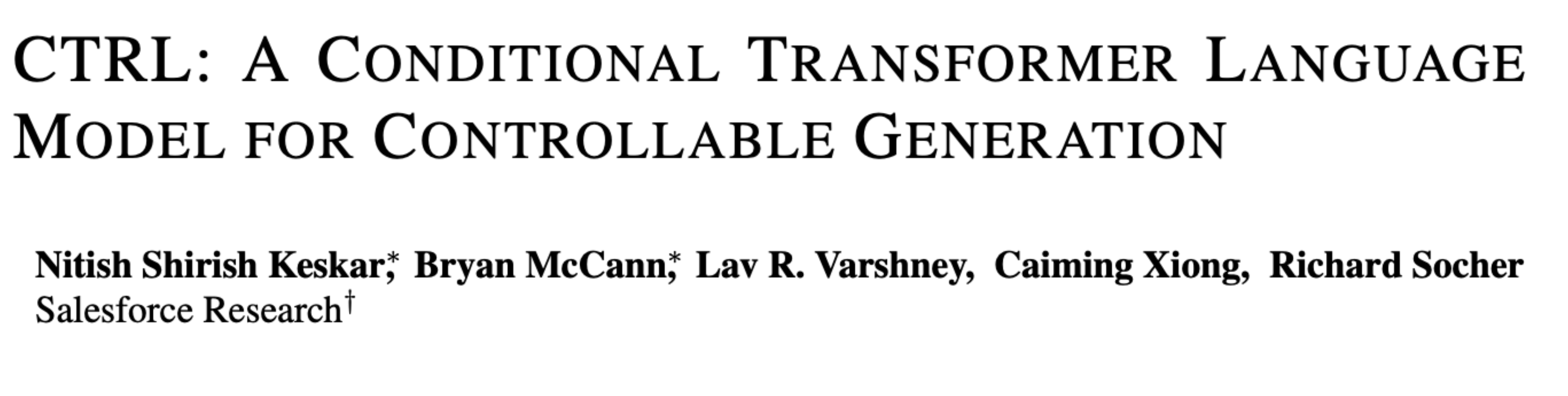
Generative AI Models
ECE 598 LV – Lecture 19

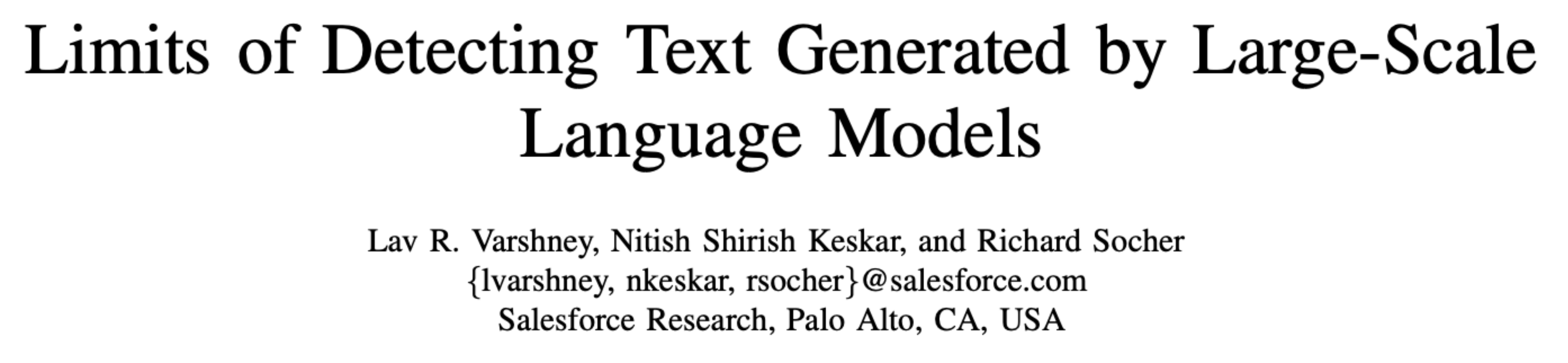
Lav R. Varshney

31 March 2022



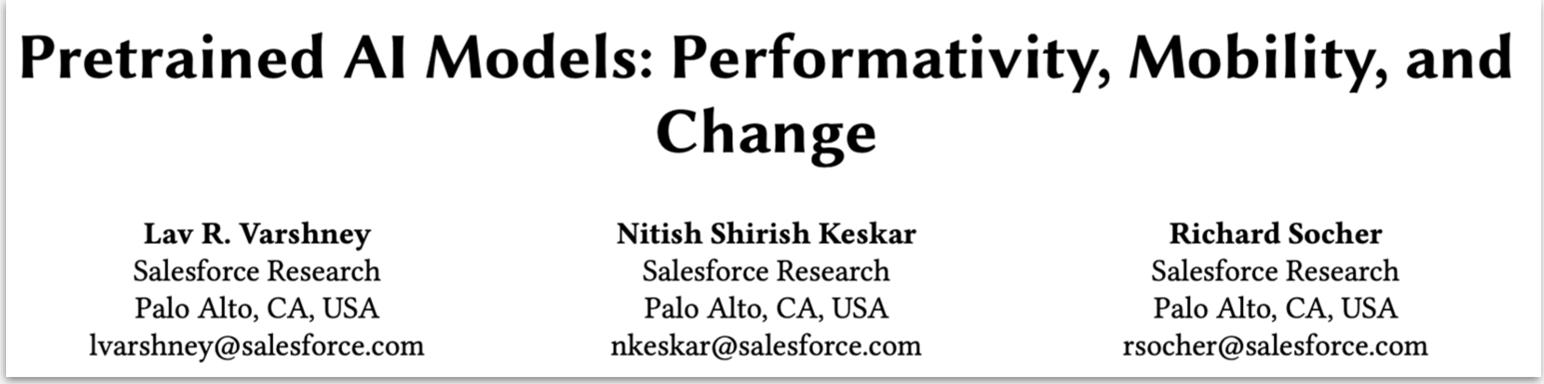
CTRL: A Conditional Transformer Language Model for Controllable Generation

Nitish Shirish Keskar,[*] Bryan McCann,[*] Lav R. Varshney, Caiming Xiong, Richard Socher
Salesforce Research[†]



Limits of Detecting Text Generated by Large-Scale Language Models

Lav R. Varshney, Nitish Shirish Keskar, and Richard Socher
{lvarshney, nkeskar, rsocher}@salesforce.com
Salesforce Research, Palo Alto, CA, USA



