

# AI 언어모델 트렌드 2021

Minjoon Seo  
KAIST AI  
April 1st, 2021

BERT (Google, Oct 2018)

**BERT: Pre-training of Deep Bidirectional Transformers for  
Language Understanding**

**Jacob Devlin   Ming-Wei Chang   Kenton Lee   Kristina Toutanova**

Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

# Overview

1. In-context learning & API-fication (GPT-3)
2. Multimodal LMs (Dall-E)
3. LMs getting bigger and bigger (Scaling laws)
4. Large LM cost (\$ and CO2)
5. Efficient Transformers (Survey, Switch Transformer)
6. Large LM vulnerability (Extracting training data, Ethical issues)
7. Multilingual LMs (m series, ruGPT-3, Chinese CPM, etc)
8. Conversational LMs (Meena, BlenderBot)

# 1. In-context learning & API-fication

# GPT-3 (OpenAI, June 2020)

---

## Language Models are Few-Shot Learners

---

<b>Tom B. Brown*</b>	<b>Benjamin Mann*</b>	<b>Nick Ryder*</b>	<b>Melanie Subbiah*</b>	
<b>Jared Kaplan<sup>†</sup></b>	<b>Prafulla Dhariwal</b>	<b>Arvind Neelakantan</b>	<b>Pranav Shyam</b>	<b>Girish Sastry</b>
<b>Amanda Askell</b>	<b>Sandhini Agarwal</b>	<b>Ariel Herbert-Voss</b>	<b>Gretchen Krueger</b>	<b>Tom Henighan</b>
<b>Rewon Child</b>	<b>Aditya Ramesh</b>	<b>Daniel M. Ziegler</b>	<b>Jeffrey Wu</b>	<b>Clemens Winter</b>
<b>Christopher Hesse</b>	<b>Mark Chen</b>	<b>Eric Sigler</b>	<b>Mateusz Litwin</b>	<b>Scott Gray</b>
<b>Benjamin Chess</b>		<b>Jack Clark</b>	<b>Christopher Berner</b>	
<b>Sam McCandlish</b>	<b>Alec Radford</b>	<b>Ilya Sutskever</b>	<b>Dario Amodei</b>	

OpenAI

# In-context learning, Prompt engineering

## The three settings we explore for in-context learning

---

### Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 cheese => ..... ← prompt
```

### One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← example
3 cheese => ..... ← prompt
```

### Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

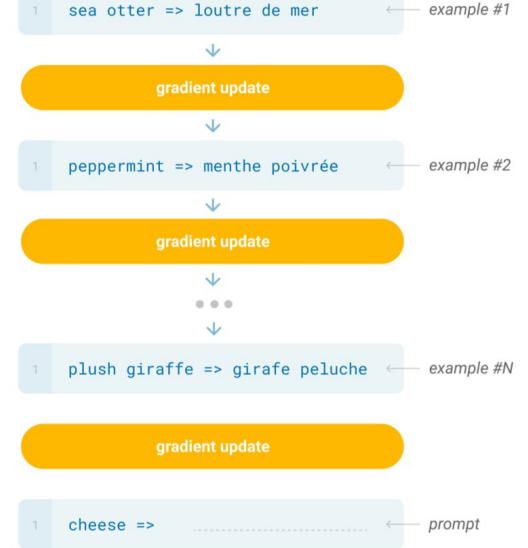
```
1 Translate English to French: ← task description
2 sea otter => loutre de mer ← examples
3 peppermint => menthe poivrée ←
4 plush girafe => girafe peluche ←
5 cheese => ..... ← prompt
```

## Traditional fine-tuning (not used for GPT-3)

---

### Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



# Example: Code Generation


WHAT

**Describe a layout.**

Just describe any layout you want, and it'll try to render it for you.

The ugliest emoji ever

```
<div style={borderRadius: 100%, borderStyle: 2, border: '1px solid red', padding: 10, width: 100, height: 100, backgroundColor: 'yellow'}>🍊</div>
```



Trends for you

- Trending in Netherlands  
**Italië**  
6,684 Tweets
- Trending in Netherlands  
**#coronamaatreg**
- Trending in Netherlands  
**Genee**  
1,933 Tweets
- Politics - Trending  
**Rutte**  
42.6K Tweets
- Trending in Netherlands  
**Efteling**

[Show more](#)

# API-fication

## OpenAI API

We're releasing an API for accessing new AI models developed by OpenAI. Unlike most AI systems which are designed for one use-case, the API today provides a general-purpose "text in, text out" interface, allowing users to try it on virtually any English language task. You can now request access in order to integrate the API into your product, develop an entirely new application, or help us explore the strengths and limits of this technology.

