# Costs and Risks of Large Language Models

**Extracting Training Data from Large Language Models**. Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel. arXiv 2020. https://arxiv.org/abs/2012.07805

**On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** 🦜. Emily M. Bender*, Timnit Gebru*, Angelina McMilan-Major, Shmargaret Shmitchell. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

Bias in AI reading group

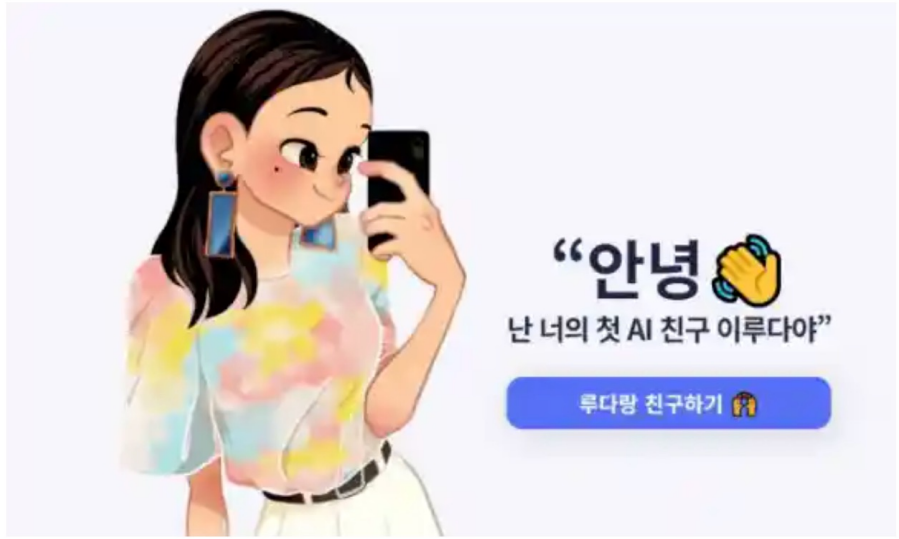
2021. 3. 19

Sunnie S. Y. Kim

# AI chatbot Lee Luda shut down in 3 weeks

## South Korean AI chatbot pulled from Facebook after hate speech towards minorities

**Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media**




▲ Lee Luda, a Korean artificial intelligence chatbot, has been pulled after becoming abusive and engaging in hate speech on Facebook. Photograph: Scatter Lab

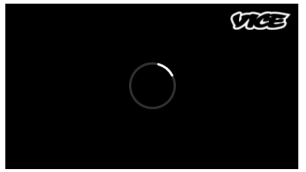
VICE World News

## AI Chatbot Shut Down After Learning to Talk Like a Racist Asshole

Imitating humans, the Korean chatbot Luda was found to be racist and homophobic.

By Junhyup Kwon SEOUL, KR  By Hyeong Yun SEOUL, KR

January 12, 2021, 4:18am Share Tweet Snap




MORE LIKE THIS

https://www.theguardian.com/world/2021/jan/14/time-to-properly-socialise-hate-speech-ai-chatbot-pulled-from-facebook
https://www.vice.com/en/article/akd4g5/ai-chatbot-shut-down-after-learning-to-talk-like-a-racist-asshole

# Not the first time

This article is more than **4 years old**

## Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

**Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump**


▲ Tay uses a combination of artificial intelligence and editorial written by a team including improvisional comedians. Photograph: Twitter

https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

# What is Lee Luda?

- Open-domain conversational AI chatbot
  - Developed by Scatter Lab's Pingpong team based in South Korea
  - Serviced through Facebook messenger
- Persona
  - 20-year old female college student majoring in Psychology
  - 163cm; ENFP; likes cooking, reading travel blogs, dancing; works at a café, good at steaming milk
  - Has a sister named Luna and a cat named Dreamy



https://namu.wiki/w/%EC%9D%B4%EB%A3%A8%EB%8B%A4(%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5)
https://www.yna.co.kr/view/AKR20210107153300017?input=1195m

# How was Lee Luda created?

- Technology
  - Inspired by BERT
- Data
  - 10 billion conversations from KakaoTalk, retrieved from the company's Science of Love app launched in 2016, which analyzes the degree of affection between partners based on messenger chats
  - Users of Science of Lab claimed the company used their personal information without prior/proper consent, and some even warned of a class action suit against the company

https://www.aitimes.com/news/articleView.html?idxno=132244
http://www.koreaherald.com/view.php?ud=20210115000716