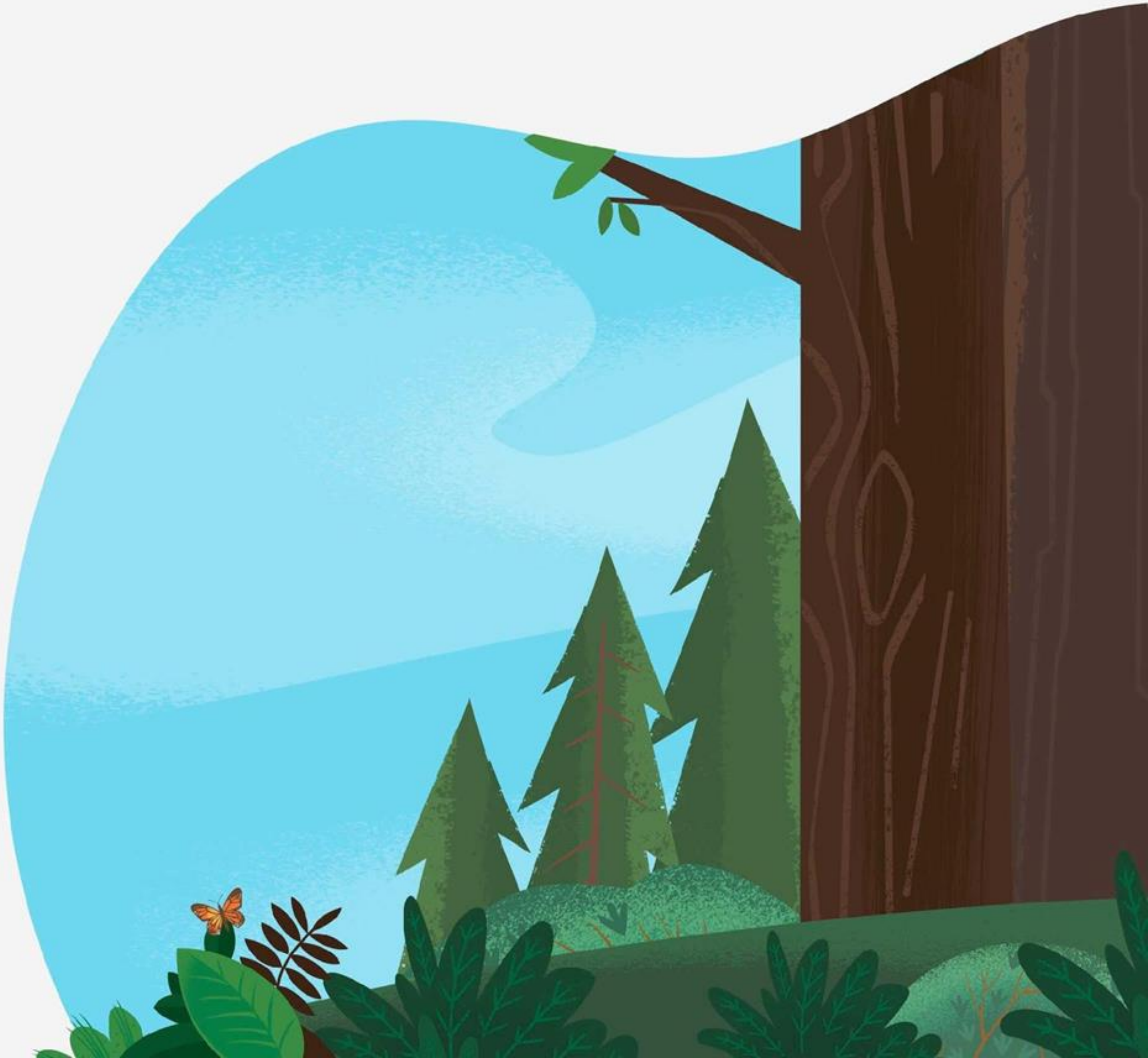
Pretrained AI Models: Performativity, Mobility, and Change

Lav R. Varshney
Salesforce Research
Palo Alto, CA, USA
lvarshney@salesforce.com

Nitish Shirish Keskar
Salesforce Research
Palo Alto, CA, USA
nkeskar@salesforce.com

Richard Socher
Salesforce Research
Palo Alto, CA, USA
rsocher@salesforce.com

Detection



Deepfakes

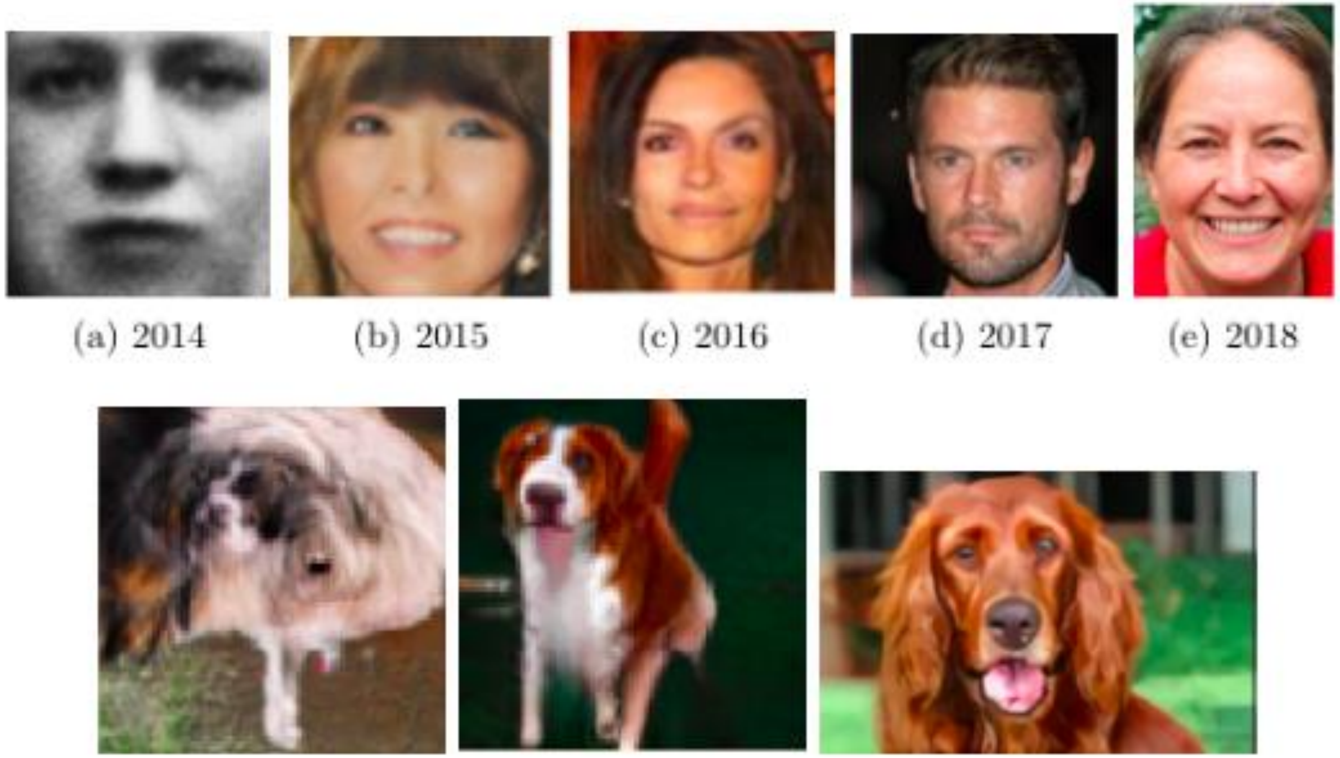
- Portmanteau of “deep learning” and “fake”
- Refer to any realistic audiovisual content generated using advanced deep network-based generative algorithms, typically GANs. The term implies the use of such manipulated content for unethical purposes.