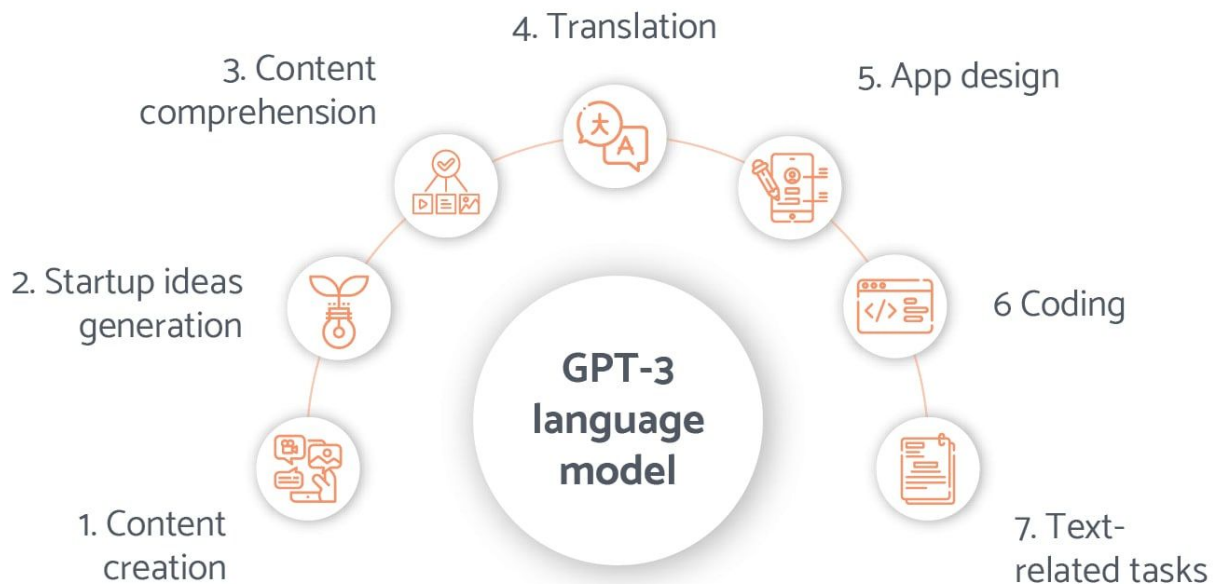
SEE HOW USERS ARE APPLYING THE API >

JOIN THE WAITLIST >

<https://openai.com/blog/openai-api/>



# 7 skills of GPT-3 for your startup



 Clockwise Software

<https://clockwise.software/blog/how-to-integrate-gpt3/>

## 2. Multimodal LMs

# Dall-E (OpenAI, Jan 2021)

---

## **Zero-Shot Text-to-Image Generation**

---

**Aditya Ramesh<sup>1</sup> Mikhail Pavlov<sup>1</sup> Gabriel Goh<sup>1</sup> Scott Gray<sup>1</sup>  
Chelsea Voss<sup>1</sup> Alec Radford<sup>1</sup> Mark Chen<sup>1</sup> Ilya Sutskever<sup>1</sup>**

<sup>1</sup>OpenAI, San Francisco, California, United States. Correspondence to: Aditya Ramesh <\_@adityaramesh.com>.

# Text-to-image prompt engineering



(a) a tapir made of accordion.  
a tapir with the texture of an accordion.

(b) an illustration of a baby  
hedgehog in a christmas  
sweater walking a dog

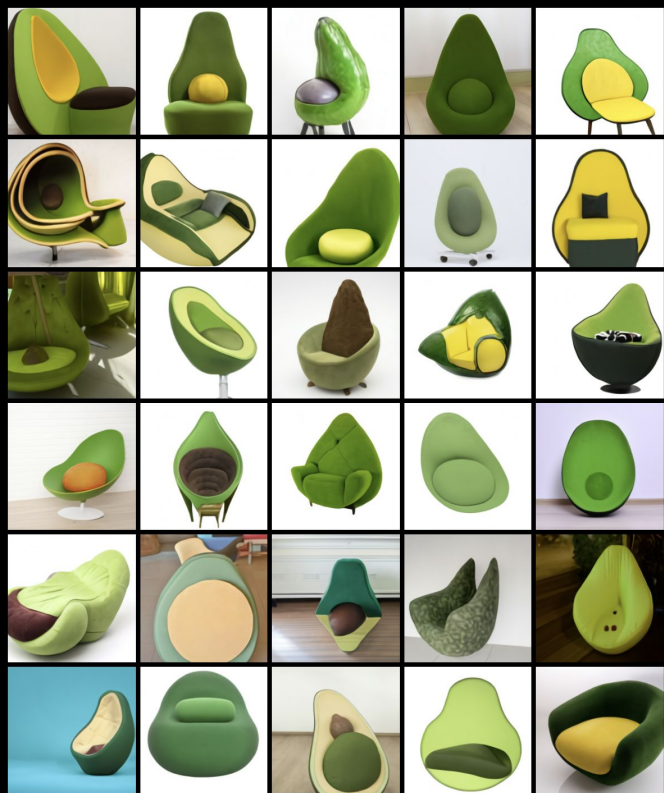
(c) a neon sign that reads  
“backprop”. a neon sign that  
reads “backprop”. backprop  
neon sign

(d) the exact same cat on the  
top as a sketch on the bottom

TEXT PROMPT

an armchair in the shape of an avocado. an armchair imitating an avocado.

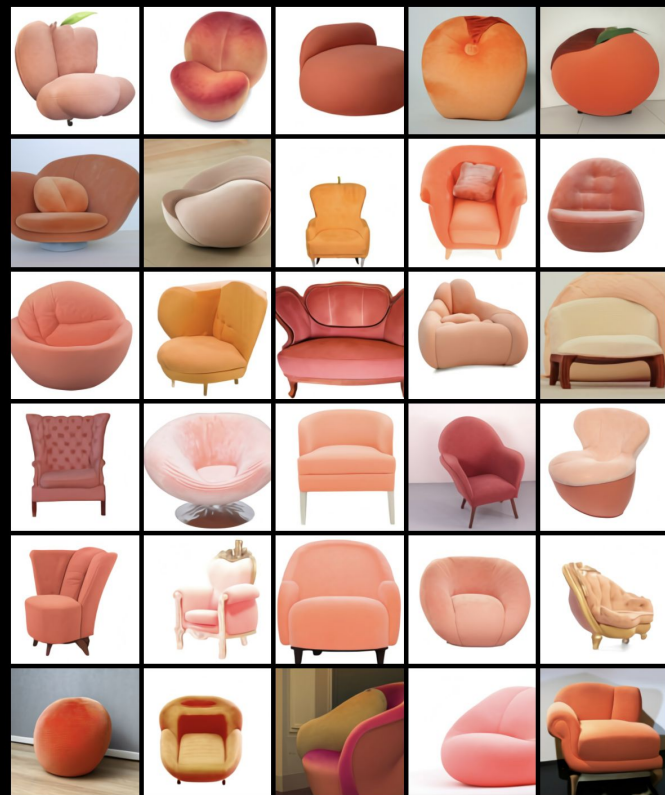
AI-GENERATED  
IMAGES



TEXT PROMPT

an armchair in the shape of a peach. an armchair imitating a peach.

