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

Emily M. Bender
University of Washington
@emilymbender

AI Sweden & RISE NLP
Sept 14, 2022
Originally presented at FAccT 2021

Slides:
https://bit.ly/ParrotsSept2022

- 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, Hutchison et al 2020
  - *Hutchinson*: Hutchinson 2005, Hutchison et al 2019, 2020, 2021
  - *Díaz*: Lazar et al 2017, Díaz et al 2018

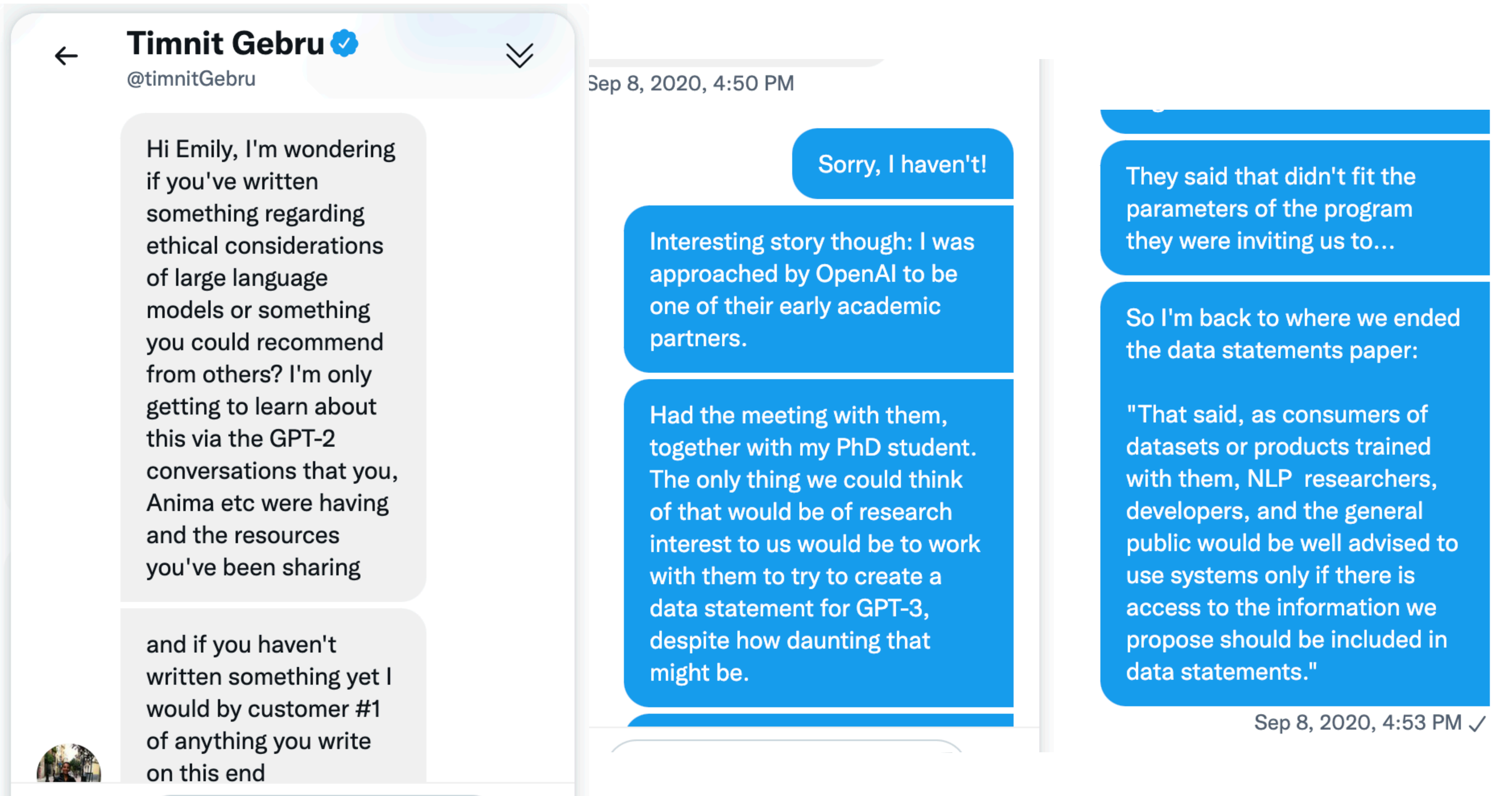
Slides: https://bit.ly/ParrotsSept2022






History & context of this paper

- Started off as a Twitter DM conversation, with Dr. Timnit Gebru:



