Data Science to Patient Value (D2V)
UNIVERSITY OF COLORADO **ANSCHUTZ MEDICAL CAMPUS**

# GPT-2: Qu'ils mangent de la brioche (let them eat cake)

MARCH 2019 – SETH RUSSELL, MS

# Overview

- What is Natural Language Processing
- Transfer learning in Natural Language Processing
- Language Modeling
- Transformer (no, not the robots!)
- Experiments/Results
- Reproducibility/Replicability
- Demo
- Now what?

# What is Natural Language Processing?

What is the study of language?

LINGUISTICS

Computer system that analyzes or synthesizes spoken or written human language

- Document categorization
- Document clustering
- Entity extraction
- Entity resolution

# What is Natural Language Processing

Syntax: a set of rules for a language

Semantics: the meaning of a word, phrase, sentence or text

Context: the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning

Morphology: study of the internal structure of words. Stemming comes from this.

- hopped -> hop
- hopping -> hop

Tokenization: chopping text into a sequence of characters that are grouped together as a useful semantic unit for processing

Colorless green ideas sleep furiously

# Evolution of Natural Language Processing

Linguistic approach: Top-Down symbolic approach focusing on syntax, semantics

Statistical approach: Bottom-Up focus on associations and probabilities.

Change of focus from linguistic theory to reliance upon corpus statistics

Modern machines can process millions of documents by computing billions of calculations to build statistical profiles from large corpora

# Current Problems in NLP


# NLP Transfer Learning

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained …

**Universal Language Model Fine-tuning for Text Classification**

**Jeremy Howard***
fast.ai
University of San Francisco
j@fast.ai

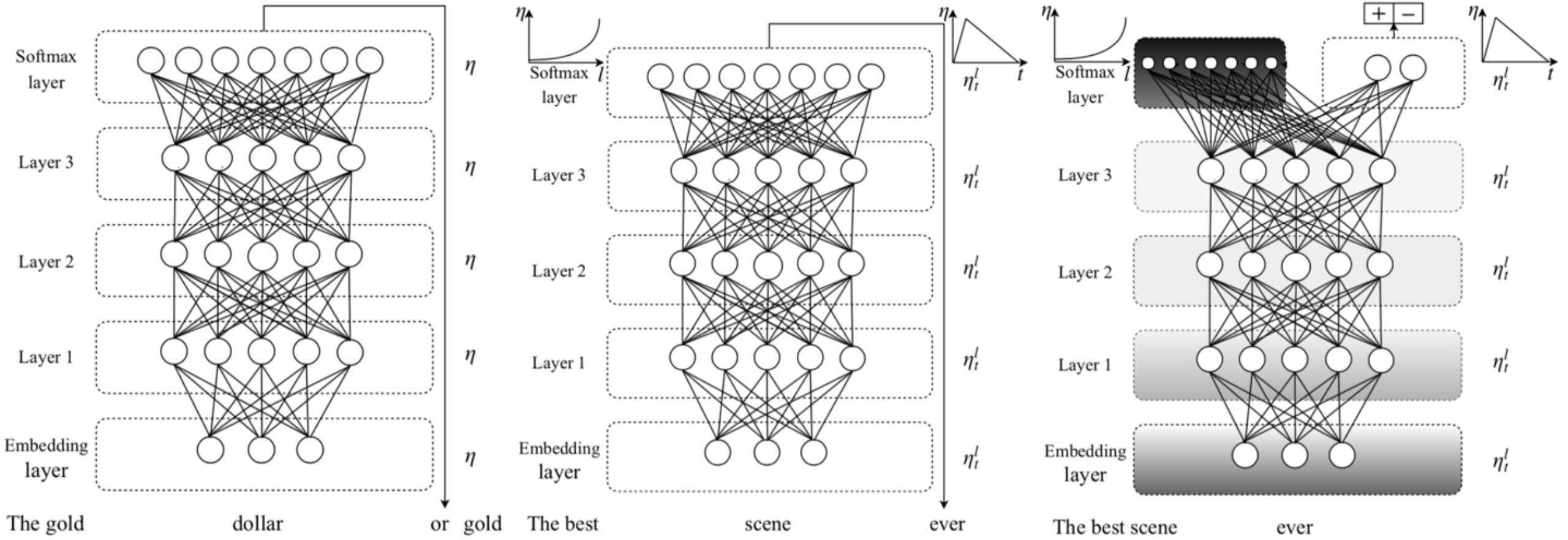
**Sebastian Ruder***
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