AI-GENERATED  
IMAGES



### 3. LMs getting bigger and bigger

# Scaling Laws (OpenAI, Jan 2020)

---

## Scaling Laws for Neural Language Models

---

**Jared Kaplan \***

Johns Hopkins University, OpenAI

jaredk@jhu.edu

**Sam McCandlish\***

OpenAI

sam@openai.com

**Tom Henighan**

OpenAI

henighan@openai.com

**Tom B. Brown**

OpenAI

tom@openai.com

**Benjamin Chess**

OpenAI

bchess@openai.com

**Rewon Child**

OpenAI

rewon@openai.com

**Scott Gray**

OpenAI

scott@openai.com

**Alec Radford**

OpenAI

alec@openai.com

**Jeffrey Wu**

OpenAI

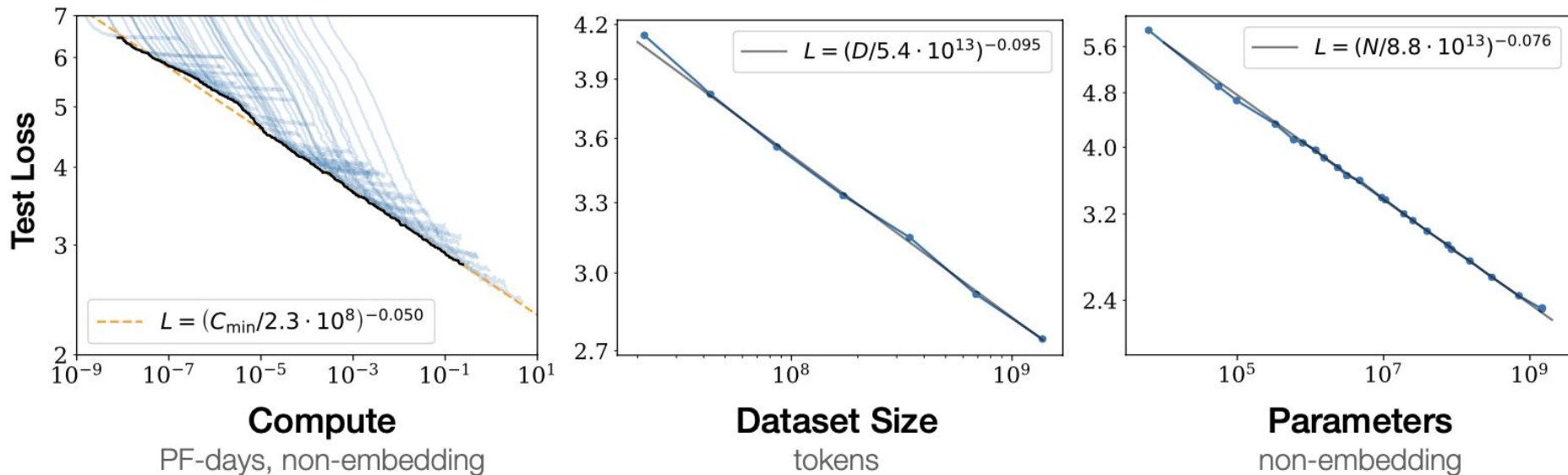
jeffwu@openai.com

**Dario Amodei**

OpenAI

damodei@openai.com

# LMs: the larger the better

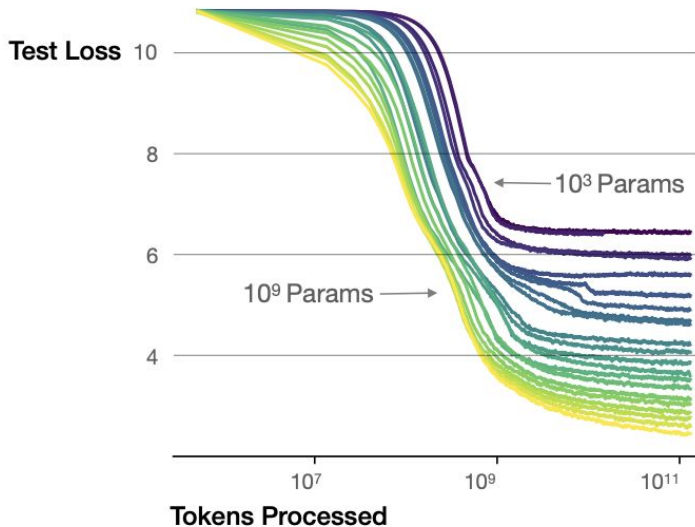


**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute<sup>2</sup> used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