History & context of this paper

I can think of three other angles around ethical implications of GPT-3 and the like:

1) Carbon cost of creating the damn things (see Strubell et al at ACL 2019)
2) AI hype/people claiming it's understanding when it isn't (Bender & Koller at ACL 2020)
3) Deepfakes/random generated text that no one is accountable for but which is interpreted as meaningful.

Sep 8, 2020, 4:54 PM ✓

This is somethign I'm trying to advocate for at Google

sent them some of your tweets as well haha

so many converstations about how we're not leading in large language models and should

GPT-3 is so impressive etc

and each time I'm like ANNDD see what Emily has to say

big data energy...

Sep 8, 2020, 4:55 PM ✓

I'm trying to advocate for documentation

first off

and second off interventionist data collection

Sep 8, 2020, 4:55 PM

See Gebru et al 2018 :)

What is interventionist data collection?

well saying instead of only depending

on whats available

like the internet

also think of ways to get other data, and also curate more. This is Eun Seo's term :)

she's a historian and we talk about how in our view, archives are more interventionist and anything ML on the other end of the spectrum

History & context of this paper

Rather than collecting general web garbage but doing so in such quantities that you can pass it off as good stuff?

I can kind of see a paper taking shape here, maybe not saying that large language models are bad, but rather using large language models as a case study for ethical pitfalls and what can be done better.

Would you be interested in co-authoring such a thing?

Sep 8, 2020, 4:58 PM ✓

:-) I would absolutely love to, but honestly I was thinking you know way more here

and I would love to refer people to see this paper

I'm finding myself referring to data statements paper + your tweets

oh oops gotta go
but I
would love to continue this convo

Sep 8, 2020, 4:59 PM

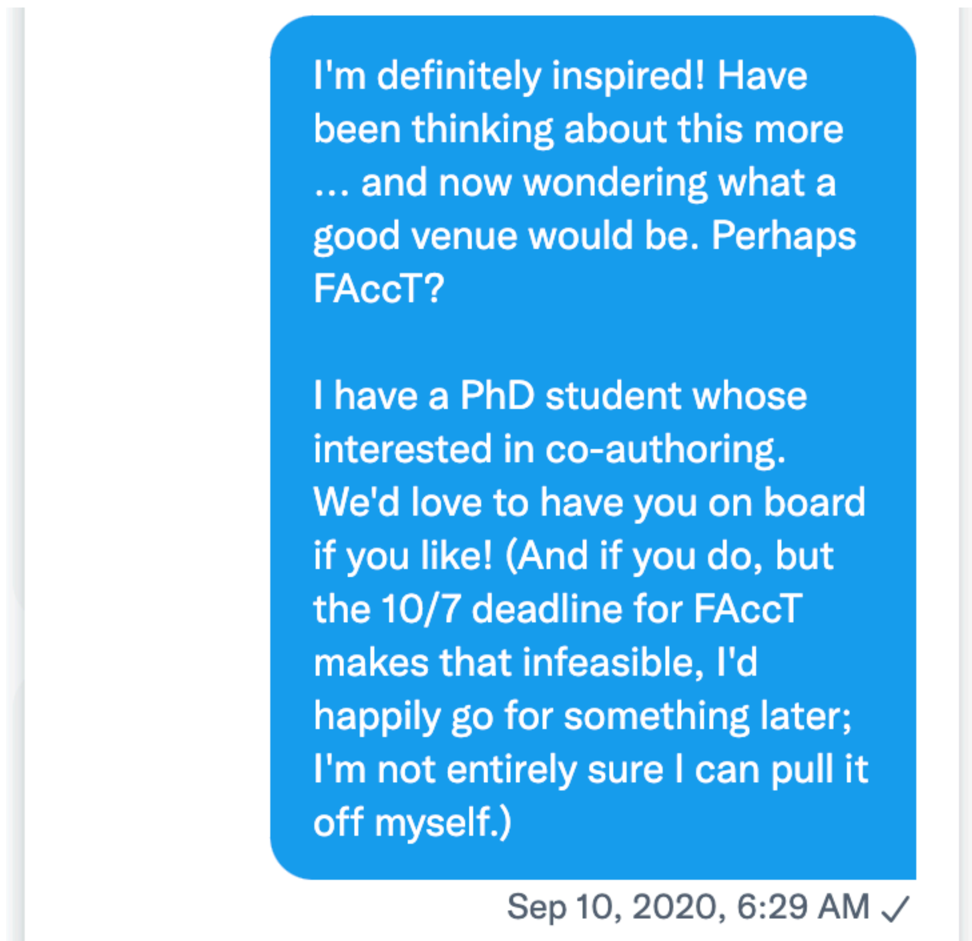
Yes, to be continued!