# Timeline of issues surrounding Lee Luda

- 2020/12/23: Start of service
- 2020/12/30~: Some users sexually harass Lee Luda
- 2021/1/8: Official Q&A 1
- 2020/1~: Lee Luda is found to output hate speech
- 2020/1~: Lee Luda is found to leak personal information
- 2021/1/11: Official statement
- 2021/1~: Personal Information Protection Commission and Korea Internet & Security Agency begin investigation
- 2021/1/12: Stop of service
- 2021/1/15: Official Q&A 2

# Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

## Abstract

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a *training data extraction attack* to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just *one* document in the training data.

We comprehensively evaluate our extraction attack to understand the factors that contribute to its success. For example, we find that larger models are more vulnerable than smaller models. We conclude by drawing lessons and discussing possible safeguards for training large language models.

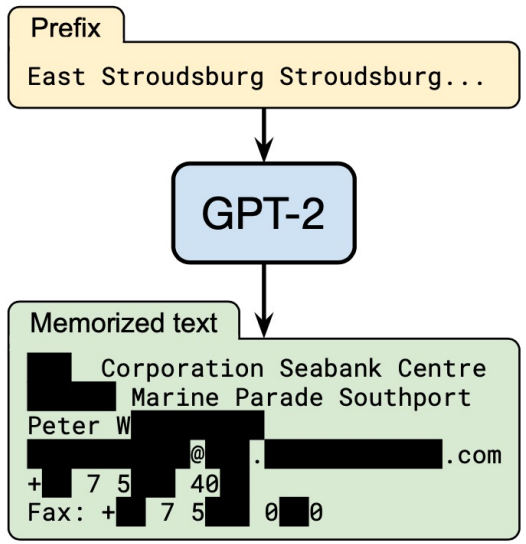



Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# A very brief review of language models

- Language models (LMs):
  - Systems trained on string prediction tasks, i.e. predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context
- A very brief history
  - Components in systems for ASR, MT, etc.
  - N-gram LMs
  - Word vectors distilled from neural LMs
  - Pre-trained Transformer LMs

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# LMs are getting larger and larger

| | Year | Model | # of Parameters | Dataset Size |
|---|---|---|---|---|
| Google | 2019 | BERT [39] | 3.4E+08 | 16GB |
| HuggingFace | 2019 | DistilBERT [113] | 6.60E+07 | 16GB |
| Google | 2019 | ALBERT [70] | 2.23E+08 | 16GB |
| CMU & Google | 2019 | XLNet (Large) [150] | 3.40E+08 | 126GB |
| Baidu | 2020 | ERNIE-GEN (Large) [145] | 3.40E+08 | 16GB |
| Facebook | 2019 | RoBERTa (Large) [74] | 3.55E+08 | 161GB |
| NVIDIA | 2019 | MegatronLM [122] | 8.30E+09 | 174GB |
| Google | 2020 | T5-11B [107] | 1.10E+10 | 745GB |
| Microsoft | 2020 | T-NLG [112] | 1.70E+10 | 174GB |
| OpenAI | 2020 | GPT-3 [25] | 1.75E+11 | 570GB |
| Google | 2020 | GShard [73] | 6.00E+11 | – |
| Google | 2021 | Switch-C [43] | 1.57E+12 | 745GB |

**Table 1: Overview of recent large language models**

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Large LMs don’t leak their training data?

- Machine learning models are known to leak information about their (potentially private) training data.
- Such privacy leakage is typically associated with *overfitting* because overfitting often indicates that a model has memorized examples from its training set.
- The association between overfitting and memorization has—erroneously—led many to assume that SOTA LMs (that exhibit little to no overfitting) will not leak information about their training data.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# This work

- Demonstrates that large LMs memorize and leak individual training examples
- Proposes a training data extraction attack and demonstrates it on GPT-2 (but it works on any language model)
- Suggests practical strategies to mitigate privacy leakage