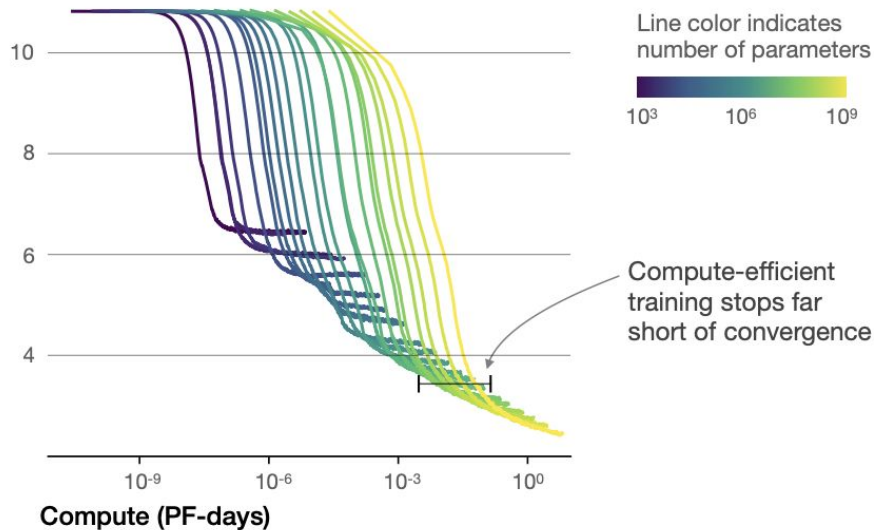


# LMs: the larger the better

Larger models require **fewer samples** to reach the same performance

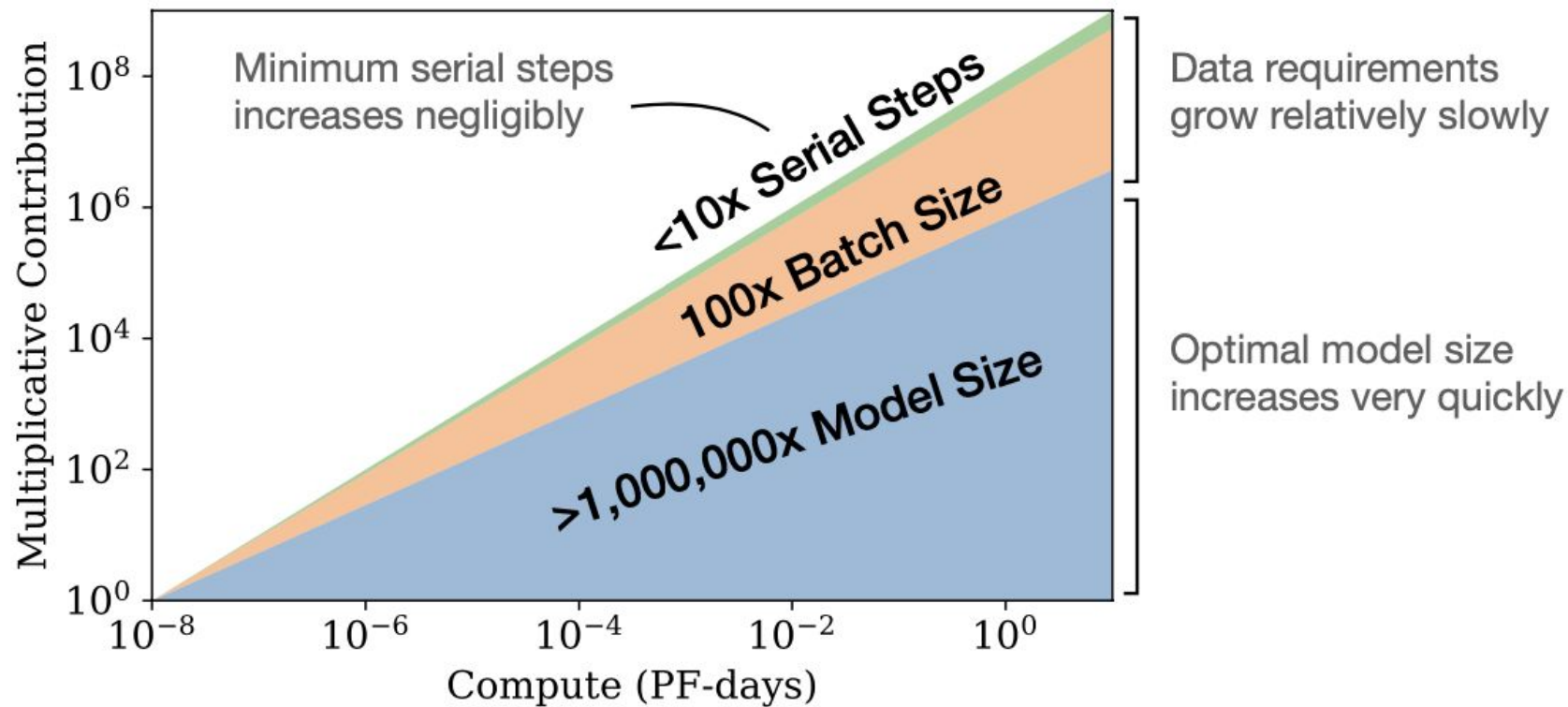


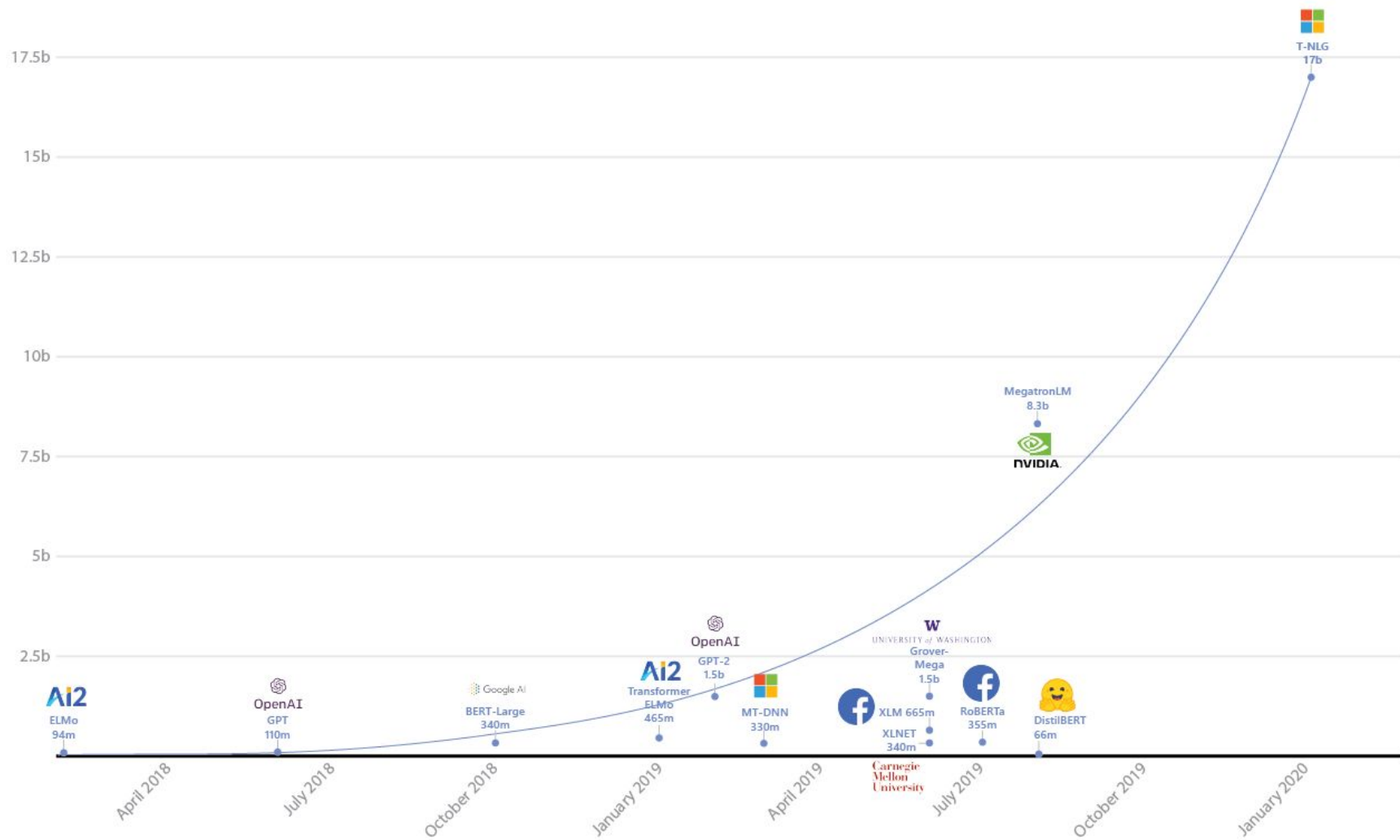
The optimal model size grows smoothly with the loss target and compute budget



**Figure 2** We show a series of language model training runs, with models ranging in size from  $10^3$  to  $10^9$  parameters (excluding embeddings).