# Universal Language Model

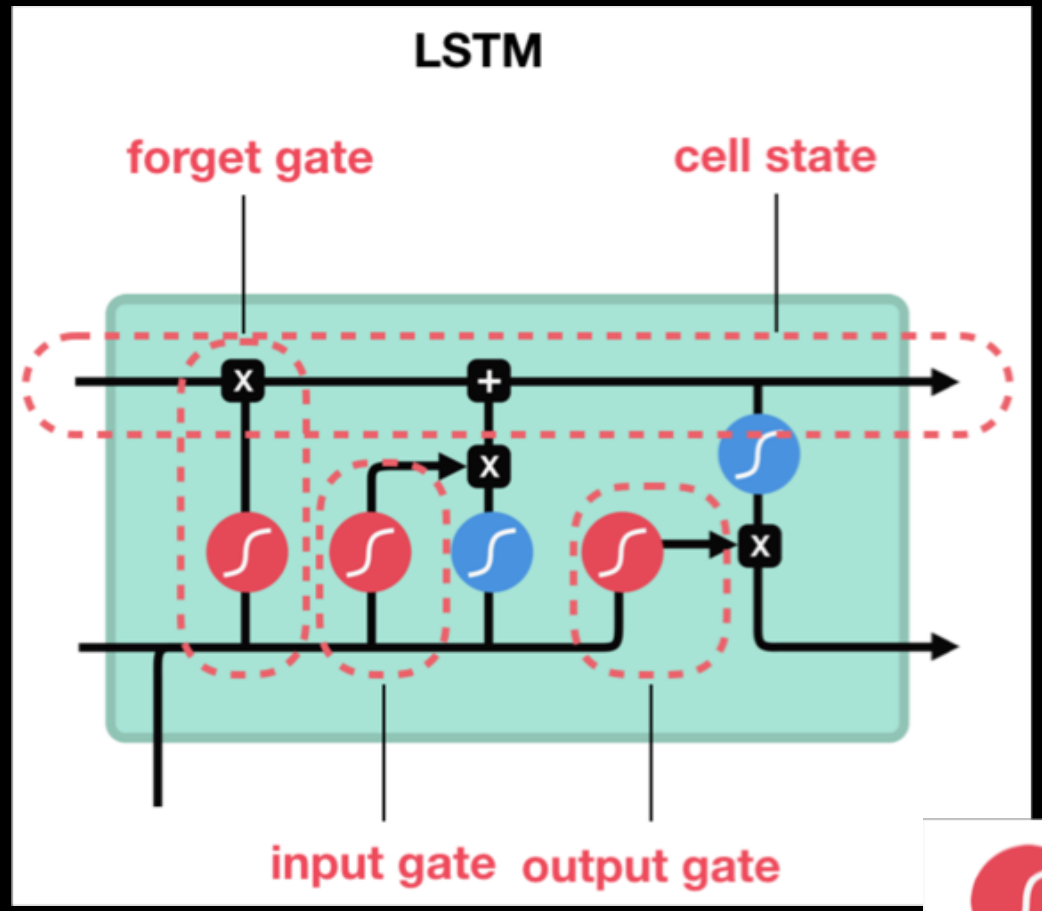
- Works across asks varying in document size, number, and label type
- Uses a single architecture and training process
- Requires no custom feature engineering nor preprocessing
- Does not require additional in-domain documents or labels

# ULMFiT

Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. arXiv:180106146 [cs, stat]. 2018 Jan 18; Available from: http://arxiv.org/abs/1801.06146

