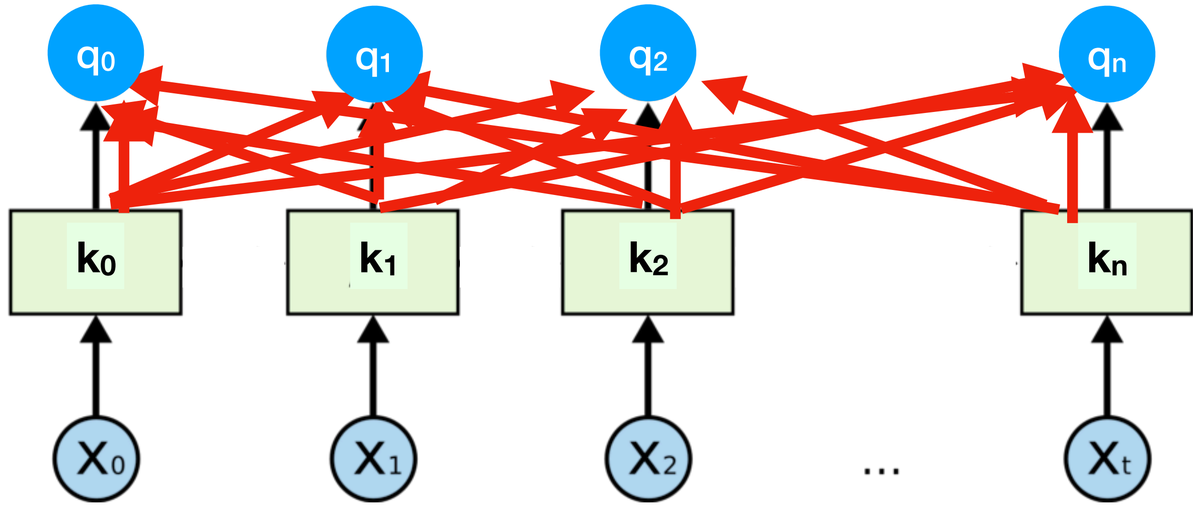
33%

	WINOG. Val. (ID)	*WSC* (OOD)
100% train	$79.7_{0.2}$	$86.0_{0.1}$
random	$73.3_{1.3}$	$85.6_{0.4}$
ambiguous	$\mathbf{78.7}_{0.4}$	$\mathbf{87.6}_{0.6}$

Transformers

Vaswani et al., 2017

- **The** method for text representation
  - Also for vision, speech, combio, …
- Each word *attends* to all other words
- O(n^2) complexity in the sentence length n
- Fatal for long sequences
  - Books, articles, etc.