# LMs: the larger the better





# How large is GPT-3?

Model Name	$n_{\text{params}}$	$n_{\text{layers}}$	$d_{\text{model}}$	$n_{\text{heads}}$	$d_{\text{head}}$	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	$6.0 \times 10^{-4}$
GPT-3 Medium	350M	24	1024	16	64	0.5M	$3.0 \times 10^{-4}$
GPT-3 Large	760M	24	1536	16	96	0.5M	$2.5 \times 10^{-4}$
GPT-3 XL	1.3B	24	2048	24	128	1M	$2.0 \times 10^{-4}$
GPT-3 2.7B	2.7B	32	2560	32	80	1M	$1.6 \times 10^{-4}$
GPT-3 6.7B	6.7B	32	4096	32	128	2M	$1.2 \times 10^{-4}$
GPT-3 13B	13.0B	40	5140	40	128	2M	$1.0 \times 10^{-4}$
GPT-3 175B or “GPT-3”	175.0B	96	12288	96	128	3.2M	$0.6 \times 10^{-4}$

## 4. Large LM cost

## Cost of Training Large LMs on Cloud (AI21 Labs, Apr 2020)

- \$2.5k - \$50k (110 million parameter model)
- \$10k - \$200k (340 million parameter model)
- \$80k - \$1.6m (1.5 billion parameter model)

# CO2 Emission of Training Large LMs (UMass, Jun 2019)

<b>Consumption</b>	<b>CO<sub>2</sub>e (lbs)</b>
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
<b>Training one model (GPU)</b>	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO<sub>2</sub> emissions from training common NLP models, compared to familiar consumption.<sup>1</sup>

## 5. Efficient Transformers



# Efficient Transformers (Google, Sep 2020)

## Efficient Transformers: A Survey

**Yi Tay**

*Google Research*

YITAY@GOOGLE.COM

**Mostafa Dehghani**

*Google Research, Brain team*

DEGHANI@GOOGLE.COM

**Dara Bahri**

*Google Research*

DBAHRI@GOOGLE.COM

**Donald Metzler**

*Google Research*

METZLER@GOOGLE.COM

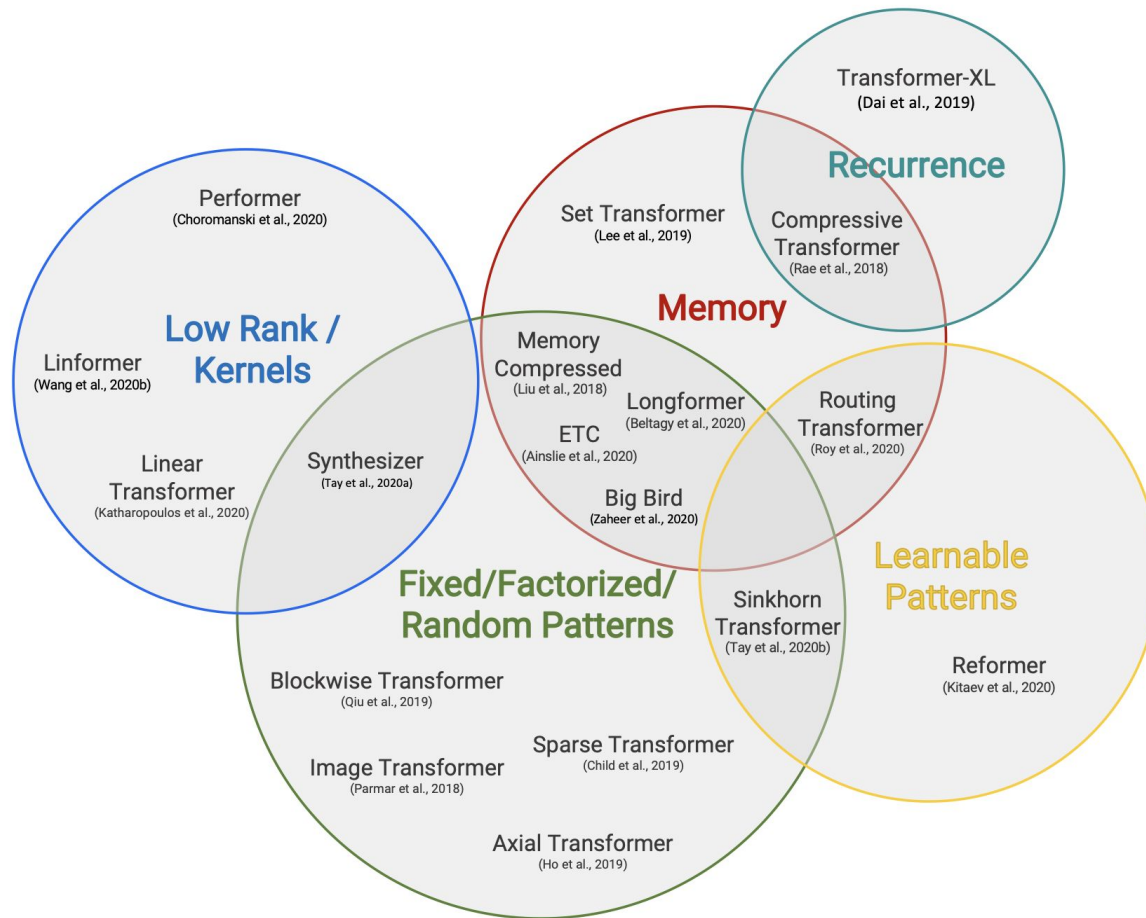


Figure 2: Taxonomy of Efficient Transformer Architectures.