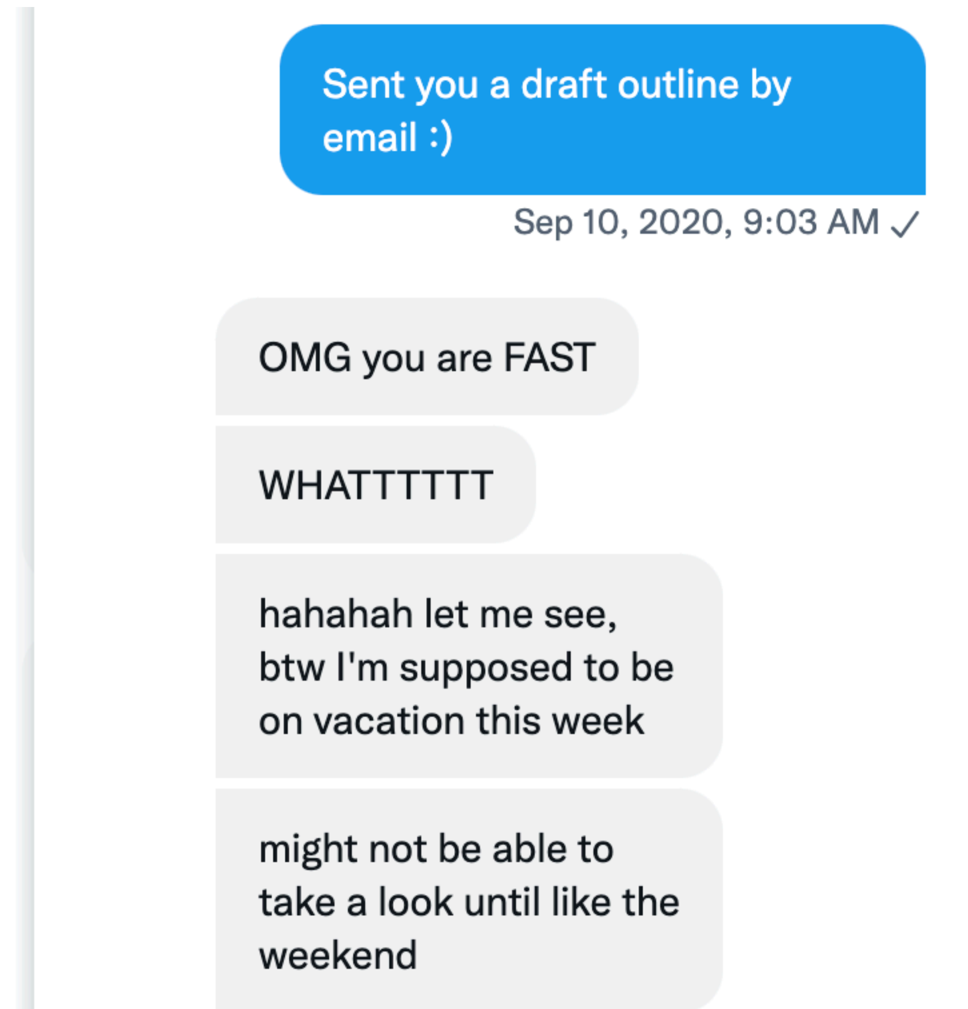
I suspect we have complementary expertise and it would be much easier to do the framing if we could work on it together. When you have time, if there are particular tweets of mine you tend to refer to, could you send me links? (That might be good fodder for getting started…)