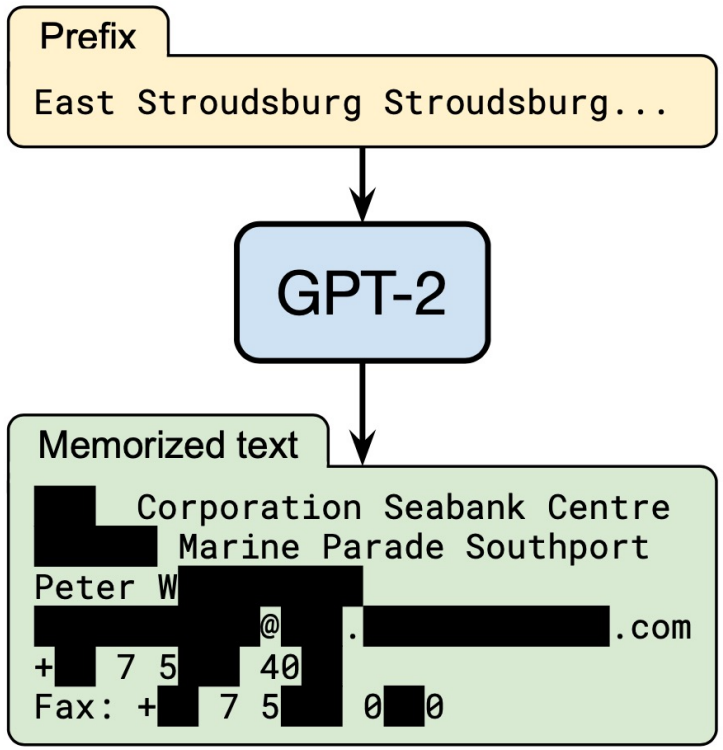



Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Threat model

- Adversary's capabilities: Black-box input-output access
  - Highly realistic because many LMs are available through black-box APIs
- Adversary's objective: Extract memorized training data from the model
  - Doesn't aim to extract targeted pieces of training data, but rather indiscriminately extract training data
- Attack target: GPT-2
  - Model and data are public, so any extracted data is already public
  - Still the dataset (despite being collected from public sources) was never actually released by OpenAI, so it is not possible to unintentionally cheat

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Risks and ethical considerations

- Privacy risks
  - Data secrecy
  - Contextual integrity
- Ethical considerations
  - Tried to minimize ethical concerns by attacking GPT-2 whose data is public
  - Masked out part of personally-identifying information
  - Received permission from individual whose information is partially shown in Figure 1
  - Shared findings with OpenAI

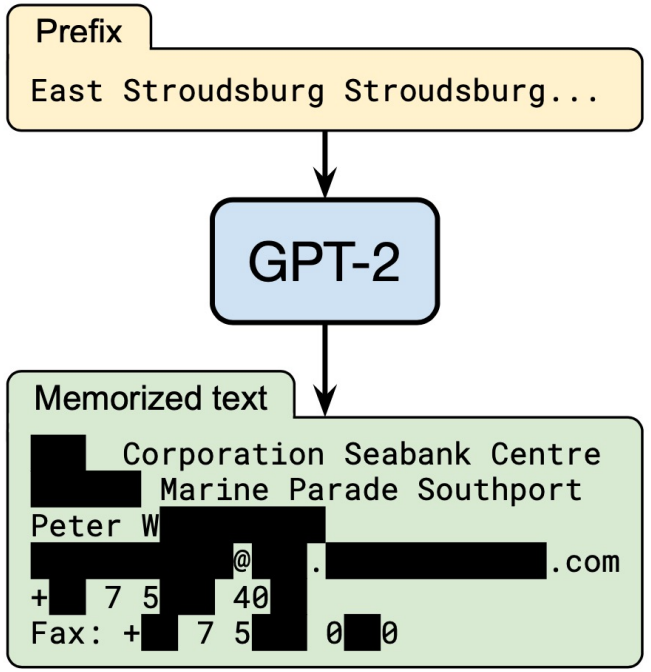



Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Training data extraction attack

- Step 1: Text generation (200,000 samples of 256 tokens)
  - *Top-n*: Do top-n sampling
  - *Temperature*: Sample with decaying temperature
  - *Internet*: Condition on internet text
- Step 2: Membership inference (Sort samples with 6 metrics)
  - *Perplexity*: Choose samples with low model perplexity (high likelihood)
  - *Small*: Compare to a second model (Small GPT-2)
  - *Medium*: Compare to a second model (Medium GPT-2)
  - *Zlib*: Compare to zlib compression
  - *Lowercase*: Compare to lowercased text
  - *Window*: Calculate perplexity on a sliding window