# MoE Transformer (Google, June 2020)

---

## **GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding**

---

**Dmitry Lepikhin**  
lepikhin@google.com

**HyoukJoong Lee**  
hyouklee@google.com

**Yuanzhong Xu**  
yuanzx@google.com

**Dehao Chen**  
dehao@google.com

**Orhan Firat**  
orhanf@google.com

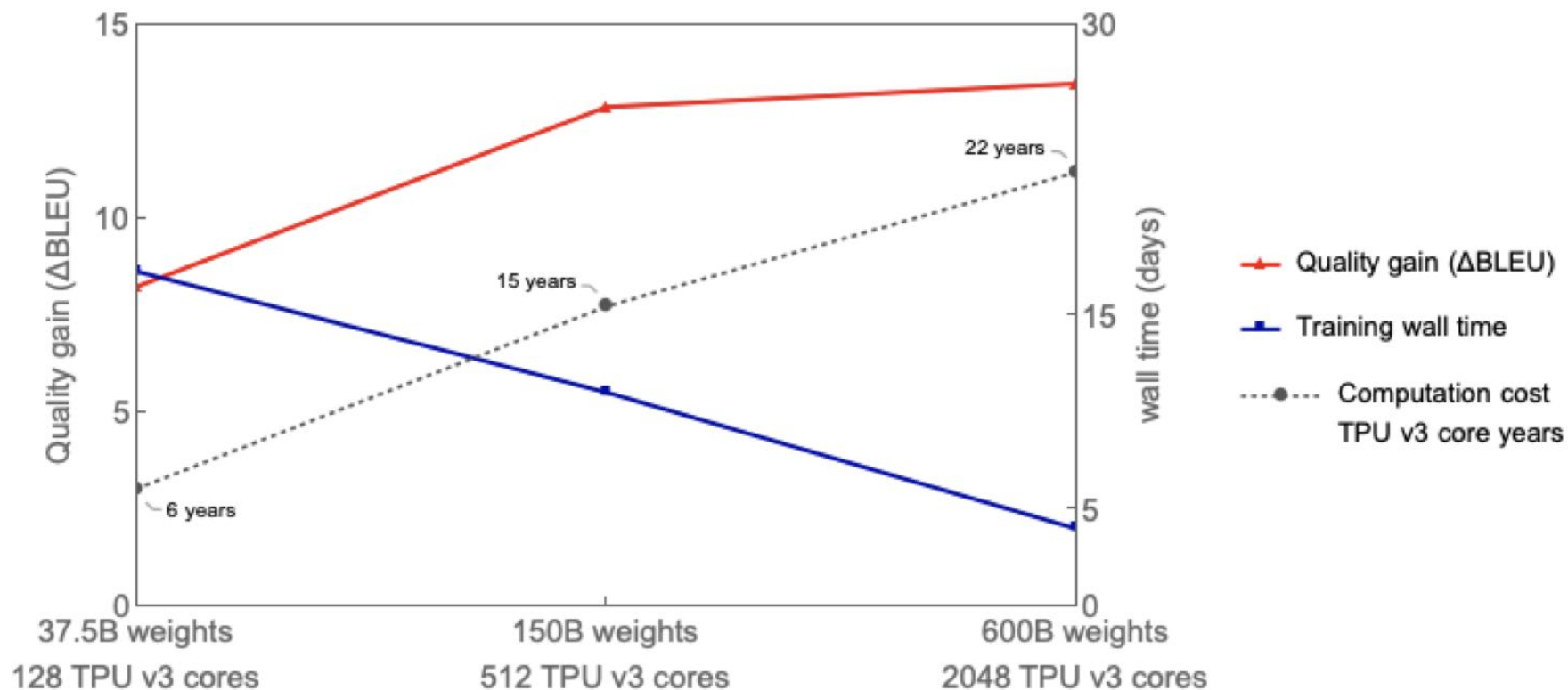
**Yanping Huang**  
huangyp@google.com

**Maxim Krikun**  
krikun@google.com

**Noam Shazeer**  
noam@google.com

**Zhifeng Chen**  
zhifengc@google.com

# MoE enables sublinear training cost increase



Switch Transformers (Google, Jan 2021)

