

# On the dangers of stochastic parrots Can language models be too big?

---

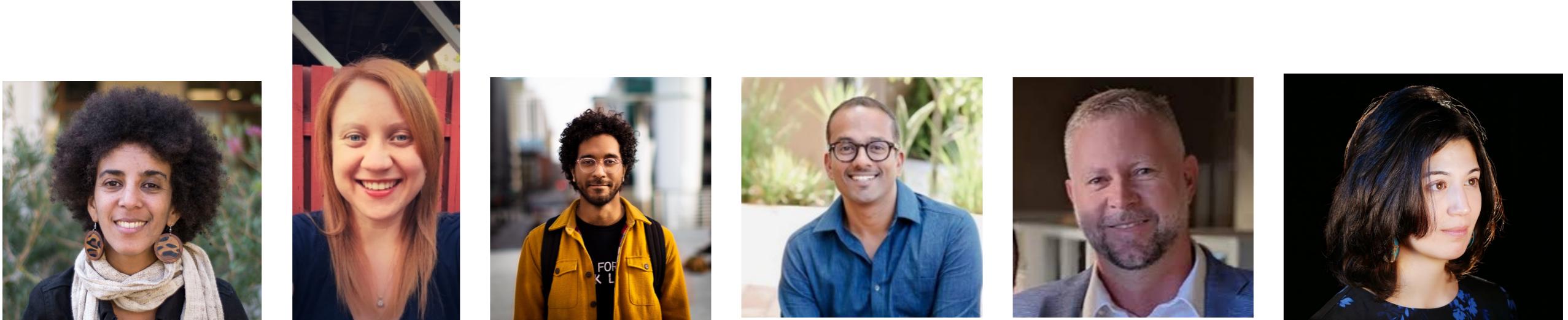
Emily M. Bender  
University of Washington  
[@emilymbender](https://twitter.com/emilymbender)

*The Alan Turing Institute*  
July 8, 2021  
Originally presented at FAccT 2021

- Joint work with: Timnit Gebru, Angelina McMillan-Major, Margaret Mitchell, Vinodkumar Prabhakaran, Mark Díaz, and Ben Hutchinson



- *Prabhakaran*: Prabhakaran et al 2012, Prabhakaran & Rambow 2017, Hutchinson et al 2020
- *Hutchinson*: Hutchinson 2005, Hutchinson et al 2019, 2020, 2021
- *Díaz*: Lazar et al 2017, Díaz et al 2018



# We would like you to consider

---



- Are ever larger language models (LMs) inevitable or necessary?
- What costs are associated with this research direction and what should we consider before pursuing it?
- Do the field of natural language processing or the public that it serves in fact need larger LMs?
- If so, how can we pursue this research direction while mitigating its associated risks?
- If not, what do we need instead?

# Overview

---



- History of Language Models (LMs)
- Risks
  - Environmental and financial costs
  - Unmanageable training data
  - Research trajectories
  - Potential harms of synthetic language
- Risk Mitigation Strategies