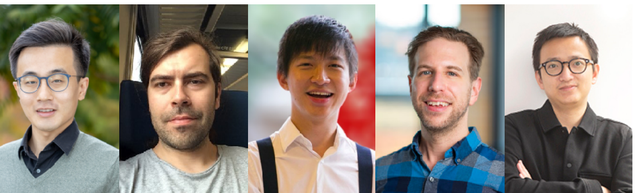
Random Feature Attention

Peng, Pappas, Yogatama, S., Smith, & Kong, ICLR 2021

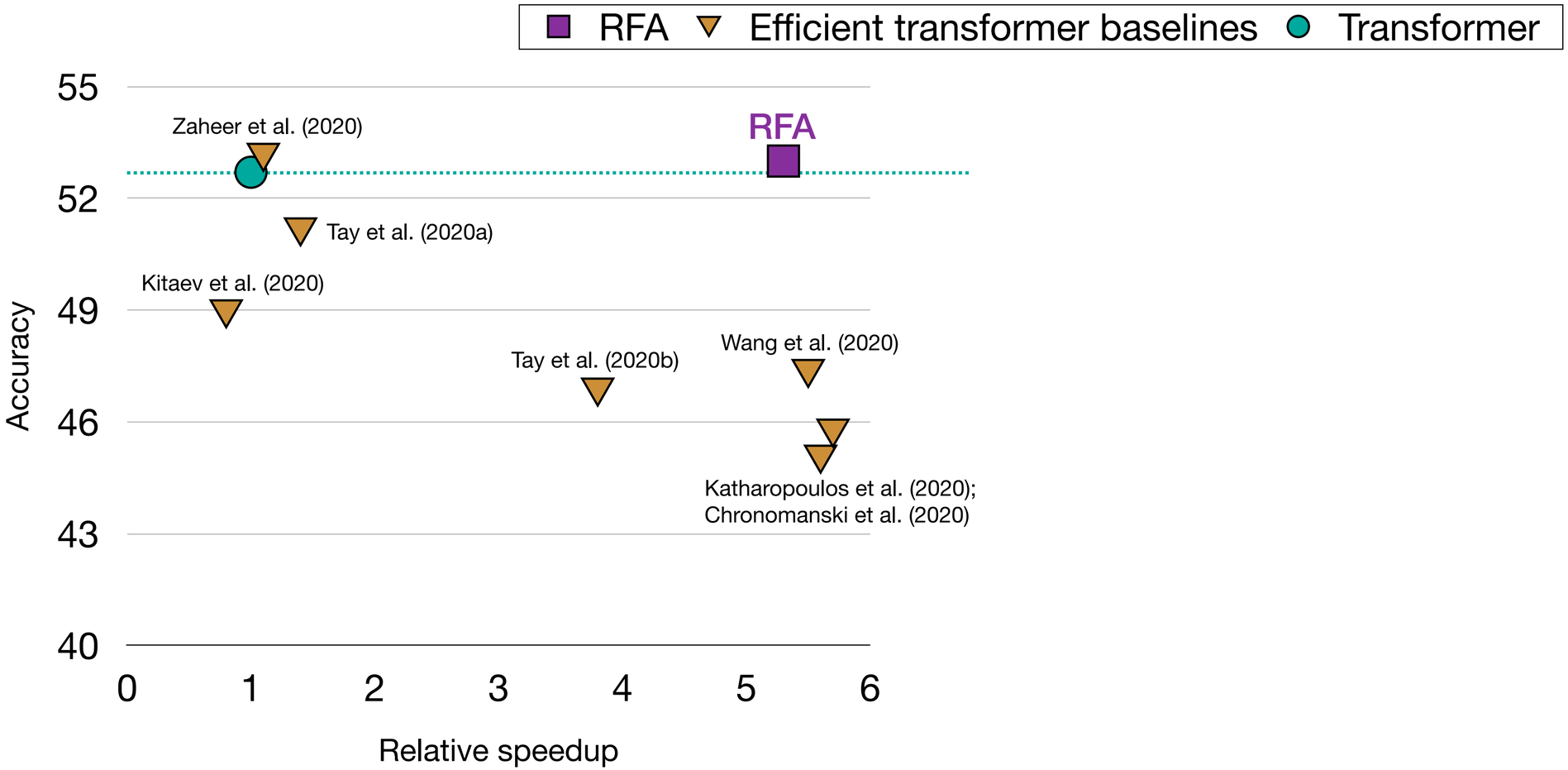
spotlight presentation

- **Key idea**: approximate the attention function using random Fourier features
  - Rahimi and Recht (2007)
- Some math
- Linear runtime and memory requirements

- *Define*
  - $\phi(\mathbf{x}) = \sqrt{1/D}[\sin\left(\mathbf{w}_1 \cdot \mathbf{x}\right), \ldots, \sin\left(\mathbf{w}_D \cdot \mathbf{x}\right), \cos\left(\mathbf{w}_1 \cdot \mathbf{x}\right), \ldots, \cos\left(\mathbf{w}_D \cdot \mathbf{x}\right)]^\top$
  - Where $w_i \sim N(0,1)$
- Then
  - $E[\phi(\mathbf{q})^T \phi(\mathbf{k})] = \exp(\mathbf{q}^T \cdot \mathbf{k})$
- $$\text{attn}\left(\mathbf{q}_t, \{\mathbf{k}_i\}, \{\mathbf{v}_i\}\right) = \sum_i softmax(\mathbf{q}_t, \mathbf{k}_i)\mathbf{v}_i^\top = \sum_i \frac{\exp\left(\mathbf{q}_t \cdot \mathbf{k}_i\right)}{\sum_j \exp\left(\mathbf{q}_t \cdot \mathbf{k}_j\right)}\mathbf{v}_i^\top$$
- $$\approx \sum_i \frac{\phi(\mathbf{q}_t)^T \phi(\mathbf{k}_i)}{\sum_j \phi(\mathbf{q}_t)^T \phi(\mathbf{k}_j)}\mathbf{v}_i^\top$$
- $$= \frac{\phi(\mathbf{q}_t)^T \sum_i \phi(\mathbf{k}_i) \otimes \mathbf{v}_i^\top}{\phi(\mathbf{q}_t)^T \sum_j \phi(\mathbf{k}_j)}$$