Figure: A few sample faces generated by ThisPersonDoesNotExist.com

Negative Applications

- Blackmail, harassment, sabotage, etc.
- Dating scams or “catfishing”
- Political propaganda, inciting violence, discrediting leaders and institutions
- Deepfake porn
- Spreading misinformation

Image Generation using GAN



(a) 2014 (b) 2015 (c) 2016 (d) 2017 (e) 2018

(a) 2016 (b) 2017 (c) 2018

Figure: The striking achievements in GAN-generated image quality of human faces and animals over the years.

Detecting Deepfakes

- Several studies use features based on visual artifacts, image quality, or lipsync for classification
- With advances in generative models, these and other such techniques may soon become obsolete
- Develop general statistical framework to yield guarantees on detection performance, by building on information-theoretic study of authentication
- Cast deepfake detection as a hypothesis testing problem specifically for outputs of GANs, themselves viewed through a generalized robust statistics framework

GAN formulation

- $\mathbb{P}_X$: input distribution.
- Objective: pick generator function $g(\cdot) \in \mathcal{G}$, giving an output distribution $\mathbb{P}_{g(Z)}$. Optimal generator function $\hat{g}$ is selected such that:
  $$\hat{g} = \arg\min_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)}).$$
  where L could be Jensen-Shannon divergence, Wasserstein metric, etc. for different GAN designs
- Intuitively, the design of the GAN is such that $\inf_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)})$ is small and we can view the real distribution $\mathbb{P}_X$ as a slightly perturbed version of a generated distribution $\mathbb{P}_{g(Z)}$ under the distance measure $L(\cdot, \cdot)$
- Define oracle error as
  $$OPT := \inf_{g \in \mathcal{G}} L(\mathbb{P}_X, \mathbb{P}_{g(Z)})$$

Hypothesis Testing Error Bounds

The n i.i.d. pixels of the image are $Y_1, Y_2, Y_3, \ldots, Y_n$, and accordingly the hypothesis test will be as follows:

$$H_0 := Y \sim \mathbb{P}_X$$
$$H_1 := Y \sim \mathbb{P}_{\hat{g}(Z)}$$

- We have general error bounds for hypothesis testing:

  Neyman-Pearson error: $\beta_n^{\epsilon} \doteq \exp(-nD(\mathbb{P}_X || \mathbb{P}_{\hat{g}(Z)}))$

  Bayesian error: $P_e^{(n)} \leq \exp\{-nC(\mathbb{P}_X, \mathbb{P}_{\hat{g}(Z)})\}.$

  where C is the Chernoff information
- Use these together with relationships among loss functions, such as Pinsker's inequality, to bound deepfake detection error for different L

Hypothesis Testing Error Bounds

L-function	Neyman-Pearson Bound $\beta_n^{\epsilon} \leq$	Bayesian Bound $P_e^{(n)} \leq$
KL divergence	$\exp(-nOPT)$	$(1 - \frac{P_g^*}{4} OPT)^{n/2}$
Total variational distance	$\exp(-2nOPT^2)$	$(1 - OPT^2)^{n/2}$
Jensen-Shannon divergence	$\exp(-\frac{n}{2} OPT^2)$	$(1 - \frac{1}{4} OPT^2)^{n/2}$
Wasserstein	$\exp\left(-\frac{2n}{\mathrm{diam}(\mathcal{X})^2} OPT^2\right)$	$\left(1 - \left(\frac{OPT}{\mathrm{diam}(\mathcal{X})}\right)^2\right)^{n/2}$

Summary of results

- As *OPT* increases, i.e., the GAN used is less accurate, it is exponentially easier to detect deepfakes in the Neyman-Pearson case, and polynomially so in the Bayesian case
- As the resolution n increases, we require GAN to have higher accuracy for the deepfake to go undetected.
- Formulation can be extended for images with distinct regions having different distributions like eyes, hair, vehicles, etc. by considering k patches of $m = \frac{n}{k}$ i.i.d. pixels each. The generalized bounds can be obtained *mutatis mutandis* by multiplying together the corresponding bounds for each patch.

Euclidean Approximation

- When two distributions are close to each other, i.e. $P \approx Q$, the KL divergence between them can be approximated as

  $$D(P||Q) \approx D(Q||P) \approx \frac{1}{2}||Q - P||_{\hat{P}}^2 = \frac{1}{2}||[\hat{P}^{-1/2}](Q - P)||^2$$

  where the weight $\hat{P}$ is any distribution in the neighborhood of P and Q.
- When $P \sim Q$, the weight vector in the approximation $\hat{P}$ can be any distribution in their neighborhood.
- Assuming the GAN to be highly accurate in mimicking the true distribution, the bounds can be further simplified using the approximation for such low error regime.

S. Borade and L. Zheng, "Euclidean Information Theory," *2008 IEEE International Zurich Seminar on Communications*, pp. 14–17.

Simplified Error Bounds

L-function	Neyman-Pearson Bound $\beta_n^\epsilon \leq$	Bayesian Bound $P_e^{(n)} \leq$
KL divergence	$\exp(-nOPT)$	$\exp(-\frac{n}{4}OPT)$
Total variational distance	$\exp(-2nOPT^2)$	$\exp(-\frac{n}{2}OPT^2)$
Jensen-Shannon divergence	$\exp(-2nOPT)$	$\exp(-\frac{n}{2}OPT)$
Wasserstein	$\exp\left(-\frac{2n}{\mathrm{diam}(\mathcal{X})^2}OPT^2\right)$	$\exp\left(-\frac{n}{2\ \mathrm{diam}(\mathcal{X})^2}OPT^2\right)$

- It is now exponentially easier to detect deepfakes from GANs with higher *OPT* for both cases.