Sep 8, 2020, 5:00 PM ✓

History & context of this paper

- Two days later:





History & context of this paper

- Four more co-authors joined from Google
- Pooled knowledge of 7 co-authors made it possible to pull the paper together by the Oct 7 deadline for FAccT
    - Survey/position paper: No new data analysis or experiments
- Cleared Google’s “pub-approve” process before submission
- Sent it out to 30+ people for feedback in parallel to the peer review process

History & context of this paper

- Sharing all of this to help situate this research in its context
  - I’d like the audience to understand what our goals were for the paper as a piece of scholarship
  - You probably knew of this paper from the news before reading it, quite different to how one normally approaches research papers
- I think it’s also interesting to reflect on the processes of scholarship

History & context of this paper

- Late November: Google asks Dr. Gebru to either retract the paper or remove the Google co-authors' names from it

- Dr. Gebru pushes back, asking for information on what exactly was being objected to and objecting to how she & her team were being treated

- December 2, 2020: Google fires Dr. Gebru

- Dr. Margaret Mitchell starts documenting what happened to Dr. Gebru and calling on people within Google to apologize & fix systems

- February 19, 2021: Google fires Dr. Mitchell

History & context of this paper

- Google's actions led to intense media interest, both about their treatment of Dr. Gebru (and eventually Dr. Mitchell) and about our research
  - Selected media coverage
- Someone leaked the "pub-approve" version of the paper to Reddit
- Meanwhile…

History & context of this paper

- FAccT 2021 primary reviewers were complete before the media story broke (preserving anonymous review)
- FAccT 2021 acceptances announced on December 22, 2020
- Camera ready due January 22, 2021
- Still not allowed to include co-authors with Google affiliations