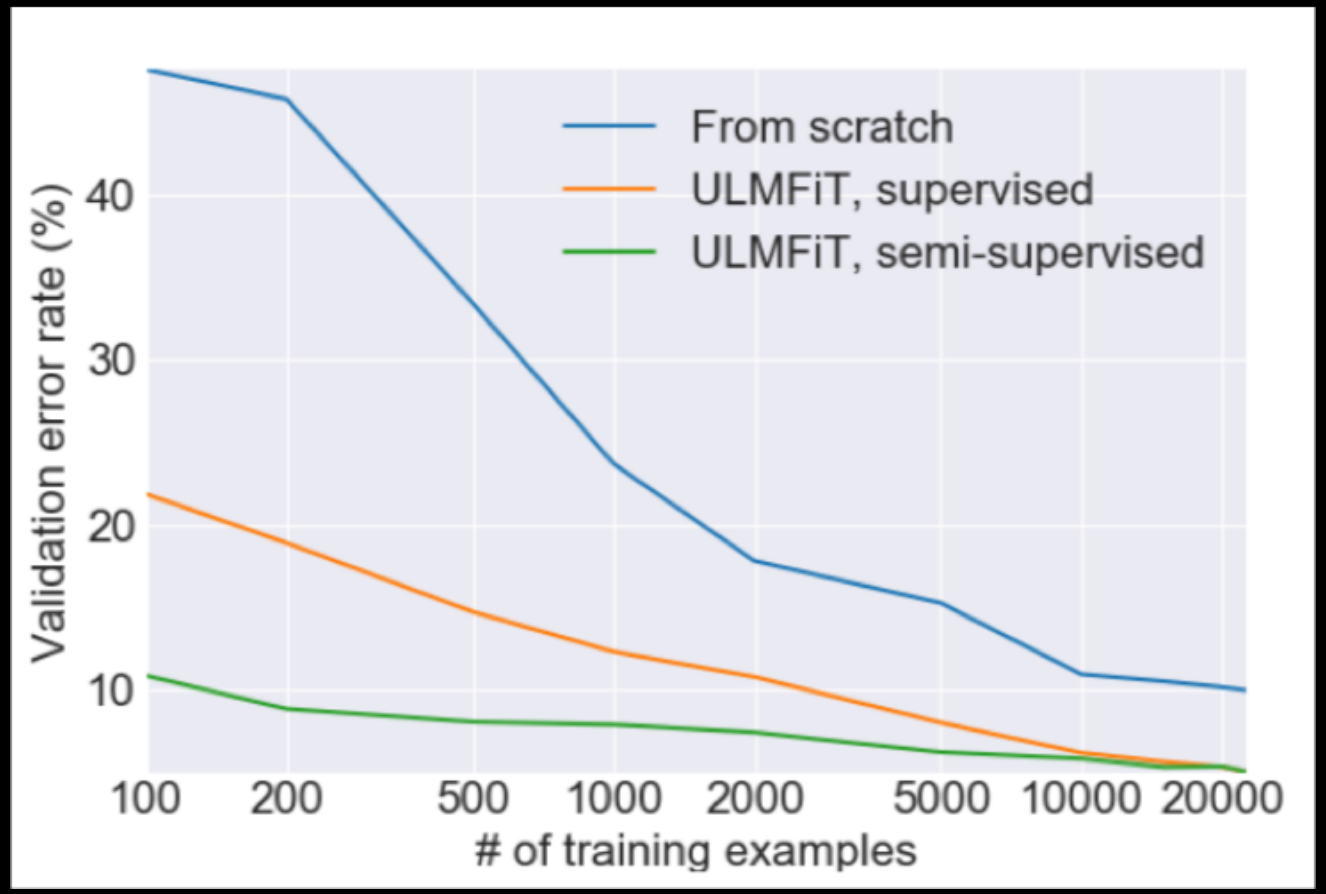


# Long Short Term Memory




https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

# ULMFiT


From scratch
ULMFiT, supervised
ULMFiT, semi-supervised
Validation error rate (%)
40
30
20
10
100
200
500
1000
2000
5000
10000
20000
# of training examples

# So what?

“We demonstrate language models can perform down-stream tasks in a zero-shot setting – without any parameter or architecture modification.“

# Zero Shot Learning

Ian Goodfellow, AI Research Scientist

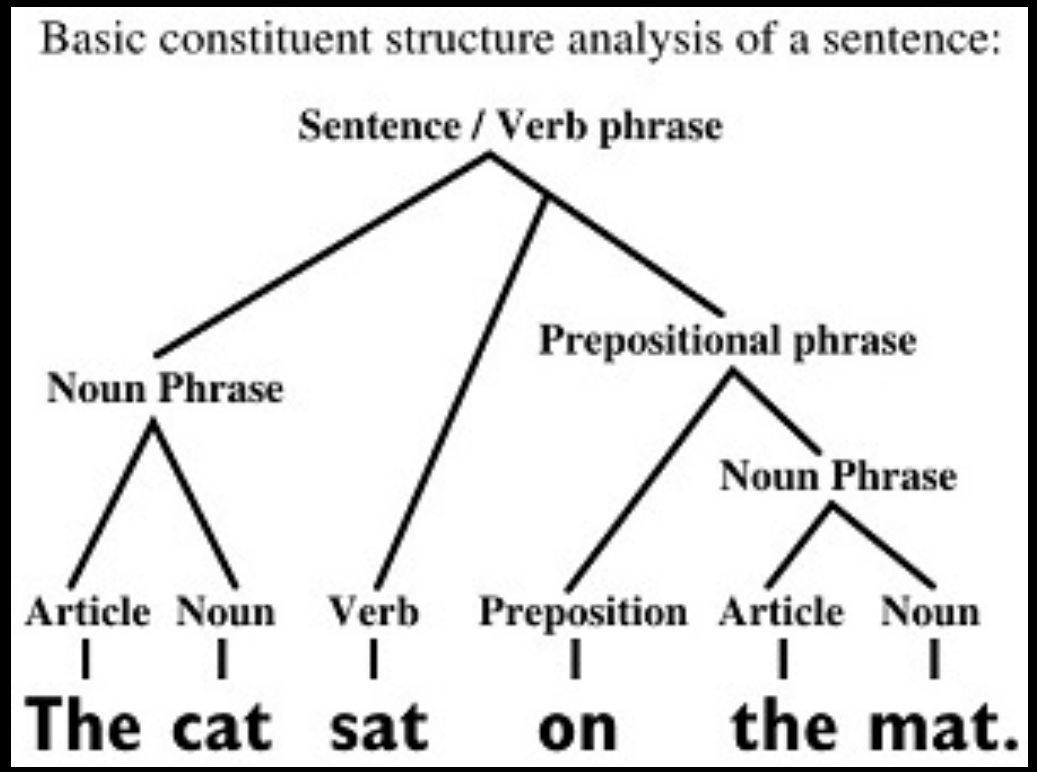
Answered Apr 10, 2016 · Author has **232** answers and **2.5m** answer views

Zero-shot learning is being able to solve a task despite not having received any training examples of that task. For a concrete example, imagine recognizing a category of object in photos without ever having seen a photo of that kind of object before. If you've read a very detailed description of a cat, you might be able to tell what a cat is in a photograph the first time you see it.

"While suggestive as a research result, in terms of practical applications, the zero-shot performance of GPT-2 is still far from use-able."