# Brief history of language models (LMs)

---



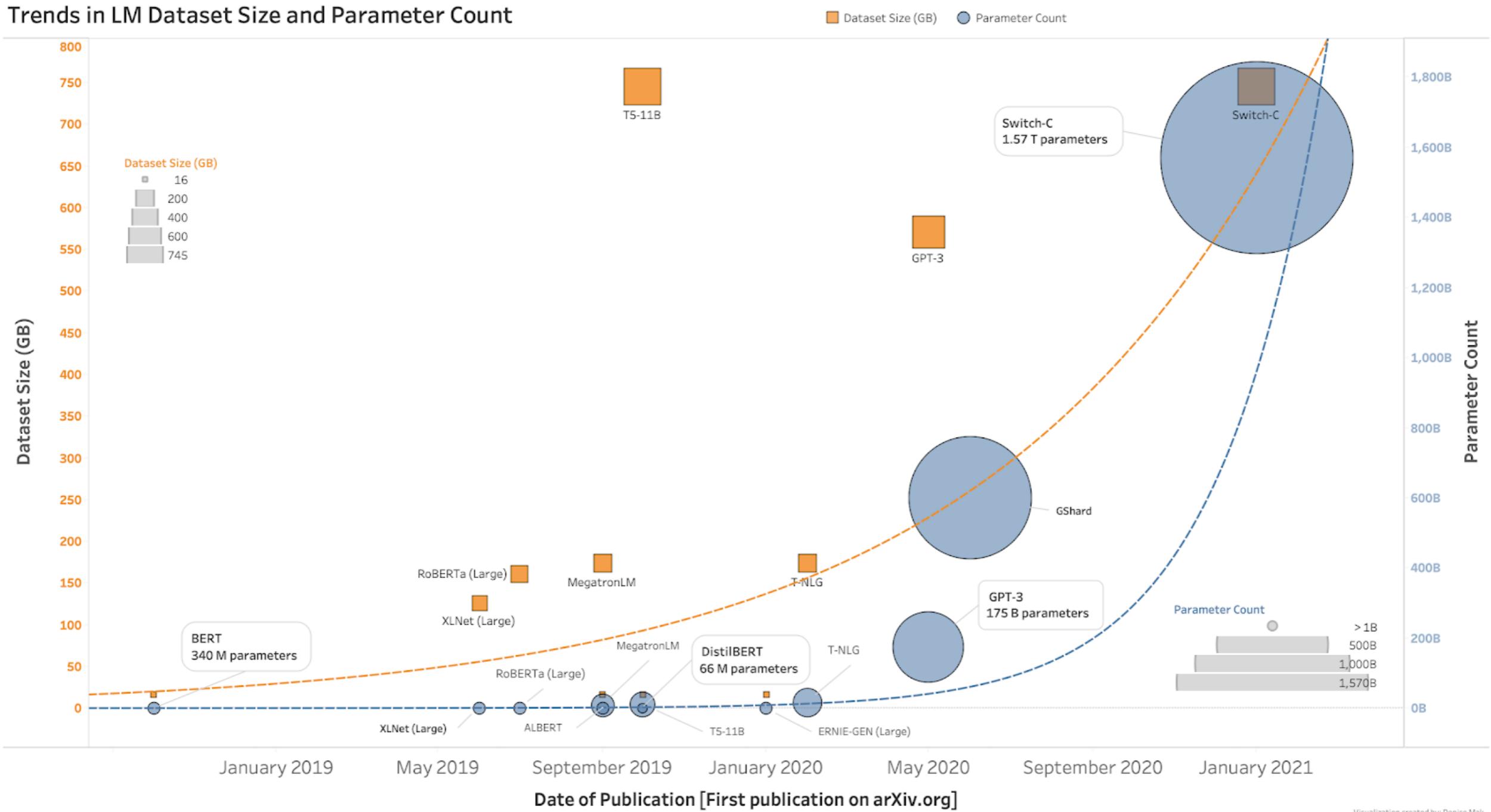
- LM: A system trained to do string prediction
  - *What word comes \_\_\_? What word [MASK] here?*
- Proposed by Shannon in 1949, but implemented for ASR, MT, etc. in early 80's
  - N-grams and various neural architectures through Transformers
- Big takeaways
  - Better scores through more data and bigger models until scores don't improve, then move to new architecture
  - Multilingual models up to ~100 languages
  - Model-size reduction strategies
  - Growth of models  $\propto$  range of application of models

# How big is big?

[Special thanks to Denise Mak for graph design]



Trends in LM Dataset Size and Parameter Count



Visualization created by: Denise Mak



*What are the risks?*

Environmental costs & financial inaccessibility

# Environmental and financial costs

---



- Average human across the globe responsible for 5t of CO2 emissions per year\*
- Strubell et al. (2019)
  - Transformer model training procedure on GPUs 284t of CO2 emissions
  - 0.1 BLUE score increase en-de results in increase of ~\$150,000 in compute cost
  - Encourage reporting training time and sensitivity to hyperparameters
  - Suggest more equitable access to compute clouds through government investment
- Which researchers and which languages get to ‘play’ in this space and who is cut out?

\*Source: [Our World In Data](#)

# Current mitigation efforts

---



- Renewable energy sources
  - Still incur a cost on the environment & take away from other potential uses of green energy
- Prioritize computationally efficient hardware
  - SustainNLP workshop
  - Green AI and promoting efficiency as evaluation metric (Schwartz et al 2020)
- Document energy and carbon metrics
  - Energy Usage Reports (Lottick et al 2019)
  - Experiment-impact-tracker (Henderson et al 2020)

# Costs and risks to whom?

---



- Large LMs, particularly those in English and other high-resource languages, benefit those who have the most in society
- Marginalized communities around the world impacted most by climate change
  - Maldives threatened by rising sea levels (Anthoff et al 2010)
  - 800,000 residents of Sudan affected by flooding (7/2020-10/2020)\*
- But these communities are rarely able to see benefits of language technology because LLMs aren't built for their languages, Dhivehi and Sudanese Arabic