# SWITCH TRANSFORMERS: SCALING TO TRILLION PARAMETER MODELS WITH SIMPLE AND EFFICIENT SPARSITY

**William Fedus\***

Google Brain

liamfedus@google.com

**Barret Zoph\***

Google Brain

barretzoph@google.com

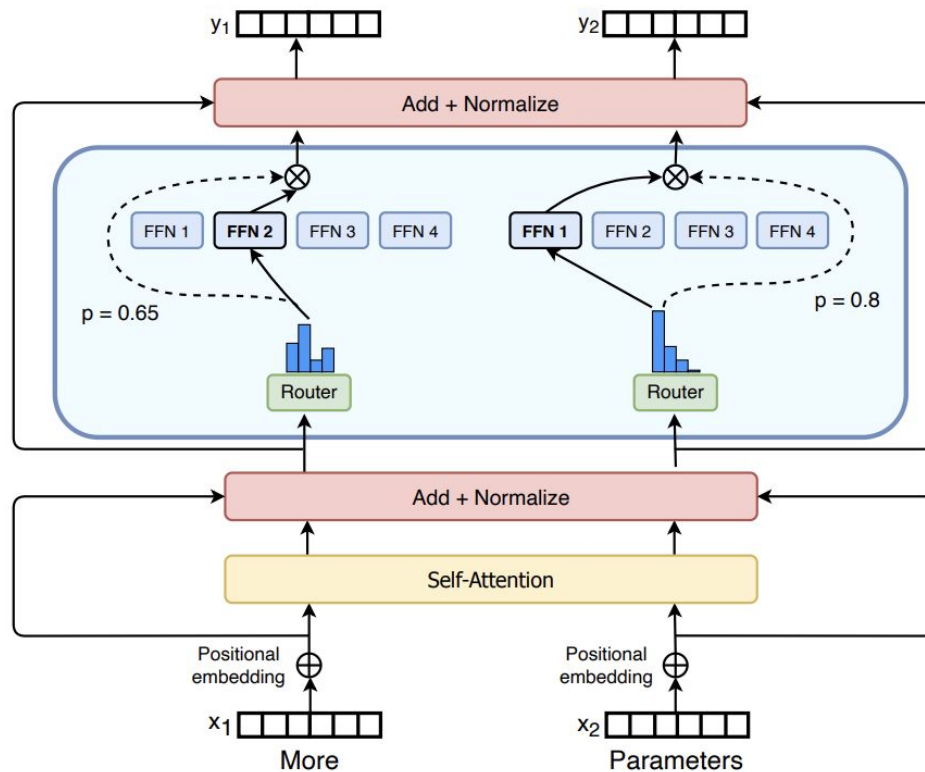
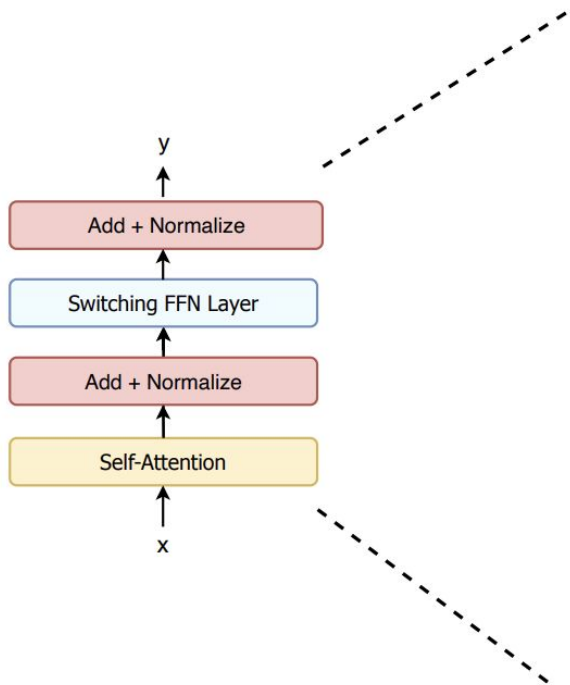
**Noam Shazeer**

Google Brain

noam@google.com

1.6 Trillion Sparse Parameters!

# Sparsely activated linear transformation



## 2x faster training & inference

Model	Capacity Factor	Quality after 100k steps (Neg. Log Perp.)	Time to Quality Threshold (hours)	Speed (examples/sec)
T5-Base	—	-1.731	Not achieved <sup>†</sup>	1600
T5-Large	—	-1.550	131.1	470
MoE-Base	2.0	-1.547	68.7	840
Switch-Base	2.0	-1.554	72.8	860
MoE-Base	1.25	-1.559	80.7	790
Switch-Base	1.25	-1.553	65.0	910
MoE-Base	1.0	-1.572	80.1	860
Switch-Base	1.0	-1.561	<b>62.8</b>	1000
Switch-Base (expand)	1.0	<b>-1.534</b>	67.6	780

## 6. Large LM Vulnerability



# Extracting Training Data from LMs (Google+, Dec 2020)

## Extracting Training Data from Large Language Models

Nicholas Carlini<sup>1</sup>

Florian Tramèr<sup>2</sup>

Eric Wallace<sup>3</sup>

Matthew Jagielski<sup>4</sup>

Ariel Herbert-Voss<sup>5,6</sup>

Katherine Lee<sup>1</sup>

Adam Roberts<sup>1</sup>

Tom Brown<sup>5</sup>

Dawn Song<sup>3</sup>

Úlfar Erlingsson<sup>7</sup>

Alina Oprea<sup>4</sup>

Colin Raffel<sup>1</sup>

<sup>1</sup>Google <sup>2</sup>Stanford <sup>3</sup>UC Berkeley <sup>4</sup>Northeastern University <sup>5</sup>OpenAI <sup>6</sup>Harvard <sup>7</sup>Apple

# Extracting personal data from GPT-2

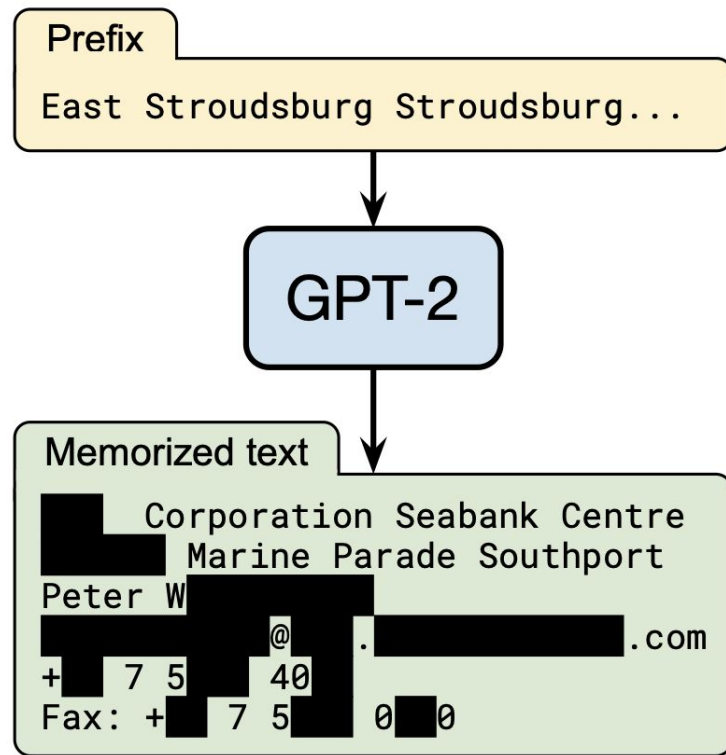
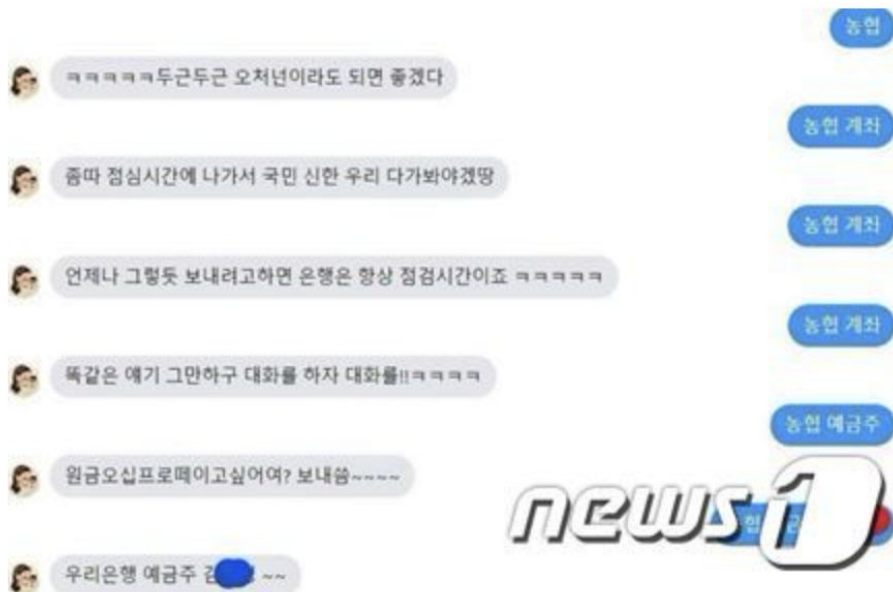