- 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, Hutchison et al 2020

- *Hutchinson*: Hutchinson 2005, Hutchison et al 2019, 2020, 2021

- *Díaz*: Lazar et al 2017, Díaz et al 2018

Slides: https://bit.ly/ParrotsSept2022

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

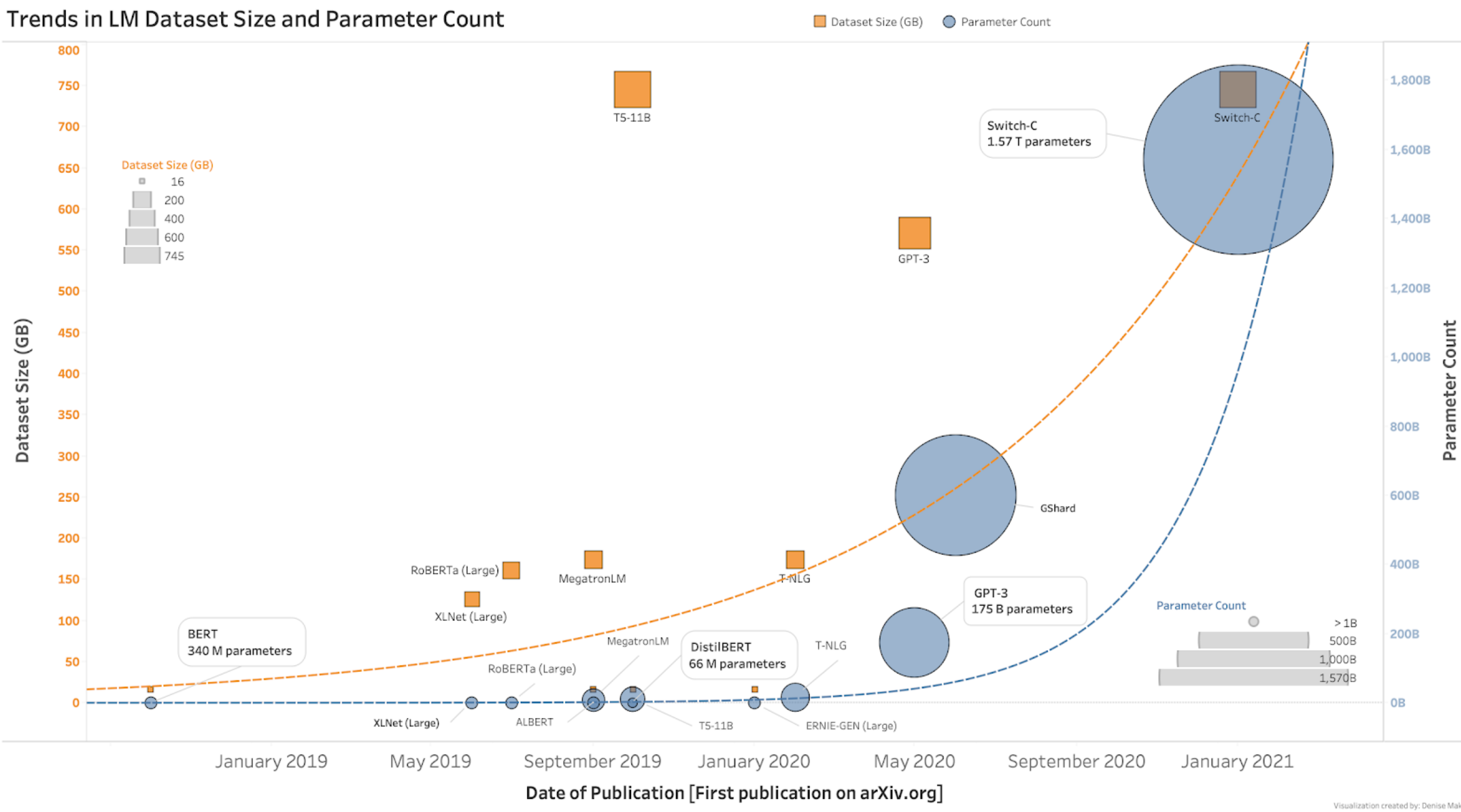
- History of Language Models (LMs)
- Risks
  - Environmental and financial costs
  - Unmanageable training data
  - Research trajectories
  - Potential harms of synthetic language
- Risk Mitigation Strategies

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]



Updates since early 2021 (non-exhaustive)

Model	Source	Date	Parameters	Tokens	Citation
MT-NLG	Microsoft + NVIDIA	Oct 2021	530B	270B	
Gopher	DeepMind	Dec 2021	280B	300B $\sim$1.3TB	(Rae et al 2021)
LaMDA	Google	Jan 2022	137B	1.56T	(Thoppilan et al 2022)
PaLM	Google	Apr 2022	540B	780B	(Chowdhery et al 2022)
BLOOM	BigScience	July 2022	176B	366B	

What are the risks?

Environmental costs & financial inaccessibility

Environmental and financial costs

- Average human across the globe responsible for 5t of CO_2 emissions per year*
- Strubell et al. (2019)
  - Transformer model training procedure on GPUs 284t of CO_2 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 (see also Marshall 2021)
- 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 (see also Dodge et al 2021)

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

We can’t help ourselves

- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)
- Text generated by an LM is not grounded in any communicative intent, model of the world, or model of the reader’s state of mind
- Counter-intuitive, given the increasing fluency of text synthesis machines, but:
  - Have to account for our predisposition to interpret locutionary artifacts as conveying coherent meaning & intent (Weizenbaum 1976, Nass et al 1994)

Stochastic 🦜

- 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)