Better Efficiency-Accuracy Tradeoff

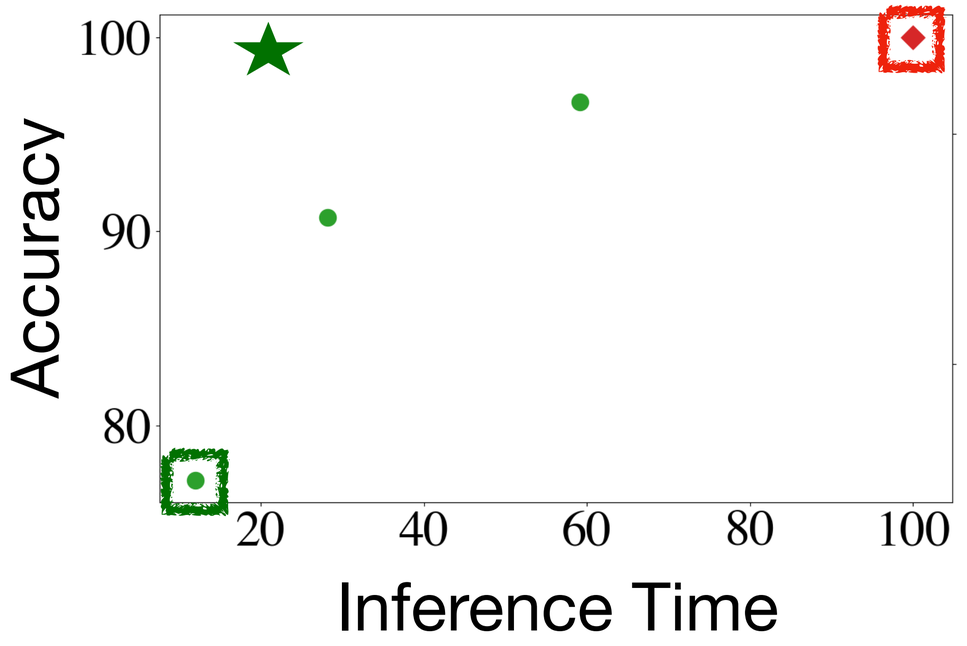


Matching Model and Instance Complexity

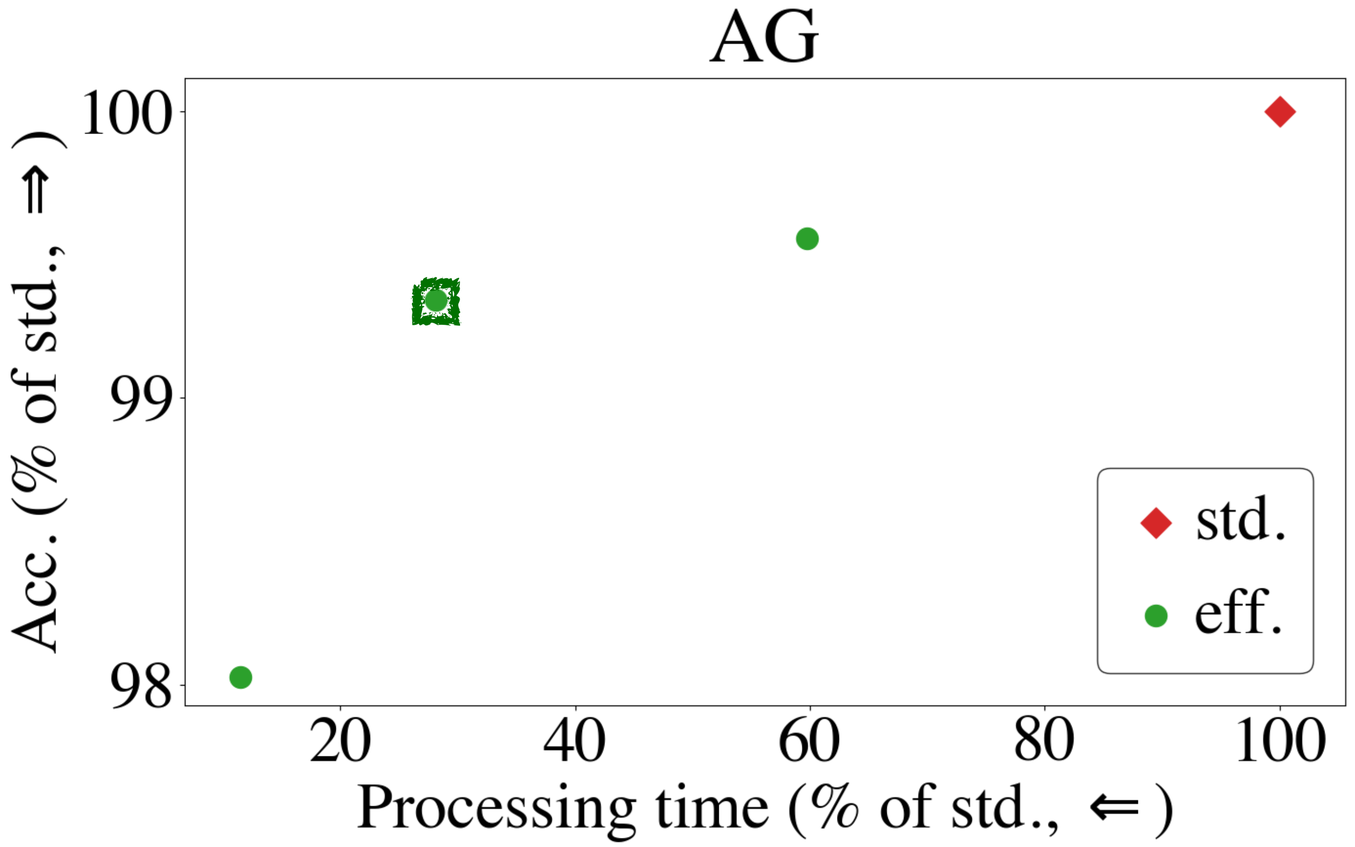
S. et al., ACL 2020

Run an efficient model on "easy" instances,

and an expensive model on "hard" instances

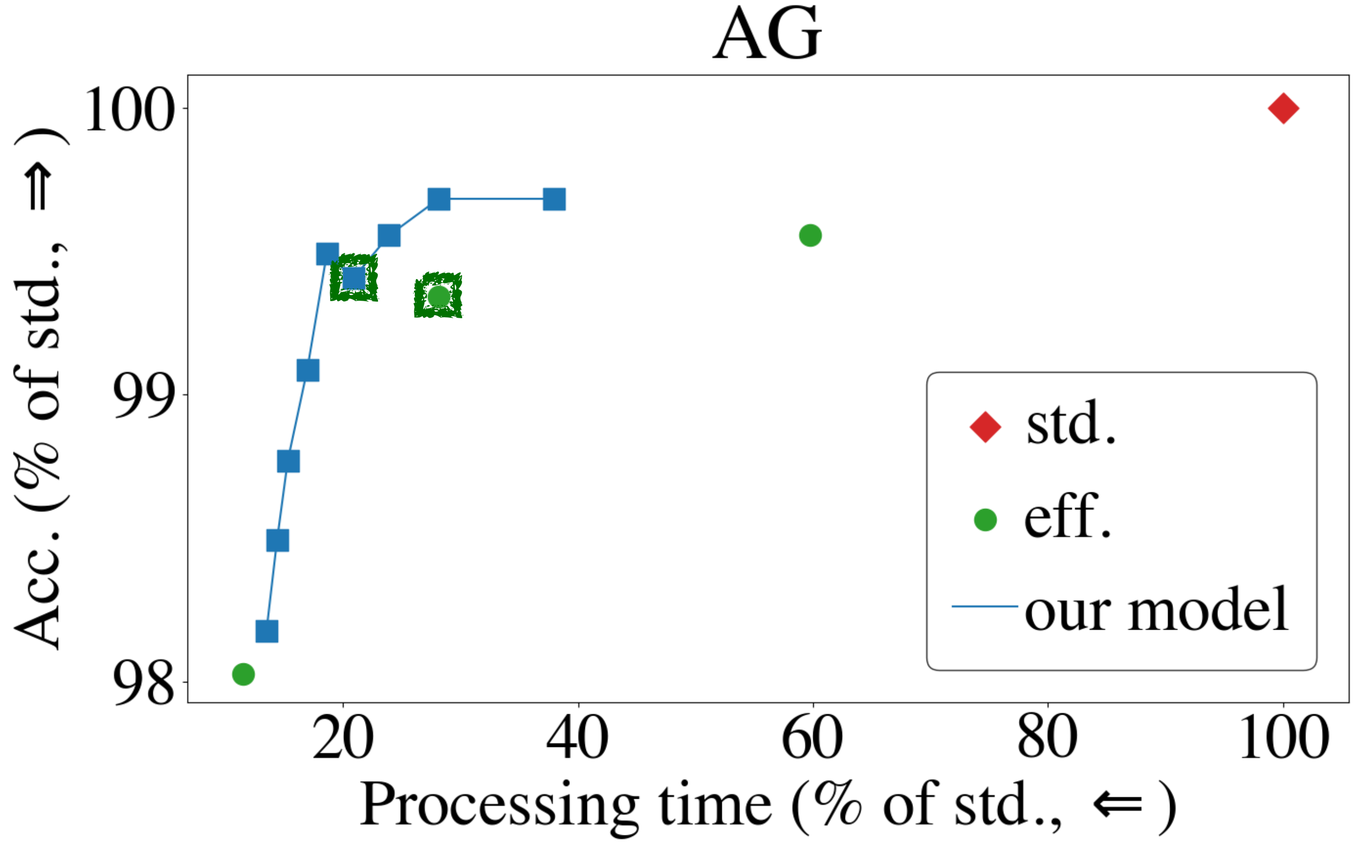


Strong Baselines!



3 times faster, within 1% of full model

Better Speed/Accuracy Tradeoff



5 times faster, within 1% of full model

Think Green!

- Red AI
  - Problems: inclusiveness, environment
- Green AI
  - Enhance reporting of computational budgets
    - Add a *price-tag* for scientific results
  - Promote efficiency as a core evaluation for AI
    - **In addition to** accuracy