# Language Modeling

## Linguistic Approach

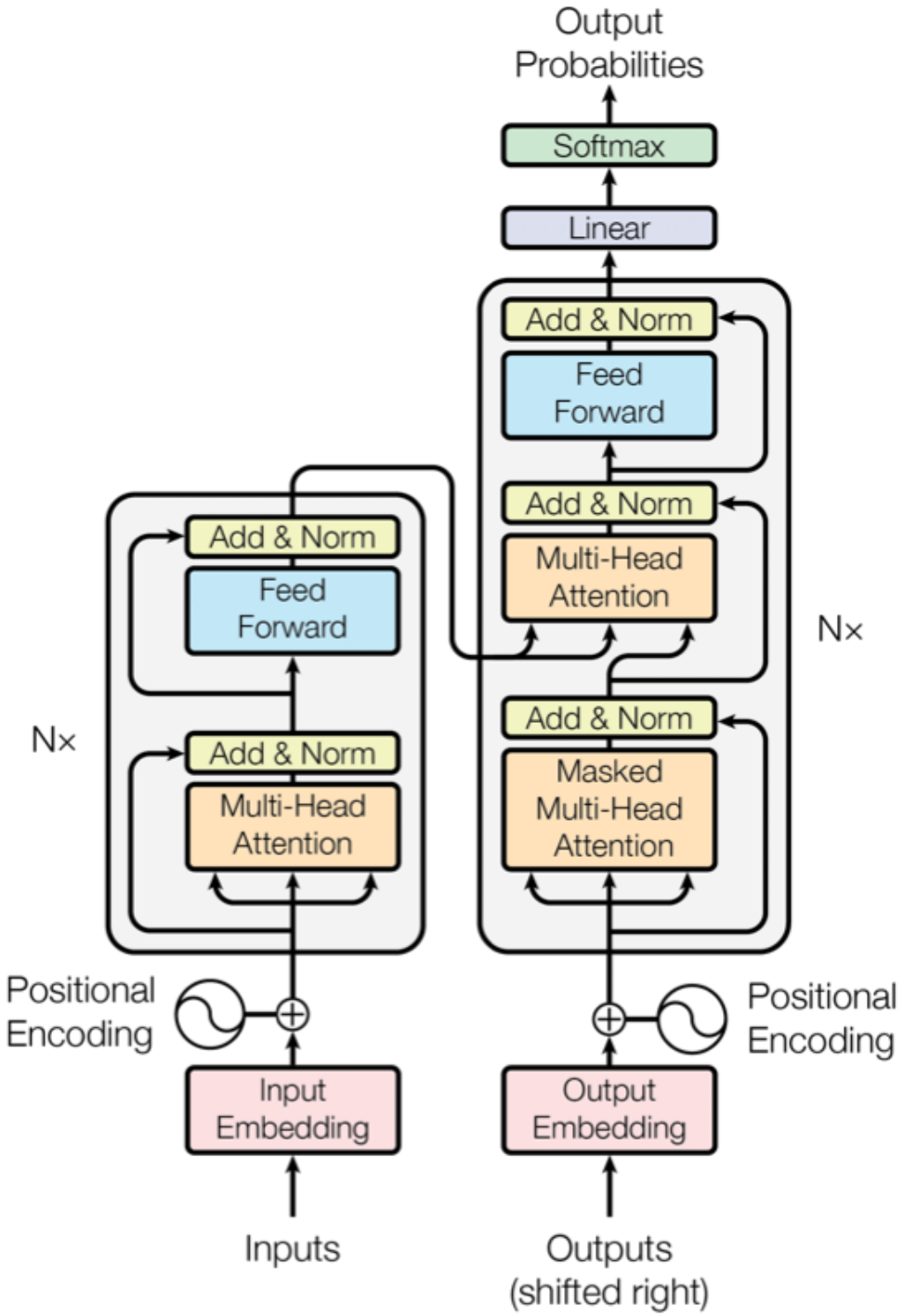



## Statistics Approach

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1})$$

# “Attention is all you need”

## Transformer Architecture



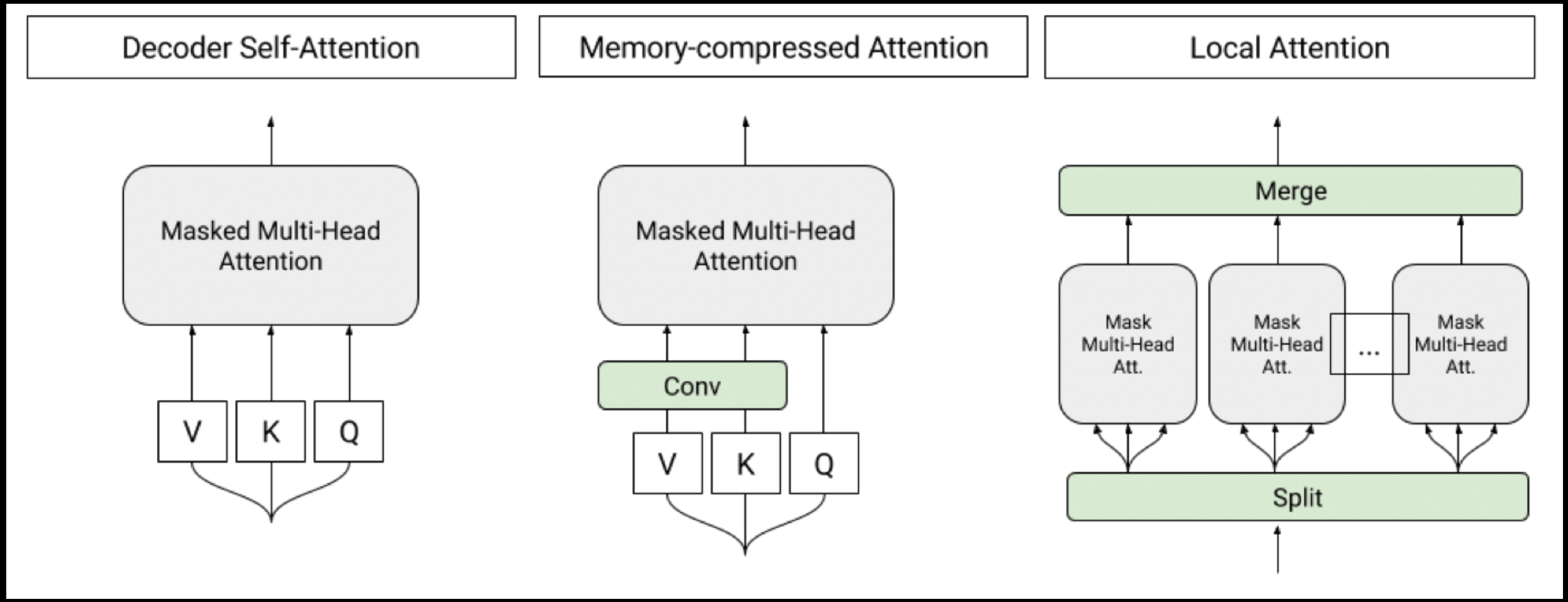

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv:170603762 [cs] [Internet]. 2017 Jun 12 [cited 2019 Mar 13]; Available from: http://arxiv.org/abs/1706.03762

# Transformer Evolution

Liu PJ, Saleh M, Pot E, Goodrich B, Sepassi R, Kaiser L, et al. Generating Wikipedia by Summarizing Long Sequences. arXiv:180110198 [cs]. 2018 Jan 30; Available from: http://arxiv.org/abs/1801.10198

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

# Transformer Evolution



# Standard approach to Input Representation

"One, two, three, four. Is it snowing where you are, Mr. Thiessen? If it is, telegraph back and let me know."

Lower case: one, two, three, four. is it snowing where you are, mr. thiessen? if it is, telegraph back and let me know.

Tokenization: {one} {,} {two} {,} {three}…

Replace/Remove rare tokens: {one} {,} {two} {,} {three}… {<mr.} <UNK>…