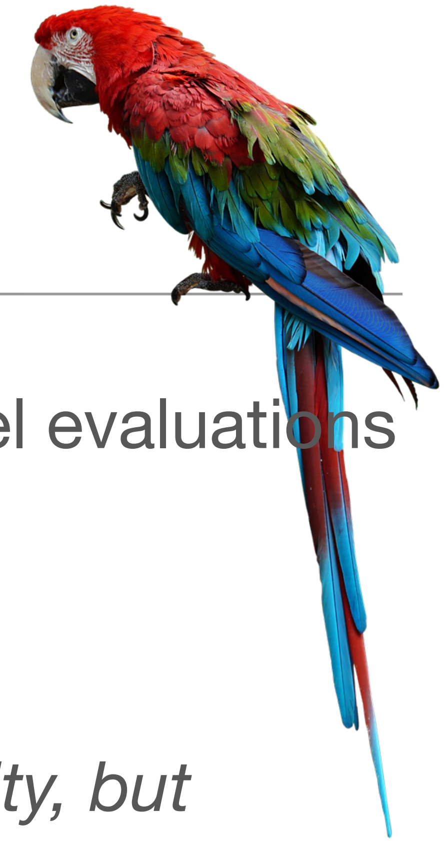
- LMs as hidden components can influence query expansion & results (Noble 2018)

Potential harms

- These harms largely stem from the interaction of the ersatz fluency of today's language models + human tendency to attribute meaning to text
- Deeply connected to issue of accountability:
    - Synthetic text can enter conversations without anyone being accountable for it
- Accountability key to responsibility for truthfulness and to situating meaning
- Maggie Nelson (2015): "Words change depending on who speaks them; there is no cure."

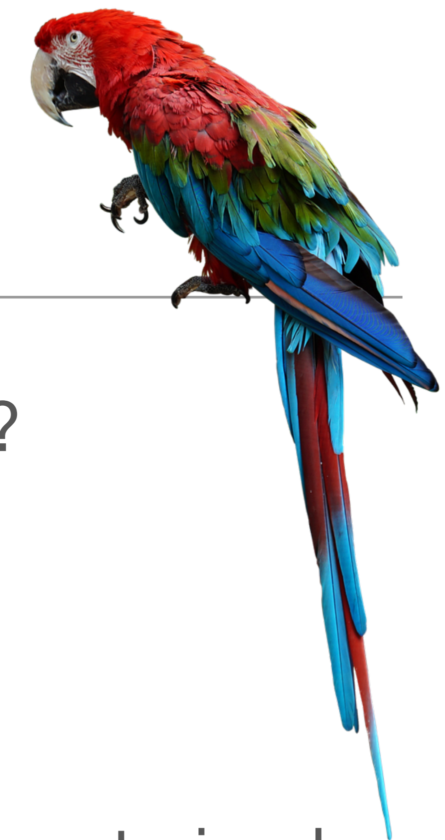
Risk management strategies

Allocate valuable research time carefully



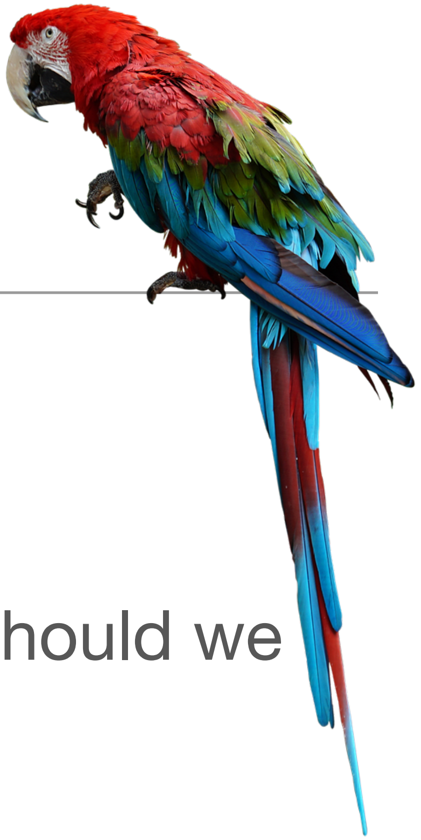
- Incorporate energy and compute efficiency in planning and model evaluations
- Select datasets intentionally
    - ‘*Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.*’ (Birhane and Prabhu 2021, after Ruha Benjamin)
- Document 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?