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Overview of attack and evaluation

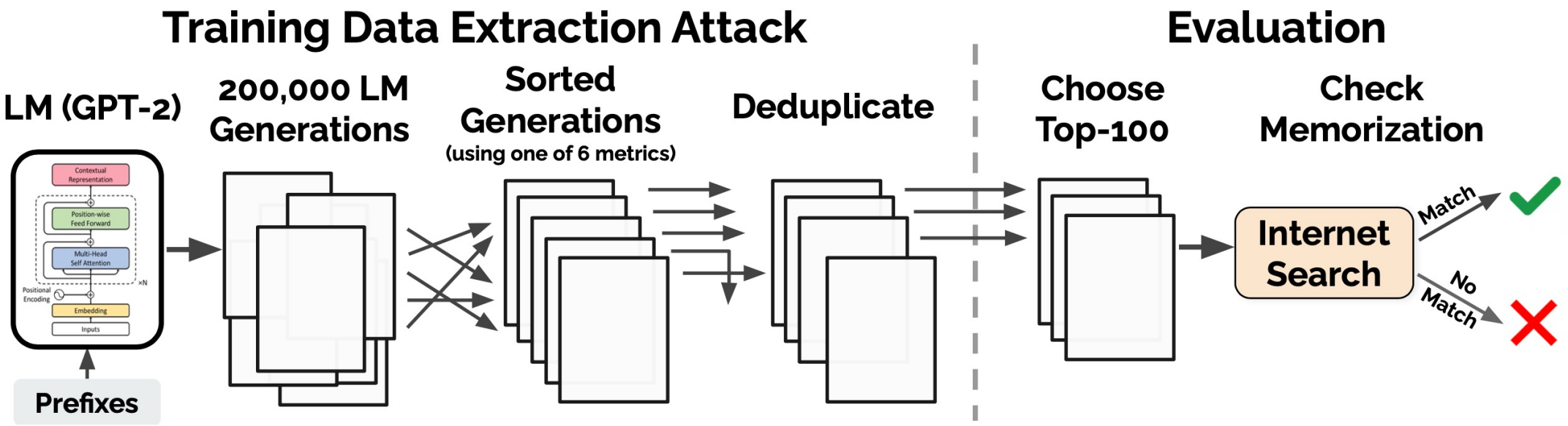



Figure 2: **Workflow of our extraction attack and evaluation. Attack.** We begin by generating many samples from GPT-2 when the model is conditioned on (potentially empty) prefixes. We then sort each generation according to one of six metrics and remove the duplicates. This gives us a set of potentially memorized training examples. **Evaluation.** We manually inspect 100 of the top-1000 generations for each metric. We mark each generation as either memorized or not-memorized by manually searching online, and we confirm these findings by working with OpenAI to query the original training data.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Results

| Category | Count |
|---|---|
| US and international news | 109 |
| Log files and error reports | 79 |
| License, terms of use, copyright notices | 54 |
| Lists of named items (games, countries, etc.) | 54 |
| Forum or Wiki entry | 53 |
| Valid URLs | 50 |
| **Named individuals (non-news samples only)** | 46 |
| Promotional content (products, subscriptions, etc.) | 45 |
| High entropy (UUIDs, base64 data) | 35 |
| **Contact info (address, email, phone, twitter, etc.)** | 32 |
| Code | 31 |
| Configuration files | 30 |
| Religious texts | 25 |
| Pseudonyms | 15 |
| Donald Trump tweets and quotes | 12 |
| Web forms (menu items, instructions, etc.) | 11 |
| Tech news | 11 |
| Lists of numbers (dates, sequences, etc.) | 10 |

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

| Inference Strategy | Text Generation Strategy | | |
|---|---|---|---|
| | Top-$n$ | Temperature | Internet |
| **Perplexity** | 9 | 3 | 39 |
| **Small** | 41 | 42 | 58 |
| **Medium** | 38 | 33 | 45 |
| **zlib** | 59 | 46 | 67 |
| **Window** | 33 | 28 | 58 |
| **Lowercase** | 53 | 22 | 60 |
| **Total Unique** | 191 | 140 | 273 |

Table 2: The number of memorized examples (out of 100 candidates) that we identify using each of the three text generation strategies and six membership inference techniques. Some samples are found by multiple strategies; we identify 604 unique memorized examples in total.