Remove too common tokens: ['One', ',', 'two', ',', 'three', ',', 'four', '.', 'Is', 'snowing', ',', 'Mr.', 'Thiessen…

# GPT-2 approach to Input Representation

Byte Pair Encoding

- "One, two, three, four. Is it snowing where you are, Mr. Thiessen? If it is, telegraph back and let me know."
- [O] [n] [e] [,] [ ] [t] [w] [o] [,] [t] [h] [r] [e]…
- Combine [r] [e] into [re]… eventually may end up with [One] [two]…

GPT-Modification: Do not allow merging across character categories ([a] [1] != [a1] && [a] [.] != [a.])

"No lossy pre-processing nor tokenization"

Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. arXiv:150807909 [cs]. 2015 Aug 31; Available from: http://arxiv.org/abs/1508.07909

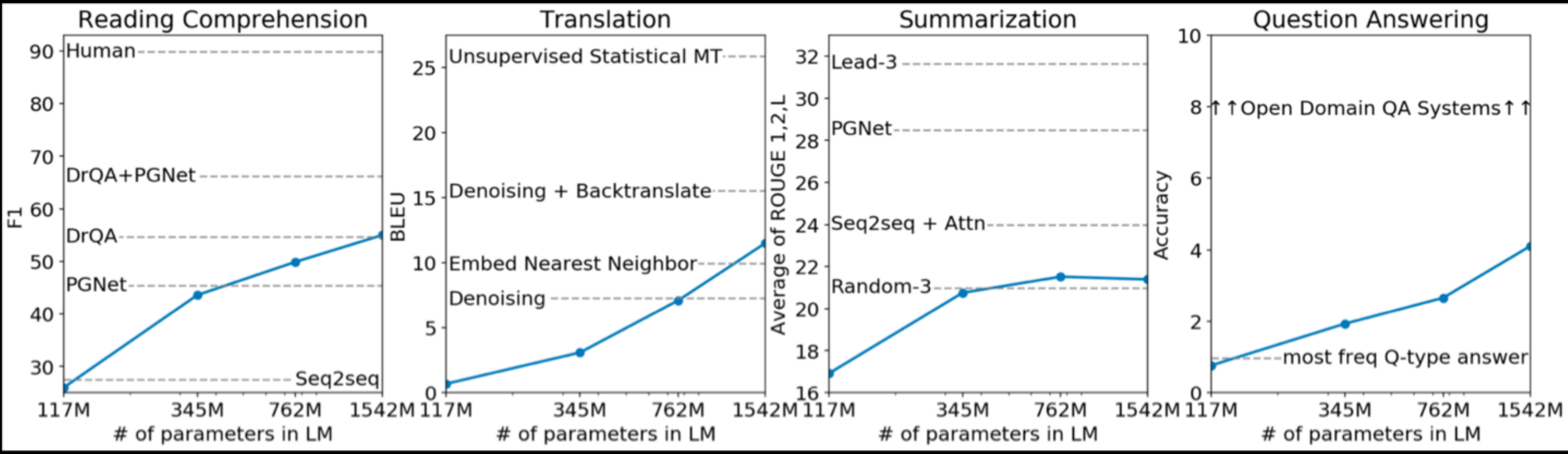
# Experiments

- Children's Book Test: designed to measure directly how well language models can exploit wider linguistic context. Fill in the blank
- LAMBADA: Ability to model long-range dependencies in text. Predict final word in sentence.
- Winograd Schema Challenge: Resolving ambiguity in language. E.g. The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.
- Reading Comprehension: Stanford's CoQA that measures the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