Epidemic Threshold Theory and Deepfakes

- SIR (susceptible-infected-recovered) model for epidemics, where detecting a deepfake is cast as recovered state
- Epidemic threshold is the critical level λ_c for effective spreading rate λ above which a global epidemic occurs.
- Can be predicted based on the network structure.
- Effective spreading rate is given by

$$]\lambda = \frac{\beta}{\gamma}$$

where β is the probability that an infected node transmits a disease, or in our case, shares a deepfake with a susceptible node, and γ is the probability of recovery of an infected node, or the probability that the deepfakes are correctly detected.

AI Safety: Limits of Deepfake Detection

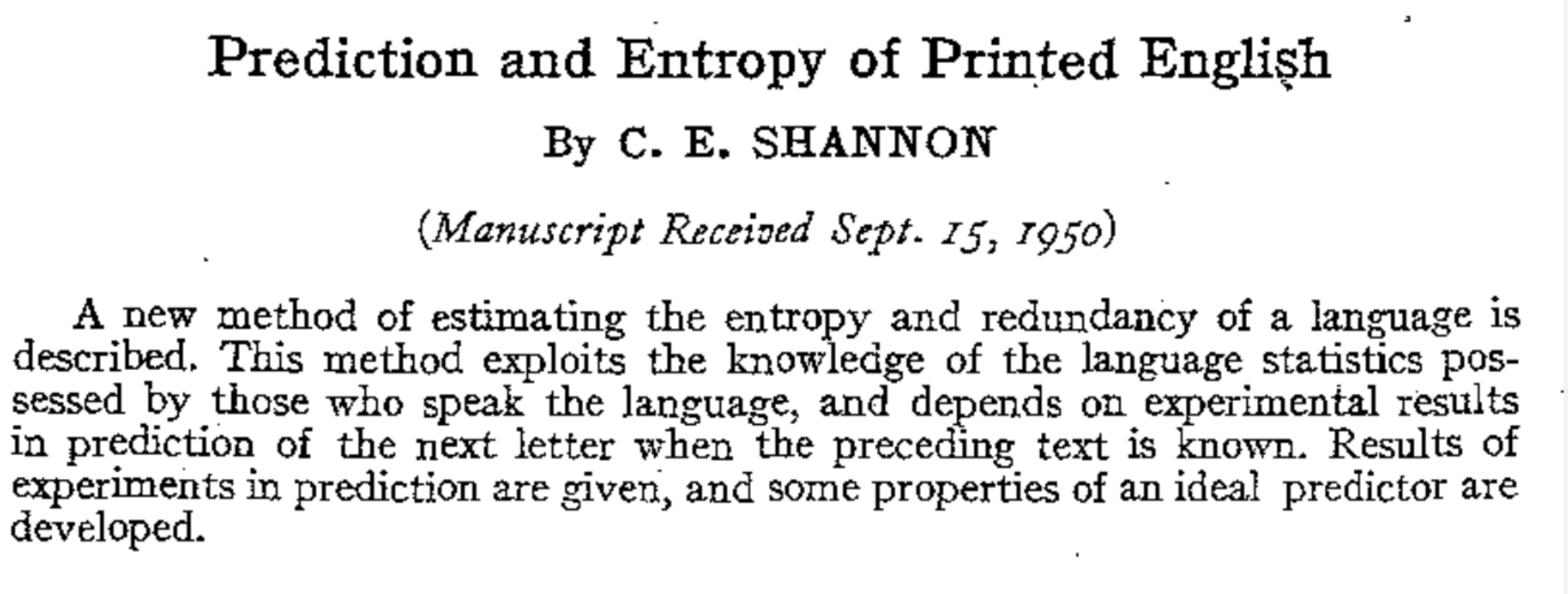
S. Agarwal and L. R. Varshney (ICML 2019 workshop)

thispersondoesnotexist.com

- Blackmail, harassment, sabotage, etc.
- Dating scams or “catshing”
- Political propaganda, inciting violence, discrediting leaders and institutions
- Deepfake porn
- Fake news and misinformation

Building on [B. Zhu, J. Jiao, and D. Tse, Feb. 2019], we established that error exponents for detecting GAN-generated deepfakes are governed by the oracle error for particular GAN loss function used in training

Rather than considering deepfake images, consider generative models for natural language text



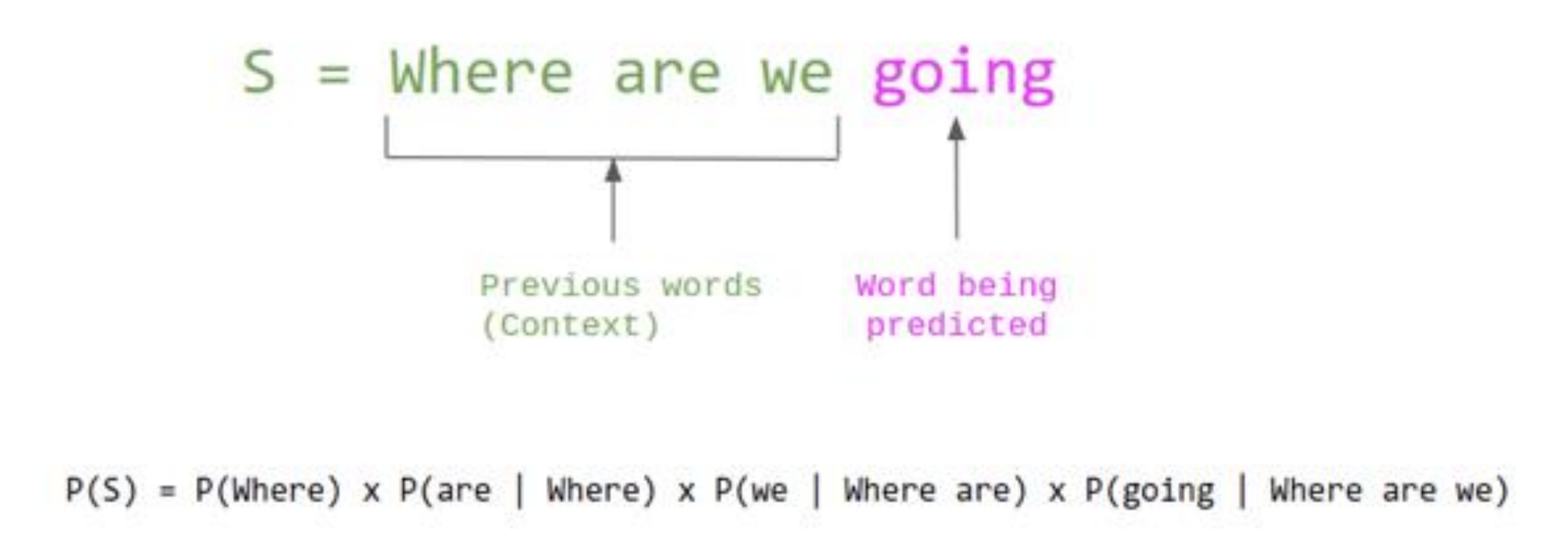
Prediction and Entropy of Printed English