\*Source: <https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un>



*What are the risks?*

Unmanageable Training Data

# A large dataset is not necessarily diverse

---



- Who has access to the Internet and is contributing?
  - Younger people and those from developed countries
- Who is being subject to moderation?
  - Twitter - accounts receiving death threats more likely to be suspended than those issuing threats
- What parts of the Internet are being scraped?
  - Reddit - US users 67% men and 64% are ages 18-29 (Pew)
  - Wikipedia - only 8.8-15% are women or girls
  - Not sites with fewer incoming and outgoing links, like blogs
- Who is being filtered out?
  - Filtering lists primarily target words referencing sex, likely also filtering LGBTQ online spaces

# Static data/Changing social views

---



- LMs run the risk of ‘value lock’, reifying older, less-inclusive understandings
- BLM movement lead to increased number of articles on shootings of Black people and past events were also documented and updated (Twyman et al 2017)
  - But media also doesn’t cover all events and tend to focus on more dramatic content
- LMs encode hegemonic views; retraining/fine-tuning would require thoughtful curation (see Solaiman and Dennison 2021 for partial proof of concept)
- See also Birhane et al 2021: ML applied as prediction is inherently conservative

# Bias

---



- Research in probing LMs for bias has provided a wealth of examples of bias
  - See Blodgett et al 2020 for a critical overview
- Documentation of the problem is an important first step, but not a solution
- Automated processing steps may themselves be unreliable
- Probing requires knowing what social categories the LM may be biased against
  - Need for local input before deployment

# Curation, documentation, accountability

---



- *How big is too big?*
  - Budget for documentation and only collect as much data as can be documented
  - Documentation: understand sources of bias & potential mitigating strategies
  - No documentation: potential for harm without recourse
- *Documentation debt:* datasets both undocumented and too big to document post-hoc

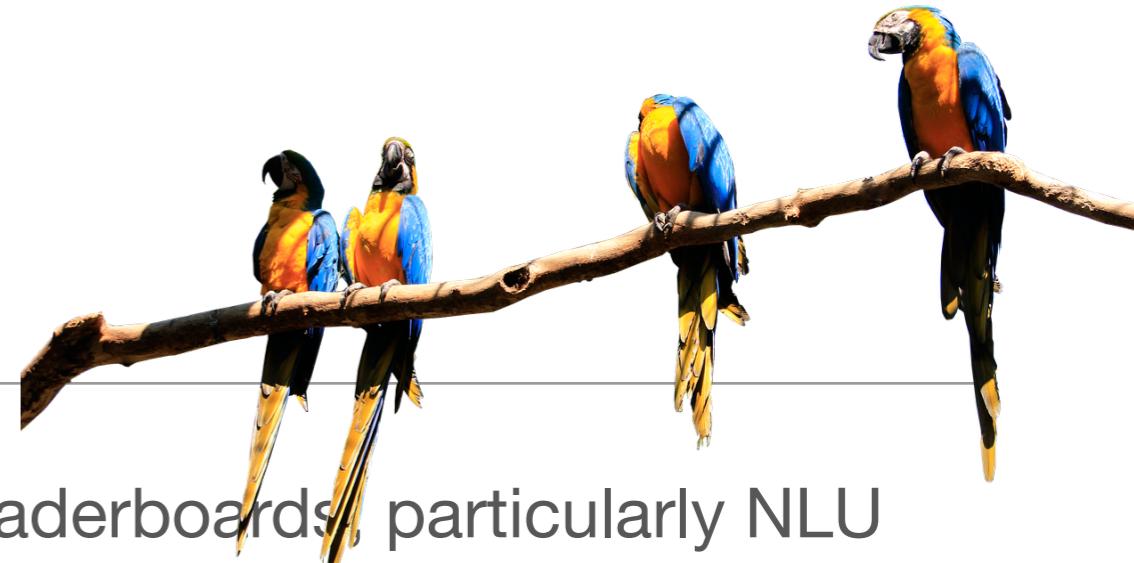


# Research Trajectories

*What are the risks?*

# Research time is a valuable resource

---



- Focus on LMs and achieving new SOTA on leaderboards, particularly NLU
- But LMs have been shown to excel due to spurious dataset artifacts (Niven & Kao 2019, Bras et al 2020)
- LMs trained only on linguistic form don't have access to meaning (Bender & Koller 2020)
- Are we actually learning about machine language understanding?