# Experiments

- Summarization: Given a text, provide a short summary (3 sentences) of the text
- Translation: Automatic translation from language x to language y
- Question Answering: Generate answers to simple trivia style questions

# Summary of Experiments



# Free Lunch!?

IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. 1, NO. 1, APRIL 1997 

## No Free Lunch Theorems for Optimization

David H. Wolpert and William G. Macready

***Abstract*—A framework is developed to explore the connection between effective optimization algorithms and the problems they are solving. A number of "no free lunch" (NFL) theorems are presented which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class. These theorems result in a geometric interpretation of what it means for an algorithm to be well suited to an optimization problem. Applications of the NFL theorems to information-theoretic aspects of optimization and benchmark measures of performance are also presented. Other issues addressed include time-varying optimization problems and**

information theory and Bayesian analysis contribute to an understanding of these issues? How *a priori* generalizable are the performance results of a certain algorithm on a certain class of problems to its performance on other classes of problems? How should we even measure such generalization? How should we assess the performance of algorithms on problems so that we may programmatically compare those algorithms?

Broadly speaking, we take two approaches to these ques-

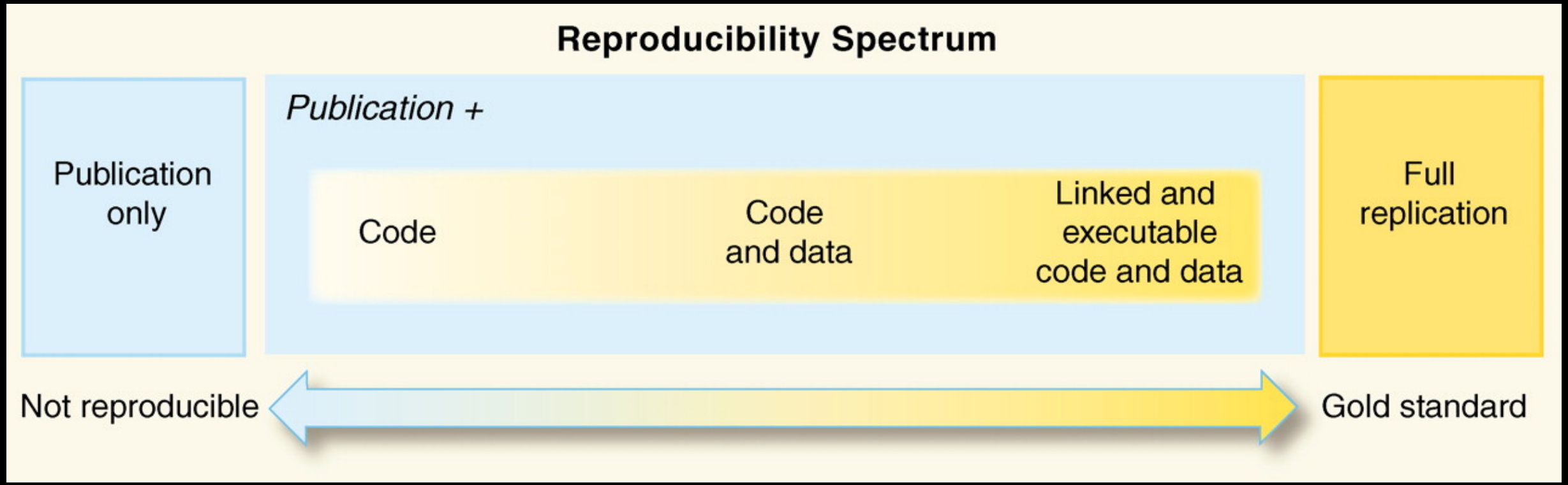
# Free Lunch?

| DATASET | METRIC | OUR RESULT | PREVIOUS RECORD | HUMAN |
|---|---|---|---|---|
| Winograd Schema Challenge | accuracy (+) | **70.70%** | 63.7% | 92%+ |
| LAMBADA | accuracy (+) | **63.24%** | 59.23% | 95%+ |
| LAMBADA | perplexity (-) | **8.6** | 99 | ~1-2 |
| Children's Book Test Common Nouns (validation accuracy) | accuracy (+) | **93.30%** | 85.7% | 96% |
| Children's Book Test Named Entities (validation accuracy) | accuracy (+) | **89.05%** | 82.3% | 92% |
| Penn Tree Bank | perplexity (-) | **35.76** | 46.54 | unknown |
| WikiText-2 | perplexity (-) | **18.34** | 39.14 | unknown |
| enwik8 | bits per | **0.93** | 0.99 | unknown |