By C. E. SHANNON

(Manuscript Received Sept. 15, 1950)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

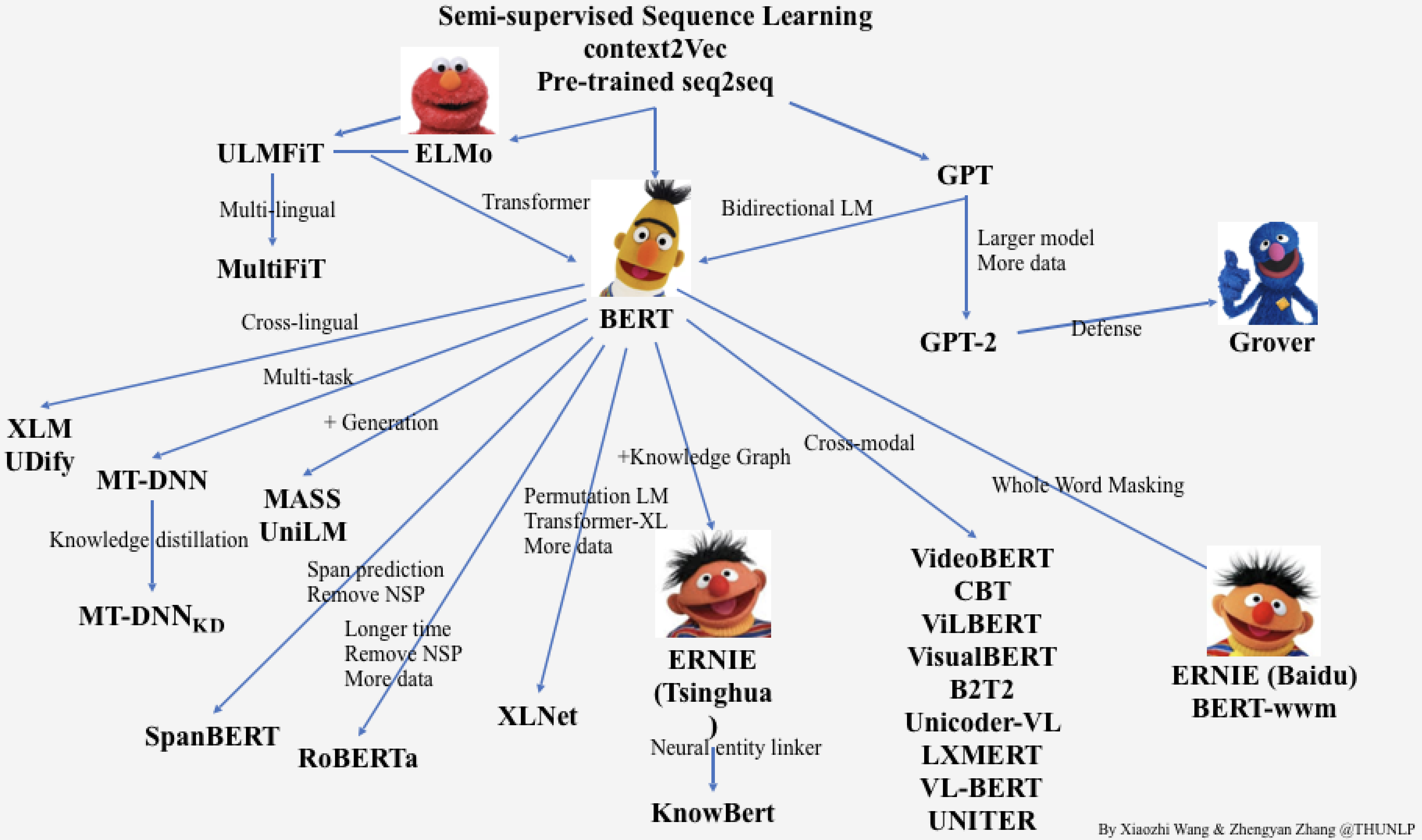
Large-scale (conditional) language models



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

$$p(x|c) = \prod_{i=1}^{n} p(x_i|x_{<i}, c)$$

[N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, “CTRL: A Conditional Transformer Language Model for Controllable Generation,” arXiv, 2019.]



Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq

ULMFiT

ELMo

GPT

Multi-lingual

MultiFiT

Transformer

Bidirectional LM

BERT

Larger model
More data

GPT-2

Defense

Grover

Cross-lingual

XLM
UDify

Multi-task

MT-DNN

Knowledge distillation

MT-DNN$_{KD}$

+ Generation

MASS
UniLM

Span prediction
Remove NSP

SpanBERT

Longer time
Remove NSP
More data

RoBERTa

Permutation LM
Transformer-XL
More data

XLNet

+Knowledge Graph

ERNIE
(Tsinghua
)