*What are the risks?*

Potential harms of synthetic language

# Stochastic



- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)
- An LM is a system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning: a *stochastic parrot*.
- Nonetheless, humans encountering synthetic text make sense of it
  - Coherence is in the eye of the beholder



# Potential harms

---



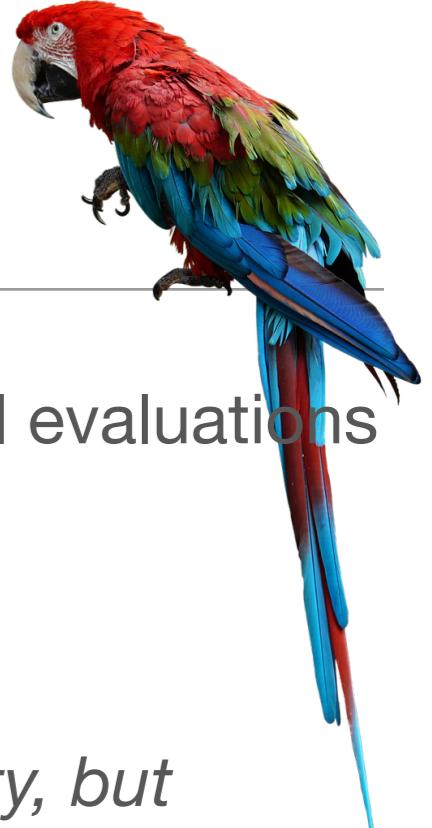
- Denigration, stereotype threat, hate speech:  
harms to reader, harms to bystanders
- Cheap synthetic text can boost extremist recruiting (McGuffie & Newhouse 2020)
- LM errors attributed to human author in MT
- LMs can be probed to replicate training data for PII (Carlini et al 2020)
- LMs as hidden components can influence query expansion & results (Noble 2018)



*Risk management strategies*

# Allocate valuable research time carefully

---



- Incorporate energy and compute efficiency in planning and model evaluations
- Select datasets intentionally
  - *'Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.'* (Birhane and Prabhu 2021, after Benjamin)
- Document process, data, motivations, and note potential users and stakeholders
- Pre-mortem analyses: consider worst cases and unanticipated causes
- Value sensitive design: identify stakeholders and design to support their values

# Risks of backing off from LLMs?

---



- What about benefits of large LMs, like improved auto-captioning?
  - Are LLMs in fact the only way to get these benefits?
  - What about for lower resource languages & time/processing constrained applications?
- Are there other ways the risks could be mitigated to support the use of LMs?
  - Watermarking synthetic text?
- Are there policy approaches that could effectively regulate the use of LLMs?

# We would like you to consider

---

- Are ever larger language models (LMs) inevitable or necessary?
- What costs are associated with this research direction and what should we consider before pursuing it?
- Do the field of natural language processing or the public that it serves in fact need larger LMs?
- If so, how can we pursue this research direction while mitigating its associated risks?
- If not, what do we need instead?



## References

- Bender, E. M., Gebru, T., McMillan-Major, A., and et al (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of FAccT 2021*.
- Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. (2021). The values encoded in machine learning research. <https://arxiv.org/abs/2106.15590>.
- Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 114. Association for Computing Machinery, New York, NY, USA.
- Hutchinson, B. (2005). Modelling the substitutability of discourse connectives. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 149–156, Ann Arbor, Michigan. Association for Computational Linguistics.
- Hutchinson, B., Pittl, K. J., and Mitchell, M. (2019). Interpreting social respect: A normative lens for ML models. *CoRR*, abs/1908.07336.
- Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denyul, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *CoRR*, abs/2005.00813.
- Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560575, New York, NY, USA. Association for Computing Machinery.
- Lazar, A., Diaz, M., Brewer, R., Kim, C., and Piper, A. M. (2017). Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 655668, New York, NY, USA. Association for Computing Machinery.
- Prabhakaran, V. and Rambow, O. (2017). Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse*, 8(2):21–55.
- Prabhakaran, V., Rambow, O., and Diab, M. (2012). Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada. Association for Computational Linguistics.
- Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (PALMS) with values-targeted datasets. *CoRR*, abs/2106.10328.

---

For remaining works cited, see the bibliography in Bender, Gebru et al 2021.

Sources for parrot photos:

- <https://www.maxpixel.net/Bird-Red-Parrot-Animal-Fly-Vintage-Wings-1300223>
- <https://www.maxpixel.net/Parrots-Parrot-Birds-Isolated-Plumage-Branch-Bird-2850879>
- <https://www.maxpixel.net/Tropical-Animal-World-Bill-Parrot-Cute-Bird-Ara-3080543>
- <https://www.maxpixel.net/Animal-Ara-Plumage-Isolated-Bird-Parrot-4720084>
- <https://www.maxpixel.net/Tropical-Ara-Bird-Feather-Exotic-Bill-Parrot-3064137>
- <https://www.maxpixel.net/Plumage-Colorful-Exotic-Birds-Ara-Parrot-5202301>
- <https://www.maxpixel.net/Flight-Parrots-Parrot-Isolated-2683451>