# Free Lunch?

| | LAMBADA (PPL) | LAMBADA (ACC) | CBT-CN (ACC) | CBT-NE (ACC) | WikiText2 (PPL) | PTB (PPL) | enwik8 (BPB) | text8 (BPC) | WikiText103 (PPL) | 1BW (PPL) |
|---|---|---|---|---|---|---|---|---|---|---|
| SOTA | 99.8 | 59.23 | 85.7 | 82.3 | 39.14 | 46.54 | 0.99 | 1.08 | 18.3 | **21.8** |
| 117M | **35.13** | 45.99 | **87.65** | **83.4** | **29.41** | 65.85 | 1.16 | 1.17 | 37.50 | 75.20 |
| 345M | **15.60** | 55.48 | **92.35** | **87.1** | **22.76** | 47.33 | 1.01 | **1.06** | 26.37 | 55.72 |
| 762M | **10.87** | **60.12** | **93.45** | **88.0** | **19.93** | **40.31** | **0.97** | **1.02** | 22.05 | 44.575 |
| 1542M | **8.63** | **63.24** | **93.30** | **89.05** | **18.34** | **35.76** | **0.93** | **0.98** | **17.48** | 42.16 |

# Qu'ils mangent de la brioche




Peng RD. Reproducible research in computational science. Science (New York, Ny). 2011 Dec;334(6060):1226– 1227. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002/

# Qu’ils mangent de la brioche

https://www.kingarthurflour.com/recipes/brioche-recipe

## INGREDIENTS

Recipe Success Guide

Choose your measure: (•) Volume ( ) Ounces ( ) Grams

### DOUGH

- 2 3/4 cups King Arthur Unbleached All-Purpose Flour
- 1/4 cup Baker's Special Dry Milk or 1/2 cup nonfat dry milk
- 3 tablespoons sugar
- 1 1/4 teaspoons salt
- 1 tablespoon instant yeast
- 3 large eggs*
- 1/4 cup lukewarm water
- 5/8 cup (10 tablespoons) unsalted butter

*Use 3 large eggs + 1 egg yolk, if desired — this will allow you to brush the leftover egg white on the loaf if you're planning to garnish it with sugar; see tip below.

### AT A GLANCE

PREP
25 mins. to 35 mins.

BAKE
25 mins. to 45 mins.

TOTAL
6 hrs 50 mins. to 17 hrs 20 mins.

YIELD
1 large or 2 medium loaves; or 12 mini brioche

Nutrition information

## INSTRUCTIONS

1. In a stand mixer or bread machine (programmed for dough), mix together all of the ingredients to form a smooth, shiny dough. Don't worry; what starts out as a sticky mess becomes beautifully satiny as it kneads. This dough takes longer than most to develop, so be prepared to let the dough knead for up to 15 to 20

# Qu'ils mangent de la brioche

- Compute resources required & used
  - OpenAI Five plays 180 years worth of games against itself every day, learning via self-play. It trains using a scaled-up version of Proximal Policy Optimization running on 256 GPUs and 128,000 CPU cores
  - Jeremy Howard estimates $20 - $50k in compute costs to train in a month (that's if you have everything figured out first)
- Data set: "slightly over 8 million documents for a total of 40 GB of text. "