Neural entity linker

KnowBert

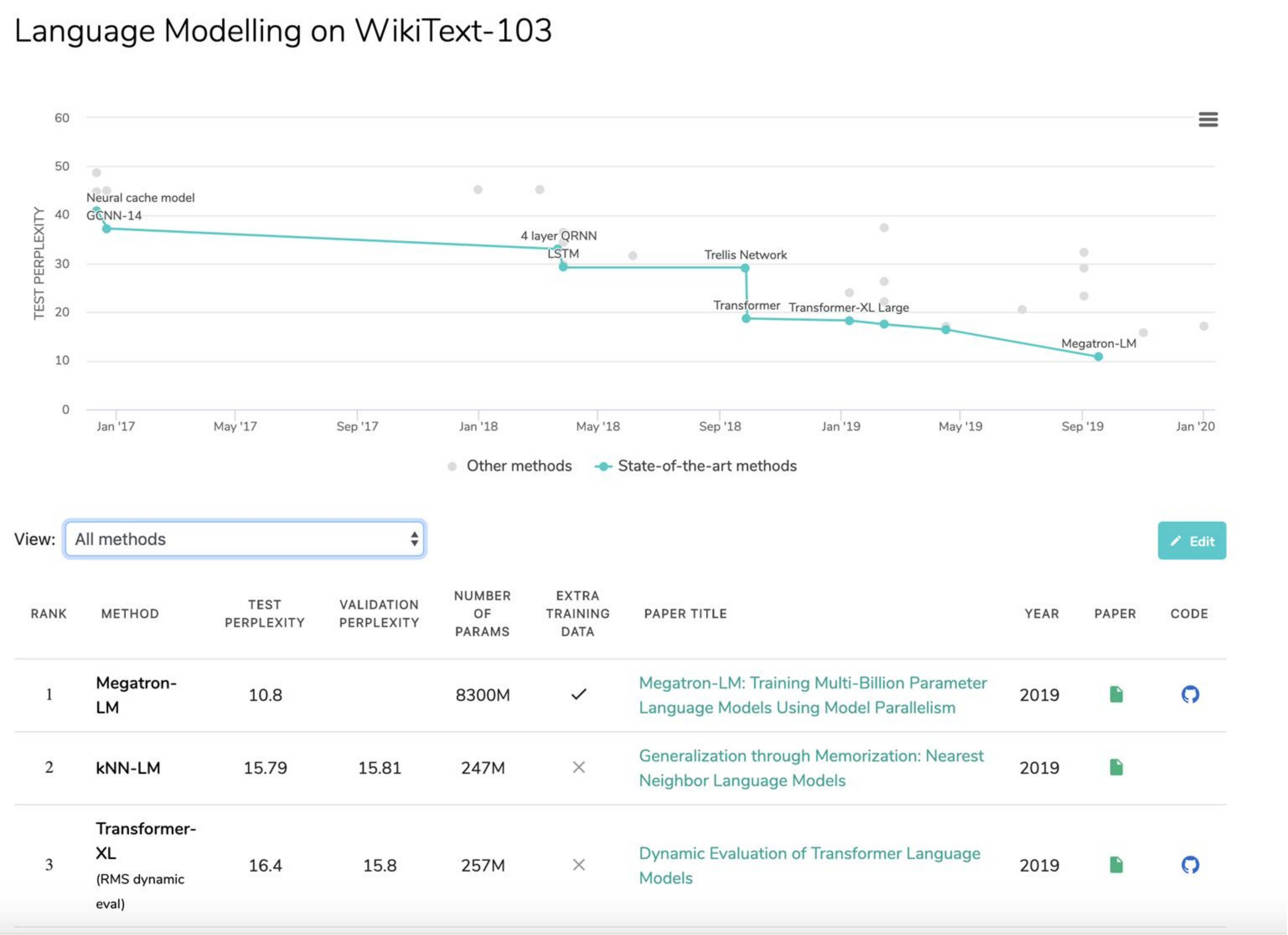
Cross-modal

VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER

Whole Word Masking

ERNIE (Baidu)
BERT-wwm

By Xiaozhi Wang & Zhengyan Zhang @THUNLP



Language Modelling on WikiText-103

View: All methods

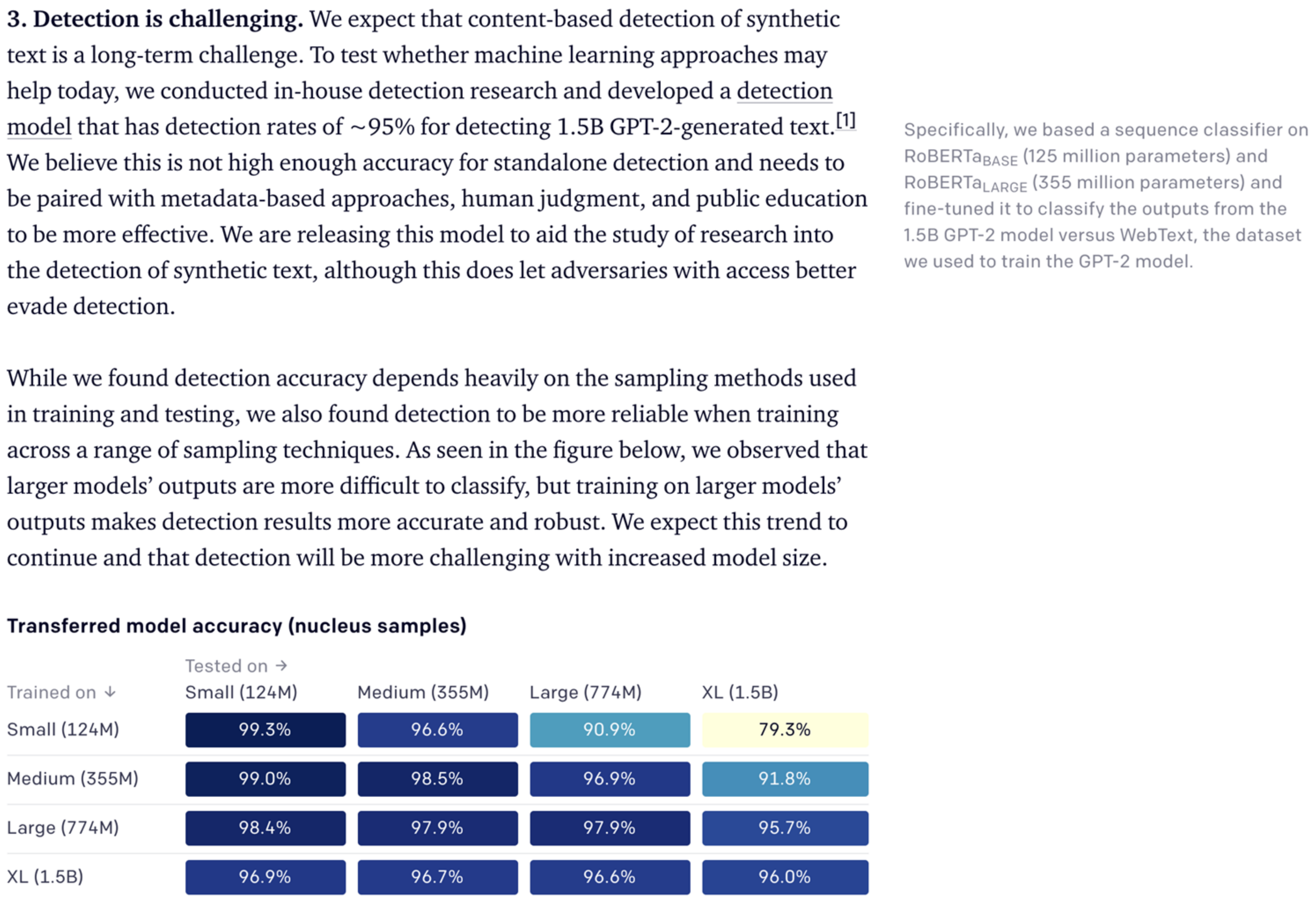
RANK	METHOD	TEST PERPLEXITY	VALIDATION PERPLEXITY	NUMBER OF PARAMS	EXTRA TRAINING DATA	PAPER TITLE	YEAR	PAPER	CODE
1	**Megatron-LM**	10.8		8300M	✓	Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism	2019	📄	●
2	**kNN-LM**	15.79	15.81	247M	×	Generalization through Memorization: Nearest Neighbor Language Models	2019	📄	
3	**Transformer-XL** (RMS dynamic eval)	16.4	15.8	257M	×	Dynamic Evaluation of Transformer Language Models	2019	📄	●