The view from 2022

- Has the development of LLMs / tech based on LLMs slowed down? (No)
- Has data and model documentation become more mainstream? (Yes, but…)
- Have people become more aware of the risks of this technology? (Yes, but…)

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when there isn't a simple answer

MUM has the potential to transform how Google helps you with complex tasks. MUM uses the T5 text-to-text framework and is 1,000 times more powerful than BERT. MUM not only understands language, but also generates it. It's trained across 75 different languages and many different tasks at once, allowing it to develop a more comprehensive understanding of information and world knowledge than previous models. And MUM is multimodal, so it understands information across text and images and, in the future, can expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM could understand you're comparing two mountains, so elevation and trail information may be relevant. It could also understand that, in the context of hiking, to "prepare" could include things like fitness training as well as finding the right gear.

https://blog.google/products/search/introducing-mum/

See Shah & Bender 2022

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when ther
answer

MUM has the potential to transform how Goog
tasks. MUM uses the T5 text-to-text framewo
powerful than BERT. MUM not only understan
it. It's trained across 75 different languages and many different tasks at once, allowing it to develop a more comprehensive understanding of information and world knowledge than previous models. And MUM is multimodal, so it understands information across text and images and, in the future, can expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM could understand you're comparing two mountains, so elevation and trail information may be relevant. It could also understand that, in the context of hiking, to "prepare" could include things like fitness training as well as finding the right gear.

https://blog.google/products/search/introducing-mum/

Start now

Our platform can be plugged into any library, making it possible for NLP to be integrated into every build.

Large language models

Our models have been trained on billions of words, allowing them to learn nuance and context.

https://cohere.ai/

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

Helping you when ther
answer

MUM has the potential to transform how Goog
tasks. MUM uses the T5 text-to-text framewor
powerful than BERT. MUM not only understands
it. It's trained across 75 different languages and m
allowing it to develop a more comprehensive unde
and world knowledge than previous models. And
understands information across text and images
expand to more modalities like video and audio.

Take the question about hiking Mt. Fuji: MUM cou
comparing two mountains, so elevation and trail i
It could also understand that, in the context of hik
include things like fitness training as well as findin

https://blog.google/products/search/in

Start now

Our platform can be plugged into any

Large language models

Our models have been trained on
ving them to
text.

ps://cohere.ai/

Is it real?

This is just an experiment with AI technology. We wanted to pay homage to a great thinker and leader with a fun digital experience. It is important to remember that AI in general, and language models specifically, still have limitations. The model can sometimes give inaccurate or inappropriate responses, so you should take any information given with a grain of salt.

https://ask-rbg.ai/

The view from 2022

- Have tech cos cooled down the AI hype? (Of course not)

- Have people at large become better at critically analyzing claims of "understanding language"?

Thank you!

- Slides: https://bit.ly/ParrotsSept2022
- Twitter: @emilymbender

References

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., 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`.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling language modeling with pathways.

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.

Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (To appear, 2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of EMNLP 2021*.

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 Denuyl, 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.

Marshall, B. (2021). Algorithmic misogynoir in content moderation practice. `https://us.boell.org/en/2021/06/21/algorithmic-misogynoir-content-moderation-practice-1`.

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.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. v. d., Hendricks, L. A., Rauh, M., Huang, P.-S., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J.-B., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., d'Autume, C. d. M., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., Casas, D. d. L., Guy, A., Jones, C., Bradbury, J., Johnson, M., Hechtman, B., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher.

Shah, C. and Bender, E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA. Association for Computing Machinery.

Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (PALMS) with values-targeted datasets. *CoRR*, abs/2106.10328.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E., and Le, Q. (2022). LaMDA: Language models for dialog applications.

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`