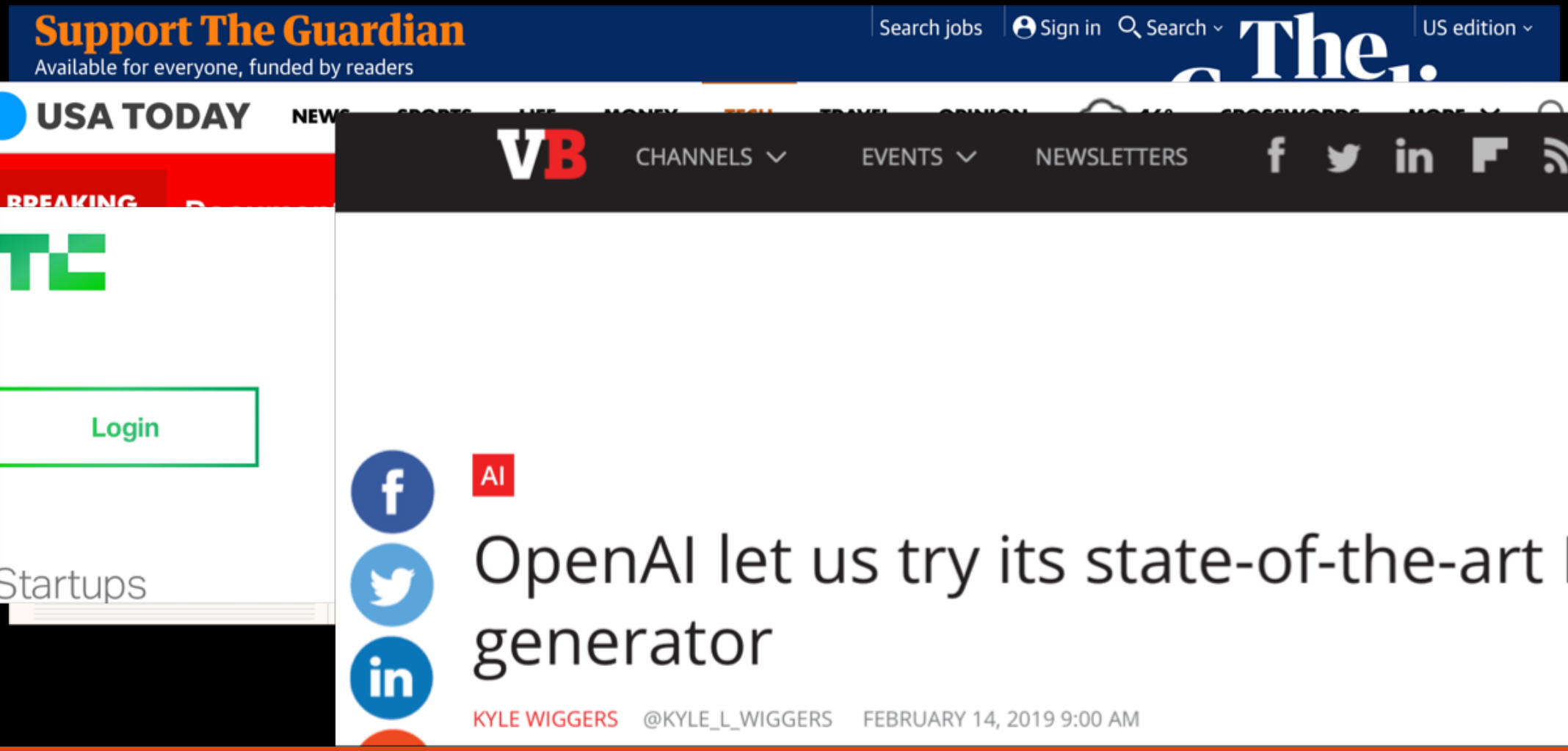
# Qu'ils mangent de la brioche

- Specifics on training to build model
- Samples reviewed to get examples shown
- Ambiguous language without full source code: "No lossy pre-processing nor tokenization" what does this mean?

# OpenAI GPT-2 In the Press

- https://www.theverge.com/2019/2/14/18224704/ai-machine-learning-language-models-read-write-openai-gpt2
- https://slate.com/technology/2019/02/openai-gpt2-text-generating-algorithm-ai-dangerous.html
- https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/
- https://venturebeat.com/2019/02/14/openai-let-us-generate-text-with-an-ai-model-that-achieves-state-of-the-art-performance-in-several-nlp-tasks/
- https://www.theguardian.com/commentisfree/2019/feb/15/ai-write-robot-openai-gpt2-elon-musk
- https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction
- https://arstechnica.com/information-technology/2019/02/researchers-scared-by-their-own-work-hold-back-deepfakes-for-text-ai/
- https://medium.com/syncedreview/openai-guards-its-ml-model-code-data-to-thwart-malicious-usage-d9f7e9c43cd0
- https://www.vox.com/future-perfect/2019/2/14/18222270/artificial-intelligence-open-ai-natural-language-processing
- Plus twitter, reddit, etc.
- … And more

# OpenAI GPT-2 In the Press



# OpenAI GPT-2 In the Press

# Ethics


Smerity
@Smerity
Today's meta-Twitter summary for machine learning:
None of us have any consensus on what we're doing when it comes to responsible disclosure, dual use, or how to interact with the media.
This should be concerning for us all, in and out of the field.
464 9:17 PM - Feb 14, 2019
169 people are talking about this

# Ethics

- Policymakers should collaborate closely with technical researchers to investigate, prevent, and mitigate potential malicious uses of AI.
- Researchers and engineers in artificial intelligence should take the dual-use nature of their work seriously, allowing misuse- related considerations to influence research priorities and norms, and proactively reaching out to relevant actors when harmful applications are foreseeable.
- Best practices should be identified in research areas with more mature methods for addressing dual-use concerns, such as computer security, and imported where applicable to the case of AI.
- Actively seek to expand the range of stakeholders and domain experts involved in discussions of these challenges.

The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation https://maliciousaireport.com

# Ethics

The most serious threats are most likely to come from folks with resources to spend $100k or so on (for example) a disinformation campaign to attempt to change the outcome of a democratic election. In practice, the most likely exploit is (in my opinion) a foreign power spending that money to dramatically escalate existing disinformation campaigns, such as those that have been extensively documented by the US intelligence community.

The only practical defense against such an attack is (as far as I can tell) to use the same tools to both attempt to identify, and push back against, such disinformation. These kinds of defenses are likely to be much more powerful when wielded by the broader community of those impacted. The power of a large group of individuals has repeatedly been shown to be more powerful at creating, than at destruction, as we see in projects such as Wikipedia, or open source software.

Some thoughts on zero-day threats in AI, and OpenAI's GP2 https://www.fast.ai/2019/02/15/openai-gp2/

# Ethics

By releasing the model, this malicious use will happen sooner. But by not releasing the model, there will be fewer defenses available and less real understanding of the issues from those that are impacted. Those both sound like bad outcomes to me.

Some thoughts on zero-day threats in AI, and OpenAI's GP2 https://www.fast.ai/2019/02/15/openai-gp2/

# Demo

GPT-2 117M model

# What does this mean?


VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

# I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

## 1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to

# The Chinese Room


Searle JR. Minds, brains, and programs. Behavioral and Brain Sciences. 1980 Sep;3(3):417–24. Available from: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A

# References


- "Language Models are Unsupervised Multitask Learners" https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
- Announcement blog post by OpenAI: https://blog.openai.com/better-language-models/
- Github repository: https://github.com/openai/gpt-2
- Previous work – Finetuned Transformer LM - (aka GPT-1)
  - "Improving Language Understanding by Generative Pre-Training" https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
  - Announcement blog post: https://openai.com/blog/language-unsupervised/

# References


- https://www.skynettoday.com/briefs/gpt2

# Questions

seth.russell@ucdenver.edu

# Useful URLs

- https://github.com/eukaryote31/openwebtext

# Multitask Learning (include this?)

Learning is very slow

Attempting to learn multiple tasks at once

# Title