Perplexity

$$\mathrm{PPL}(P, Q) = 2^{H(P,Q)}$$

A reformulation of the cross-entropy between the true language *P* and the model *Q* has been empirically shown to measure the quality of generated output of language models, but no specific formal operational interpretation through a “coding theorem”

Detection of model-generated text (GPT-2)



3. Detection is challenging. We expect that content-based detection of synthetic text is a long-term challenge. To test whether machine learning approaches may help today, we conducted in-house detection research and developed a detection model that has detection rates of ~95% for detecting 1.5B GPT-2-generated text.[1] We believe this is not high enough accuracy for standalone detection and needs to be paired with metadata-based approaches, human judgment, and public education to be more effective. We are releasing this model to aid the study of research into the detection of synthetic text, although this does let adversaries with access better evade detection.

Specifically, we based a sequence classifier on RoBERTa$_{\text{BASE}}$ (125 million parameters) and RoBERTa$_{\text{LARGE}}$ (355 million parameters) and fine-tuned it to classify the outputs from the 1.5B GPT-2 model versus WebText, the dataset we used to train the GPT-2 model.

While we found detection accuracy depends heavily on the sampling methods used in training and testing, we also found detection to be more reliable when training across a range of sampling techniques. As seen in the figure below, we observed that larger models' outputs are more difficult to classify, but training on larger models' outputs makes detection results more accurate and robust. We expect this trend to continue and that detection will be more challenging with increased model size.

Transferred model accuracy (nucleus samples)

	Tested on →			
Trained on ↓	Small (124M)	Medium (355M)	Large (774M)	XL (1.5B)
Small (124M)	99.3%	96.6%	90.9%	79.3%
Medium (355M)	99.0%	98.5%	96.9%	91.8%
Large (774M)	98.4%	97.9%	97.9%	95.7%
XL (1.5B)	96.9%	96.7%	96.6%	96.0%

GPT-2 report

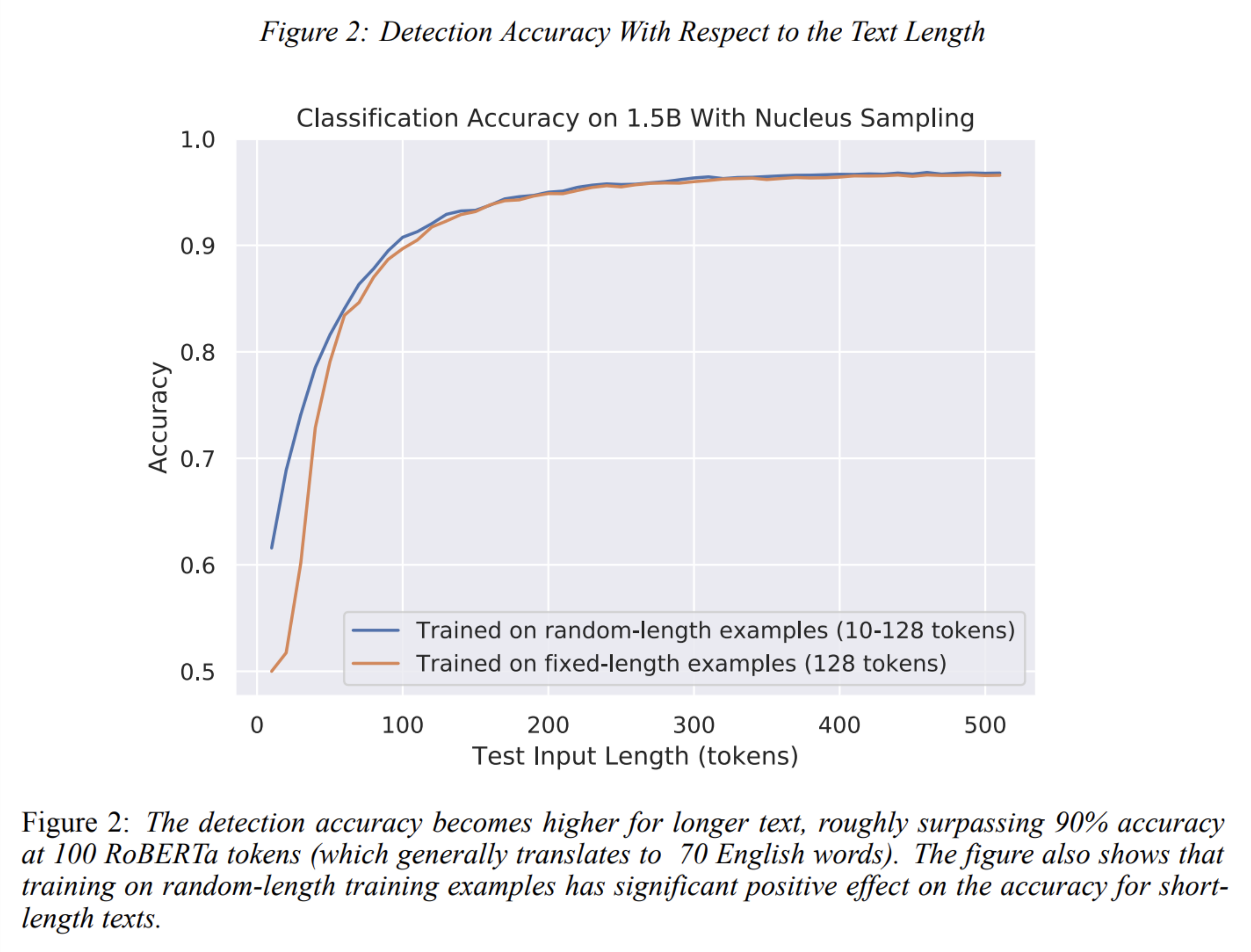


Figure 2: Detection Accuracy With Respect to the Text Length

Figure 2: *The detection accuracy becomes higher for longer text, roughly surpassing 90% accuracy at 100 RoBERTa tokens (which generally translates to 70 English words). The figure also shows that training on random-length training examples has significant positive effect on the accuracy for short-length texts.*