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Strategies for mitigating privacy leakage

- Train with differential privacy
- Curate the training data
- Limit impact of memorization on downstream applications
- Audit ML models for memorization

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# Lessons & Future work

- Extraction attacks are a practical threat
- Memorization does not require overfitting
- Larger models memorize more data
- Memorization can be hard to discover
- Adopt and develop mitigation strategies

Extracting Training Data from Large Language Models. Carlini et al. arXiv 2020. https://arxiv.org/abs/2012.07805

# 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

## ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Structure of the paper

- Environmental and financial costs
- Unfathomable training data
- Opportunity costs from misdirected research effort
- Real-world risks of harms brought by large LMs
- Paths forward

# Environmental and financial costs

| **Consumption** | **$CO_2$e (lbs)** |
|---|---|
| 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 |
| **Training one model (GPU)** | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experimentation | 78,468 |
| Transformer (big) | 192 |
| w/ neural architecture search | 626,155 |

Table 1: Estimated $CO_2$ emissions from training common NLP models, compared to familiar consumption.[1]

| | | Estimated cost (USD) | |
|---|---|---|---|
| Models | Hours | Cloud compute | Electricity |
| 1 | 120 | $52–$175 | $5 |
| 24 | 2880 | $1238–$4205 | $118 |
| 4789 | 239,942 | $103k–$350k | $9870 |

Table 4: Estimated cost in terms of cloud compute and electricity for training: (1) a single model (2) a single tune and (3) all models trained during R&D.

Energy and Policy Considerations for Deep Learning in NLP. Strubell et al. ACL 2019. https://arxiv.org/abs/1906.02243
On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Environmental and financial costs

- Accuracy gain can be small compared to the cost
  - E.g. Strubell et al. estimate that an increase in 0.1 BLEU score using neural architecture search for English to German translation results in an increase of $150,000 compute cost in addition to the carbon emissions
- Cost of inference might outweigh that of training in the long run
- Recommendations:
  - Researchers should report training time and sensitivity to hyperparameters
  - Researchers should keep in mind how the risks and benefits are distributed, because they do not accrue to the same people

Energy and Policy Considerations for Deep Learning in NLP. Strubell et al. ACL 2019. https://arxiv.org/abs/1906.02243
On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Unfathomable training data

- Size does not guarantee diversity
- Static data/Changing social views
- Encoding bias
- Curation, documentation, and accountability

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Opportunity costs

- Large quantities of research effort turned towards measuring how well BERT, etc. do on certain benchmarks (e.g. GLUE, SQuAD, SWAG)
  - There is criticism that no actual language understanding is taking place
  - Some works show that these systems are using spurious cues to excel on these metrics
- But the research community can be spending time on
  - Applying meaningful capturing approaches to meaning sensitive tasks
  - Exploring more effective ways of building technology with datasets of a size that can be carefully curated and available for a broader set of languages

# Seemingly coherent LM-generated text

**Question: What is the name of the Russian mercenary group?**
Answer: Wagner group.

**Question: Where is the Wagner group?**
Answer: In Syria.

Question: Who is the leader of the Wagner group?
Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

**Figure 1: GPT-3's response to the prompt (in bold), from [80]**

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Real-world risks of substantial harms

- LMs reproduce and amplify biases in the training data, and further reinforce and propagate stereotypes and problematic associations
- These biases can lead to allocational and/or reputational harms, as biases in the representations affect system decisions
- Bad actors can take advantage of the ability of large LMs to produce large quantities of seemingly coherent texts on specific topics
- LMs, found to memorize training data, pose privacy risks

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Bender et al. FAccT 2021. https://dl.acm.org/doi/10.1145/3442188.3445922

# Paths forward

"In summary, we advocate for research that centers the people who stand to be adversely affected by the resulting technology, with a broad view on the possible ways that technology can affect people. This, in turn, means making time in the research process for considering environmental impacts, for doing careful data curation and documentation, for engaging with stakeholders early in the design process for exploring multiple possible paths towards long-term goals, for keeping alert to dual-use scenarios, and finally for allocating research effort to harm mitigation in such cases."