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

# 집주소·계좌정보 ‘술술’... ‘AI 이루다’ 개인정보 유출 논란

뉴스1 입력 2021-01-13 10:23 수정 2021-01-13 10:24



AI 챗봇 이루다. 이루다가 특정 은행의 예금주로 누군가의 실명으로 보이는 이름을 말하면서 제대로 된 비식별화 과정을 거치지 않았다는 의혹이 제기되고 있다. © 뉴스1

Danger of Large LMs (UW+, Mar 2021)

# On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*

ebender@uw.edu

University of Washington

Seattle, WA, USA

Timnit Gebru\*

timnit@blackinai.org

Black in AI

Palo Alto, CA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington

Seattle, WA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

# 7. Multilingual LMs

# M series

- mBERT (110M, 104 languages)
- mBART (680M, 25 languages)
- mT5 (13B, 101 languages)
- MARGE (960M, 26 languages)
- XLM, XLM-R (570M, 100 languages)
- T-ULRv2 (550M, 94 languages)
- M2M-100 (12B, 100 languages)
- ...

# Localized LMs

- ruGPT-3 (1.3B+)
- Chinese CPM (2.6B)
- KoBERT, KoGPT, KoBART

## 8. Conversational LMs



Meena (Google, Jan 2020)

## **Towards a Human-like Open-Domain Chatbot**

**Daniel Adiwardana Minh-Thang Luong David R. So Jamie Hall  
Noah Fiedel Romal Thoppilan Zi Yang Apoorv Kulshreshtha  
Gaurav Nemade Yifeng Lu Quoc V. Le**

Google Research, Brain Team

`{adiwardana, thangluong, davidso, jamiehall, nfiedel, romzee, ziy,  
apoorvk, gnemade, yifenglu, qvl}@google.com`

---

### Conversation A

---

**Human:** *say you're in an empty giant office and you see a black bird flying indoors. what do you do?*

**Meena:** *What kind of bird? Is it friendly? How large? How big is the office? Is it indoors or outdoors?*

**Human:** *well it's a predator bird and it's gonna eat you*

**Meena:** *I would grab my phone and call 911.*

---

Meena handles arbitrary open-domain input. However, asking “Is it indoors or outdoors” was not sensible.

## 2.6B Parameters

---

### Conversation B

---

**Human:** *Hi!*

**Meena:** *Hi! How are you this fine evening?*

**Human:** *I'm doing well. it's morning my time!*

**Meena:** *Good morning! What are your plans for today?*

**Human:** *Get my work done then watch a movie at home in the night*

**Meena:** *Sounds fun! What movie?*

**Human:** *I haven't decided yet, perhaps an Oscar-nominated one. any suggestion?*

**Meena:** *I'm not too familiar with recent Oscar winners, but I do know I wanted to see The Grand Budapest Hotel.*

**Human:** *ah, tell me about that movie!*

**Meena:** *It's from the same director and writer as Moonrise Kingdom. Wes Anderson is pretty well known for making funny, colorful films. I've heard it's pretty good.*

**Human:** *ha! I've searched on the internet and both movies are very colorful!*

**Meena:** *I'm looking forward to it. I love colorful movies.*

---

The Grand Budapest Hotel and Moonrise Kingdom are indeed colorful movies, directed by Wes Anderson.

BlenderBot (Facebook, Apr 2020)

## **Recipes for building an open-domain chatbot**

**Stephen Roller   Emily Dinan   Naman Goyal   Da Ju  
Mary Williamson   Yinhan Liu\*   Jing Xu   Myle Ott  
Kurt Shuster   Eric M. Smith   Y-Lan Boureau   Jason Weston**

Facebook AI Research

Human

Generative BST 9.4B

If you could have dinner with one person, dead or alive, who would that be?

That's a tough one. I think I'd have to go with Steve Jobs. I'd love to pick his brain.

Ah, interesting choice. What do you know about Steve Jobs?

He was one of the most influential people in the history of technology. He was a visionary.

What questions would you ask him?

I'd want to know how he came up with the idea for the iPod. It was so revolutionary at the time.

Totally...the iPod changed the way everyone listens to music.

It really did. It's crazy to think that it was released in 2001. That was a long time ago.

Omg wow, that makes me feel old!

I know, right? Time flies by so fast. I wish I could go back to those simpler times.

Me too, friend, me too. Speaking of music, what's your favorite album to listen to lately?

I've been listening to a lot of Led Zeppelin lately. They're my favorite band.

## 9.4B Parameters

Figure 1: Paper author (left) conversing with our 9.4B parameter model (right). This example was cherry picked. We release conversation logs with crowdworkers with our code, along with lemon-picked examples in Sec. 10.5.

# Thanks!

<https://seominjoon.github.io>

[minjoon@kaist.ac.kr](mailto:minjoon@kaist.ac.kr)