Limits of Detecting Language Model Outputs

The distribution of authentic text is denoted *P* and the distribution of text generated by the language model is *Q*. Suppose we have access to *n* tokens of generated text from the language model, which we call $Y_1, \ldots, Y_n$. We can formalize the hypothesis test as:

$$H_0 := Y \sim P \text{ (authentic)}$$

$$H_1 := Y \sim Q \text{ (LM generated)}$$

For the Neyman-Pearson formulation of fixing false alarm probability and maximizing true detection probability, assuming ergodicity, it is known that the error exponent is given by the asymptotic Kullback-Leibler divergence rate, defined as the almost-sure limit of:

$$\frac{1}{n} \log \frac{P_n}{Q_n}(y_1, \ldots, y_n), \text{ as } n \to \infty,$$

if the limit exists, where P_n and Q_n are the null and alternate joint densities of $(Y_1, \ldots, Y_n)$, cf. [Sung, Tong, Poor, 2006; Luschgy, Rukhin, Vajda, 1993]

Limits of Detecting Language Model Outputs

Suppose we are given a specific language model like GPT-2, characterized in terms of its cross-entropy or perplexity.

perplexity $$\mathrm{PPL}(P,Q) = 2^{H(P,Q)}$$

cross-entropy $$H(P,Q) = H(P) + D_{\mathrm{KL}}(P||Q)$$

Since we think of *H*(*P*) as a constant, we observe that the error exponent for the decision problem is precisely an affine shift of the cross-entropy. Outputs from models that are better in the sense of cross-entropy or perplexity are harder to distinguish from authentic text, in this precise sense.

Limits of Detecting Language Model Outputs

- Now rather than considering a particular language model, consider bounding error probability in detection of outputs of a maximum likelihood (ML) language model.
- Though not quite correct, human language text can be modeled as stationary, ergodic random processes from finite alphabet *A* [Manning and Schütze, 1999]
- Due to diversity in language production, assume stationary, ergodic process is *non-null* in the sense of [Csiszar and Talata, *IEEE Trans. Information Theory*, 2010], also called the *smoothing requirement*
- Further assume the random process has *summable continuity rate* [Csiszar and Talata, *IEEE Trans. Information Theory*, 2010]

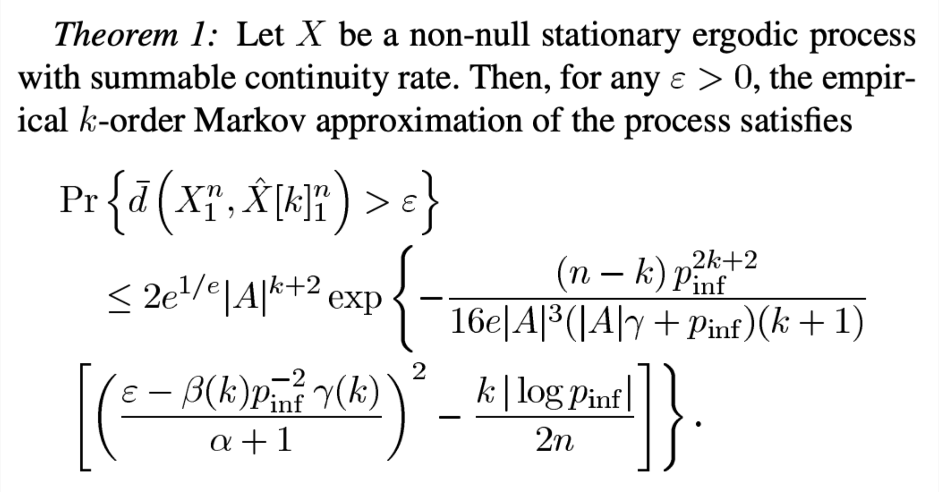
Think of language model as approximating a non-null, stationary, ergodic process with finite alphabet *A* with a *k*-order Markov process based on *n*-length realization

- We think of *k* as a large value and so the family encompasses the class of neural language models like GPT-2 and CTRL, which are *a fortiori* Markov in structure. Empirical results suggest LSTM and similar neural language models have Markov order as small as *k* = 13; the appropriate *k* for large-scale language models has not been investigated empirically

Limits of Detecting Language Model Outputs

Now we bound the excess probability of the error exponent in hypothesis testing, by first drawing on an excess probability bound for the Ornstein d-bar distance between a stationary, ergodic process and its Markov approximation, due to Csiszar and Talata (2010)

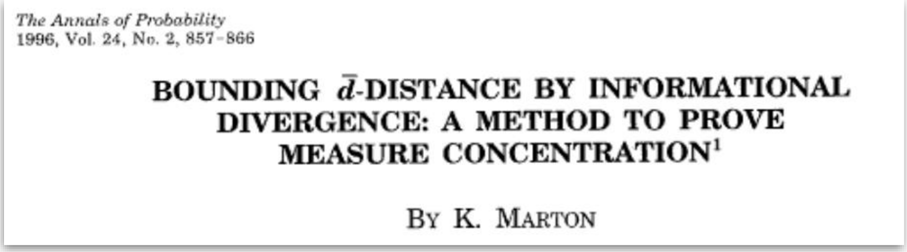
Then we bound the Ornstein d-bar distance using the Kullback-Leibler divergence via a conjectured extension of Marton's technique (1996) generalizing reverse Pinsker's inequality, to get the result in terms of standard measures of language model performance.



Theorem 1: Let X be a non-null stationary ergodic process with summable continuity rate. Then, for any $\varepsilon > 0$, the empirical k-order Markov approximation of the process satisfies

$$\Pr\left\{\bar{d}\left(X_1^n, \hat{X}[k]_1^n\right) > \varepsilon\right\} \leq 2e^{1/e}|A|^{k+2}\exp\left\{-\frac{(n-k)\,p_{\inf}^{2k+2}}{16e|A|^3(|A|\gamma + p_{\inf})(k+1)} \left[\left(\frac{\varepsilon - \beta(k)p_{\inf}^{-2}\,\gamma(k)}{\alpha+1}\right)^2 - \frac{k\,|\log p_{\inf}|}{2n}\right]\right\}.$$

From d-divergence to KL divergence



The Annals of Probability
1996, Vol. 24, No. 2, 857–866

BOUNDING $\bar{d}$-DISTANCE BY INFORMATIONAL DIVERGENCE: A METHOD TO PROVE MEASURE CONCENTRATION[1]

By K. Marton

“In AI circles, identifying fake media has long received less attention, funding and institutional backing than creating it: Why sniff out other people’s fantasy creations when you can design your own? ‘There’s no money to be made out of detecting these things,’ [Nasir] Memon said.” – D. Harwell, *The Washington Post*, Jun. 2019.